Long Video Content Analysis: Learning to Summarize Wireless Capsule Endoscopy Videos

by

Sodiq Adewole

A dissertation submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
to the
School of Engineering and Applied Sciences
University of Virginia
Charlottesville, VA

December 2021

Committee Approval:

We, the undersigned committee members, certify that we have reviewed this dissertation and approve it in partial fulfillment of the requirements of the degree of Doctor of Philosophy in Systems and Information Engineering.

Donald E. Brown, Ph.D. (Engineering Systems and Environment) Date
Advisor

Laura Barnes, Ph.D. (Engineering Systems and Environment) Date
Committee Chair

Michael D. Porter, Ph.D. (Engineering Systems and Environment) Date
Committee Member

Sana Syed, MD, MS, MSDS (School of Medicine) Date
Committee Member

Afsaneh Doryab, Ph.D. (Engineering Systems and Environment) Date
Committee Member

Certified by:

Jennifer L. West, Dean, School of Engineering and Applied Science Date

Abstract

The overall goal of this dissertation is in two parts; 1) minimizing experts' review time on Capsule Endoscopy (CE) videos via video summarization; and 2) developing models that captures the temporal and topological relationship between frames in the videos as against an independent image analysis that has been addressed in literature.

With an estimated 70 million Americans affected by different digestive tract diseases each year, physicians use VCE as a nonsurgical procedure to examine the entire digestive tract without the invasiveness associated with the traditional upper and lower endoscopy procedures. While VCE helps ease diagnosis of many digestive tract diseases, a single capsule endoscopy study can last between 8 - 11 hours generating up to 100,000 images of various sections of the digestive tract. Even when up to fifty thousand (50,000) images are obtained in a typical small bowel VCE study, it is possible for pathology of interest to be present in as few as one single frame. Physicians have to review the entire video in order to identify frames with the pathology of interest.

Many researchers have proposed different techniques to automate analysis of CE frames, however, large proportion of the proposed techniques require fully labelled video frames for each class of abnormality in the video. Meanwhile, collecting frame-level annotation for medical video is not an easy task. In this dissertation, we developed novel models with three (3) levels of supervision to mitigate this problem. Our goal is to generate summaries with selected representative frames that captures the regions of abnormality in the GI tract thereby saving the physician the time and effort required to review the entire video.

The first model in this dissertation is an unsupervised video shots boundary detector used for efficient temporal segmentation of the VCE video. The key novelty is in the efficient representation of the video frame features with a lower 1-dimensional embedding. It is prohibitively expensive to temporally segment a video using the high-dimensional frame features extracted from a CNN model. Therefore, we projected the frame features to a 1-dimensional embedding space to minimize the computational cost of detecting shots boundaries. Our experiments with multiple embedding algorithms shows that encoding the video features using PCA achieved the best performance in shot boundary detection on the videos.

Secondly, we developed a weakly-supervised temporal segmentation technique using Graph-based representation learning. We believe the topological relationship between the frames is better captured using a GCNN model as it relaxes the hard assumption of temporal dependence between the frames as well the implicit frames independence assumption in traditional CNN model. In addition, while a short video may follow a temporal correlation assumption, multiple scene and

events in long videos may not. The goal of our GCNN model is to learn to map the nodes in the graph into binary class-agnostic categories. During testing, we use the categories of the nodes to segment the videos into normal and abnormal segments. To achieve this, we represented each video segment as a graph and each frame as the nodes in the graph. We trained the graph in a class agnostic manner to separate normal from abnormal nodes. We represented the relationship between the frames as the edge weights of the graph and the model acts as a binary classifier to classify each frame into abnormal or normal frame. Chaining this prediction together allows us to temporally detect scene change in the video and segment the video into an homogeneous identifiable pathological unit.

Lastly, leveraging the boundary detection technique above, we developed an end-to-end weakly supervised abnormality localization model where we applied video-level labels, to localize the frames where the relevant disease is captured. A GCNN model was trained to generate an embedding for each video segment and then classifies the video into binary category of abnormal or normal. We considered full video as a graph, each video segment as a sub-graph and the frames as the nodes. The model was divided into two parts - graph classification and abnormality localization. The graph classification model, trained based on cross-entropy loss classifies each sub-graph (video segment) into binary disease-agnostic classes and the disease localization selects relevant frames from each abnormal video segment that contains the respective disease. An extension of this framework, which we describe in our future work, would be an end-to-end localization of the full long video with multiple abnormalities.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

BoVW	Bag of Visual Word
BS	Binary Segmentation
CAE	Convolutional Autoencoder
CD	Celiac's Disease
CE	Capsule Endoscopy
CNN	Convolutional Neural Network
CPD	Change Point Detection
EHFC	Energy and High Frequency Content
GCNN	Graph Convolution Neural Network
GI	Gastrointestinal
GNN	Graph Neural Network
HMM	Hidden Markov Models
HSI	Hue-Saturation-Intensity
IFB	Inflammatory Bowel Disease
IID	Independent Identically Distributed
KL	Kullback–Leibler
LSTM	Long Short Term Memory
MIL	Multi - Instance Learning
NB	Naive Bayes
OGIB	Obscure GI Bleeding
OIC	Outer-Inner-Contrastive Loss
OP	Optimal Partitioning
PCA	Principal Component Analysis
PELT	Pruned Exact Linear Time
SBC	Small Bowel Crohn's

SIFT	Scale Invariant Feature Transform
SN	Segment Neighbourhood
SURF	Speeded-Up Robust Features
SVD	Singlular Value Decomposition
SVM	Support Vector Machine
TAL	Temporal Abnormality Localization
TSNE	T-Stochastic Neighborhood Embedding
TSN	Temporal Segment Network
VS	Video Summarization
WCE	Wireless Capsule Endoscopy

List of Symbols

V^n	Complete video for individual patient
f_i	Video frame at index i
x_i	CNN features for frame f_i
v_i^n	Video segment i for patient n
P_{f_k}	Probability density of frame f_k features
T	Total length or number of frames in a video
p	Dimension of video features x_i
$\forall$	For all
τ	Detected change points in video sequence
θ_i	Parameters of frame feature vectors in segment i
m	Number of detected segment boundaries
μ	Mean of video segment feature vectors
λ_i	1-d feature vector after projection
G	Undirected graph
E	Edges between nodes of graph G
N	Nodes representing frames of graph G
e_{ij}	Edge weights between two nodes $\{i, j\}$ of G
n	Number of frames in graph G
A	Adjacency matrix for graph G
D	Degree matrix of graph G
W_l	Parameters of GCNN network at layer l
Γ	Graph convolution operation in GCNN network
h_i^l	Hidden state of GCNN for node i at layer l
z_i	Embedding features for node i at last layer
y_i	Class of disease for node i in G
N_{ab}	Number of abnormal nodes/frames in G
c_i	Indicator function for detected abnormal node/frame
L_{div}	Diversity loss function
L_{recon}	Reconstruction loss function

Chapter 1

Introduction

Analysis of videos encompasses tasks such as object detection, object recognition, tracking, action localization and general understanding of objects behavior in a video. Video structured data is becoming more and more popular across multiple domains including surveillance, medical diagnosis, human social behavior and entertainment. The rate at which video data is generated across these and many more domains keeps increasing. Meanwhile, while collection of video data keeps getting easier, analysis of videos, particularly VCE videos, is extremely tedious, time-consuming and also prone to error.

The key to analysis of video structured data is leveraging both spatial (images) and temporal information in the data. Analysis of VCE videos has received significant attention for more than two (2) decades. However, significant portion of the prior works have limited the task to detection of objects/abnormalities in the video frames while paying little to no attention to the temporal and topological relationship between the frames. We believe the frames in VCE videos have *spatio-temporal* as well as *topological* structure that needs to be considered when building any automated system for the analysis. Capturing the relationship between the sequence of frames, would also allow us track the dynamics of any abnormality in the GI tract. Factoring such relationship into models would, in addition, offer useful information that could guide the development of more robust, end-to-end automated VCE video analysis system.

Research efforts on automating analysis of videos have been on for many years and many interesting techniques have been developed over the years with promising results (see literature review section). However, many of the systems focus on short video clips with fixed number of frames or clips containing only one target event or activity. Meanwhile, real videos, in offline settings, are either longer with multiple actions or events or constantly being streamed. Very little attention has be

given to an end-to-end offline analysis of long videos nor the computational cost that is required for such task. Therefore, adapting any off-the-shelf video analysis model to a real world long video application would require significant modification or some form of manual preprocessing for it to be effective. This modification could be very expensive, particularly when an expert attention is required, which may not be readily available. The work in this dissertation is motivated by this real world application and we will be describing our contribution on developing an end-to-end system to automate analysis of long VCE videos.

Furthermore, VCE videos have unique properties that, if not properly factored-in, significantly degenerates the performance of any generic image and video analysis technique. For example, poor illumination, occlusion by food, unstable camera motion resulting in frequent camera flip in the digestive tract, and inter-patient variability creates a wide gap between models for ordinary video analysis and VCE video analysis. In order to understand the motivation for this dissertation, it is critical to mention the above which defines the structure of the research work that was conducted. We strove to strike a balance between generalization of the approach in this work as well as the clinical relevance when the specific characteristics VCE video data is considered.

This dissertation tackles two challenging problems in long video analysis: (1) Video summarization; and (2) Abnormality localization. The main focus of the first solution is to develop a system that reduces the computational cost of generating video summaries from long videos by exploiting both spatial and temporal structure of the data. The second solution focuses on leveraging the topological relationship between the frames to generate a summarized form by localizing abnormalities from video-level information to frames containing the abnormalities. Our solutions is applied to VCE videos collected during standard clinical procedure.

This chapter provides an overview of the dissertation, highlights key contributions and discusses the contributions in our other works that have both been published. Section 1.4 gives the full list of our previous publication as well as brief explanation of the contribution in each paper.

1.1 Video Capsule Endoscopy (VCE)

Endoscopy is the non-surgical procedure used to visualize and examine the stomach, upper small bowel and colon of a person (see fig. 1.1). Using an endoscope, a flexible tube which carries light by fibreoptic bundles with attached camera, the physician is able to view pictures of the digestive tract on a color TV monitor. Traditionally, three main endoscopy procedures include gastroscopy, small-bowel endoscopy and colonoscopy. During gastroscopy, also known as the upper endoscopy, an endoscope is easily passed through the mouth and throat and into the esophagus,

thereby allowing the physician to view the esophagus and stomach [1].The small bowel endoscopy advances further and allows visibility into the upper part of the small intestine. Colonoscopy involves passing endoscopes into the colon through the rectum to examine the colon. Small bowel endoscopy is especially limited by how far it can advance into the small bowel, thereby limiting the extent of the physicians' examination. All three traditional methods are also limited due to the invasiveness and discomfort that accompanies them. While there has not been a complete replacement for these traditional procedures, especially when a biopsy (removal of tissue) is necessary, VCE has innovatively changed the approach to endoscopy to make the procedure a lot less invasive and uncomfortable.

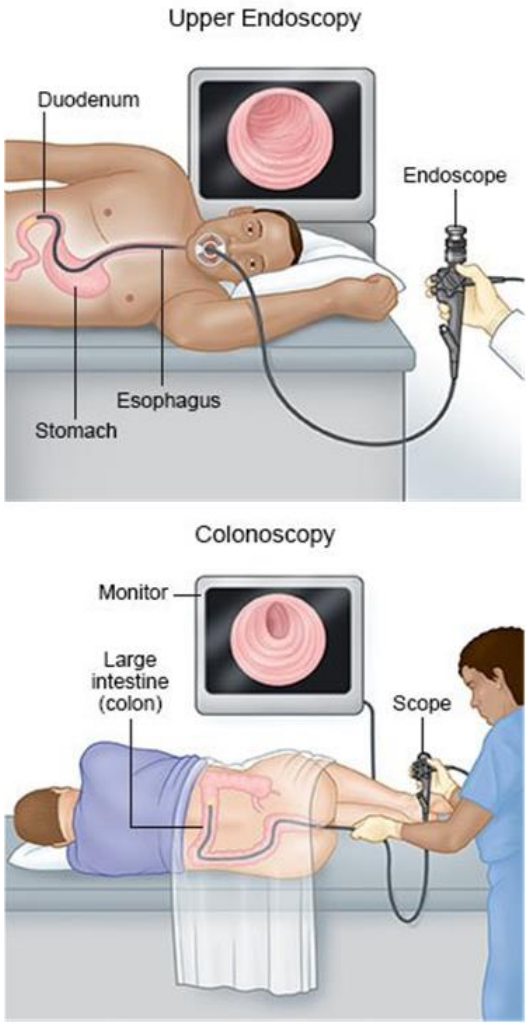


Figure 1.1: Upper and Lower Endoscopy Procedure

VCE was invented in 2000 by Iddan G. [2] to allow for painless imaging of the whole small bowel. And for more than two decades, it has gained popularity as the less invasive alternative to other

traditional endoscopy procedures. Having now become the standard procedure for visualization and diagnostics of several gastrointestinal (GI) tract diseases, transforming the traditional gastroscopy, small-bowel endoscopy and colonoscopy to allow non-invasive visualization of the entire GI tract. The small bowel region of the GI tract could contain multiple abnormalities including Inflammatory Bowel Disease (IFB) [3], Obscure GI Bleeding (OGIB) [4], Small Bowel Crohn's (SBC) [5], Celiac Disease (CD) [5], ulcer [6], and polyps [7]. Traditional upper and lower GI endoscopy allows a gastroenterologist to visualize up to the proximal duodenum and ascending colon respectively [8] but are not able access and visualize the distal duodenum, jejunum, and ileum of the small bowel. These together is 5.5-6 meters in length [9, 10]. VCE innovatively allows visualization of all regions of the entire GI tract, thereby facilitating prompt and easy diagnosis.

During the VCE procedure, the patient swallows a tiny capsule camera (see fig. 1.2) which is propelled down the gastrointestinal (GI) tract through peristaltic movement of the intestinal walls. The capsule camera captures images at about 2 to 6 frames per second (fps) as it navigates through the entire digestive system. A single VCE procedure can last between 8 - 11 hours generating up to 100,000 images [11]. The collected images are transmitted to a recorder [1] attached to the patient and subsequently transferred to a work station where they are reviewed and analysed by a human expert gastroenterologist. The video is manually reviewed and analysed - frame-by-frame - in order to identify regions of the GI tract with lesioned tissues and/or other abnormalities. In a single CE study, up to fifty thousand (50,000) images may be obtained in the small bowel region, while as few as one single frame may capture pathology of interest. Given the large volume of images generated in single VCE procedure, coupled with the high redundancy rate in the frame distribution, this review process can be very tedious, time-consuming, and error prone. The detection rate of VCE depends on abnormality indication but was 56% - 61% [9] in a pooled analysis. While capsule endoscopy procedure is superior in many respects to alternative imaging, it has a significant miss rate of 5.9% for vascular lesions, 0.5% for ulcers, and 18.9% for neoplasm, many of which are due to inherent limitations in human readability [12].

[1] https://www.medtronic.com/covidien/en-us/products/capsule-endoscopy/pillcam-sb-3-system.html

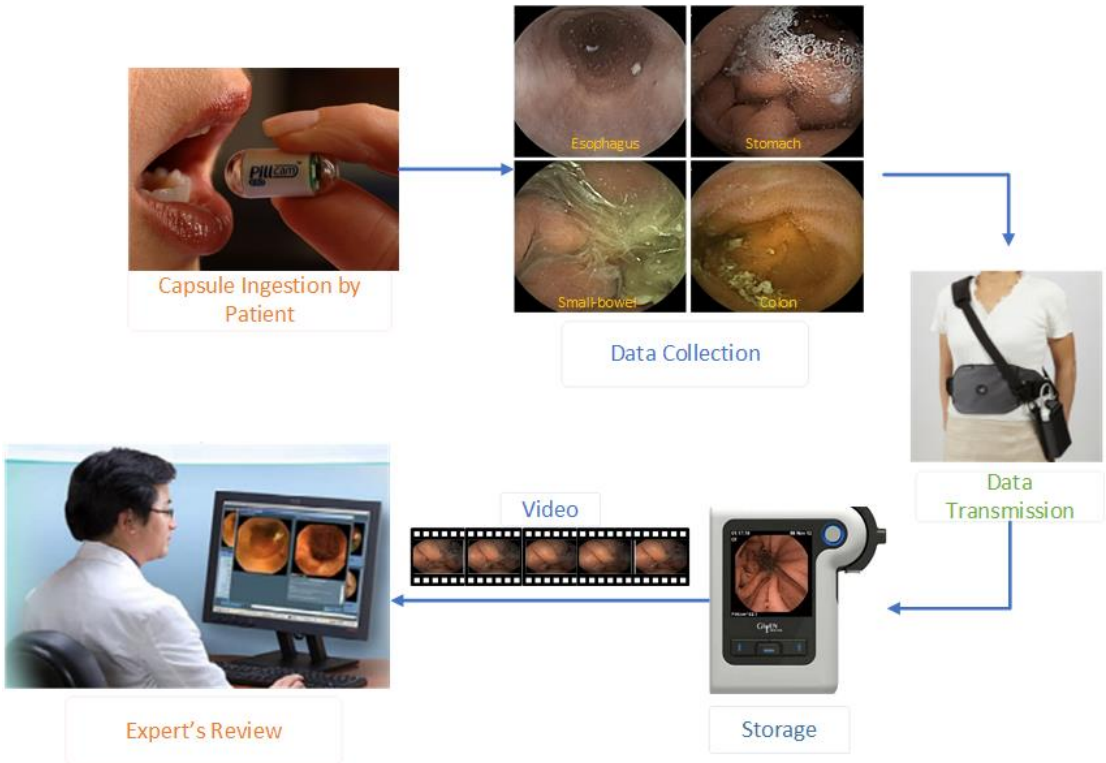


Figure 1.2: Video Capsule Endoscopy Procedure

1.2 VCE Video Analysis

In this section, we discuss some of the motivation for this dissertation and also provide overview for the work in the remainder of the dissertation.

Analysing VCE videos encompasses disease or abnormality recognition, quantifying severity of a disease, approximating location of any identified abnormality, making proper follow-up diagnosis and also decisions on intervention. For close to two decades, researchers have proposed different techniques to automate this process by leveraging both classical machine learning and image analysis techniques [7, 13] and as well as more recent and advanced deep learning methods [14, 15, 16]. These prior works fall into three broad categories which include 1) Detecting or recognizing specific object or abnormality such as bleeding in VCE frames; 2) Anomaly or outlier detection where frames with abnormalities are consider outliers; and 3) VCE video summarization by selecting key representative frames from the entire video.

To motivate the work in this dissertation, we will discuss each of these categories, the problem formulations and their limitations to pave the way for the work done in the remaining chapters of this woek.

Firstly, approaches aimed at detecting or recognizing specific disease, objects or abnormalities in VCE video frames have been proposed by many researchers. These models are trained to recognize only a specific disease / abnormality in the video frames and using a binary classifier, they identify only those frames containing the disease of interest. This approach tends to work well when the categories of abnormality captured in a each endoscopy study are limited but spread over the entire video [13]. However, VCE videos exhibit distribution that is usually skewed towards the normal categories. So, every category that is different from the normal frames will have far fewer samples. This characteristic of CE video makes gathering and annotating sufficient example frames of every abnormality very difficult. Therefore, models that identify specific lesions will suffer lack of robustness on any new class of lesion encountered in an unseen video. Consequently, this will necessitate time-to-time retraining of an already trained model to generalize to new classes. Secondly, given a very large labelled dataset, Deep Convolutional Neural Network (DCNN) models have demonstrated superior performance on different object detection and other image analysis tasks across multiple domains including detection of diseased frames in VCE frames. However, deep learning models are notoriously sample inefficient and training them to generalize to multiple diseases will require very large labelled examples of each disease across multiple patients. This will impact classification based models as its impracticable to gather enough samples of all the classes of the GI abnormalities to take advantage of the performance of these DCNN models. Furthermore, just applying a DCNN model on video structured data, with frames captured in sequence, assumes complete temporal independence between the frames, thereby ignoring the correlation between them. Other sequence-based models such as variants of Recurrent Neural Network - LSTM and GRU - have also shown promising results on multiple task when applied on text and video data. However, their performance degenerates significantly when applied on a very long sequence, which is typical of VCE videos.

Another formulation of VCE video analysis is outlier detection methods [3,15,17]. Outlier detectors identifies rare objects, events or observations which raise suspicion by differing significantly from other data points [18]. However, outlier detectors do not provide detailed information of the detected outlier, employing this approach on CE data will lead to flagging large number of frames as outlier, most of which will contain similar redundant information of the same disease/abnormality. This leads to very little time saving for the physician. Both single or multiple lesion detection model and outlier detection models do not account for any associative relationship between the sequence of frames. Each frame instance is assumed to be independent of the other frames in the video.

The third approach involved key frame extraction on VCE videos [19,20]. Other literature termed this task video summarization [16,21,22,23,24,25,26,27] where the goal is to reduce the amount of data that must be examined in order to retrieve the desired information in the video. This allows reviewer to only examine few selected key frames thereby distilling the information contained in

the entire video. Video summarization techniques generally assume the videos have already been manually segmented into shots (a continuous sequences of frames taken over a short period of time) and therefore extract key frames from within each shot. One major challenge with this traditional approaches to key frame extraction is that the length of the video summary must be set depending on the number of shots in the video. Secondly, the generated summary does not guarantee that the selected frames will not be correlated and determining the time interval between shots may be difficult. However, video summarization techniques generally take into account the information contained between frame sequence, eliminates redundancy and consequently reduces the required time for review and analysis of videos.

With only about five percent (5%) or less of VCE video frames capturing useful and informative content that aids the gastroenterologists' diagnosis, the motivation for this work is borne out of effort towards minimizing time spent by expert gastroenterologist in reviewing VCE video. In many cases, in order to save time, expert gastroenterologist ask a junior medical researcher for an initial review and to extract summarized and more informative frames for secondary review.

1.3 Dissertation Outline

This work investigates three key solutions to long VCE video summarization, figure 1.3 captures the overall structure of the content of the chapters and below, we give more detailed overview of each chapter.

1.3.1 Chapter 2: Literature Review

Chapter two (2) covers review of techniques from literature that have been developed to solve various problems in VCE and other video analysis. The review covers temporal shot boundary detection techniques, video summarization and video activity localization.

1.3.2 Chapter 3: Shot Boundary Detection and Temporal Segmentation in Long Videos

The work presented in chapter (3) has been submitted to the 2021 ***IEEE International Conference on Bioinformatics & Biomedicine*** and it is ***currently under review***. In this paper, we developed a model for automatic video shot boundary detection with minimal computational cost. Detecting temporal boundaries allows us to automatically segment long VCE videos into short,

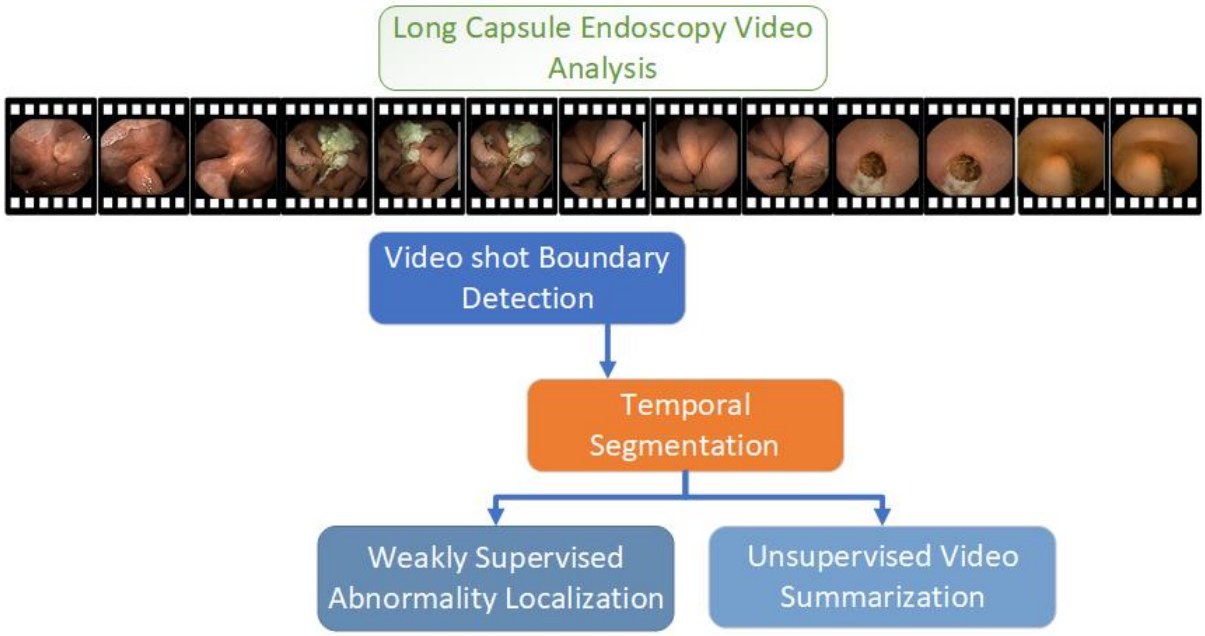


Figure 1.3: Overall Dissertation Outline

meaningful, homogeneous and identifiable video clips.

We extracted frame-features matrix from a model pretrained on large imageNet data and then fine-tune on our VCE video frames. Thereafter, we projected the frame-features into a 1-dimensional manifold space with the sequence for the entire video appearing like a single time series data. Projecting from p-dimensional video features reduces the computational cost of segmenting the video from $\mathcal{O}(n^p)$ or $\mathcal{O}(np)$ to linear $\mathcal{O}(n)$. We conducted experiment using different embedding methods and then applied the Pruned Exact Linear Time statistical change point detection technique to detect points at which there is a distributional shift in the sequence.

Many open dataset used in video analysis research have already been manually segmented into short video clips with fixed frame counts, therefore many video analysis techniques, especially deep learning based models, are designed to operate mostly on short video clips. Manually segmenting long video into clips have two (2) main problems: 1) The sequence of frames contained in different video shots cannot be guaranteed uncorrelated. Manually segmenting long videos, therefore, will not yield an homogeneous and identifiable segment that can lead to optimal summarization output; 2) When a non-homogeneous video segment is to be summarized, there is a chance of selecting a non-key frame as the representative frame, leading to higher miss-rate in the diagnosis.

1.3.3 Chapter 4: Weakly Supervised Temporal Segmentation of Long Capsule Endoscopy Video Using Graph Neural Network

The work presented in this chapter will be submitted to the 2021 *IEEE International Symposium on Biomedical Imaging (ISBI)* . Following our unsupervised temporal shot boundary detection model in chapter 3 in this work, we developed a weakly supervised class-agnostic model for temporal segmentation of the long CE videos. Detecting boundary is sensitive to visual change in the sequence of frames which may not necessarily coincide with a pathological event. Frequent camera flip due to the unstable peristaltic movement of the bowel leads to frequent visual change in the sequence of frames.

To mitigate this problem, we developed a weakly-supervised boundary detection for our VCE temporal segmentation. We trained a class-agnostic GCNN model for binary classification. Each frame is predicted to belong to either the abnormal or normal category [17]. Abnormal frames contain any type of disease or content that physician would be interested in. To benchmark this approach, we trained a baseline CNN model on same dataset. Using such binary classifier as our temporal segmentation model, we are able to more accurately segment the video into homogeneous and pathologically identifiable video segments for our summarization. The performance of the summarization model therefore depends on the performance of this classifier. The novelty of this model is in striking a balance between extreme assumption of temporal correlation using time-series model and complete independence assumption using traditional outlier detection technique that do not factor in the temporal information. GCNN takes advantage of topological relationship between the nodes in the graph through message passing, to learn a representation for each node also known as embedding.

1.3.4 Chapter 5: Weakly Supervised Abnormality Localization

The work presented in this chapter has been submitted to the 2021 ***IEEE International Conference on Big Data*** and is currently ***under review***. The main contribution of this work lies in mitigating the high cost of frame-level annotation for VCE videos. Obtaining frame level label for VCE videos is a very difficult task. With the sparsity of frames containing abnormality in the entire video, and the expertise needed to assign labels, the task can be highly time consuming and also expensive. However, using other patients' medical record and self-reported information, physicians are able to retrieve global video level information, though with some uncertainty. In this we, we used this video level label to develop a weakly supervised localization model using GCNN. The GCNN model localizes the video labels to the frames containing the abnormality. The GCNN model has two components - Graph Classification and Abnormality Localization - after training to classify video

segments into binary categories, the model tries to localize the frames capturing the abnormality based on the node-feature activation score.

A single long video contains an average of three (3) to four (4) different categories of abnormality. Since each video is more likely to contain more than one disease or abnormalities; a set of video of videos will contain different categories abnormalities. First we apply shot boundary detection technique described in chapter 3 to partition the videos into short segments and during training, the model classifies each video segment into binary abnormal / normal video segments. During testing, we applied a temporal pool layer over the network to select the top k-nodes in the graph to be representative of each video segment. We demonstrated the effectiveness of this model on multiple VCE videos.

1.3.5 Chapter 6: Video Summarization Using Encoder-Decoder LSTM for Key Frame Selection

Following the work in chapter 3 where we temporally segment the videos into short video segments, here we developed a key / most representative frame selection from each video segment to serve as the summary of the long VCE video. We trained an encoder-decoder LSTM model using the trio of diversity, sparsity and reconstruction losses. Details of the model as well as our experimentation is covered in chapter 6.

1.4 List of Publications

- Adewole, Sodiq, Michelle Yeghyayan, Dylan Hyatt, Lubaina Ehsan, James Jablonski, Andrew Copland, Sana Syed, and Donald Brown. "Deep Learning Methods for Anatomical Landmark Detection in Video Capsule Endoscopy Images." In Proceedings of the Future Technologies Conference, pp. 426-434. Springer, Cham, 2020.
  Abstract: Video capsule endoscope (VCE) is an emerging technology that allows examination of the entire gastrointestinal (GI) tract with minimal invasion. While traditional endoscopy with biopsy procedures are the gold standard for diagnosis of most GI diseases, they are limited by how far the scope can be advanced in the tract and are also invasive. VCE allows gastroenterologists to investigate GI tract abnormalities in detail with visualization of all parts of the GI tract. It captures continuous real time images as it is propelled in the GI tract by gut motility. Even though VCE allows for thorough examination, reviewing and analyzing up to eight hours of images (compiled as videos) is tedious and not cost effective. In order to pave way for automation of VCE-based GI disease diagnosis, detecting the location of the capsule

would allow for a more focused analysis as well as abnormality detection in each region of the GI tract. In this paper, we compared four deep Convolutional Neural Network models for feature extraction and detection of the anatomical part within the GI tract captured by VCE images. Our results showed that VGG-Net has superior performance with the highest average accuracy, precision, recall and, F1-score compared to other state of the art architectures: GoogLeNet, AlexNet and, ResNet.

- Adewole, Sodiq, Philip Fernandez, James Jablonski, Michelle Yeghyayan, Michael Porter, Andrew Copland, Sana Syed, and Donald Brown. "Lesion2Vec: Deep Meta Learning For Few Shots Multiple Lesions Recognition In Video Capsule Endoscopy Video." This work has been accepted to the Future Technology Conference for publication.
  Abstract: Effective and rapid detection of lesions in the Gastrointestinal (GI) tract plays a critical role in how fast gastroenterologist can respond to life-threatening diseases. Capsule Endoscopy (CE) has revolutionized traditional endoscopy procedure by allowing physician visualize the entire GI tract non-invasively. Once the tiny capsule is swallowed, it captures sequence of images as it is propelled down the GI tract. A single video can last up to 8 hours producing between 30,000 to 100,000 images. Automating the detection of frames containing specific lesion in CE video would relieve gastroenterologists of the arduous task of reviewing the entire video before making diagnosis. Convolutional Neural Network (CNN) based models have been very successful in various image classification tasks. However, they suffer excessive parameters, are sample inefficient and rely on very large amount of training data. Deploying a CNN classifier for lesion detection task will require time-to-time fine-tuning to generalize to any unforeseen category. In this paper, we propose a meta-learning framework followed by a few-shot lesion recognition in CE video. Meta-learning framework is designed to establish similarity or dissimilarity between concepts while few-shot learning (FSL) aims to identify new concepts from only a small number of examples. We train a feature extractor to learn a representation for different small bowel lesions using meta-learning. At the testing stage, the category of an unseen sample is predicted from only a few support examples, thereby allowing the model to generalize to a new category that has never been seen before. We demonstrated the efficacy of this method on real patient CE images.We conducted experiments to evaluate the impact of the number of support samples and compared performance across multiple CNN networks. Our experiment showed that this approach performs competitively with baseline models and is effective in few-shot lesion recognition in CE images.

- Adewole, Sodiq, Philip Fernandez, James Jablonski, Michelle Yeghyayan, Michael Porter, Andrew Copland, Sana Syed, and Donald Brown. "Unsupervised Temporal Segmentation of Long Capsule Endoscopy Videos." This work has been submitted to the 2021 ***IEEE***

International Conference on Bioinformatics and Biomedicine (BIBM) and is ***under review***. The technique proposed in this work is covered in chapter 3. The work covers computational efficient algorithm for temporal boundary detection and segmentation of long CE videos. We investigated feature extraction techniques that captures pathological information of the sequence of frames and the most effective method to project the high dimensional feature extracted from the frames by the CNN network to a lower dimensional 1-D embedding space.
Abstract Physicians use Capsule Endoscopy (CE) as a non-invasive and non-surgical procedure to examine the entire gastrointestinal (GI) tract for diseases and abnormalities. A single CE examination could last between 8 to 11 hours generating up to 80,000 frames which is compiled as a video. Physicians have to review and analyze the entire video to identify abnormalities or diseases before making diagnosis. This review task can be very tedious, time consuming and prone to error. While only as little as a single frame may capture useful content that is relevant to the physicians' final diagnosis, frames covering the small bowel region alone could be as much as 50,000. To minimize physicians' review time and effort, this paper proposes a novel unsupervised and computationally efficient temporal segmentation method to automatically partition long CE videos into a homogeneous and identifiable video segments. However, the search for temporal boundaries in a long video using high dimensional frame-feature matrix is computationally prohibitive and impracticable for real clinical application. Therefore, leveraging both spatial and temporal information in the video, we first extracted high level frame features using a pretrained CNN model and then projected the high-dimensional frame-feature matrix to lower 1-dimensional embedding. Using this 1-dimensional sequence embedding, we applied the Pruned Exact Linear Time (PELT) algorithm to searched for temporal boundaries that indicates the transition points from normal to abnormal frames and vice-versa. The key novelty of this work is in three (3) folds - first, the automated detection of temporal boundaries in long CE video has not been previously considered. Secondly, the reduction in the computational cost of the temporal boundary detection search by using a lower dimensional frame feature embedding; and lastly, the entire temporal segmentation of the CE videos requiring no supervision from medical expert is a new concept. The output of our model can be easily integrated into any CE video summarization model where physicians only need to review a selected sample frame from each video segment. We experimented with multiple real patients' CE videos and our result showed PCA was superior in capturing the transition between pair of normal and abnormal frames in the video. We also bench-marked with expert provided label, and our system achieved an AUC of 66% on multiple test videos.

- Adewole, Sodiq, Philip Fernandez, James Jablonski, Michelle Yeghyayan, Michael Porter, Andrew Copland, Sana Syed, and Donald Brown. "Weakly Supervised Temporal Segmentation of Long Capsule Endoscopy Videos Using Graph Convolutional Neural Network". This work covers the work presented in chapter 4 and will be submitted to the 2021 ***IEEE***

International Symposium on Biomedical Imaging (ISBI). In this paper, we employed Graph Convolutional Neural Network (GCNN) model for unsupervised temporal segmentation. We represented the whole CE video as a graph while the frames in the video are the nodes in the graph. GCNN model gives more flexibility to capture more complex as well as simple relationship in the edge weights of the graph. As against hard temporal dependence assumption, the nodes in the graph is connected to every other node and this connection is weighted based on a defined similarity metric. Essentially, the GCNN is a binary classifier that learns to embed normal and abnormal frames and discriminate between the two classes. This model is more powerful than any CNN model as the nodes receives messages from other nodes in the network and is able to map the embedding for similar nodes to the closer embedding space than a CNN network that assumes independence of the frames in the video.

- Adewole, Sodiq, Philip Fernandez, James Jablonski, Michelle Yeghyayan, Michael Porter, Andrew Copland, Sana Syed, and Donald Brown. "Graph Convolution Neural Network For Weakly Supervised Abnormality Localization In Long Capsule Endoscopy Videos". This work has been submitted to ***IEEE International Conference on Big Data - 2021*** and currently ***under review***. The work covers what is presented in chapter 5.

  Abstract Temporal abnormality localization in long Wireless Capsule Endoscopy (WCE) videos is an important problem. The cost of obtaining frame level label for long WCE videos is prohibitive. In this paper, we propose an end-to-end temporal abnormality localization for long WCE videos using only weak video level labels. Physicians use Capsule Endoscopy (CE) as a non-surgical and non-invasive method to examine the entire digestive tract in order to diagnose diseases or abnormalities. While CE has revolutionized traditional endoscopy procedures, a single CE examination could last up to 8 hours generating as much as 100,000 frames. Physicians must review the entire video, frame-by-frame, in order to identify the frames capturing relevant lesion or abnormality. This, sometimes could be as few as just a single frame. Given this very high level of redundancy, analysing long CE videos can be very tedious, time consuming and also error prone. This paper presents a novel multi-step method for an end-to-end localization of target frames capturing abnormalities of interest in the long video using only weak video labels. First we developed an automatic temporal segmentation using change point detection technique to temporally segment the video into uniform, homogeneous and identifiable segments. Then we employed Graph Convolutional Neural Network (GCNN) to learn a representation of each video segment. Using weak video segment labels, we trained our GCNN model to recognize each video segment as abnormal if it contains at least a single abnormal frame. Finally, leveraging the parameters of the trained GCNN model, we replaced the final layer of the network with a temporal pool layer to localize the relevant abnormal frames within each abnormal video segment. We experimented with multiple real patients' endoscopy videos and achieved an accuracy of 89.9% on the graph

classification task and a specificity of 97.5% on the abnormal frames localization task.

Chapter 2

Literature Review

In this chapter, we review prior works on VCE video analysis viz-a-viz the proposed techniques, limitations as well as general challenges in developing automated system for VCE video analysis. Thereafter, we focus on techniques proposed across other domains on video shot boundary detection, video summarization, and video action localization, providing a good context for the contributions made in this dissertation.

2.1 Video Capsule Endoscopy Video Analysis

First we consider how developing automated system for analysis of long VCE videos has evolved, focusing on the problem formulation by different researchers, the techniques and the limitations of the formulations in building a realistic system. The discussion in this chapter will serve as the motivation for subsequent chapters in this dissertation.

For more than two (2) decades VCE has become a routine, first line investigational tool for many small bowel pathologies [28] and the task of reviewing the videos by the physician after the endoscopy process is very tedious, time consuming and also error prone. Leveraging the distributional structure of the frames in the video, where there are far more normal frames than frames with abnormality, researchers have framed automatic analysis of capsule endoscopy video in three ways: this include lesion detection methods [29, 30], outlier or abnormality detection using disease-agnostic models [13, 15, 17] and key frame selection or video summarization [3, 4, 7, 11, 14, 16, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44]. Many of these proposed methods aim to automatically detect different lesions or abnormalities while others capture high-level information by only differentiating abnormal from normal frames.

Videos are spatial-temporal data with information captured in both spatial (image), temporal (sequence) and topological structure of the frames in the video. This means that in formulating analysis of VCE video data, capturing spatial elements as well as the temporal and topological relationship between the frames will benefit any model or solution. However, many prior works [4, 11, 32, 38] on VCE video analysis considered each frame in the video as an independent and identically distributed data point with no relationship with the other frames in the video. This approach over-simplifies the problem of video analysis and diminishes the applicability of the models in real-world clinical setting. In the next few sections, we will consider some of the formulations and methods that researchers have proposed in solving the problem of automated review and analysis of VCE video data. First we review, in more detail, these three (3) main categories of prior research on VCE video analysis by considering the problem formulation as well as their limitations. This will form the basis for the remainder of this dissertation. In the next two subsections, we will discuss the framework of independent disease recognition and anomaly detection. We will discuss works on video summarization in section 2.3.

2.1.1 Single or Multiple Lesion Detection

Many techniques have been proposed to detect specific lesion such as bleeding in [4], polyp in [29, 38, 39, 40], ulcer [44], and angioectasia [41, 42, 43]. In [39] Mamonov et al. proposed a model for colorectal polyp detection based on a binary classification using geometric analysis and texture content of the frames. Their model achieve 47% sensitivity and 90% specificity. Similarly, Hwang et al. [29] proposed a polyp detection model by first segmenting the affected region using Gabor texture features before applying K-means clustering algorithm. The resulting geometric information is then used to identify frames containing polyp. Yixuan et al. [40] proposed a bag of feature (BoF) technique using integration of multiple features such as texture features, Scale-Invariant Feature Transform (SIFT), Complete Local Binary Pattern (CLBP) with visual words to automatically detect polyp in CE video frames. While SIFT remains the baseline feature for traditional image analysis, CNN based models achieve superior performance in complex geometric and lighting conditions which is typical of CE frames. Akiyoshi et al. [41] used a Single Shot Multibox Detector (SSMD) to automatically detect frames with angioectasia in CE images. Similar task was attempted in [43] using saliency-based unsupervised method. In [42], the authors combined deep learning and handcrafted features by concatenating the extracted features for a polyp detection problem.

First, one of the limitations of this formulation is that, for each class of lesion or abnormality, large labelled examples are needed to train the model in order to generalize properly. Secondly, Convolutional Neural Network (CNN) based models are, currently, the state-of-the-art models for object detection and image recognition [45, 46] task across multiple domains. However, they are

notoriously sample inefficient and, given the high class imbalance in typical VCE video frames, training CNN models will require large number of labelled examples for each class of disease. While large volume of frames can be generated in a single CE video, the distribution is always skewed, with far more normal frames than abnormal ones. This leads to very high redundancy rate in the normal samples limiting the sample size for each disease category that one can collect from a single patient's video. When we couple this fact with the cost of obtaining expert label for each frame across multiple patients, given that they still have to sift through all the redundant normal frames in order to identify the few abnormal ones. This can be prohibitively expensive. Therefore, any technique that can help minimize this cost by filtering out the normal frames leaving the few abnormal ones for annotation would save significant amount of time and effort. One attempt to alleviate this problem was proposed in [11]. Here the authors proposed a meta learning framework using Siamese Neural Network (SNN) that learns to project each abnormality into a embedding region such that the distant embedding features for same abnormality class are closer than other abnormality classes.

Lastly, the limitation of models that are trained to only detect specific diseases or lesions also include their inability to generalize to new unseen class of disease without having to retrain the model to capture the new category. One solution to this challenge is formulating the overall detection as an outlier or anomaly detection problem where any frame with disease are separated as outliers from normal frames without any disease. This will help the physician minimize the time that could have been spent reviewing the entire video and they only have to focus on the abnormal frames but the abnormal / normal categorization lacks granularity that may be useful to most physician's use cases.

2.1.2 Abnormal / Outlier Frame Detection

Anomaly detection involves identifying a sample data point that shows significant statistical difference from other data points [18]. In VCE video analysis, some prior research efforts considered the task an abnormal / outlier frames detection problem, classifying any frame that differs from the normal class as abnormal. In the context of CE videos, an abnormal/outlying frame may not differ significantly when compared with other frames in the video. This is mainly because some abnormality such as bleeding gradually decrease in intensity as you move away from the source. This framework demonstrate promising real-world clinical applicability as it allows the physician focus only of the selected abnormal frames. However, most prior formulations of this problem have neglected the temporal and topological relationship between the frames in the video. This is the motivation for the work done in chapter 6 where we combine temporal segmentation with binary classification in a weakly supervised manner. The temporal segmentation allow us to further capture the temporal

dependence between the frames thereby minimizing the redundancy even within the abnormal frame categories. Some of the proposed techniques in this area include, [34], where Sivakumar et al. proposed using Bag-of-Visual Words (BOVW) technique to extract feature and then applying Naive Bayes (NB) classification model to detect abnormal frames with bleeding. Similarly, [3] Miaou et al. proposed a four-stage classification model based on low-level Hue-Saturation-Intensity (HSI) features followed by fuzzy-C means clustering analysis to separate images carrying different lesions. The final stage is a neural network model that discriminate normal from abnormal frames. In [36] the authors applied similar multi-stage technique to extract quality frames by removing over-/under-expose images as well as images with significant non-tissue areas. Using color histogram of images, [37] employed fuzzy neural model which combines fuzzy systems and artificial neural networks to detect lesions in CE images. In [3] Miaou et al. propose a four-stage classification model based on low-level Hue-Saturation-Intensity (HSI) features followed by fuzzy-c means clustering analysis to separate images carrying different abnormalities in a step-wise manner. The final stage is a neural network model that discriminate normal from abnormal frames. In [36] Mewes et al. applied similar multi-stage technique to extract quality frames by removing over-/under-expose images as well as images with significant non-tissue areas.

Using color histogram for representation, [37] employed fuzzy neural model which combines fuzzy systems and artificial neural networks to detect lesions in CE images. In [4], the authors proposed to detect bleeding regions in frames by computing statistical features of the first order histogram probability of the three color channels (RGB) in the images before passing the computed features to a neural network to discriminate bleeding from non-bleeding frames. In [15], the authors applied semi-supervised CNN-based model to detect frames with abnormality. The model was trained only on normal images and subsequently applied to flag outliers based on a determined parameter threshold. Outlier frames such as bleeding may span multiple frames capturing the same content, outlier detection models do little to minimize this redundancy. Limitation of outlier detection approach is ineffectiveness in minimizing review time spent by the gastroenterologist in reviewing the video since multiple abnormal frames with same information are captured as outliers.

The chapter 4 of this work is motivated by the limitation of outlier detection problem where prior works fail to capture the temporal correlation between the frames. Considering the relationship between the frames will allow further reduction in the redundancy among the abnormal categories.

2.1.3 VCE Video Summarization

By mainly leveraging the distribution of the frames, works that have been proposed under video summarization [16,19,20,21,22,23,24,25,26,27] or key frame extraction tries to reduce any redundancies

in the video. Although they do not necessarily provide granular annotation of the specific lesion in the frames, they significantly cut down the time it takes to review a complete VCE video while still guaranteeing same information coverage that would have been achieved had the complete video been reviewed. In addition, it helps the gastroenterologist focus only on frames with few informative frames capturing disease or lesions.

The primary goal of VCE is to detect mucosal abnormalities such as blood, ulcer, polyp etc in the gastrointestinal tract. With close to 100,000 frames, as few as a single frame of the total video could be relevant for the physician diagnosis [34]. Therefore developing techniques to automatically reduce the number frames to only relevant ones would have very significant clinical implication. One technique used to minimize review time spent by the gastroenterologist on CE video is to extract only informative/key frames from the entire VCE video using both low [19, 20, 21] and high [16] level features.

Prior proposed techniques on VCE video summarization can be broadly categorised based on the level of supervision. In [16] Chen et al. applied the Siamese Neural Network (SNN) framework where the CNN model learns the features of each frame based on a distance metric from the neighbouring frames. Using a similarity matrix, every pair of frames in the sequence were assigned binary labels by medical expert. Scaling this task to many more videos can be prohibitively expensive and tedious. Training the SNN based on contrastive loss function, the extracted features is passed to a Support Vector Machines (SVM) classifier to identify video temporal segments. While a Siamese Neural Network feature extractor is a very laudable approach, using contrastive loss function requires getting a label for each pair of frames in the video as mentioned above. This can be really tedious, time consuming and also very expensive.

Ismail et al [27] proposed an unsupervised endoscopy video summarization approach where the collection of video frames is first partitioned into homogeneous categories based on their visual and temporal features. A clustering approach was guided using the frames' temporal information before generating a possible membership score for each frame in the subset. Other works apply clustering on the video frames before selecting representative frames from each cluster [25, 26]. Other works such as [22, 24], performed key frames extraction from VCE video using different low level features. Main shortcomings of the unsupervised summarization technique based on clustering include: 1) Having to manually specify the number of clusters, in which case, the temporal information in the sequence of frames is not taken into consideration; 2) In the cases where temporal information is considered, having a specific number of frames in each video segment will result into non-homogeneous video segment with huge impact on the extent to which the model can reduce the redundancy in the data. On the other hand, annotating VCE videos, for a supervised model, is also a very difficult task mainly due to the volume of frames generated in each VCE study.

2.2 Shot Boundary Detection and Temporal Segmentation

Serving as the motivation for the work presented in chapter 3, this section discusses related works on automated temporal shot boundary detection in long videos.

2.2.1 Shot Boundary Detection in Long Videos

Partitioning a video sequence into shots is the first step toward video-content analysis and content-based video browsing and retrieval. Shot boundary detection (SBD) is the process of automatically detecting the boundaries between shots in long videos [47]. These boundaries are used to temporally segment the video into short homogeneous segments. A video shot is defined as a series of inter-related consecutive frames taken contiguously by a single camera and representing a continuous action in time and space [48]. The problem of temporal segmentation in videos structured data is not new and has attracted much attention since video data became available digitally. This area has been a core research area for more than two (2) decades and also popular among researchers working on video analysis. SBD is an essential pre-processing step to almost all video analysis, indexing, summarization, search, and other content-based operations.

Various methods of automatic shot boundary detection have been proposed and claimed to perform reliably [49, 50]. In [51], the authors proposed model using color histogram for boundary detection. Their method was able to differentiate abrupt temporal boundaries by the analysis of color histogram differences and smooth temporal boundaries by temporal color variation. The aim of their method is to provide a simple and fast algorithm that is able to work in real-time with reasonable high performances in a video indexing tool. Their results showed reduced computational cost as well as an overall precision of 84.7% and a recall of 80.6%.

Similarly, in [52], the authors proposed a model-based shot boundary detection technique using frame transition parameters. They formulated a frame estimation scheme using the previous and the next frames. And instead of using properties of frames itself, frame transition parameters and frame estimation errors based on global and local features were used for boundary detection and classification. The Local features include scatter matrix of edge strength and motion matrix before finally classifying the frames as no change, abrupt change, or gradual change frames using a multi-layer perceptron network.

2.2.2 Temporal Segmentation of Long Videos

Temporal segmentation is usually the first step when trying to automate analysis of long videos. The goal is to divide the video stream into a set of meaningful segments known as *shots*. In conventional videos, *shots* transition are two types: *abrupt* or *gradual*. While abrupt transitions are easier to detect due to higher gradient between the two boundary frames in the sequence, gradual transitions are much more difficult to detect. Different models have been proposed for shot transition detection in conventional videos, however, they do not work well for CE videos [33]. While little to know attention has been devoted to shot boundary detection and temporal segmentation on VCE videos, one prior work used digestive peristalsis and image analysis techniques for shot boundary and organ boundary detection. In [30], Vu et al. proposed a coherent three-stage procedure to detect intestinal contractions. They utilized changes in intestinal edge structure of the intestinal folds for contraction assessment. The output is contraction-based shots. Another limitation that has received less attention is the computational cost of boundary detection on high dimensional features.

Another attempt at shot boundary detection scheme based on digestive organs was proposed by Mackiewicz et al. in [32]. The authors utilized three dimension LBP operator, color histogram, and motion vector to classify every 10th image of the video. The final classification result was assessed using a 4-state hidden Markov model for topographical segmentation. In [31], two color vectors that were created with hue and saturation components of HSI model were used to represent the entire video. Spectrum analysis was applied to detect sudden changes in the peristalsis pattern. Chen et al. assumed that each organ has a different peristalsis pattern and hence, any change in the pattern may suggest an event in which a gastroenterologist may be interested. Energy and High Frequency Content (HFC) functions are used to identify such change while two other specialized features aim to enhance the detection of duodenum and cecum. Zhao et al. [17] proposed a temporal segmentation approach based on adaptive non-parametric key-point detection model using multi-feature extraction and fusion. The aim of their work was not only to detect key abnormal frames using pairwise distance, but also to augment gastroenterologist's performance by minimizing the miss-rate and thus, improving detection accuracy.

2.3 Video Summarization

Video Summarization (VS) is critical to video semantic analysis, browsing, and retrieval. VS involves effectively extracting important information from video data while removing redundant ones. It also refers to the process of eliminating redundant information in a video by selecting frames that

captures information that is considered most representative of the entire video. In [53], the authors applied singular value decomposition (SVD) on the feature-matrix first to derive the refined feature space and then cluster visually similar frames. Using the degree of visual changes computed based on the amount of visual content contained in each frame cluster, they found the most static frame cluster, define it as the content unit, and use the context value computed from it as the threshold to cluster the rest of the frames. They generated optimal set of key-frames and a summarized motion video with the user specified time length from the video. Similar techniques was adopted on VCE video in [20]. Another work [54] used visual co-occurrence by exploiting the visual co-occurrence across multiple videos. In [55], the authors propose technique base on user attention model by estimating the attentions viewers may pay to video contents

2.4 Video Abnormality Localization

Despite the high cost of frame-level annotation for VCE videos, little to no attention has been given to weakly supervised model for temporal abnormality localization. In [56], the authors proposed as weakly supervised lesion detection in a 2D VCE frame. To the best of our knowledge, this is the first work to considered temporal abnormality localization on VCE video data. We present this idea in chapter 5 where our model generate proposal score sequence for each frame in the video. This score is considered the degree of representativeness of the frame for the video. The score sequence, similar to Class Activation Mapping presented in [57] for image data with only one dimension. In the next few paragraphs, we will consider some of the prior works on weak supervision and temporal action localization.

Recent work on temporal action localization in video analysis include [58, 59, 60, 60, 61, 62, 63, 64, 65, 66]. Reduce cost of storage and ease of video data collection has recently been the main motivation for recent research interest in this area.

For example, in [67], Shou et al. proposed a weakly supervised temporal action localization (TAL) method focus on generating good Class Activation Sequence (CAS) over time and conduct thresholding on CAS to localize actions. Similarly, [68, 69] proposed Temporal Segment Network which employs two-stream network to model the long-range temporal structure in video by combining sparse temporal sampling strategy and video-level supervision. However, using softmax over action proposal may not be effective when applied to CE video where there are multiple instances of a particular diseases separated by instances of normal frames.

Most of these works on temporal video action localization proposed techniques requiring full supervision with frame level annotation. Supervised approaches construct predictive models where each training

example has a label indicating ground truth. While many supervised learning models have achieved great success, video annotation on a frame-by-frame basis, particularly VCE video with large volume of frames and high redundancy, remains a challenge. Expertise required to annotate large number of VCE videos is not readily available and may be too expensive, thereby making frame-level annotation impracticable in real world environment. In addition, training a model on VCE video dataset requires training over a wide range of patients' videos so as to minimize patient bias. Lastly, strong frame-level labels for VCE video are not common as even gastroenterologist sometimes have doubt as to the true identity of lesion in a frame. The leads to noisy labels and consequently, high cost of data-labeling.

Weakly supervised models require global object level annotation that can be localized to the components of the object. Weakly supervised learning techniques [70] are in three (3) broad categories: *incomplete supervision*, where only a subset of training data is given with labels; *inexact supervision*, where the training data are given with only coarse-grained labels; and *inaccurate supervision*, where the given labels are not always ground truth. In chapter 5, we limit our task to the problem of of inexact supervision with video-level labels to localize frames with highest class-activation in our VCE video.

Temporal action localization using weakly supervised methods have recently started gaining attention of the computer vision research community [60, 63, 64, 65, 67, 71]. TAL [67, 72, 73] is an extension of weakly supervised segmentation on video structured data. Some of these works focus on task such as semantic segmentation [74, 75], video captioning [76], and visual relation detection [77].

Methods such as structured segment network [59], contextual relation learning [78], multi-stage CNN [58], temporal association of frame-level action detection [61], and techniques using recurrent neural networks [62, 79]. Action proposals [66, 80, 81] in action localization is an extension of object proposal for object detection.

In [82], the authors proposed a weakly supervised framework where they randomly hide patches in the training image, thereby forcing the network to seek other relevant parts when the most discriminative part is hidden. Their proposed model that do not only localize the most discriminative parts of an object, rather than all relevant parts. Several other works have proposed to solve the problem of temporal action localization in long untrimmed videos [83].

As a counterpart to weakly supervised temporal object localization, recent application of weakly supervised CNN based object localization was proposed in [71, 84] with promising results. Given a training video, in TAL methods, several segments are randomly sampled and are then fed into a network together to yield a video-level class prediction. During testing, the trained network is slided over time to produce the classification score sequence of being each action over time. The score

sequence is similar to the Class Activation Map [85] in one dimension. This is referred to as Class Activation Sequence (CAS). A simple thresholding is thereafter applied on the CAS to localize each action instance in terms of the start and end time. Models for object detection has been significantly improved via combining Multiple Instance Learning (MIL) [86] and deep networks [73,87,88,89,90] with most techniques are built upon Fast-RCNN [91]. These methods first generated candidate proposals beforehand; then they employed deep networks to classify each proposal and the scores from all proposals were fused together to obtain one label prediction for the whole image to be compared with the image-level label. In, [92], the authors proposed a recurrent neural networks to model relationships between time segments in a video. However, relationships between time segments that are temporally distant, or that belong to different videos cannot be modeled with this approach. Conversely, GCNN-based model is not restricted by temporal proximity when modeling similarity and dissimilarity relationships between time segments. Other works employed boundary regression model to learn to predict more accurate boundaries [64,65].

Chapter 3

Unsupervised Shot Boundary Detection and Segmentation for Long Capsule Endoscopy Videos

3.1 Introduction

In this chapter, we developed an unsupervised, domain-agnostic shot boundary detection and temporal segmentation for long VCE videos. Analysis of video structured data have received significant attention from the research community [47, 93] including the medical domain [94]. With the recent exponential rate at which video data is generated, developing models to automatically segment long videos has never been more important.

Analyses of videos encompasses tasks such as summarization [95, 96, 97], learning representation [98, 99], anomalous event detection [100], video classification [46, 101] and video retrieval [102]. Each of these tasks have multiple applications across different domains such as action recognition [99] and analysis of the content of the video [103]. Approaches to video analysis have mainly been applied on temporal video clips as most researchers benchmark their model on open dataset such as TRECVID videos [47, 104], VSUMM [95], and the open video project [105]. These videos have already been manually segmented using fixed frame count. Other works on long videos proposed to divide the video into N clips with equal number of frames / duration [69]. In analysis of short video clips, researchers only have to deal with one activity or a single event [63, 106]. However, in long videos, the problem gets much more complicated [107] but more realistic. Examples include work on streaming video [108] from surveillance and offline long video sources such as VCE [16, 25]. Long

videos have peculiar characteristics which include multiple scene changes and multiple objects.

Two main approaches for analysing long videos are the offline [109] and online [108, 110] methods. Typically, Offline analysis will require temporal segmentation into independent homogeneous temporal unit by detecting temporal boundaries in the video [47, 48, 49, 50, 51, 52]. Each member frame within a segment are correlated and have visual similarity while each segment will exhibit independence characteristic. Temporal segmentation of long videos is a very challenging problem mainly due to the high dimensionality of the frames. The problem can be further aggravated, as in the case of VCE video, when visual change in scene may not mean pathological event. Ground truth labels for temporal boundaries is very subjective and depends on type of transition between the scenes, leading to noise and subjective labels. In the case of VCE videos, collecting annotation frame-by-frame requires an expert to determine both pathological and visual change points in the sequence. Such medical expertise are hardly available or expensive to get.

In this chapter, we developed an unsupervised algorithm for temporal segmentation of long VCE videos. We leveraged prior works in time series domain for change point detection (CPD) in a sequence of observation [111, 112, 113]. CPD methods have been successfully applied on time-series data in one dimension with linear computational time [111]. However, video frame features are usually in higher dimensions, therefore, exponentially increasing the computational cost. In our model, we applied the Prune Exact Linear Time (PELT) algorithm [111] to detect the temporal change point with linear computational time without requiring any label. The novelty of this work is that no any form of annotation is required for the temporal segmentation of the videos, thereby saving experts significant amount of time. To the best of our knowledge, this is the first work to approach VCE video analysis using concept from CPD model to exploit the temporal information in the sequence of frames. This work has the potential to minimize redundant frames in any VCE video and saving expert time when annotating VCE video data.

3.2 Related Work

Detection of Change Points on sequence data has been considered in solving sequence segmentation problems across various applications such as medical condition monitoring [114, 115, 116, 117], climate change detection [118, 118, 119], audio activity segmentation and boundary recognition for silence in speech [120, 121], speaker segmentation, scene change detection, human activity analysis [122, 123, 124] as well as medical imaging [111]. Other areas where detection and location of distributional changes in data arises include online sequential time series analysis [125, 126, 127, 128]. These tasks involved partitioning a sequence into several homogeneous segments.

Some prior works proposed a probabilistic sequence models such as Hidden Markov Models (HMM) [129] or the discriminative counterpart such as Conditional Random Fields [130] for this task. These probabilistic models require a good knowledge of the transition structure between the segments and require careful pre-training to yield a competitive performance. This may not be practicable for online applications where data are acquired online.

Parametric approaches model the distribution before and after the change based on maximum likelihood framework [112] while non-parametric methods [131] have been mostly limited to univariate data. Other change point detection techniques in time series 1-dimensional data leveraged weakly-supervised learning methods where the number of change points is known and provided to the model before hand. The model, therefore, tries to optimize the locations of these change points within the sequence.

Kernel-based methods use maximum kernel fisher discriminant ratio as a measure of homogeneity between segments [113] and have been applied on multivariate, high-dimensional data. The approach used a regularized kernel-based test statistic to determine if: 1) there is a change point in the data and 2) the location/instant of the change point, if there is one [113,132]. In [69] the authors presented a sampling method to trim long videos using temporal change points. Their model was based on HOG features for each frame and then they calculated the HOG feature difference between adjacent frames. The absolute value of this difference was used to measure the change of visual content based on a pre-determined threshold [69]. In summarizing VCE, [16] proposed to find transition boundaries in the video using pair-wise similarity between the sequence of frames. A threshold parameter is used to determine the boundaries based on the similarity score between frame pairs.

Other algorithms such as Bayesian change point detection [133], two-sample homogeneity test based on Wilcoxon rank statistic [134]; and kernel based methods [113, 135, 136]. These methods can achieve good results for moderately multidimensional data or in specific situations (e.g., if the data lie on a low-dimensional manifold). They lack robustness when moving to larger dimensions. Particularly, kernel-based methods are not robust with respect to the presence of contaminating noise and to the fact that the changes in the detected points may only affect a subset of the components of the high-dimensional data. Below we discuss some of the algorithms in detail.

Algorithms such as binary change point [137] and dynamic programming [137], can identify locations where there are significant changes in the distribution of a sequence of data through recursive search. However, in order to use these techniques, one has to already know the number of change point instances in the sequence. The algorithms only try to recursively find the location of these points using maximum likelihood estimation.

Binary Segmentation (BS) search is the most established in literature. BS is an approximate

method with an efficient computational cost of $\mathcal{O}(n\log n)$, where n is the number of data points. The algorithm works by iteratively applying a single change point method to the entire sequence to determine if a split exists or not. If a split is detected, then the sequence splits into two sub-sequences. The same process is then applied to both sub-sequences [138].

Dynamic Programming (DP) search method is an exact search method, with a computational cost of $\mathcal{O}(Q_n^2)$, where Q is the max number of change points and n is the number of data points [137]. DP can also be applied using different kernels such as the linear or Gaussian kernels.

Window-based Searching is an approximate search method. The window-based search computes the discrepancy between two adjacent windows that move along with signal y. When the two windows are highly dissimilar, a high discrepancy between the two values occurs, which is indicative of a change point. Upon generating a discrepancy curve, the algorithm locates optimal change point indices in the sequence [137].

Pruned Exact Linear Time (PELT) PELT is an unsupervised CPD technique where no prior knowledge of the number of change point is necessary. Rather the model finds the optimal location as well as count of the change points in the series based on a penalty parameter that can be set by the user.

3.2.1 Problem formulation

Shot boundary detection in unlabelled sample sequence of frames use change point analysis to 1) test whether there is a change in the distribution within the sample and 2) If a change occurs, estimating the change point instant after which the distribution of observation switches from one distribution to another different distribution.

Shot boundary detection problem involves testing hypothesis where for every time step, we consider the null hypothesis - H_0 - there is a change point and H_a - there is no change point. We use available data to determine whether to reject the null hypothesis.

Let $f_1, ..., f_T$ be a sequence of frames in a sample CE video V. The shot boundary detection analysis of a sample video consist of;

Step 1: $\boldsymbol{H}_0 \rightarrow P_{f_1} = ... = P_{f_k} = ... = P_{f_T}$

$\boldsymbol{H}_A \rightarrow \exists\, 1 < k^* < T : \quad P_{f_1} = ... = P_{f_k} \neq P_{f_{k^*+1}} = ... = P_{f_T}$

Step 2: Estimate k^* from the sample if $\boldsymbol{H}_A$ is true

Figure 3.1 shows the end-to-end shots boundary detection pipeline.

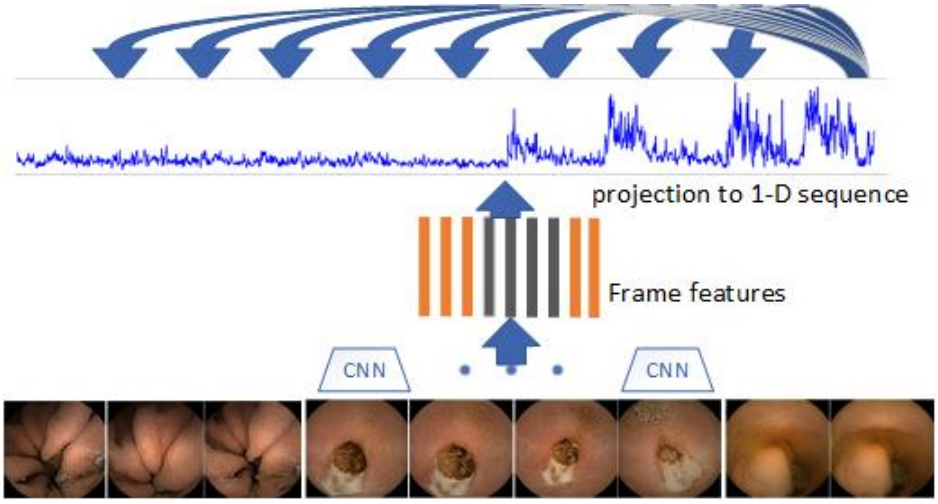


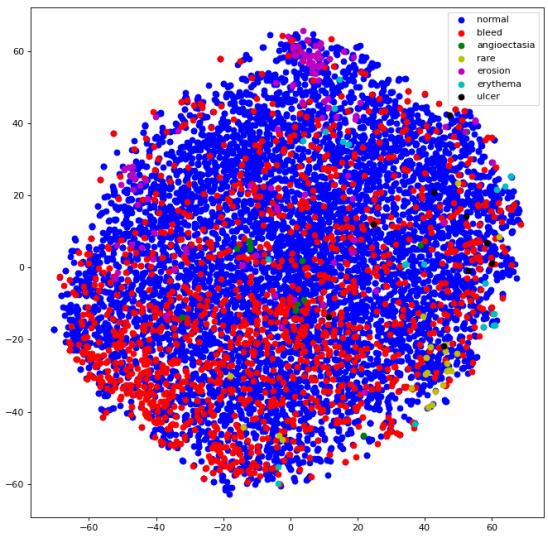
Figure 3.1: Proposed Unsupervised VCE Video Temporal Segmentation Pipeline

3.3 Methodology

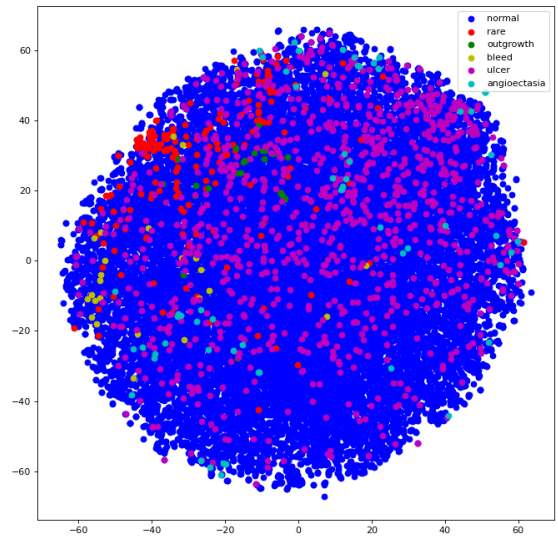
3.3.1 Feature Extraction

First we consider feature extraction to learn a representation for the frames in the video. There are different techniques for image-level feature extraction categorised as low and high-level techniques. Low level feature extraction techniques include Local Binary Pattern (LBP) [139], Scale Invariant Feature Transform (SIFT) [140], Speeded-Up Robust Features (SURF) [141] and Bag of Visual Words [40, 142]. convolutional Neural Networks have shown superior performance in high level feature extraction from 2-dimension image data. These can be done as a unsupervised feature representation learning using Convolutional Auto-Encoder (CAE) [143] as well as using supervised pre-trained models such as those trained on very large ImageNet data [144]. Convolutional Feature Extraction uses sequence of convolution, pooling, activation and batch-normalization to extract meaningful features from a 2-dimensional image. These extracted features are referred to as high-level features. We performed feature extraction on each frame by testing multiple CNN architectures [145, 146, 147, 148, 149, 150]. Figures 3.2 and 3.3 shows the visualization of the different architecture on a sample video indicating their representation capabilities. Based on the separation between the classes shown in the figure, we selected the architecture with the most distinction between the classes.

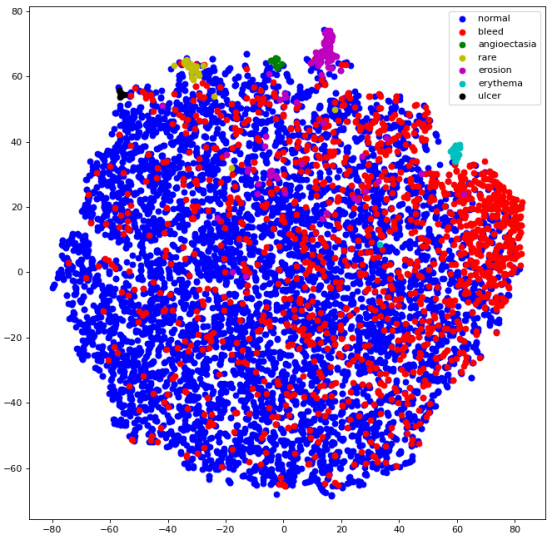
The figure 3.2 shows the 2-dimensional TSNE visualization of our feature extraction experiment using different CNN architecture pretrained on Large ImageNet data and then fine-tuned on VCE images. We selected the most representative architecture based the VCE dataset classes of the CE image data for this rest of this work.



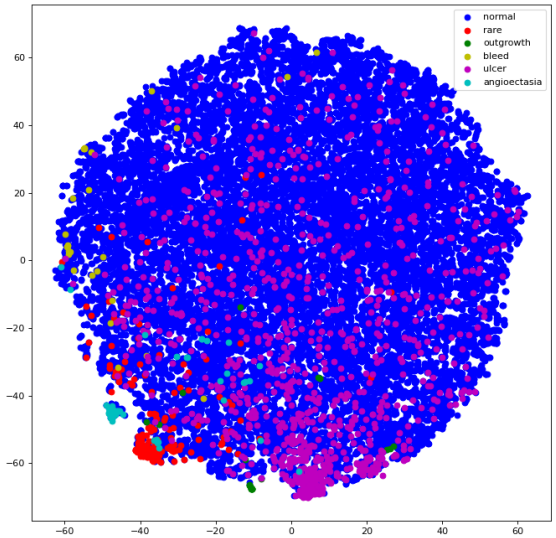
(a) Mobile-Net: Video 2



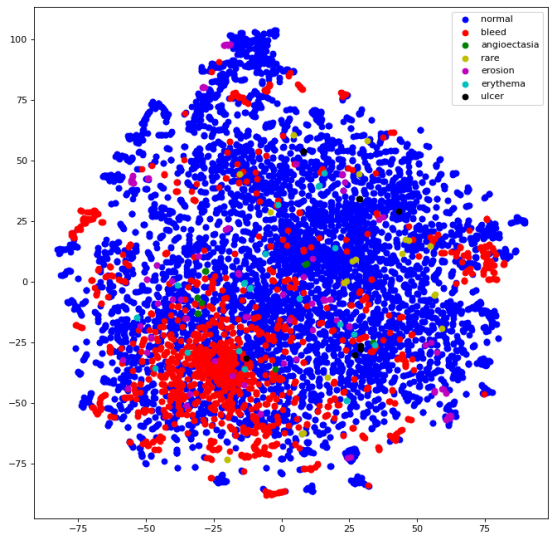
(b) Mobile-Net: Video 3



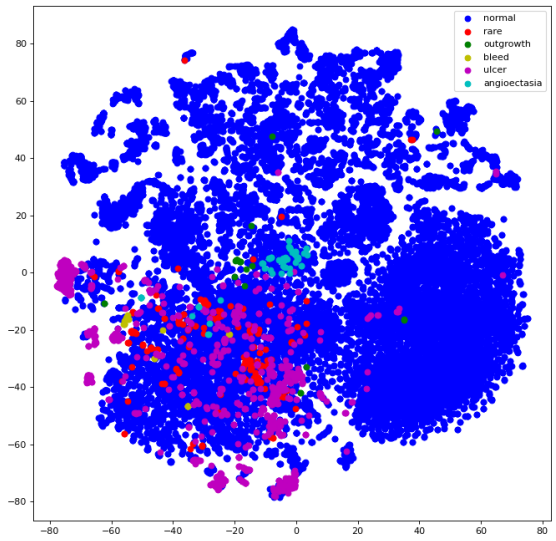
(c) ResNet-152: Video 2



(d) ResNet-152: video 3

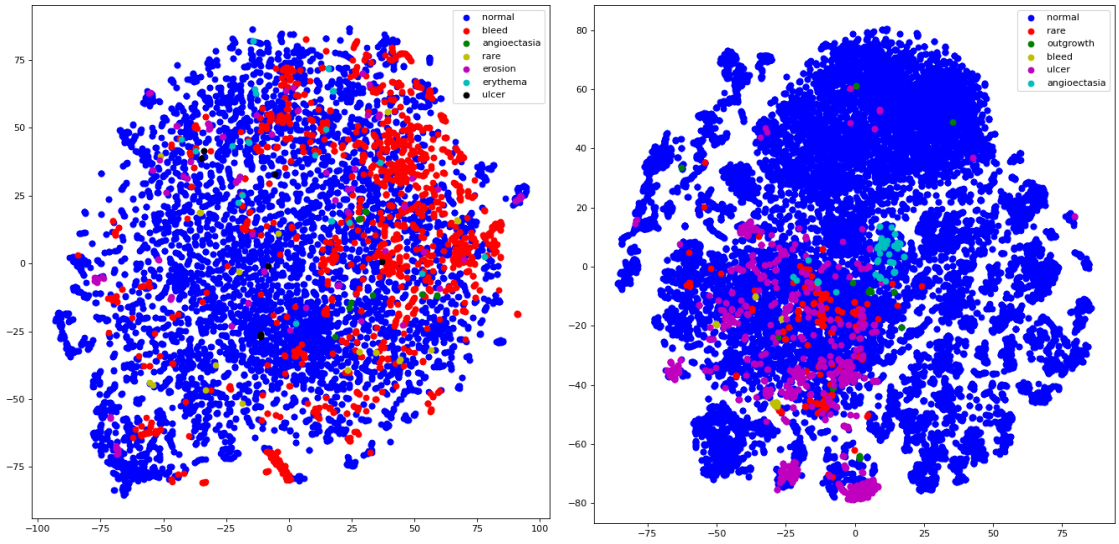


(e) VGG-19: Video 2



(f) VGG-19: video 3

Figure 3.2: 2-D plot of Video Features Using Different CNN Architectures



(a) Alex-Net: Video 2 (b) Alex-Net: Video 3

Figure 3.3: 2-D plot of Video Features Using Different CNN Architectures

3.3.2 Class Oversampling

VCE video is highly imbalance with far more normal frames than frames containing diseases or lesion. To train a feature extraction model, the disproportionately large normal frame samples, if not adjusted for, impacts the models ability to distinguish between normal and diseased frames in the embedding space. Figure 3.4 shows the distribution of five sample videos used in all our training including training the feature extractor. We applied oversampling technique by placing more sampling weight on the classes with fewer samples to ensure the model learns a good representation of each class. We applied the inverse proportion of each class in the entire dataset as a sampling weight. While there are other techniques such as the SMOTE, weighted loss and under sampling to account for imbalance data distribution, we opted for the oversampling due the ease of implementation and effectiveness on VCE structured data.

3.3.3 Feature Embedding

Once a good representation for the frames is learnt, in this section, we describe our approach for encoding the extracted features to a 1-dimensional embedding. We applied this technique to improve the computational efficiency of finding the temporal boundaries in the video sequence. We approached this by projecting the high dimensional frame feature vector to a lower 1-dimensional embedding space. As described in section 3.3.1, we extracted features for each frame in the video using the model with the most discriminative representation between classes of the frames. In

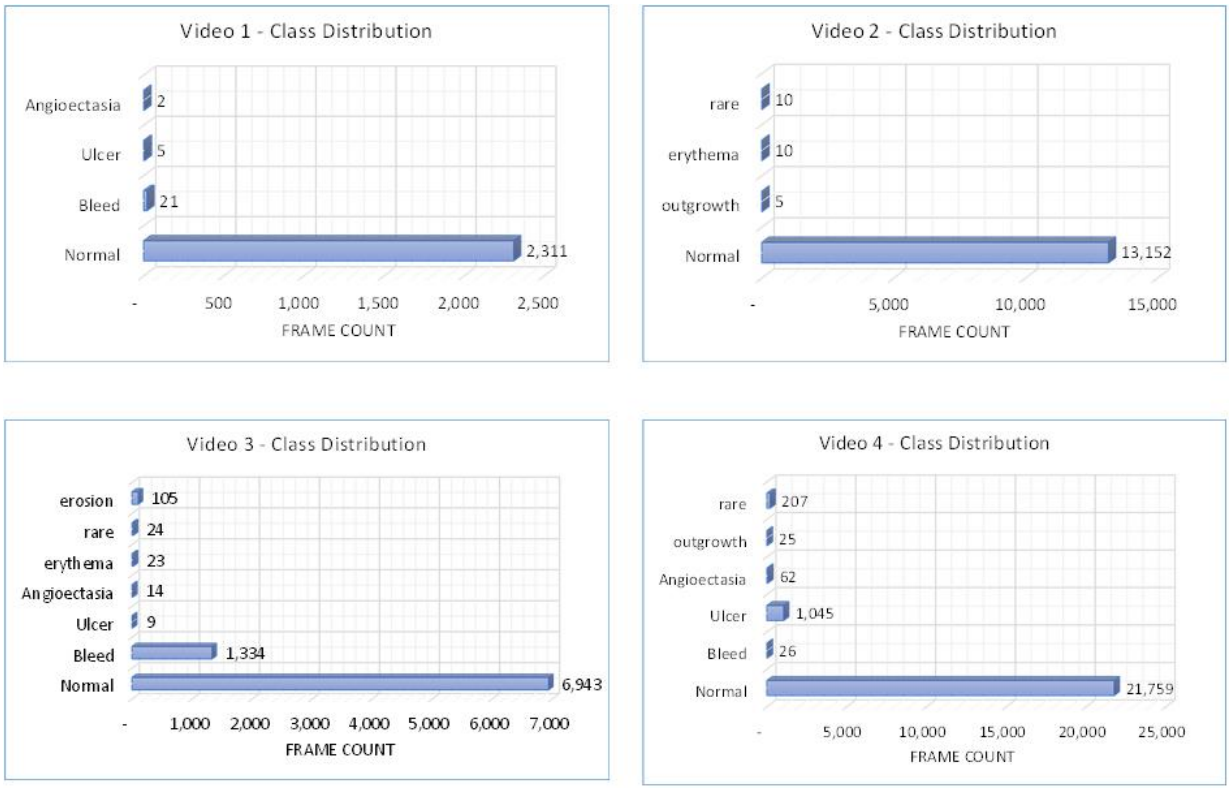


Figure 3.4: Frame Distribution for 4 Video Samples

determining the shot boundaries, first, we experimented with detecting change boundaries using this high dimensional feature vectors, however, our experiment showed that this was impractical and computationally prohibitive as it took several days to segment a single video using the Pruned Exact Linear Time (PELT) algorithm. In order to minimize computation cost of detecting boundaries between the sequence of frame features, we projected the high-dimensional frame features to a 1-dimensional manifold space. We compared multiple manifold learning frameworks such as PCA, Auto-encoder, TSNE, Kernel-PCA with different kernels and LSTM-encoder.

In figure 3.5, we show the visualization of the 1-dimensional plot for one (1) test video using different encoding method.

Principal Component Embedding (PCE)

The principal component of a feature matrix extracts the dominant patterns in the matrix in terms of a complementary set of score and loading plots [151]. PCA is used to decompose a multivariate dataset in a set of successive orthogonal components that captures maximum variance in the data. PCA is a linear dimensionality reduction technique that uses Singular Value Decomposition (SVD)

of the data to project it to a lower dimensional space. The input data is centered but not scaled for each feature before applying the SVD [152]. The computational efficiency and speed of PC method makes it a very popular option in most data analysis. However, the linearity assumption between the higher dimensional and lower dimensional space makes it less effective on most data that are structurally non-linear. Below we consider variants of the PCE algorithm by replacing the linear kernel with other kernels. We used PCA to reduce the dimensionality of the frame features to a single component (1-dim). See figure 3.5 for the visualization of a sample video projected on the 1-dimension that explains most variance using 4096-dimensional feature vector extracted from VGG-19.

Kernel Principal Component Embedding (KPCE)

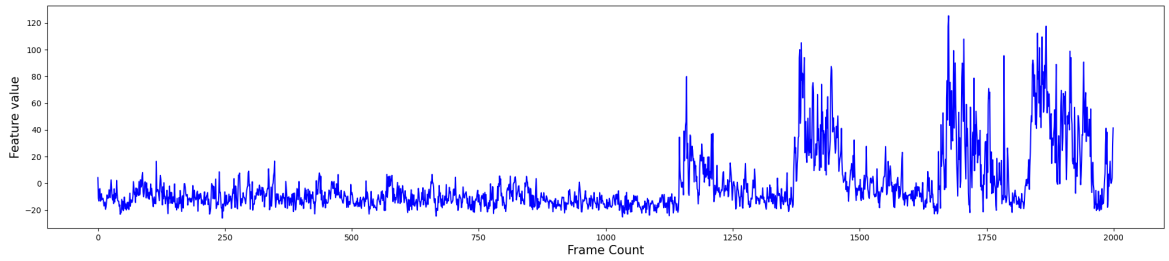
In order to capture the non-linearity between the high-dimensional feature vector and the lower-dimensional embedding space, we applied kernel principal component which achieves non-linear dimensionality reduction through the use of kernels. Kernels are measure of similarity. PCA uses a linear kernel $k(x, y) = X^T y$ to construct the eigen-decomposition of the covariance matrix of the data. Kernel PCA uses the kernel trick by mapping the data to a hyperplane with the original linear eigen-decomposition performed in a reproducing kernel hilbert space [153]. We experimented with three different kernels - gaussian and cosine kernels. Figure 3.5 shows the 1-d plot using different kernels.

Figures 3.5 shows the visualization of a sample video after projecting to a 1-dimensional manifold space.

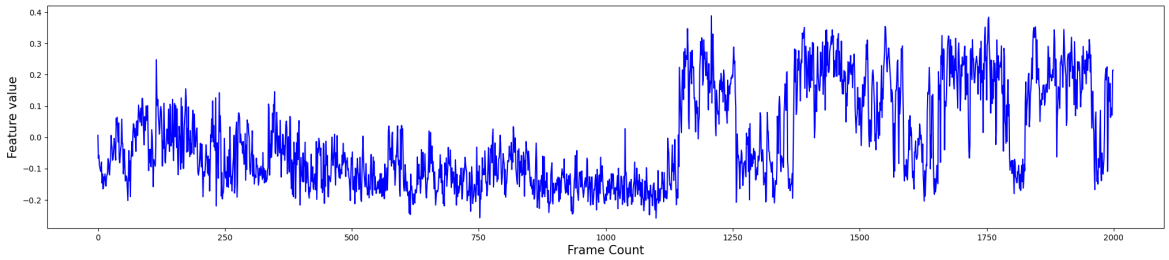
The cosine kernel compute the using cosine distance metrics $d(x, y) = \frac{x^T y}{||x|| \cdot ||y||}$. Two objects that are exactly alike have zero distance. The gaussian kernel is an exponential function of the gamma scaled quadratic distance between any two points $k(x, y) = exp(-\gamma||x - y||^2$. The aim of comparing multiple kernels as shown in figure 3.5 is to understand the impact on the sensitivity of the change point algorithm as we shall discuss below.

Auto-Encoder

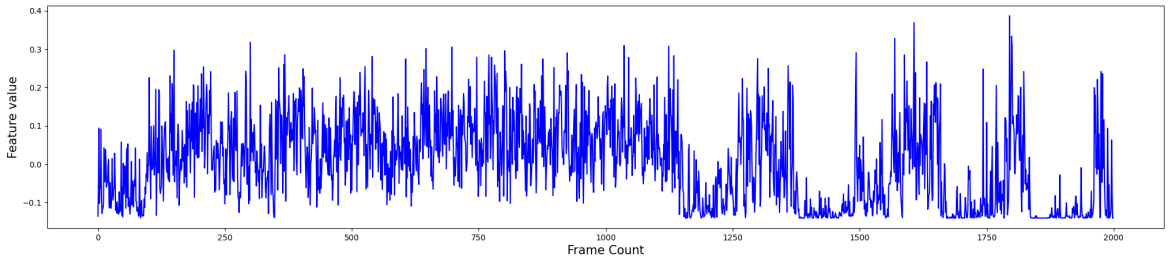
Auto-encoders learns useful representation with little or no supervision. The goal of an autoencoder is to learn a mapping from high-dimensional observations to a lower-dimensional representation space such that the original observations can be reconstructed (approximately) from the lower-dimensional representation. It is a parametric model that is trained using an encoder-decoder neural network architecture without any supervision. The parameters are optimized using mean



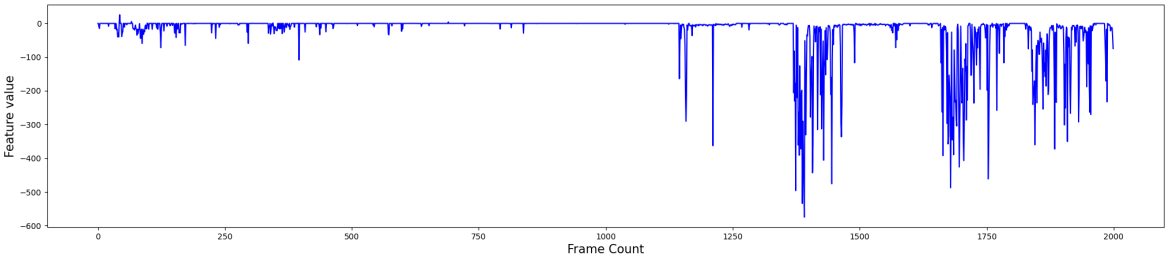
(a) Kernel PCA - Linear kernel



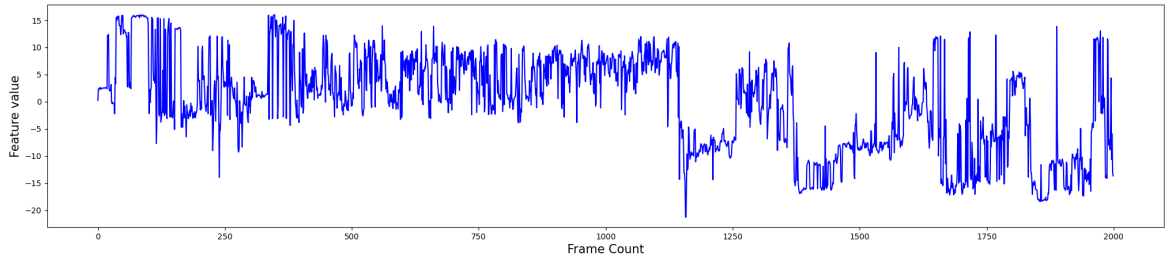
(b) Kernel PCA - Cosine kernel



(c) Kernel PCA - Gaussian kernel



(d) Autoencoder Embedding



(e) TSNE

Figure 3.5: 1-D Plot of Sample Video Using VGG-19 Feature Extractor

squared loss function [154,155]. We pretrained our autoencoder and optimized the parameters using the mean squared error loss. We saved and used the trained model to encode the extracted videos features to a 1-dimensional sequence. The model parameters were optimized to capture as much information as required to reconstruct the original feature vector. In training the model, we also applied the oversampling technique as described in 3.3.2.

T-Stochastic Neighborhood Embedding (TSNE)

TSNE [156] uses a probabilistic model to minimize the KL-divergence between the high dimensional input Gaussian distributed feature vector and the lower dimensional t-distributed manifold. We also applied this model to encode the extracted video features to a 1-dimensional manifold. TSNE fits the data by minimizing the KL-divergence between the higher dimensional gaussian distributed input feature and the lower dimensional t-distributed embedding. We set the perplexity parameter to 50 which is similar to the number of nearest neighbour that is used in other manifold learning. Figure 3.5 shows the 1-d embedding plot for our test video. TSNE is computationally intensive and does not scale well for very large data.

3.3.4 Video Shot Boundary Detection

In this section, we outline the shot boundary detection problem and describe strategy for building shot boundary analysis in long VCE video. Video temporal boundaries or visual change points are sudden visual variations in the sequence of frames leading to statistical variation in the vector representation of the frames. These variations typically represent transitions occurring between states in a process that generates the data. Change Point Detection on sequence data has been applied in solving sequence segmentation problems across various applications [129,130,157] including video shot boundary detection [50, 51, 52]. In our video analysis, we define boundary detection as detecting change in both visual property as well as a pathological event in the sequence of observed frames. As discussed above, different techniques exist in literature to optimize detection of change point. The techniques are typically applied on lower dimensional time series data. However, little to no prior work has been proposed to optimize similar technique for high-dimensional video data [158]. Some of the algorithms used on time series sequence data include, Binary Segmentation (BS) algorithm, the Segment Neighbourhood (SN) algorithm, Optimal Partitioning (OP) algorithm, and the Pruned Exact Linear Time (PELT) algorithm [111]. The PELT algorithm is based on the OP algorithm but involves a pruning step within the dynamic program to minimize the computational cost. The pruning reduces the computational cost without affecting the exactness of the resulting segmentation making it an ideal candidate for high dimensional video data.

The PELT method is an exact method, and generally produces quick and consistent results. This algorithm solves the penalized detection problem when the number of change points in the sequence is unknown. It tries to minimize the log-likelihood cost function (see 3.1) by estimating both the number of change points as well as location of the change in a sequence of data. The algorithm has a computational cost of $\mathcal{O}(n)$, where n is the number of data points [159]. The PELT algorithm can solve the change point detection problem using different kernels. The commonly used ones are the linear and Gaussian kernels.

Similar to time series data, video data are a sequence of measurement over time describing the visual behaviour of objects being captured. Therefore, any CPD technique can be considered applicable after the manifold projection. In our experimentation, we applied the PELT algorithm on the VCE video data after projecting to a 1-dimensional manifold for temporal segmentation task.

3.3.5 Temporal Segmentation of VCE Video

To divide the sequence of frames into segments, we assume that a sequence of observed frames $\{x_1, x_2, ..., x_T\}$ can be divided into non-overlapping, homogeneous segments $\{\theta_1, ...\theta_\tau\}$. The delineations between partitions are called the temporal boundaries. We further assume that for each partition θ, the data within it are temporally correlated and also comes from the same distribution $P(x_t|\theta_i)$ while each segment are independent.

While temporally segmenting long can be very challenging, the problem is much more complicated in VCE videos. VCE videos have peculiar non-uniform characteristics and inter-frame variations. Such variations may be due to poor lighting in a particular region, inter-patients variations as well as instability of the camera motion due peristaltic movement of bowel. Often times, this leads to highly degraded and poor quality video. Also, detecting change points in a video needs to be in line with the objective of the analysis as visual change points does not necessarily indicate a pathological event. For CE videos, change in the visual property of the sequence of frames due to the camera flip, may not necessarily represent a pathological event.

Applying the PELT algorithm on an ordered sequence of frames features $x_1, ..., x_T$, our CPD model will have m change points with their positions $\tau_{1:m} = \{\tau_1, ..., \tau_m\}$; where $1 \leq m \leq T-1$. We specify $\tau_0 = 0$ and $\tau_{m+1} = T$ and assume change points are ordered such that $\tau_i < \tau_j$. The m change points will split the data into $m+1$ segments with the ith segment containing $x_{(\tau_{i-1}+1):\tau_i}$

The algorithm begins by first conditioning on the last point of change, it then iteratively relates the optimal value of the cost function to the cost for the optimal partition of the data prior to the last change point plus the cost for the segment for the last change point to the end of the data [111].

Let $\tau = \{\tau : 0 = \tau_0 < \tau_1 < \cdots < \tau_m < \tau_{m+1} = T\}$ be set of possible vectors of change points for the video. Set $F(0) = -\beta$. The optimal partition is defined as:

$$\begin{aligned} F(s) &= \min_{\tau \in \mathcal{T}_s} \sum_{i=1}^{m+1} [C(x_{(\tau_{i-1}+1):\tau_i}) + \beta] \\ &= \min_{t} \left\{ \min_{\tau \in \mathcal{T}_t} \sum_{i=1}^{m} [C(x_{(\tau_{i-1}+1):\tau_i}) + \beta] + C(x_{t+1} : n) + \beta \right\} \\ &= \min_{t} \left\{ F(t) + C(x_{t+1} : n) + \beta \right\} \end{aligned} \tag{3.1}$$

Where C is a cost function for the i^{th} segment; β_m is a penalty to guard against over fitting which essentially determines how many change points the algorithm will find. The higher the specified β_m the less the number of detect change points forcing the algorithm to reduce the False Positives (FP). Experimenting with this to make sure increasing the penalty β_m is not jeopardising the ability to detect true change points (TP).

$$C(x_{(\tau_{i-1}+1):\tau_i}) = (\tau_i - \tau_{i-1}) \left(\log(2\pi) + \log \left(\frac{\sum_{j=(\tau_{i-1}+1)}^{\tau_i} (x_i - \mu)^2}{\tau_i - \tau_{i-1}} \right) + 1 \right) \tag{3.2}$$

C is chosen as twice the negative log-likelihood as in 3.2 and the minimum segment length $\tau_{i-1} - \tau_i \geq 1$. Temporal segmentation algorithm:

Data: VCE video with frames $1 : T$; $V = \{f_1, f_2, ..., f_T\}$
Result: video shot boundary $\theta = (\theta_1, ..., \theta_k)$
begin
  for $f_i \in V$ **do**
    Extract Features using CNN: $X_i \leftarrow G(f_i)$
    Project each feature vector $\boldsymbol{x_i}$ to 1-D manifold $\lambda_i \leftarrow \boldsymbol{x_i}$
  Concatenate manifold projections $\forall f \in V$; $\boldsymbol{\lambda} = \{\lambda_i, ..., \lambda_T\}$
  Compute change points $\{\tau_i, ..., \tau_m\}$
  Get segments for V; $\{v_j\}_{j=1}^{k}$

Video	Training samples	Testing samples
Video 1	12,303	-
Video 2	13,177	-
Video 3	8,452	-
Video 4	23,124	-
Video 5	32,181	-
Video 6	-	8,701
Video 7	-	16,909
Video 8	-	10,037

Table 3.1: Data summary for training and test videos

3.4 Experiments

3.4.1 Dataset and Pre-processing

We conducted experiments using eight (8) VCE videos. In review and analysis of VCE, physicians are only interested in the small bowel region which can only be accessed through VCE and not through any of the upper and lower endoscopy procedures.

We extracted the videos from the RapidReader software program and pre-processed each video into frames. The eight (8) videos were collected from different patients during a clinical endoscopy procedure using the SB3 Given Imaging PillCam capsules. The capsules were equipped with 576 x 576 pixel camera. For each complete video, the small bowel transit time corresponds to about 3.93 ± 1.43hr [35]. In order to isolate the small bowel region, each video was reviewed and annotated by two endoscopy research scientists. After the annotation, the total number of frames in the videos is summarized in table 3.1 and the class distribution based on the content is show in figure 3.4.

We randomly selected 5 videos for pre-training our feature extraction model and also to perform other pre-training. The remaining three (3) videos were reserved for testing the models. These three test videos were not at any point shown to the model. Using videos from different patient during testing helps minimize any bias and ensures our approach will generalize to any new unseen video data.

3.4.2 Implementation

During the preprocessing stage, we trimmed the frames to 500 x 500 to remove the black boundary region. We developed our entire system using the Pytorch framework on NVIDIA GTX2080 machine. We ensured that all our experiment was run on the same configuration for consistency

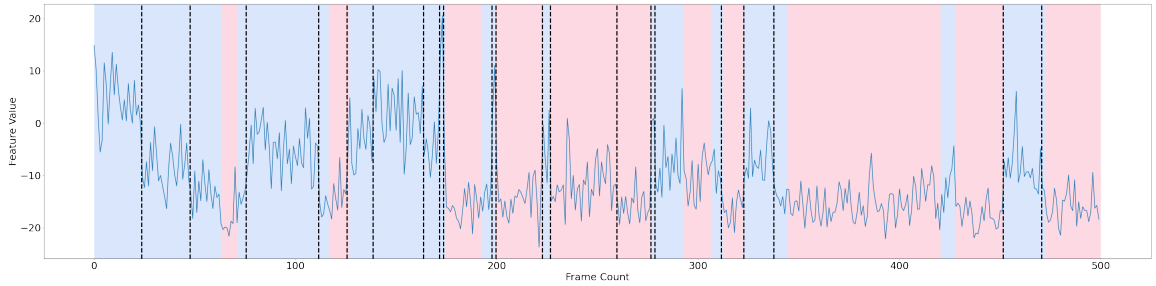


Figure 3.6: Detected Boundaries vs Ground Truth using PCA @ beta=150

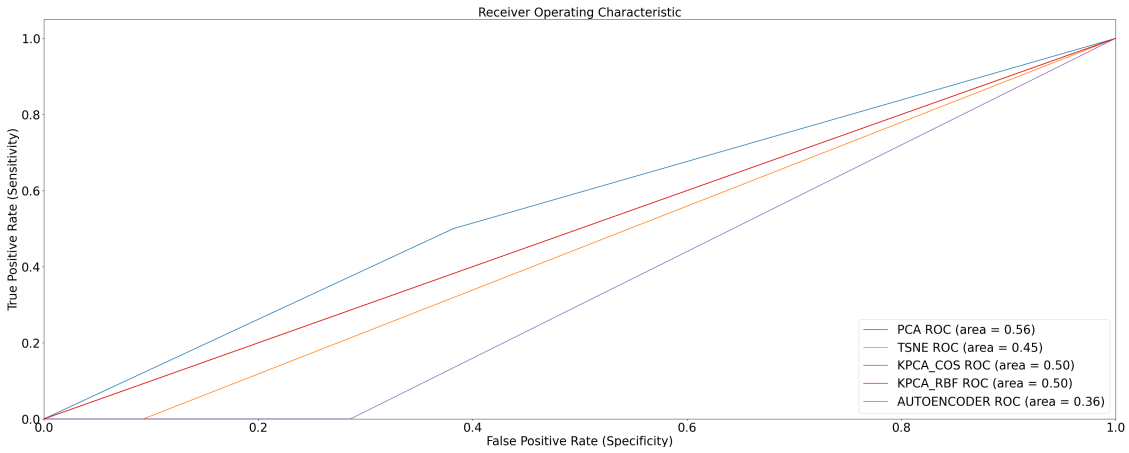
across the compared techniques. Each of the feature extractors were trained for 30 epochs using 0.001 as learning rate and Stochastic Gradient Descent optimization. We also trained the autoencoder to embed the frame-features for 50 epochs. During of the training, we over-sampled the minority classes based on the inverse of their population in the data. This gave a significant boost to the representation learned by the model.

Evaluation

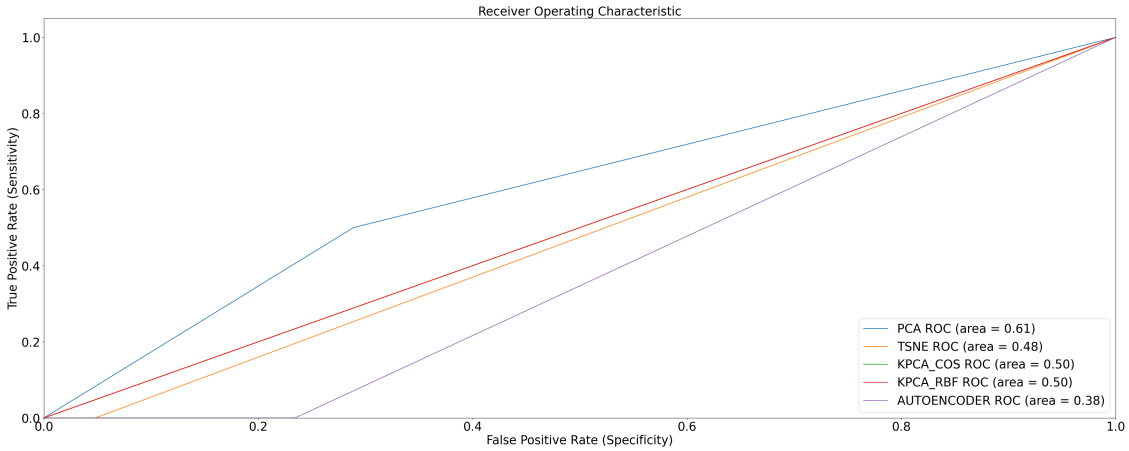
We evaluated the performance of this method based on the AUC-ROC. At each time step t, the model predicts whether t is a change point or not. A change point is defined when the class of frame at $t-1$ is different from the frame at t. Using the predicted output, we computed the True Positive and False Positive rates and we applied this in computing the ROC. Each change point is considered to be a pathological change point and so we benchmarked against the ground truth label provided by the medical experts. This is, obviously a very challenging problem as the change point detector has no prior information on the statistical properties of any pathology.

3.5 Results and Discussion

Figure 3.6, shows experimental results of detected boundaries using PCA embedding and the PELT change point algorithm. Each of the alternating pink-colored intervals are sections of some pathological abnormality. There are points where visually one can observe changes but are not really pathological events. These points are due to the camera rotation and flips as it is propelled down the GI tract through peristalsis.



(a) ROC Plots @ beta=10



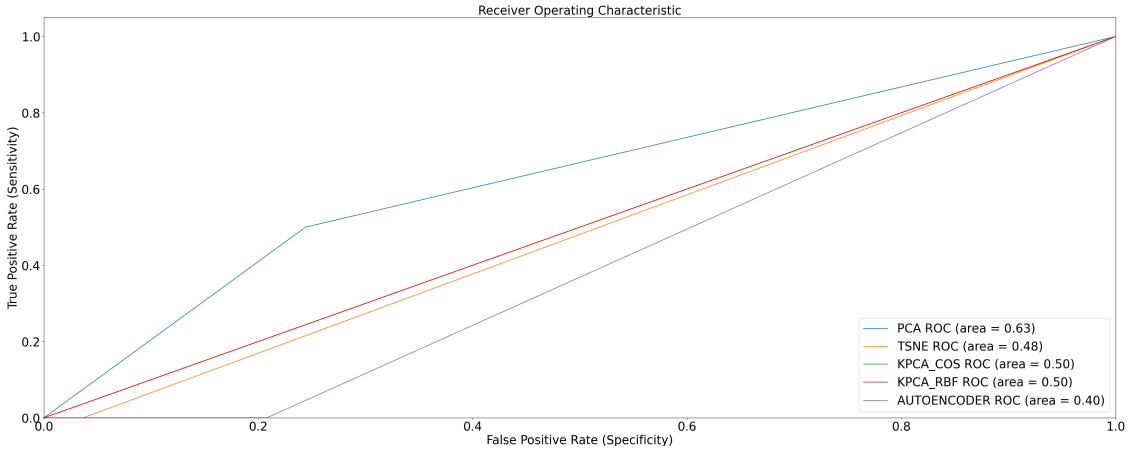
(b) ROC Plots @ beta=50

Figure 3.7: ROC Plots @ beta=10 & 50

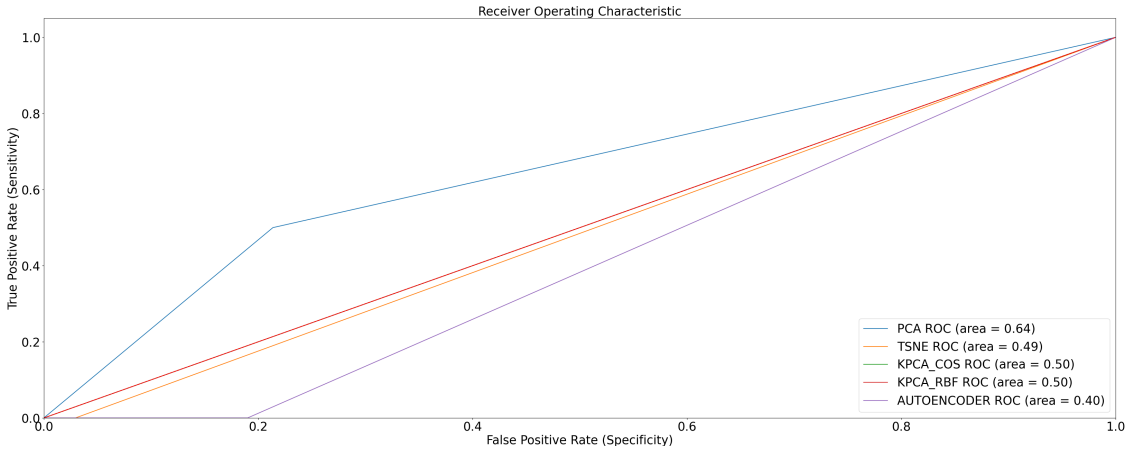
Experiments on feature extraction also showed that feature extraction capability of the base CNN model is critical to what the boundary detector is able to locate. How well the base CNN is able to encode the lesioned-frames, different from the normal frames will impact the performance of the boundary-detector model. In addition, different CNN architectures showed varying performance when applied on different classes of lesion. Lesions show significant difference both geometrically and in terms of color, texture as well as the surrounding lighting condition. This indicates that the base CNN capabilities are not universal and some architectures better capture some structure than others.

Figure 3.7 - 3.9 shows comparative results using different parametric and non-parametric embedding techniques. Parametric representation frameworks such as auto-encoder are very difficult to train, but are able to capture some non-linearity in the data wherever they train successfully. Similar to training the feature extractor, we also adjusted for the data imbalance when training the auto-

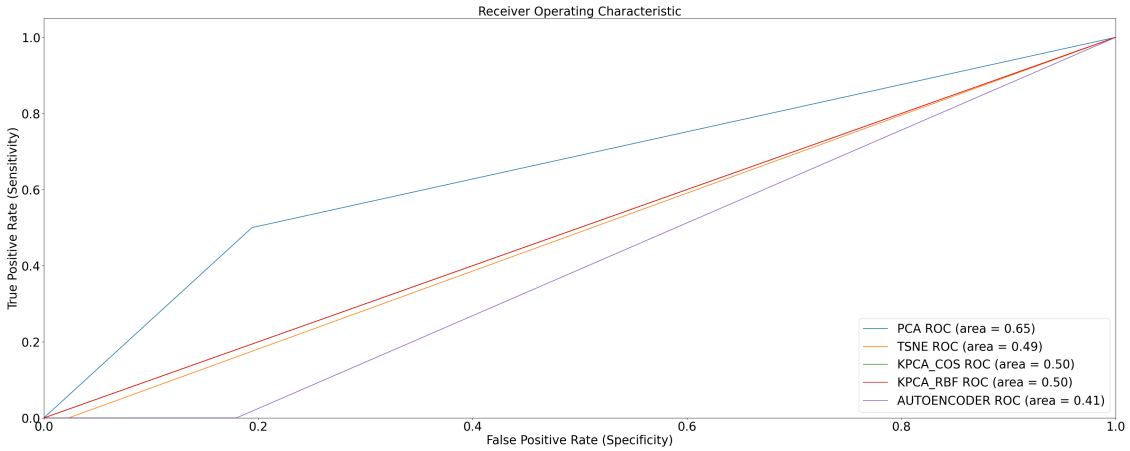
encoder to avoid encoding bias into the network.



(a) ROC Plots @ beta=100

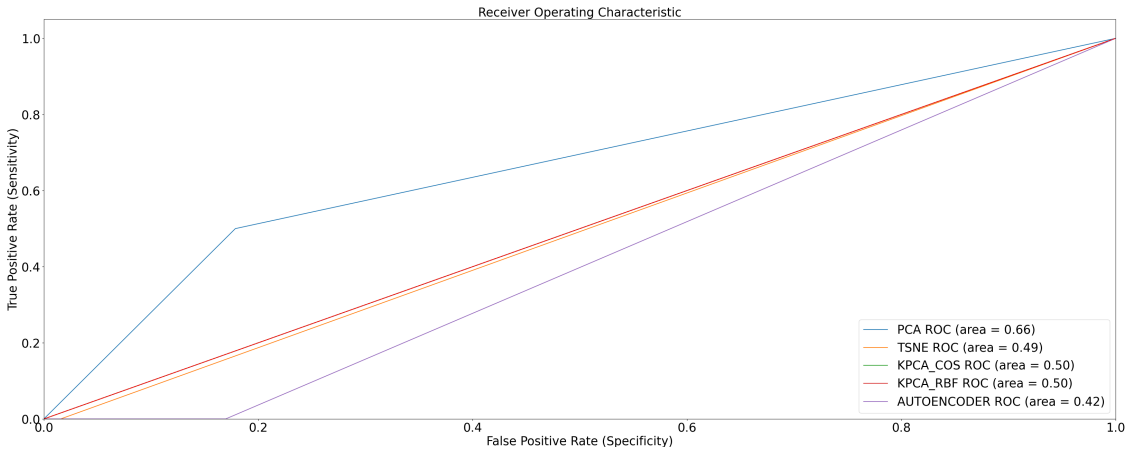


(b) ROC Plots @ beta=150

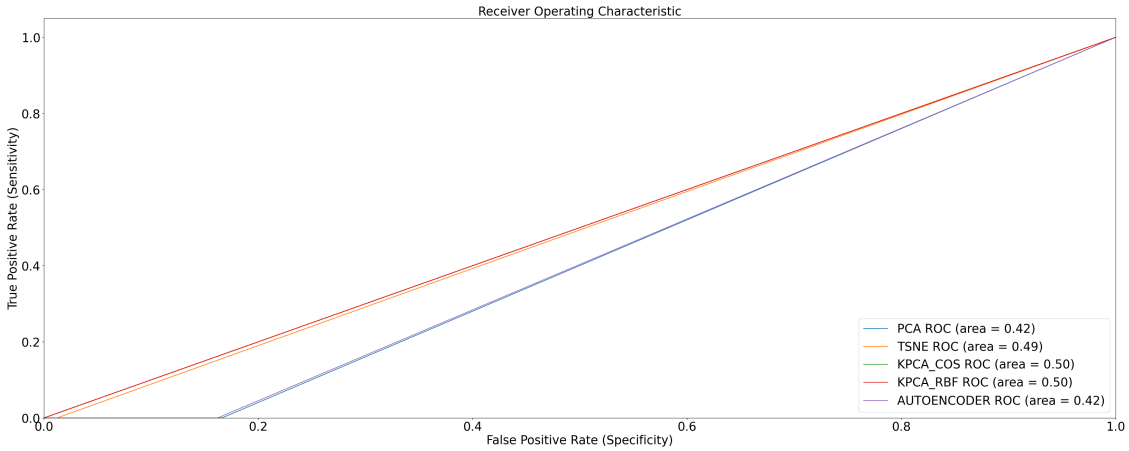


(c) ROC Plots @ beta=200

Figure 3.8: ROC Plots @ beta = 100, 150 & 200



(a) ROC Plots @ beta=250



(b) ROC Plots @ beta=300

Figure 3.9: ROC Plots @ beta=250 & 300

Figure 3.7 - 3.9 compares the receiver operating characteristics of different embedding techniques on a test video using different penalty hyper-parameter. From the figure, PCA with linear kernel is able to encode the video to capture more abnormal boundaries than other kernels as well as autoencoder and tsne.

While VGG-19 was able to encode the frames and separate diseased frames from the normal frames, the final pool layer of the model has 4096 features. Encoding 4096-d to 1-d is very challenging due to the complexity of the higher dimensional space and doing this without supervision is even harder.

Detected Video Boundaries in a Sample Test Video

Figure 3.10 below show the detected change points in the sequence of frames.

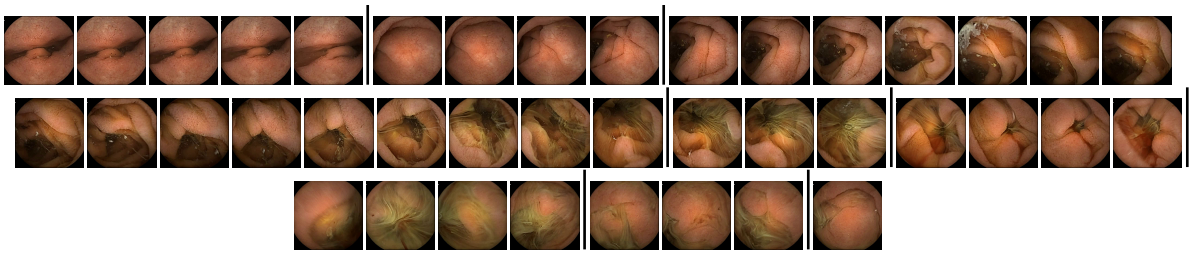


Figure 3.10: Visual Illustration of Detected Video Boundaries

As shown in figure 3.10, the actual boundaries detected in the video frames does not necessarily indicate pathological event. However, very similar frames are captured in the same temporal boundaries. Clearly detecting pathological boundaries in VCE videos is not trivial and also a very challenging problem. Developing a model, similar to binary segmentation, that would only require broad abnormality label - normal/abnormal - without pathology level annotation would help mitigate the challenge of completely unsupervised when adapted for temporal segmentation task. This is the objective of the work covered in chapter 5. The model developed in chapter 5 is trained on pathology-agnostic basis to flag the boundary between sequence of frames based on binary classification.

Chapter 4

Weakly Supervised Temporal Segmentation of Long Capsule Endoscopy Video Using Graph Neural Network

4.1 Introduction

Graph Convolutional Neural Network (GCNN) allows more flexibility in capturing the topological relationship between frames in a video. GCNN strikes a balance between extremely rigid temporal dependence relationship between the frames and also highly independence frame analysis that has been mostly adopted in VCE video abnormality recognition. Therefore, GCNN allows us to perform non-sequential search for an temporal boundary in a video frame sequence by predicting a binary category for each node in the graph without the restriction of temporal correlation.

Weakly-supervised learning for video temporal segmentation based on binary categories can generate a higher quality video summaries than unsupervised approach that are blind to the categories of activities in the video. Motivated by the work in [160], this chapter focuses on temporal segmentation of VCE video into semantically-consistent segments, delimited not only by temporal boundaries but also pathological change points.

VCE video data differ from conventional video structured data with predictable temporal dependence between the sequence of frames. The unstable peristaltic movement of the bowel leads to frequent

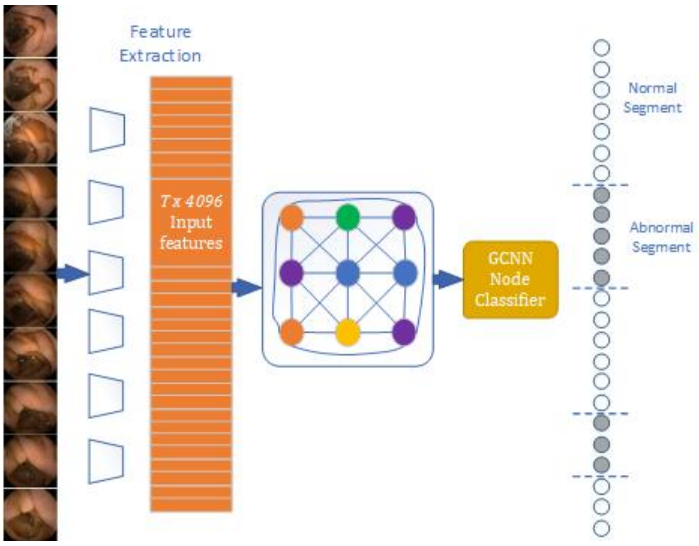


Figure 4.1: VCE Video Network Representation

camera flip that impacts the temporal relationship between the sequence of frames. Intermittent movement of the intestinal walls and sudden transition between different regions of the GI tract are other characteristics that are peculiar to VCE video. This characteristics leads to unusual temporal relationship between the frames, even though they are captured in sequence. Any pair of frames captured in the sequence may not be correlated. Therefore, change in visual property in the frame sequence may not correspond to a pathological event. This peculiar property makes applying traditional video analysis or sequence-to-sequence model to VCE videos unrealistic.

In this chapter, we consider relationship between frames in a CE video as an undirected independent graph $G = (N, E)$ where the frames are represented as the nodes of a graph and the entire VCE video is a complete graph. As against the assumption of temporal correlation between the sequence of frames, with this formulation, we are able to capture relationship between the frames in the video based similarity which was captured into the edge weights between the nodes of the graph. We applied this formulation in disease-agnostic manner by formulating the problem as an abnormal vs normal frames problem. This formulation is similar to outlier detection problem in [3, 34, 37]. However, we leverage additional frame relationship information by using Graph Convolutional Neural Network (GCNN) to model a CE video (fig 4.1). We argue that CE video frames with graph structure performs better that modelling the frames independently using traditional CNN. The VCE video will, therefore form a graph network with frames containing abnormalities sharing similar properties while normal frames also share similar properties. The unique capability of graphical model on VCE video data allows capturing the structural relations among the frames thereby allowing us to harvest more insights between pairs of frames compared to independent or open temporal dependence assumption.

Graph Convolutional Neural Networks (GCNN) [161] have recently gained popularity among deep learning researchers [162, 163]. They are an efficient variant of Convolutional Neural Networks (CNNs) on graph structured data [164]. Previously, deep learning based model have been successful on spatial structured data, such as images [45]. And also on sequential data using variants of Recurrent Neural Network (RNNs) [165, 166]. However, they have not been able to generalize to graph structured data. Recent advances in GCNN allows leveraging advantages of deep models on graph structured data. In CNN, the trainable local filters enable the automatic extraction of high-level features. The computation with filters requires fixed number of ordered units in the receptive fields. However, the number of neighboring units is neither fixed nor are they ordered in generic graphs (see fig. 4.1). GCNN stack layers of learned first-order spectral filters followed by a nonlinear activation function to learn node and graph representations.

Motivated by binary segmentation [138], we developed a disease agnostic GCNN model to classify nodes (frames) of VCE video into normal / abnormal, for each time step in the video, our model predict the binary category for each node in the graph. The output of our model forms a complete VCE video sequence with corresponding boundary marker using the prediction from our GCNN model. Each contiguous homogeneous segment are considered temporal regions of either abnormal or normal frames. Finally, we select a representative frame within each boundary as the corresponding video summary. This technique generalizes to many unseen new videos as GCNN can be trained both transductively and inductively.

4.2 Related Work

4.2.1 Anomaly Detection in VCE Images

Our propose approach can also be formulated as an outlier detection model which have been applied on problems such as intrusion detection [167] and traffic speed forecasting [168]. Many researchers have also applied semi-supervised concept to anomaly detection on VCE videos in [3, 13, 15]. The techniques for detecting abnormal frames or outlying frames have leveraged the skewed distribution of VCE video data where there is far more normal frames than abnormal. Learning the characteristics of the normal frames allowed the models to flag any frames that are out of the range of the learned characteristics. Wide variations exist between different patients' videos in endoscopy studies. Therefore, an attempt to capture characteristics of all possible normal may be infeasible. We propose to mitigate this problem by learning both transductively and inductively using GCNN model then applying the model in a summarization context to generate representative frames as video summaries of an entire long VCE video. Our model differs from traditional outlier

detection model as GCNN takes advantage of topological relationship between the nodes in the graph as against assuming independence of individual node. This tends to balance the extreme assumption of temporal correlation using time-series assumption and complete independence using traditional outlier detection technique.

4.2.2 Graph Convolutional Neural Network (GCNN)

GCNNs extends existing neural network methods for processing data represented in graph and graph-based applications. They can generally be divided into graph-focused and node-focused applications [169]. In graph-focused applications, the function Γ is independent of the node n and implements a classifier or a regressor on a graph structured dataset. The mapping $\Gamma(G)$ may be used to determine overall category of G. In node-focused applications, Γ depends on the node n, so that the classification (or the regression) depends on the properties of the node. This is very applicable in object detection and localization [170]. GCNN has also been applied on video data. For example, in [171], GCNN was applied to localize action in short videos. Similarly, [172] applied spatial-temporal graph convolutional network (ST-GCNN) in recognizing human action in a video. Another work [173] propose stacked spatio-temporal graph convolutional networks for action segmentation in video data. In [174], the authors applied graph convolutional neural network for video question answering. [175] combines LSTM network with Graph Convolutional Network in a dynamic fashion to model applications where the relationship between edges of the graph changes over time. Other applications of Graph Convolution Network include text classification [176]. In [177], the authors proposed a graph convolution tracking for visual tracking of objects in videos.

4.2.3 Problem Formulation and Notations

We represent a CE video as a graph $G = (N, E)$ where, N represents the vertex set consisting of nodes $\{x_1, ..., x_n\}$ representing the frames' features, and $E \in \mathcal{R}^{\text{nxn}}$ is a symmetric adjacency matrix where e_{ij} denotes the edge weight between nodes pair $\{x_i, x_j\}$. A missing edge is represented with $e_{ij} = 0$. We also define the degree matrix $D = \text{diag}(d_1, ..., d_n)$ as a diagonal matrix where $d_{ij} = \sum_j e_{ij}$. Each node x_i has a $p-$dimensional feature vector $\mathbf{x}_i \in \mathcal{R}^p$. For a complete video, the feature matrix $\mathbf{X} \in \mathcal{R}^{\text{nxp}}$ stacks n feature vectors on top of one another, $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]^T$.

4.3 Methodology

4.3.1 Feature Extraction

Our feature extraction was a follow up on the procedure described in section 3. We extracted visual features using pretrained CNN model which was fine-tuned on the VCE video data as described in Chapter 3. We selected the VGG-19 network for our model based on it's more representative capability.

4.3.2 Graph Convolutional Neural Network Classification

We employed the GraphSAGE network proposed in [163] to learn the embedding for each video frame from the extracted features. The GCNN method leverage the neighborhood of each node and aggregate the representation in each layer over the entire network.

Node Embedding

The embedding generation (forward propagation) assumes that the model has already been trained with fixed parameters $\boldsymbol{W^l}$ where l is the number of layers in the network. $\boldsymbol{W}$ performs the linear transformation of the node embeddings from the input dimension to a specified output dimension.

To generate embedding for each node i in the graph, we aggregate the representations of nodes in its immediate neighborhood $N(i)$ to the neighborhood vector $h^l_{N_i}$, which depends on the representations generated in the previous iteration:

Given a graph $G = (N, E)$ with input features $\{\boldsymbol{x}_i, \forall i \in V\}$ with l layers each with weight matrices $\boldsymbol{W}^l$, The GCN network iteratively computes the embedding for each layer-l as follows:

Starting from the initial input features

$$\boldsymbol{h}_i^0 = \mathbf{x}_i, \quad \forall i \in \mathcal{V}; \tag{4.1}$$

Aggregator Function In contrast to our baseline model where we learn over the features of the frames without any neighborhood contribution, the aggregator function operates over an un-ordered set of vectors and account for message passing between nodes and layers of the network. An ideal aggregator function would be symmetric (i.e. invariant to permutations of its inputs) while still being

trainable and maintaining high representational capacity [163]. Motivated by [163], we considered three (3) aggregator functions in our training.

$$\boldsymbol{h}^{L}_{N(v)} = \text{AGGREGATE}\Big(e_{ij}h^{l}_{j},\ \ \forall j \in N(i)\Big); \tag{4.2}$$

Mean Aggregator This is an element-wise mean of the input vectors in every layer. The mean aggregator is nearly equivalent to the convolutional propagation rule applied in transductive GCNN.

$$\boldsymbol{h}^{l+1}_{N(i)} = \frac{1}{N-1}\sum_{j}\big(e_{ij}h^{l}_{j}\big); \tag{4.3}$$

LSTM Aggregator This is a more complex aggregator compared to the mean aggregator as it has the advantage of larger expressive capability. However, our configuration for the LSTM architecture differs from the mean and pool aggregator. This is because LSTM is not permutation invariant as they process input in a sequential manner. Rather than use the similarity based function on the edges of the graph, we specified a chain function where every x_t is connected to the previous node x_{t-1} and x_{t+1}

$$h^{l+1}_{i} = \sigma\Bigg(\mathbf{W}\cdot LSTM\Bigg(\{\mathbf{h}^{l}_{i}\}\cup\{\mathbf{h}^{l+1}_{j}, \forall j \in \mathcal{N}(i)\}\Bigg)\Bigg) \tag{4.4}$$

The LSTM aggregation steps are as follows:

$$z_l = \sigma\big(W_z\cdot[h^{l}_{i}, e_{ij}h^{l}_{j}],\ \ \forall j \in N(i)\big); \tag{4.5}$$

$$r_l = \sigma\big(W_r\cdot[h^{l}_{i}, e_{ij}h^{l}_{j}],\ \ \forall j \in N(i)\big); \tag{4.6}$$

$$\tilde{h}_l = tanh\big(W\cdot[r_l * h^{l}_{i}, e_{ij}h^{l}_{j}],\ \ \forall j \in N(i)\big); \tag{4.7}$$

$$\boldsymbol{h}^{l+1}_{N(i)} = (1-z_l) * h^{l}_{i} + z_l * \tilde{h}_l \tag{4.8}$$

Max-Pool Aggregator Similar to a pool layer on a convolutional network, the pool aggregator performs an element-wise max-pool function across neighbor vectors.

$$\boldsymbol{h}^{l+1}_{N(i)} = \max_{j}\big(e_{ij}h^{l}_{j}\big); \tag{4.9}$$

The node's current representation h_i^l is concatenated with its aggregated neighborhood vector $h_{N(i)}^l$, which is later fed into a fully connected layer with a nonlinear activation function σ. This is then used for the next layer representations:

$$\boldsymbol{h}_i^{l+1} = \sigma\Big(\boldsymbol{W}^l \cdot \big(h_i^l \cup h_{N(i)}^{l+1}\big)\Big) \tag{4.10}$$

We experimented with each aggregator function and found LSTM to outperformed the others such as mean and pool aggregation functions. The embedding at the last layer is used as the representation for each node.

$$\boldsymbol{z}_i = \boldsymbol{h}_i^L \tag{4.11}$$

where $\boldsymbol{z}_i$ is the learned embedding for node i.

In contrast to [163] we train the parameters of the network using weak disease-agnostic labels provided by the expert research gastroenterologist. This formulation allows the network to generalize to any unseen category of disease. The network parameters were trained using cross-entropy loss function.

4.3.3 Temporal Segmentation

We consider any homogeneous segment of the CE video with a certain number of frames as independent with member frames identically distributed. The relationship between members of a segment can be captured by a similarity model such as nearest neighborhood. For our model, we applied the cosine similarity function to capture this neighborhood relationship.

4.4 Experiment

We ran our experiments using eight (8) CE videos. We trained the GCNN on five (5) videos to learn the k $\mathbf{W}$ linear transformation parameters and evaluated on three (3) separate videos. The model was trained using back propagation and cross entropy loss function. Our temporal segmentation task was modeled as a binary classification problem. Essentially, for every time step, the model determines if $\mathbf{x}_t$ is normal or abnormal frame. We applied this model in detecting a change point in the sequence of frames based on whether the current frame contains an abnormal tissue or not.

4.4.1 Dataset and Preprocessing

A total of 8 VCE videos were collected from different patients during a clinical endoscopy procedure using the SB3 Given Imaging Pillcam capsules. The capsules were equipped with 576 x 576 pixel camera. The videos have been diligently annotated and verified by two medical gastroenterology experts. We randomly selected five (5) videos of training while three (3) remaining videos were used as the test set. Each frame in the video was trimmed to 500 x 500 to remove the black boundary region. We employed pre-trained VGG-19 CNN network for our feature extraction as described in chapter 3. The input to our model is the frame-feature matrix $X \in \mathcal{R}^{\text{nxp}}$ where n is the length of the video and p is the dimension of the features for each frame.

4.4.2 Implementation

Our model was developed using the Pytorch framework on NVIDIA P100 machine. The model was trained for 50 epochs with a learning rate of 0.001.

4.4.3 Evaluation

The models are evaluated as a binary classifier based on precision, recall, f-score and accuracy for each binary class. The metrics are defined as follows:

$$Precision = \frac{\text{TP}}{\text{TP + FP}} \tag{4.12}$$

$$Recall = \frac{\text{TP}}{\text{TP + FN}} \tag{4.13}$$

$$F1 - Score = \frac{2 \cdot \text{Precision * Recall}}{\text{Precision + Recall}} \tag{4.14}$$

$$Accuracy = \frac{\text{TP}}{\text{TP + FN}} \tag{4.15}$$

4.5 Results and Discussion

Table 4.1 shows experimental results on a sample test video. The test video was never seen by the model during training. We experimented with different aggregator function at each layer of

Table 4.1: Results of Abnormality Classification

Method	Metric	Class	
		Normal (34,508)	**Abnormal (1,139)**
GCNN (mean)	Accuracy	**0.970**	
	Precision	0.971	**0.171**
	Recall	**0.997**	0.011
	F1 score	**0.984**	0.021
GCNN (pool)	Accuracy	0.955	
	Precision	0.971	0.111
	Recall	0.982	0.013
	F1 score	0.976	0.023
CNN-Baseline	Accuracy	0.956	
	Precision	**0.974**	0.142
	Recall	0.980	**0.076**
	F1 score	0.977	**0.099**

the GCN network and compared this with the baseline model. From the table, the GCNN model out-performs the baseline CNN model showing that the GCNN is able to leverage the neighborhood information based on message passing at each layer of the network to learn a better representation of the nodes (frames) in the video. As against the independent representation by the CNN baseline model with an accuracy of 95.6% across three test videos while the GCNN with mean aggregator outperforms the model with 97.0% accuracy. Other metrics captured are the precision, recall and the f-score. The GCNN shows significant improvement over the baseline indicating they are able to generalize better to unseen videos than traditional deep CNN models. This result is despite weighted oversampling the minority weight while training the CNN-baseline.

While we can generally classify VCE data across patients as easy and hard videos, our sample videos were never seen before by the models. With lower precision, recall and f-score for the abnormal class across all three (3) models, this indicates that oversampling the minority class in training the CNN-baseline model did little to mitigate the problem. The result also further underscores the challenge in developing robust system for VCE video analysis with minimal miss-rate for abnormalities. As against traditional video structured data in other domains, generalizing across patients is equally a difficult problem.

Furthermore, from result in table 4.1, the difference in performance between the normal and abnormal class is clear. Low recall on the normal class indicates high false negatives (FN) which will lead to the model flagging a lot of non-change points as change points when adapted to temporal segmentation problem. This means the model will not significantly reduce the redundancy in the data amounting to less time saving for the physician. Similarly, low recall on the abnormal class

equally indicates high false negative which will lead to the model ignoring a number of abnormal frames and considering them as normal. This has huge implication on the diagnosis outcome of the physician.

Finally, a direction for future work would be to consider weighted oversampling in training the GCNN model. This will expose the model more to the examples in the abnormal category to minimize the effect of the class imbalance.

Chapter 5

Graph Convolution Neural Network For Weakly Supervised Abnormality Localization In Long Capsule Endoscopy Videos

5.1 Introduction

In this chapter, we address the problem of temporal abnormality localization in long VCE videos using video-level class labels.

Activity localization or action detection [178,179,180] in a video involves identifying the region where the activation score of frames corresponding to the class of activity in the video is maximum. Activity localization in a short video has received significant attention among computer vision research community [58, 59, 59, 61, 62, 63, 64, 65, 66, 72, 78, 79, 80, 81]. However, models such as structured segment network in [59], multi-stage CNN model in [58] and boundary regression in [65] requires frame-level labels to train. Obtaining frame annotation in medical domain, particularly for CE video data is very challenging. In order to develop a model that generalizes across multiple patients and diseases, the model would require large sample of each abnormality collected across multiple patients. Furthermore, while Deep Convolutional Neural Network (DCNN) based models have demonstrated improved performance on various image recognition [144, 145, 146] and video analysis tasks [96, 97, 106, 181, 182] including medical image analysis [3, 4, 7, 44, 143], they are notoriously sample inefficient requiring large samples per training class to optimized its parameters for ease of generalizability.

The challenge of obtaining frame level label is further exacerbated, in the medical domain, when the expertise, time and effort required are not readily available. Despite the large volume of frames generated in a single CE examination, the high class imbalance, with significantly more normal frames than disease-containing frames, limits the feasibility of training a fully supervised DCNN model that generalizes across multiple abnormalities and also patients.

Prior works on CE video have mostly focused on single or multiple lesion detection on each individual and independent frames in the video [3,4,6,7,11,13,14,15,17,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44]. Despite the extreme difficulty of obtaining frame-level label for CE video frames, little to no attention has been made towards leveraging temporal or topological relationship between the frames to develop a more robust system. To the best of our knowledge, no prior work has addressed the task of temporal abnormality localization within a sequence of CE video frames. We believe that analysis of video data requires leveraging the spatial, temporal and topological relationship between the frames to achieve a system that can be deployed in real clinical environment to aid physicians in their diagnosis. The novelty of the work proposed in this chapter is in three (3) folds; Firstly, we leverage the spatial, temporal and topological relationship between the frames to develop a model to localize abnormal regions containing the disease or abnormality of interest in a full CE video. Secondly, our model uses only weak video level labels for this task, thereby obviating the need for an expert provided frame-level annotation, which is often very challenging. Thirdly, we employed Graph Convolutional Neural Network (GCNN) model, based on the GraphSage architecture [163]. This allows us to learn a robust representation of CE videos both transductively and inductively by leveraging the message passing architecture and neighborhood information aggregation.

Different techniques have been proposed for lesion segmentation within a 2-D CE video frame [17,29,32]. Similar to the high cost of obtaining pixel-level label for image segmentation, obtaining frame-level labels for CE videos is not an easy task. First, annotating individual frame is much more tedious than the normal CE video review process. Secondly, challenging conditions such as poor illumination and camera instability due to peristaltic motion of the bowel impacts the quality of frames generated in the video, leading, sometimes, to noisy and unreliable expert-provided frame-level annotation. To the best of our knowledge, this is the first work on abnormality localization in a sequence of CE video frames using weak video level labels. The model proposed in this chapter, addresses several issues in CE video analysis where our weakly supervised model requires no frame-level annotation from medical experts. In addition, by using graph-based model, we learn a more robust representation of the video through message passing and information aggregation.

Given weak labels for each video segment, we train a weakly supervised GCNN model on aggregate frame features and classify each video segment. During testing, we applied an adaptive temporal pool layer on the GCNN model to generate frames' activation score corresponding to the video

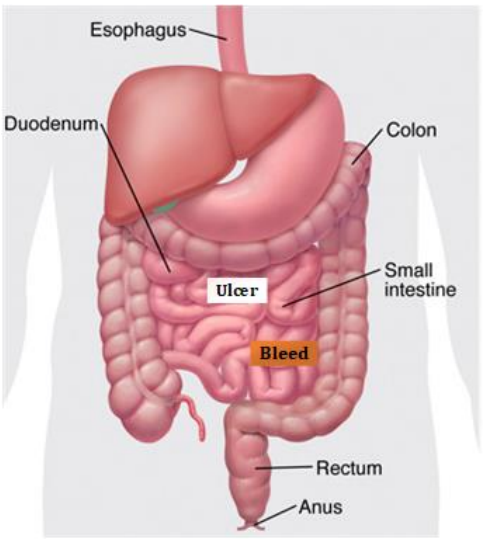


Figure 5.1: Abnormality Localization in Capsule Endoscopy Video

class activation map over the sequence of frames. The adaptive temporal pool layer ranks the frames within each segment based on the significance to identifying the segment as abnormal. This significance of this framework is in minimizing experts' review time on CE videos by generating frames relevant to abnormality of interest for review by the expert physician or gastroenterologist without the need for frame level labels.

Long videos typically differs from short videos based on the duration and also the number of actions contained in the sequence. Since short videos usually contain one object or activity of interest, activity localization within a short video involve detecting a single high energy region in the sequence. Meanwhile, long videos pose additional challenge with multiple energy activation regions requiring temporal segmentation before localization. In addition to the novelty of the work previously mention, with the end-to-end system for long videos proposed in this chapter, we are able to generalize the concept activity localization to long videos with multiple activities within the sequence. Without requiring manual partitioning of the video into fixed frame length. Localizing action in short videos involves a temporal search for a single class activation map within the sequence while long videos with multiple activities will have multiple actions withing the sequence.

Models based on weakly supervised learning have recently gained traction among the machine learning researchers [87, 88, 89]. In this chapter, we address the problem of temporal abnormality localization in long CE videos using video-level class labels. We developed a weakly supervised Graph Convolutional Neural Network (GCNN) model for frame-level localization, based on global video level multiple labels. Figure 5.1 shows the framework of the proposed model in this chapter.

To further demonstrate the uniqueness of this work, figure 5.2 shows the comparison between activity

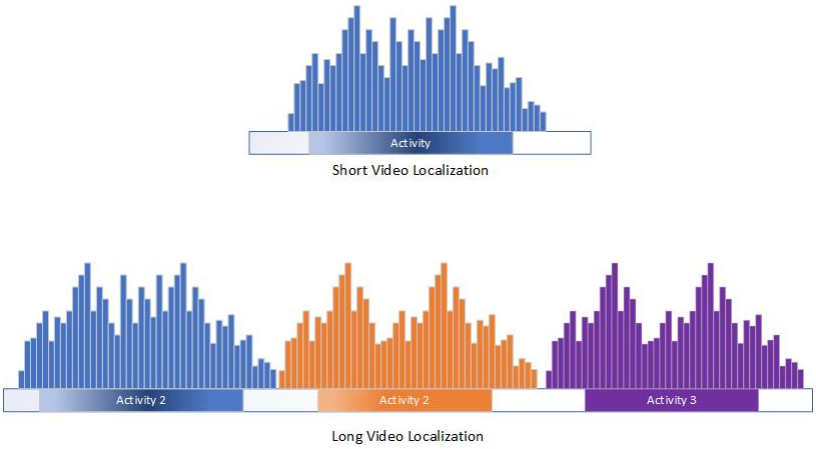


Figure 5.2: Comparing Long and Short Videos

localization in long and short videos as it applies to activity localization. Short videos usually contain one object or activity of interest while long videos, due to the extended duration, often capture multiple object or actions of interest requiring temporal segmentation before localization. The novelty of the work in this chapter is in the generalization of activity localization concept to long videos with multiple activities within the sequence. Localizing action in short videos involves a temporal search for a single class activation map within the sequence while long videos with multiple activities will have multiple actions withing the sequence.

5.2 Related Work

In this section, we discuss the prior works on analysis of CE video as well as techniques that have been developed for various disease and abnormality detection. Our review covers work on GCNN in other domains and different formulations and solutions to weakly supervised learning tasks.

5.2.1 Abnormality Detection in Capsule Endoscopy Videos

Analysis of CE videos encompasses tasks such as disease or abnormality detection, quantifying severity of identified diseases, localizing identified abnormalities, and decision making on appropriate intervention by the physician. Prior works on automating review and analysis of CE videos can broadly be categorized into three (3) - 1) detection of specific disease or lesion such as bleeding in [4], polyp [39], ulcer [44], and angioectasia [41,42]; 2) abnormal or outlier frame detection where frames with abnormalities are consider outliers [15,17]; and 3) models aimed at minimizing experts review time on CE video - video summarization. Here key frames capturing abnormalities are selected as

representative frames from the entire video [16, 20, 21, 22, 26, 27]. While obtaining frame level label for CE videos is very difficult, little to no attention has been paid to models that will leverage the relationship between the frames to mitigate this challenge. To the best of our knowledge, no prior work has considered temporal abnormality localization on CE video data. The work proposed in this chapter aligns with the concept of video summarization where, by leveraging the temporal and topological relationship between the video frames, we localize the abnormality to a more narrow temporal region. This allows us to select representative samples within each abnormal region as a video summary for the experts. In addition, our model does not require any frame level label to identify the abnormal regions and localize abnormal frames in the video.

5.2.2 Graph Convolutional Neural Network (GCNN)

Following the work in [161], GCNN continues to gain increased popularity among deep learning and machine learning researchers. GCNN extends techniques such as *Recursive Neural Networks (RNN)* [183, 184] and *Markov Chains* [185, 186] while leveraging the powerful representation power of neural networks on graph structured data. Traditional deep learning models are well developed for spatial (CNN) and sequential (RNN) data with little contribution on graph structured data. CNNs are used to learn representation on 2D spatial image data while RNNs learns to encode and represent sequential data. While CNNs and RNNs based models [16] can automatically learn the internal encoding of the graph structured data, SVM's internal representation needs to be user designed. Meanwhile, many natural interactions between objects can be represented as a graph with the relationship between the objects captured in the edges between the nodes of the graph. Graph Neural Networks (GNN) models are robust and generic enough to also accommodate spatial and sequence data by specifying the nature of the edge and node relationships.

Main operations on graph network are filtering, activation and pooling. Similar to regular convolution, Graph Convolution Network (GCNN) combines the benefit of spatial and spectral based filtering operations [161] in addition to non-linear transformation of the input features to achieve a robust representation of the graph structured data. GCNN represents features as nodes in the graph and wide range of relationships, from simple similarity (e.g. cosine similarity) to long- short term memory (LSTM) can be modelled to capture the relationship between the nodes as weighted edges. Graph filtering uses neighborhood aggregation from the previous layer to determine the representation of each node in subsequent layer [163]. [163] proposed GraphSage to leverage both inductive and transductive learning capability of GNN. For each layer of the network, the model aggregates the representation for each node in the graph based neighborhood sampling from surrounding nodes. Graph Attention Network (GAT) [187] was proposed to improve the neighborhood aggregation by ranking the neighboring nodes using an attention layer to generate better representation.

5.2.3 Weakly Supervised Localization

State-of-the-art methods address the problem of temporal action localization in long videos by applying RNN based action classifiers on sliding windows [178, 179] for action detection in a video sequence. Methods such as structured segment network in [59], multi-stage CNN model in [58] and boundary regression in [65] are some of the approaches to action detection in a sequence of video frames. However, these techniques require frame level annotation which is a very difficult to collect, particularly in medical domain. In order to mitigate this challenge, weakly supervised methods using global video level labels for activity localization has recently been gaining traction among researchers [60, 85, 188, 189, 190]. In [60] Nguyen et al., proposed sparse temporal pooling network for action localization in an untrimmed video. Using video-level class labels, their model predicts temporal intervals of human actions in a video. In [189] the authors proposed the Weakly supervised Temporal Activity Localization and Classification (W-TALC) framework using only video-level labels. They used two sub-networks - a two-stream based feature extractor network and a weakly-supervised module - trained by optimizing two complimentary loss functions. The model learns to classify the videos and also localize the region of the action within the video. Class Activation Mapping (CAM) was introduced in [85] for weakly supervised action localization in an untrimmed video. Similarly, [65] proposed a cascaded boundary regression method for temporal action localization.

In [64], Lin et al. proposed a single shot technique for temporal action detection in a video. Their model based on 1D temporal convolutional layers, skips the proposal generation step in detection by classification framework, to directly detect action instances in untrimmed videos. [63] used convolution de-convolution network to precisely localize action in untrimmed videos. The work in [71] is focused on weakly supervised localization of novel objects using the objects' appearance transfer framework. Another unique attempt at action localization was proposed in [67], where the authors temporally localized action in untrimmed videos using (Auto-loc). UntrimmedNets was proposed in [69] for temporal action recognition and detection.

While our proposed framework is motivated by [163, 188, 190], our model combines more effective GraphSage representation network with a final attention layer in the classification model. As against just simple temporal attention model used in [60], our GCNN representation is able to leverage the neighborhood information for more effective representation of each member node in the graph. However, we adapted the temporal pool layer based on [188] for the abnormal frame localization during inference. GCNN localization framework was considered in [190], our model is different in that the aim of our localization task is to select sparse representative frames in each video segment as against using similarity between time segments to determine the temporal boundaries [190]. Secondly, this chapter addresses the problem of temporal abnormality localization in long CE videos which is collected under more unstable and challenging digestive tract environment than most open

dataset. Lastly the peculiarity of this work as against other prior works on activity localization is that abnormal regions in CE videos are not usually contiguous, making frameworks developed temporal segment boundary detection ineffective. Our model, therefore aims to select sparse non-contiguous representative frames within each video segment by applying a temporal pool layer over the final GCNN activation layer. To the best of our knowledge, this is the first work using temporal information to localize abnormal frames in CE video data.

5.3 Methodology

Weakly supervised temporal abnormality localization is an extension of weakly supervised object segmentation task on 2-dimensional images. The weakly-supervised abnormality localization and classification problem addressed in this chapter can be directly mapped to Multiple Instance Learning (MIL) problem [191]. In extending this to videos, we consider a video segment as a bag of normal and abnormal frames i.e. given a video $V \in \mathrm{R}^{H\mathrm{x}W\mathrm{x}T}$ where H and W are the height and width of the frames and T is the temporal length of the video or number of video frames, we considered $\mathcal{V} = \{f_n, f_a\}$ where f_n and f_a are normal and abnormal frames respectively. A single VCE video may contain multiple abnormalities making the video a mixture of both normal and abnormal frames with different diseases while the video may also contain no abnormality at all.

Following the above, we define a graph $G = \{N, E\}$ with nodes N representing the frames in the video and edges E representing the connections between the frames. Secondly, we denote a sub-graph $g = \{v, e\}$ representing video segment and edges e representing edges between the frames in the video segment. Recall that V is a bag containing both normal and diseased frames occurring at different points in the video. Our goal are in two stages, First is the graph classification where, for any video segment containing at least one abnormal frame, we predict abnormal label. Next is the abnormality localization where we generate frame-level activation score the abnormal video segment. The goal of the video segment classification is to first learn a mapping of $\Gamma(G) \rightarrow \{y_1, ..., y_k\}$ to k categories of diseases including normal frames contained in the video. The graph G is then optimized to the multiple instances of normal and abnormal frames it contains by aggregating the embedding of the frames to predict the corresponding multiple classes. Next is to use the parameters of the learned network to score each frame based on their contribution to the graph prediction. This will be subsequently referred to as our localization step which happens only during testing.

For a long video with multiple diseases at different regions, we applied our unsupervised temporal segmentation method described in chapter 3 to split the videos into homogeneous visual segments. Each segment is then considered a bag of normal and abnormal frames with some frames containing only normal frames. A video segment is considered abnormal if it contains at least one frame with

an abnormality. Similar to a Multi-instance learning problem, an abnormal video segment contains a mix normal and abnormal frames [192, 193, 194].

5.3.1 Feature Extraction

Similar to the procedure described in Chapter 3, we compared multiple feature extraction approaches on the VCE frames, and adopted the VGG-19 [145] network for our feature extraction from each frame. The network was pretrained on five (5) VCE videos by oversampling on the minority classes to create a balanced exposure of the model for better representation. We obtained a 4096-dimensional feature vector per frame from the pool-5 layer. Each video segment is represented by feature volume of t x p_{in} where t is the length of the segment and p_{in} is the dimension of each frame extracted from the pretrained VGG-network.

5.3.2 Model Architecture

For this task, we applied the Graph Convolutional Neural Network (GCNN) model where the VCE video segment is considered a graph while the frames represents the nodes in the graph. The model architecture is shown in figure 5.3. The input to the model is the extracted frame features for each video and partitioned into sub-videos txp_{in} where t is the length of the video and video segment (i.e. number of frames) and p_{in} is the dimension of the feature vectors.

During training, we only have access to weak video-level labels. While we know there is a certain abnormality in the video, we do not have granular information as to the frames where the disease is captured nor frequency of occurrence of the disease in the entire video. A fully supervised model will utilize labels pointing to the actual frame containing the disease as localizing the frame with the disease is important in helping the physician make quick and proper diagnosis.

5.3.3 Graph Convolution Network - GCNN

Few prior works have proposed Graph Convolutional Network (GCN) model for action localization in videos [190]. However, the uniqueness of this work is the application to long videos where there may be more than action within the sequence. Secondly, this work advances other prior works through frame level localization as against localizing to temporal region or volume in the video. Lastly, rather that using a fixed length temporal segment as proposed in [171, 190], our temporal segmentation is integrated into the temporal segmentation work described in chapter 3 for an end-to-end automated summarization system. Such end-to-end segmentation and localization helps

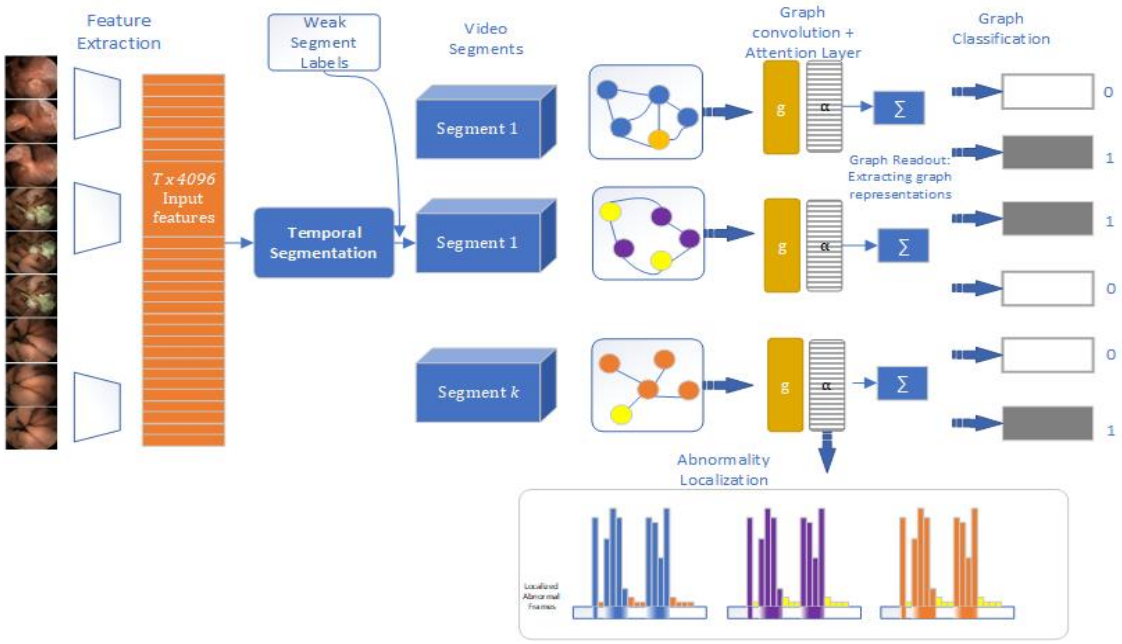


Figure 5.3: Weakly Supervised Abnormality Localization Model

mitigate against any intersection between member frames in different temporal shots.

For a long CE video with multiple diseases at different regions, we applied unsupervised temporal segmentation method using the PELT change point detection algorithm to split the videos features into homogeneous visual segments. Each segment is then considered a bag of normal and abnormal frames with some frames containing only normal frames. A video segment is considered abnormal if it contains at least one frame with an abnormality. Similar to a Multi-instance learning problem, an abnormal video segment contains a mix normal and abnormal frames [192, 193, 194].

Each unique patient's video V^n is temporally segmented into k video shots $\{v^n_{i=0}, ..., v^n_{i=k}\}$ based on the visual temporal boundaries. While the long video V^n can contain multiple diseases and therefore have multiple labels $\{y_1, ..., y_l\}$, the result of the segmentation step allows us to only capture one or no abnormality within each short video segment. The illustration in shown in figure 5.3. First, we partition the videos into uniform segments $\{v^n_{i=0}, ..., v^n_{i=k}\}$ such that not more than abnormality is present in each segment with no overlapping frames. We considered a disease agnostic framework with labels $y_i \in \{0, 1\}$ such that we only classify each segment as either abnormal or normal based on whether it contains at least an instance of an abnormal frame. This binary disease-agnostic framework will allow our model to generalize to any unseen category of abnormality in the future.

5.3.4 Graph Representation and Classification

We applied Graph-Sage convolution framework from [163]. The framework allows for inductive and transductive learning on large graphs. The model architecture is shown in figure 5.3. The input to the model is the extracted frame features for each video segment t x p where t is the length of the video segment (i.e. number of frames) and p is the dimension of the feature vectors.

During training, we only have access to weak labels for the video segments as shown in fig. 5.3. While we know there is a certain abnormality in the video, we do not have granular information of the frames where the disease is captured. Physicians also use the frequency of occurrence of a disease in multiple frames to determine its severity. A fully supervised node classification model will utilize labels pointing to the actual frame containing the disease as localizing the frame with the disease is important in helping the physician make quick and proper diagnosis. We consider a video segment as a bag of normal and abnormal frames i.e. given a video $v \in \mathrm{R}^{hxwxt}$ where h and w are the height and width of the frames and t is the number of video frames in the segment. We consider $V = \{f_n, f_a\}$ where f_n and f_a are normal and abnormal frames respectively. A single CE video may contain multiple diseases or abnormalities making the abnormal class a combination of different abnormalities or diseases. This class agnostic model makes the model generalize to other new unseen diseases in the future.

Following the above, we define a graph $G = \{V, E\}$ with nodes $\mathcal{V}$ representing the frames in the video and edges E representing the connections between the frames. Secondly, we denote a sub-graph for each video segment $g = \{v, e\}$ and edges e representing edges between the frames in the video segment. Recall that each sub-graph v is a bag containing both normal and diseased frames occurring at different points in the video. Our goal are in two stages, First is the multi-instance graph classification where, for any video segment containing at least one abnormal frame, we predict abnormal label otherwise, we predict abnormal label. Next is the abnormality localization where we generate frame-level activation score the abnormal video segment. The goal of the video segment classification is to first learn a mapping of $\Gamma(G) \rightarrow \{y_i\}_{i=0}^{1}$ to binary normal and abnormal video segment. In our case, we employed disease agnostic binary category so as to be able to generalize to any unseen videos of new patient. The Graph-Sage network, through the sequence of transformation, aggregation, attention and multi-instance classification learns to classify each video segment into binary category of normal and abnormal segment. Next is to use the parameters of the learned network and the sequence of linear transformation, aggregation and final attention layer to score the frames in the abnormal segment based on their contribution to the graph prediction. This step is our localization step which occurs only at test time.

5.3.5 Graph Convolution Network

The uniqueness of this work is the application to long videos where there may be more than action within the sequence. Secondly, this work advances other prior works through frame level localization as against localizing to temporal region or volume in the video. Lastly, rather that using a fixed length temporal video features as input to the GCNN network [171, 190], our video segment inputs have varying length based on detected shot boundary in the long video. This makes our framework a complete end-to-end localization framework which has not been previous addressed in literature. Such end-to-end automatic segmentation, classification and localization helps mitigate against any intersection and correlation between member frames in each video segment.

The graph convolution involves three main steps: 1) Neighborhood aggregation; 2) Node representation: which involve concatenation, linear transformation and non-linear activation steps; 3) graph read-out.

Steps (1) and (2) occur at each layer of the network, while step (3) occurs at the final layer of the network. For our model, we used two (2) graph convolution layer.

The input to the network are the frame feature sub-matrix where each frame-feature represents a node in the graph with the weighted edges computed as the similarity between the features. Each node is directly connected to every other nodes but the edge weights is set to be proportional to the level of similarity between the pair of nodes. Thus, each video graph is a *complete graph*. Since all frames are images of different locations of the small bowel, we allow nodes to derive message from every other nodes in the graph. Secondly, the formulation allows feature similarity and dissimilarity to be incorporated into the parameter learning process. This similarity between edges, essentially, captures the topological relationship between the frames. GCNN explicitly ensures relationship between frames is put into consideration during both training and testing as it aggregates the neighbouring nodes into each node for every layer of the network. Each frame feature vector is transformed by a weighted average of all other neighbouring frames it is connected to with weights based on learned edge strengths. In our case, all the frames in the video is a neighbour but the edges are weighted by the similarity function. We applied cosine similarity defined in 5.11 as the similarity metric between pair of the frame feature vector. Frames without any similarly will have edge weight of zero - meaning no connection between them. Other similarity function such as nearest neighbor, correlation and euclidean distance were also experimented with and we compared the results across. The only problem with using a nearest neighbor relationship is having to set the number of neighbors k, which may not be optimal for the dataset.

Feature Aggregation and Node Embedding

Fig. 5.4 shows a representation for he neighborhood feature aggregation. After the first layer, each node feature is a weighted average of all the neighboring node features.

Starting from the initial input features

$$\boldsymbol{h}_i^0 = \mathbf{x}_i, \quad \forall i \in v; \tag{5.1}$$

where i represents a nodes (frames) and v is the video segment.

the representation of the neighbors of node i at layer $l+1$ is given as the weighted aggregation of all neighboring node j features;

$$\boldsymbol{h}_{N(i)}^{l+1} = \text{AGGREGATE}\Big(e_{ij}h_j^l, \ \forall j \in N(i)\Big); \tag{5.2}$$

where j represents neighboring node to node i.

Aggregation functions such mean, max-pool and LSTM can be applied. After experimenting with the different aggregator functions, LSTM outperformed the others and also more stable to train. We used the LSTM aggregation between each pair of the nodes. Eq 5.2 becomes

$$\boldsymbol{h}_{N(i)}^{l+1} = LSTM\Big(e_{ij}h_j^l, \ \forall j \in N(i)\Big); \tag{5.3}$$

where $N(i)$ is the total number of neighbors of node i. For a complete graph, this will be one short of the total number of nodes in the graph. The LSTM aggregation steps are as follows in step eqn. 5.4:

$$z_l = \sigma\Big(W_z \cdot [h_i^l, e_{ij}h_j^l], \ \forall j \in N(i)\Big); \tag{5.4}$$

$$r_l = \sigma\Big(W_r \cdot [h_i^l, e_{ij}h_j^l], \ \forall j \in N(i)\Big); \tag{5.5}$$

$$\tilde{h}_l = tanh\Big(W \cdot [r_l * h_i^l, e_{ij}h_j^l], \ \forall j \in N(i)\Big); \tag{5.6}$$

$$\boldsymbol{h}_{N(i)}^{l+1} = (1 - z_l) * h_i^l + z_l * \tilde{h}_l \tag{5.7}$$

Next, we get the embedding for node i by concatenating neighboring nodes representation $\boldsymbol{h}_{N(i)}^{l+1}$

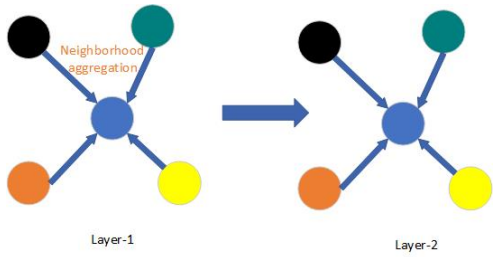


Figure 5.4: Neighborhood Aggregation

with the previous layer embedding of node i itself;

$$\boldsymbol{h}_i^{l+1} = \sigma\left(\boldsymbol{W}^k \cdot \text{CONCAT}\left(h_i^l, h_{N(i)}^{l+1}\right)\right) \tag{5.8}$$

Eq. 5.3 is the aggregation of the features from all connected neighboring nodes weighted by the edge similarity.

Graph Attention and graph aggregation Layer

After the final layer of the graph convolution operation, we applied an attention layer over the node embedding. The attention layer allows us to learn a parametric weighting of the nodes based on their importance to the graph classification. This allows the model to learn to place more weight on abnormal frames for the video segments with abnormality as also the most relevant frames for segments that are completely normal. We learn a representation of the entire GCN network at that last layer by aggregating features from all the nodes. Attention-based LSTM and GRU have been report effective in learning similar representation over sequences [195]. However, GCNN model allows additional flexibility over a wide range of representation from mean to max-pooling over the nodes to the more complex LSTM aggregation at this layer. This final graph aggregation is called the graph readout layer.

$$h_g = \frac{1}{N}\left(\sum_i \alpha_i h_i\right) \tag{5.9}$$

where h_g is the representation of the entire graph g of the video segment V_k. Other readout operations include mean, summation and max-pool over the nodes embedding learnt across the layers of the network.

Figure 5.5: Multi-Instance Graph Classification

Multi-Instance Graph Classification

Once we aggregate the graph into a single feature vector, the final graph classification layer is a fully-connected layer that maps the graph embedding to the number of categories in our dataset before applying a sigmoid layer. We predict the binary label for each graph as follows:

$$\hat{y}_{i=1}^{N} = \frac{1}{1 + e^{-h_g}} \tag{5.10}$$

Where N is the number of graphs and h_G is the learned representation of g.

Fig. 5.5 shows the illustration of the multi-instance graph classifier.

$$\boldsymbol{e}_{i,j} = \frac{\boldsymbol{x}_i^T \boldsymbol{x}_j}{||\boldsymbol{x}_i||_2 \cdot ||\boldsymbol{x}_j||_2} \tag{5.11}$$

5.3.6 Graph Localization Network

The graph localization network is the second step after training the parameters W of g. The step occurs during testing, using the trained parameters W, we replaced the final graph readout function with a temporal pool layer to allow us generate a class activation map over the sequence of frames in the video. Our localization network generates ranking for each node in the graph. We sampled the temporal pool layer to identify nodes with the abnormality. Since abnormal regions in CE video is not necessarily contiguous, temporal pool over frames better captures the localization than temporal boundary regression [65]. Non-contiguity of abnormal regions is a unique property of CE videos which differentiates it from other video structured data.

5.4 Experiments

5.4.1 Dataset Description

Our dataset consist of nine (9) long VCE videos collected during real clinical endoscopy procedure under the supervision of expert gastroenterologist. All IRB requirements and approval processes were completed prior to analyzing the data. Since physicians are more interested in the small bowel region for the CE video examination, we focused our analysis on images of the small bowel region only. Each video was carefully annotated by two (2) endoscopy research scientist and verified by an expert gastroenterologist. We fine-tuned the pretrained feature extractor network on the first five (5) videos in our dataset and used it to extract features for all other videos. Since each video is unique to each patient, we ensured separation between videos that have been previously seen by the model were not part of the test videos. This helps mitigate patient bias.

The training video and the diseases captured in training data one is shown in table 5.1

Table 5.1: Training & Test Video Data Description

	Video Content	
Train Video	***Nodes Count***	***Abnormal Categories***
Video - 1	13,177	Normal, Erythema, Outgrowth (Mass)
Video - 2	8,452	Normal, Angioectasia, Diffuse bleeding, Erosion, Erythema, Ulcer
Video - 3	23,124	Normal, Diffuse bleeding, Ulcer, Angioectasia, Outgrowth
Video - 4	12,303	Normal, Angioectasia, Outgrowth Erythema, Erosion, Clot
Video - 5	29,236	Normal, Bleeding, Ulcer Erythema
Total	**86,292**	
	Video Content	
Test Video	***Nodes Count***	***Abnormal Categories***
Video - 6	14,173	Normal, Ulcer, Angioectasia, Erythema, Erosion
Video - 7	16,909	Normal, Bleeding
Video - 8	10,037	Normal, bleeding, Angioectasia
Video - 9	19,104	Normal, Bleeding, Ulcer
Total	**60,223**	

In our proposed model (shown in Figure: 5.3), the score predicted for each frame corresponds to the

node activation sequence for the frame. Rather than using the granular class of each abnormality shown in table 5.1, we used a class-agnostic binary label for the graph classification. This allows the model to generalize to any unseen categories of abnormalities in future videos.

5.4.2 Evaluation

We evaluated our proposed framework in two folds. First, the performance of the multi-instance graph classification model was evaluated on new patients' test videos based on accuracy, sensitivity, specificity and f-score. Evaluation based on the intersection-over-union that has been employed in literature on localization does not directly apply on CE video since an abnormal temporal bound may not be contiguous. Instead, we employed the widely adopted evaluation framework on CE videos [19, 20] - *Coverage*. Which also is a measure of specificity of the model on the abnormal frames. The specificity of the abnormal classes is the most important criteria on which medical experts base the performance of machine learning models since this impacts the accuracy of their diagnosis. The coverage is defined as in equation 5.12 which is the number of selected sample frames as a proportion of all abnormal frames in the segment. We aggregate this over the entire video to report our result.

$$\mathrm{C} = \frac{\sum_{i}^{N_{ab}} c_i}{N_{ab}}; c_i = \begin{cases} 1, & \text{Abnormal frame is selected} \\ 0, & \text{otherwise} \end{cases} \tag{5.12}$$

where N_{ab} is the count of video segments with abnormality. From 5.12 The metric scores one (1) if at least one abnormal frame is selected and zero otherwise.

Implementation

Our entier model was implemented in Pytorch [196] on NVIDIA RTX2080 GPU. We trained the GCNN using stochastic gradient descent optimization algorithm using cross entropy loss function with a batch size of 8 and learning rate of 0.001. The models were trained for a minimum of 100 epochs.

5.5 Results and Discussion

Table 5.2 shows the result of the binary multi-instance graph classification task applied on four (4) different video data. The four (4) test videos are different from the training videos and have never been seen before by our model. This allows generalization of our model to new patients' videos. The total segments is the count of the abnormal and the normal video segments and the disease categories is the number of different diseases present in the complete video. Table 5.3 shows the performance on the localization task. The result in both tables is the weighted average of the metrics computed over the binary classes which accounts for the class imbalance in the dataset.

Table 5.2: Video Graph Classification Results

Metrics	**Test Video Data**			
	Video 6	**Video 7**	**Video 8**	**Video 9**
Frames Count (T)	14,173	16,909	10,037	19,104
Total Segments	770	1,124	248	1,071
Disease Categories	5	2	3	3
Accuracy	**0.899**	0.848	0.560	0.859
Sensitivity	**0.911**	0.804	0.601	0.889
Specificity	**0.899**	0.848	0.560	0.859
F-score	**0.905**	0.822	0.578	0.873

On video-1, the model achieved classification accuracy of 89.9% on 770 video segments with 5 different categories of diseases. The sensitivity, specificity and F-score are 91.1%, 89.9% and 90.5% respectively. The best performance was recorded on video-1 indicating that the performance across patients are not equal and some patients' videos may be more challenging than others. With different number of classes across each of the videos, the result of the model reflects the realistic output when a new patient's video is shown to the model. Prior to administering the capsule endoscopy, patients are advised not to eat or consume any opaque liquid that could obstruct the visibility of the camera. Occlusion and other factors in the digestive tract varies across patients leading to difference in classification performance. On the segment classification task, the model performed least on video-3 with classification accuracy of 56.0%; sensitivity of 60.1%; specificity of 56.0% and F-score of 57.8%. The performance on the other two videos are better and closer to the performance on video-1. With the highest number of disease classes in video-1, the performance on video-2 makes it rather difficult to believe the number of different abnormalities present in the video may impact the performance of the multi-instance classifier.

From table 5.3, for video-1 at k=1, by sampling a single (1) frame from each abnormal video segments the model is able to cover 92.5% of all the abnormalities in the video. Similarly, by sampling the

top-2 frames, the models covers 97.5% of all abnormalities in the video. This, however, flattens after this point which may be attributed to a number of reasons. The number of activated high energy frames in the video segments is a proportion of the total number of frames that captures abnormality and the total length of the video segment. With very few (e.g. only 1) abnormal frames and very long video segment, it may difficult for the model to identify this single frame within the segment.

The performance on the localization task does not exactly mirror the graph classification when looking across patients' videos. However, the trend is that the more the number of high energy frames selected, the higher the coverage that is obtained. While this may appear obvious, the performance varies across the videos with video-2,3 and 4 requiring a minimum of 9-high activation frames to achieve the same coverage obtained on video-1 with just 2-samples. High coverage means high true positive rate and indicates the model is able to accurately identify and rank frames leading to the output of the multi-instance graph classifier for abnormal graphs. Very high coverage will also allow the physician to only focus and examine the few selected localized frames as against having to review the entire video which would be much more time consuming. For example, in video-1, by selecting a sample frame from each abnormal video segment, physician will only have to review 40 frames to make their diagnosis as against the entire 14,173 of the small bowel region. On the other hand, a low coverage indicates high false positive (FP) leading to frames that do not contain any abnormality being selected as high energy frame. This will lead to increase sample frames that physician will have to review and analyse, thereby saving them less time and effort.

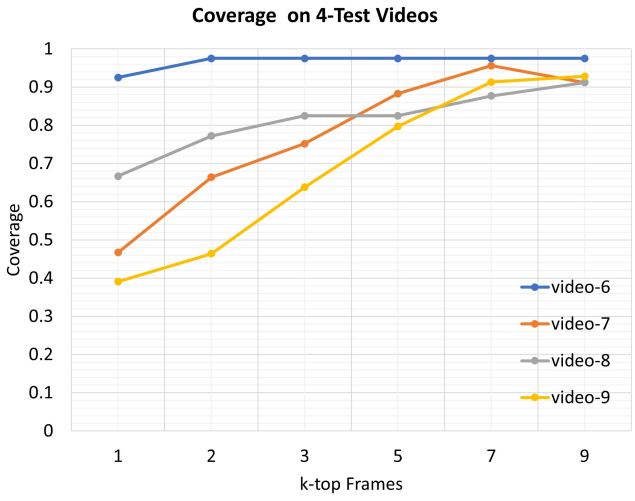


Figure 5.6: Performance on Abnormality Coverage

Table 5.3: Results of Abnormality Localization using Adaptive Temporal Pool Node Sampler

Metrics/Data	**Test Video Data**			
	Video 6	**Video 7**	**Video 8**	**Video 9**
Frames Count (T)	14,173	16,909	10,037	19,104
Abnormal Segments (N_{ab})	40	137	57	69
	Coverage ($C = \sum c_i / N_{ab}$)			
k=1	**0.925**	0.467	0.667	0.391
k=2	**0.975**	0.664	0.772	0.464
k=3	**0.975**	0.752	0.825	0.638
k=5	**0.975**	0.883	0.825	0.797
k=7	**0.975**	0.956	0.877	0.913
k=9	**0.975**	0.912	0.912	0.928

Chapter 6

Video Summarization Using Encoder-Decoder Key Frame Selection

6.1 Introduction

Following the work described in chapter 3, it is very difficult to have completely homogeneous segment can can be identified as pathology or no pathology. This is mainly because most instances, due to the unstable movement of the capsule camera causing frequent flips, there is a visual change in what the camera is seeing, which may not necessarily be a pathological change.

Our main idea in this chapter is to use Long Short Term Memory (LSTM) encoder-decoder to model selection of key frame in a short video clip. This chapter details our key frame selection technique from a non-homogeneous segment where there is a mix of normal and abnormal frames. Identifying the most representative frame within a video segment with non-homogeneous content has been considered in prior works [16, 97, 197, 198]. Our contribution in this chapter is a novel unsupervised technique for key frame selection in a VCE video segments. The input to the model in this chapter is the output of the temporal segmentation previously described in chapter 3. While the expected output of a weakly supervised temporal segmentation that we will describe in chapter 4 should be fully homogeneous class of frames, based on pathology, in some cases differentiating between normal and abnormal frames is not a very clear decision.

Key frame selection eliminates redundant and uninformative frames in a video, selecting only frames with content relevant for the physicians to make diagnostics. It is important to differential between key frame selection and video compression, since video compression [199] focuses on minimizing redundant information due to digital storage limitation, key frame selection [96, 97, 181, 182] is

focused on reducing temporal redundancy in the due to review time constraint. Either of the two tasks can leverage correlation and similarity between the frames to achieve their objective.

6.2 Related Work

In VCE literature, automating review and analysis of the long VCE videos emphasizes the task of image (frame) analysis capturing only spatial content as against actually capturing both spatial and temporal content of the data. This, potentially, can limit the clinical applicability of such models in real-life. However, few works have attempted the summarization task of VCE videos, for example [16,19,20,21,25,26,27]. The goal of the summarization task is to minimize the review time spent by the expert on the review and analysis step before making diagnosis.

Research focused on automatically generating summary for video structured data have been on for years [96, 97, 181, 182]. While the open dataset on which this techniques have been evaluated are generally short videos that have been manually segmented into fixed frame counts, our work extends these techniques to long videos with multi-stage summarization. First, following the output of the automatic temporal segmentation proposed in chapter 3, we passed the output to the summarization network to select the representative frames that captures the pathological content of the video segment. In [96], the authors proposed an encoder-decoder architecture for weakly supervised video summarization by using web-crawled data as a prior. Chen et al. [181] and [182] used reinforcement learning approach to hierarchically generate summary for videos. Other works that have also applied reinforcement learning techniques include [200] where the author proposed diversity-representativeness reward to motivate the agent to select the most representative frames in the segment. Mahsenni et al. proposed and unsupervised model using and LSTM-Encoder-Decoder architecture to generate representative frames for videos as summary.

6.3 Methodology

6.3.1 Overview of the Approach

Fig 6.1 shows the framework for our key frame selection model. From a capsule endoscopy video of the small bowel region, the first step involves extracting features from each frame of the video. Following the work presented in chapter 3, a lower dimensional representation of the feature is extracted using Principal Component Analysis (PCA) and then projected to a one-dimensional manifold space. The motivation behind this dimensionality is to minimize the computational

cost of computing the points where the statistical property of the sequence of frames changes. Following this, we employed the Pruned Exact Linear Time change point detection algorithm to determine the points that mark the change in the visual properties of the sequence. These frames are considered the change points and used in the temporal segmentation of the video into multiple homogeneous segments. The next stage uses an LSTM network for the Most Representative Frame (MRF) selection. The network selects the MRF with maximum pairwise orthogonal distance from the neighbors. The final step involve an LSTM-Encoder encoding the selected MRF to a fixed dimensional representation before using another LSTM-Decoder to reconstruct the video from the encoded representation.

The generated video summary is a sparse, diverse, and representative frames selected by the LSTM-selector for each video segment.

6.3.2 Representative Frame Selection and Encoder-Decoder Networks

As shown in 6.1, there are three (3) LSTM networks comprising the MRF selector-LSTM ($LSTM_S$), the encoder $LSTM_E$ and the decoder $LSTM_D$ networks. The CFE takes a sequence of frames $\{f_1^i, ..., f_t^i\}$ in each video clip v^i and encodes them into a fixed dimensional embedding feature. The summarizer network is an LSTM frame-selector ($LSTM_S$) with a final Sigmoid layer that scores each frame based on diversity from other members of the segment. Since each video clip contains homogeneous set of frames, we consider the frame with minimum pairwise similarity the most diverse and the most representative of the segment. The Encoder-LSTM ($LSTM_E$) takes the output of the $LSTM_S$ multiplied by the original features extracted by the CFE to generate an encoding of the most representative frame within the segment. The encoder-LSTM ($LSTM_E$) network accounts for the long term dependence between the frames and encodes the sequence into a fixed length context vector C^n. The decoder-LSTM ($LSTM_D$) takes the context vector C^n and tries to reconstruct the original input sequence of features corresponding to the input video $\hat{v}^i = \{\hat{h}_1^d, ..., \hat{h}_t^d\}$. Given a distance between the output of the $LSTM_E$ representations of the selected input frames features, our goal is to optimize the frame selector such that the similarity is minimized over training examples. The $LSTM_D$ is used to reconstruct the entire video features from the input context vector by minimizing the reconstruction loss $LSTM_D$.

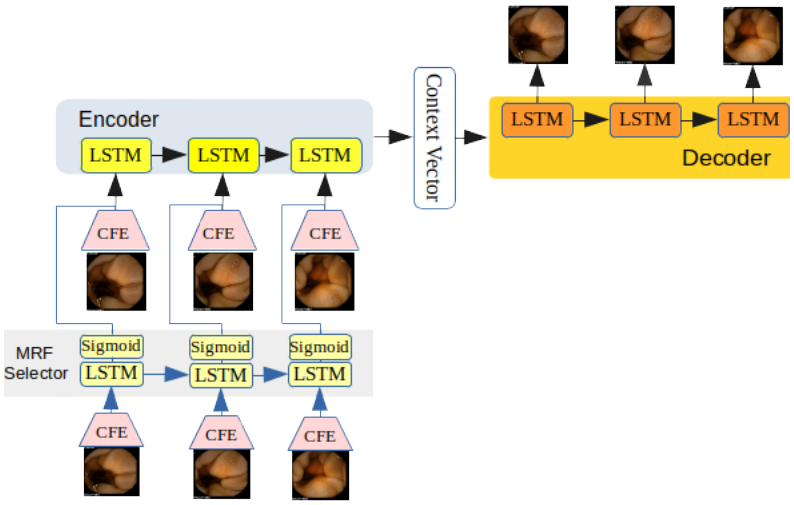


Figure 6.1: Network Architecture

Training Losses

Parameters of the $LSTM_E$ and $LSTM_D$ were optimized using the reconstruction loss function eq. 6.1 and $LSTM_S$ using the diversity loss function eq. 6.2.

$$\mathcal{L}_{recon} = \sum^{t} ||x_i - f(h_i^d)||_2; \tag{6.1}$$

where h_i^d is the latent state of the decoder network $LSTM_D$

Motivated by [200], the diversity loss ensures maximum orthogonal distance between the selected frame and the rest of the neighbourhood frames within the segment.

$$\mathcal{L}_{div} = \frac{1}{t} \sum_{i \neq j} ||h_i^s - h_j^s||_2 \tag{6.2}$$

where t is the the length of the video segment and h_i and h_j are the hidden state of $LSTM_S$. $\mathcal{L}_{div}$ ensures dissimilarity in the subset of frames selected.

Evaluation

In evaluating the performance of the model, we adopted widely used evaluation model - Coverage (C) and Compression ratio (CR). The summarized video should contain at least one representative frame from each of the abnormal findings in the original video data. The compression ratio determines the proportion of redundant frames that was eliminated from the video. We computed coverage

Video	Number of videos	Total Number of Frames
Training videos	10	421,955
Testing videos	5	43,258

Table 6.1: Data summary for training and test videos

based on diffuse bleeding, polyp and angioectasia present in our test videos.

$$\mathrm{C} = \frac{\sum_i^{N_{ab}} c_i}{N_{ab}}; c_i = \begin{cases} 1, & \text{if at least one frame from the abnormal frames is selected} \\ 0, & \text{otherwise} \end{cases} \tag{6.3}$$

$$\mathrm{CR} = \frac{t}{T} \tag{6.4}$$

where N_{ab} is the count of video segments with abnormality; t is the number of video segments generated and T is the total count of frames in the video.

6.4 Experiments and Analysis

Dataset Summary and Pre-processing

A total of 15 VCE videos were collected from 15 different patients during a clinical endoscopy procedure using the SB3 Given Imaging Pillcam capsules. The capsules were equipped with 576 x 576 pixel camera. For each complete video, the small bowel transit time corresponds to about 3.93 ± 1.43hr [35]. In order to isolate the small bowel region, each video was reviewed and annotated by two GI research scientists. After the annotation, the total number of frames in the videos is summarized in table 6.1. We randomly split the video into 70% train and 30% test set. Since there were 15 videos, we ended up with ten (10) training videos and five (5) test videos. Each frame in the video was trimmed to 500 x 500 to remove the black boundary region. Our model was developed using the Pytorch framework on NVIDIA P100 machine. The model was trained for 50 epochs with a learning rate of 0.0004. We specified a dimension of 512 for the hidden states of the LSTM models with two (2) hiddle layers while the original input features were 2048 dimension.

6.5 Results and Discussion

Fig. 3.5 in chapter 3 shows the time series plot for two selected videos with clear change in the computed manifold projection observed at different points. Figures 3.10 in chapter 3, shows a sample sequence of frames from the test videos with detected multiple change points indicated. In table 6.2, we show the compression ratio for each video in our test data as well as the coverage of the generated summary. For each video, our model achieved 100% coverage which is mainly due to the sensitivity of CPD model on the 1-D manifold representation.

Video	**Video Length**	**Compression ratio (%)**	**Coverage (%)**
Video 1	2367	81.79	100
Video 3	8577	79.88	100
Video 7	6535	83.58	100
Video 8	14532	86.62	100
Video 15	11247	80.33	100

Table 6.2: Frame Reduction Ratios on Test VCE Videos

6.6 Conclusion and Limitations

We developed a novel unsupervised summarization algorithm for Video capsule endoscopy videos. Following the techniques discussed in chapter 3, this approach provides an end-to-end unsupervised capsule endoscopy video summarization by extracting most diverse and representative frame from each homogeneous temporal segment of the video. In our experiments, We demonstrated the capability of the frame selector network to effectively pick good representatives to form the video summary based on diversity training loss function. While our model is able to reduce the number of frames that needs to be reviewed by gastroenterologist, one limitation is that the generated summary still contains frames without any abnormality or lesions from complete homogeneous segments. Since we selected representative frames from each segment, our summary will also include normal video frames that may not he useful for the physician. This is partly responsible for the very high coverage achieved by the model. Completely eliminating normal frames would save the gastroenterologist additional time and allow them to only focus on lesion containing frames. However, due to large set of possible abnormalities in the GI tract, our model generalizes better than specifically training to identify certain lesions or abnormality in the small bowel.

Chapter 7

Conclusion and Future Works

This dissertation presents our work on long video analysis with special application to video capsule endoscopy videos. The focus of the work is three fold: unsupervised and weakly supervised long temporal segmentation (Chapter 3 and 4), temporal abnormality localization (Chapter 5) and unsupervised video summarization (Chapter 6). Under each focus area, we presented a novel architectures and applied the models on long VCE videos. In this chapter, we summarize the contributions and discuss directions for future research.

7.0.1 Summary of Contribution

Unsupervised Shot Boundary Detection and Temporal Segmentation of Long Capsule Endoscopy Videos

First we developed an unsupervised temporal segmentation method for long VCE videos in linear computational time complexity. We projected the high-dimensional frame representation feature-vector of the videos to a 1-dimensional embedding space. To do this we investigated multiple embedding techniques to find the most suitable to the VCE dataset. Subsequently, we applied uni-variate time-series change point detection algorithm - Pruned Exact Linear Time algorithm (PELT) - on this lower dimensional manifold to detect the pair of frames on which the boundaries changed. We investigated this technique across multiple embedding algorithms to determine the best performing method. We conclude that While it may be easier to detect change in visual characteristics of a sequence of frames, detecting pathological events require more information for the model to do well. Detecting pathological event without any supervision is much more challenging problem. We will investigate a supervised approach to solving this problem in the next chapter. Secondly, adjusting for class imbalance is very critical to what the CNN-based feature extractor can

learn. In VCE data, there's significantly more normal frames in some videos than any abnormality, without adjusting for the class imbalance, the model will be bias towards only the normal class with little to no capability to separate the abnormal categories properly.

Weakly Supervised Temporal Segmentation of Long Capsule Endoscopy Video Using Graph Neural Network

Detecting temporal boundary that mark pathological event in the sequence of VCE video without supervision is very challenging. In this chapter, we developed a weakly-supervised disease-agnostic model for temporal segmentation of long video using graph convolutional neural network. The model is trained as a binary classifier where it learns to predict the binary category for each node in the graph based on whether the frame is abnormal or normal. This model captures the topological relationship between nodes of the graph through message passing in each layer of the network to construct an embedding for the nodes at the final layer. Our model was trained on extracted features using VGG-19 as described in chapter 3 and compared to a baseline binary CNN classifier. The GCNN model performed better than the baseline model using overall accuracy, recall and f-score for the normal class. Both model suffer significantly in properly identifying the abnormal frames with the best model yielding 14.2% precision, 7.6% recall and 9.9% f-score.

Graph Convolution Neural Network for Weakly Supervised Abnormality Localization

In this chapter, we developed a novel end-to-end temporal abnormality localization for long wireless capsule endoscopy video using only weak video level annotation. We achieved the abnormality localization in three-steps, first is the long video temporal segmentation, then video segment classification before finally localizing to the high energy frames within each segment using temporal pooling. In the classification step, our model learns to identify abnormal video segments from the aggregated embedding feature vectors using multi-instance learning framework. The localization step involves leveraging the representation of the graph to generate the high energy frames from each abnormal video segments. The end-to-end system involves, first applying an unsupervised temporal segmentation technique to partition the long WCE video into short, homogeneous segments. Thereafter, we trained a Graph Convolution Neural Network (GCNN) on each video segment to classify them into binary categories. We consider each video segment as a graph and the frame features as the nodes in the graph. We learnt a representation of the video segments using a 2-layer graph convolution. We applied attention layer on the nodes embedding before aggregating the node features at the final layer to generate the graph representation. The final layer is a multi-instance graph classifier that classifies the video segment feature vector into binary class-agnostic categories. Leveraging the parameters of the trained GCNN model, we replaced the final classifier with a temporal pool layer to select the most activated frames within the video segment which represents the highest energy elements of the graph. Similar to a video summarization model, the approach proposed in this paper for CE video abnormality localization allows physicians and gastroenterologist to quickly

focus and review identified abnormal frames that captures abnormal lesion or diseases in more detail as against having to wade through the entire long CE video with thousands of redundant normal frames.

Video Summarization Using Encoder-Decoder Key Frame Selection
Following the work on temporal segmentation, we integrated the method into a summarization model to select the most representative frame from each video segment. Our model consist of an LSTM-Encoder-Decoder architecture trained on reconstruction, sparsity and diversity losses. The model learns to select frames within the video segment that are most representative of the entire video segment. The model was evaluated based on redundancy reduction and coverage. We achieved 100% coverage for each of the test video and redundancy reduction of 80%. The proposed approach is able to reduce the number of frames that needs to be reviewed by gastroenterologist with 100% coverage in each case. However, one limitation is that the generated summary still contains frames without any abnormality or lesions. This is because some shot boundaries in the video are completely normal but was detected as change points due to change in view of the capsule camera. Completely eliminating normal frames would save the gastroenterologist additional time and allow them to only focus on lesion containing frames with little to no redundancy. However, due to large set of possible abnormalities in the GI tract, our model generalizes better than specifically training to identify certain lesions or abnormality in the small bowel.

7.1 Future Works

- Convolution feature extraction that takes into account the topological relationship between the frames as against assuming frames independence.
- Domain shift in VCE video data. With different models of the capsule cameras being regularly released such as Pillcam SB2, PillCam SB-3, PillCam Crohn's System, PillCam Colon system and PillCam UGI system. Our models were designed and tested only on PillCam SB3 with image resolution 576x576. Performance of the model may degenerate when applied on videos from a different capsule camera due to domain shift.
- Performance variability between patients and factors responsible for any performance degradation when models are tested inter-patient is an equally important research area. This is because patients have different internal GI conditions and factors.
- Model performance across different classes of abnormalities. Most work try to solve identification of single abnormality but generalizing across multiple abnormalities would be an interesting problem. Different abnormalities have different geometric, edge and coloration properties.

Abnormality such as bleeding is much easier to detected compared to mass or polyp in terms of size and how dispersed the disease is. Similarly, angioectasia can vary significantly in size and differs in color when compared to Ulcer. It could also be confused with bleeding leading to noise in labels that comes even from experts. These and many more differences between abnormalities impact the sensitivity of the models and worth investigating further.

- Investigating the impact of depth of GCNN model on representation performance on VCE videos segments.
- Occlusion is a major issue in VCE videos. There may be obstruction by food particles or shadow of other part of the GI tract itself. Estimating the tissue structure that lies under the occluded region using the surrounding tissue would be an interesting problem to consider.
- Temporal segmentation using partial or noisy annotation where the model is given partial information on the region around which there is a change point. The model can then be trained to detect the exact change point based on distance metric.
- Automated system to generate report from VCE videos using language and sequence-to-sequence would be an interesting future research area.

References

[1] P. Swain, "Wireless capsule endoscopy," *Gut*, vol. 52, no. suppl 4, pp. iv48–iv50, 2003.

[2] G. Iddan, G. Meron, A. Glukhovsky, and P. Swain, "Wireless capsule endoscopy," *Nature*, vol. 405, no. 6785, pp. 417–417, 2000.

[3] S. G. Miaou, F. L. Chang, I. K. Timotius, H. C. Huang, J. L. Su, R. S. Liao, and T. Y. Lin, "A multi-stage recognition system to detect different types of abnormality in capsule endoscope images," *Journal of Medical and Biological Engineering*, vol. 29, no. 3, pp. 114–121, 2009.

[4] S. Sainju, F. M. Bui, and K. A. Wahid, "Automated bleeding detection in capsule endoscopy videos using statistical features and region growing," *Journal of medical systems*, vol. 38, no. 4, p. 25, 2014.

[5] A. Wang, S. Banerjee, B. A. Barth, Y. M. Bhat, S. Chauhan, K. T. Gottlieb, V. Konda, J. T. Maple, F. Murad, P. R. Pfau, *et al.*, "Wireless capsule endoscopy," *Gastrointestinal endoscopy*, vol. 78, no. 6, pp. 805–815, 2013.

[6] B. Li and M. Q.-H. Meng, "Computer-based detection of bleeding and ulcer in wireless capsule endoscopy images by chromaticity moments," *Computers in biology and medicine*, vol. 39, no. 2, pp. 141–147, 2009.

[7] T. Rahim, M. A. Usman, and S. Y. Shin, "A survey on contemporary computer-aided tumor, polyp, and ulcer detection methods in wireless capsule endoscopy imaging," *Computerized Medical Imaging and Graphics*, p. 101767, 2020.

[8] J. Cohen, M. A. Safdi, S. E. Deal, T. H. Baron, A. Chak, B. Hoffman, B. C. Jacobson, K. Mergener, B. T. Petersen, J. L. Petrini, *et al.*, "Quality indicators for esophagogastroduodenoscopy," *Gastrointestinal endoscopy*, vol. 63, no. 4, pp. S10–S15, 2006.

[9] R. Trasolini and M. F. Byrne, "Artificial intelligence and deep learning for small bowel capsule endoscopy," *Digestive Endoscopy*, vol. 33, pp. 290–297, Jan 2021.

[10] J. T. Collins, A. Nguyen, and M. Badireddy, "Anatomy, abdomen and pelvis, small intestine," *StatPearls [Internet]*, 2020.

[11] S. Adewole, P. Fernandez, J. Jablonski, S. Syed, A. Copland, M. Porter, and D. Brown, "Lesion2vec: Deep metric learning for few shot multiple lesions recognition in wireless capsule endoscopy," *arXiv preprint arXiv:2101.04240*, 2021.

[12] B. Lewis, G. Eisen, and S. Friedman, "A pooled analysis to evaluate results of capsule endoscopy trials," *Endoscopy*, vol. 37, no. 10, pp. 960–965, 2005.

[13] R. Nawarathna, J. Oh, J. Muthukudage, W. Tavanapong, J. Wong, P. C. De Groen, and S. J. Tang, "Abnormal image detection in endoscopy videos using a filter bank and local binary patterns," *Neurocomputing*, vol. 144, pp. 70–91, 2014.

[14] S. Adewole, M. Yeghyayan, D. Hyatt, L. Ehsan, J. Jablonski, A. Copland, S. Syed, and D. Brown, "Deep learning methods for anatomical landmark detection in video capsule endoscopy images," in *Proceedings of the Future Technologies Conference*, pp. 426–434, Springer, 2020.

[15] Y. Gao, W. Lu, X. Si, and Y. Lan, "Deep model-based semi-supervised learning way for outlier detection in wireless capsule endoscopy images," *IEEE Access*, vol. 8, pp. 81621–81632, 2020.

[16] J. Chen, Y. Zou, and Y. Wang, "Wireless capsule endoscopy video summarization: a learning approach based on siamese neural network and support vector machine," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 1303–1308, IEEE, 2016.

[17] Q. Zhao and M. Q.-H. Meng, "An abnormality based wce video segmentation strategy," in *2010 IEEE International Conference on Automation and Logistics*, pp. 565–570, IEEE, 2010.

[18] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," *SIGMOD Rec.*, vol. 30, pp. 37–46, May 2001.

[19] S. Tsevas, D. K. Iakovidis, D. Maroulis, and E. Pavlakis, "Automatic frame reduction of wireless capsule endoscopy video," in *2008 8th IEEE International Conference on BioInformatics and BioEngineering*, pp. 1–6, IEEE, 2008.

[20] D. K. Iakovidis, S. Tsevas, and A. Polydorou, "Reduction of capsule endoscopy reading times by unsupervised image mining," *Computerized Medical Imaging and Graphics*, vol. 34, no. 6, pp. 471–478, 2010.

[21] A. Z. Emam, Y. A. Ali, and M. M. B. Ismail, "Adaptive features extraction for capsule endoscopy (ce) video summarization," in *International Conference on Computer Vision and Image Analysis Applications*, pp. 1–5, IEEE, 2015.

[22] I. Mehmood, M. Sajjad, and S. W. Baik, "Video summarization based tele-endoscopy: a service to efficiently manage visual data generated during wireless capsule endoscopy procedure," *Journal of medical systems*, vol. 38, no. 9, p. 109, 2014.

[23] S. Tsevas, D. Iakovidis, D. Maroulis, E. Pavlakis, and A. Polydorou, "Non-negative matrix factorization for endoscopic video summarization," in *Hellenic Conference on Artificial Intelligence*, pp. 425–430, Springer, 2008.

[24] J. S. Huo, Y. X. Zou, and L. Li, "An advanced wce video summary using relation matrix rank," in *Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics*, pp. 675–678, IEEE, 2012.

[25] B. Li, M. Q.-H. Meng, and Q. Zhao, "Wireless capsule endoscopy video summary," in *2010 IEEE International Conference on Robotics and Biomimetics*, pp. 454–459, IEEE, 2010.

[26] A. Mohammed, S. Yildirim, M. Pedersen, Ø. Hovde, and F. Cheikh, "Sparse coded handcrafted and deep features for colon capsule video summarization," in *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 728–733, IEEE, 2017.

[27] M. M. B. Ismail, O. Bchir, and A. Z. Emam, "Endoscopy video summarization based on unsupervised learning and feature discrimination," in *2013 Visual Communications and Image Processing (VCIP)*, pp. 1–6, IEEE, 2013.

[28] R. Eliakim, "Video capsule endoscopy of the small bowel," *Current opinion in gastroenterology*, vol. 29, no. 2, pp. 133–139, 2013.

[29] S. Hwang and M. E. Celebi, "Polyp detection in wireless capsule endoscopy videos based on image segmentation and geometric feature," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 678–681, IEEE, 2010.

[30] H. Vu, T. Echigo, R. Sagawa, K. Yagi, M. Shiba, K. Higuchi, T. Arakawa, and Y. Yagi, "Detection of contractions in adaptive transit time of the small bowel from wireless capsule endoscopy videos," *Computers in biology and medicine*, vol. 39, no. 1, pp. 16–26, 2009.

[31] Y.-j. Chen, W. Yasen, J. Lee, D. Lee, and Y. Kim, "Developing assessment system for wireless capsule endoscopy videos based on event detection," in *Medical Imaging 2009: Computer-Aided Diagnosis*, vol. 7260, p. 72601G, International Society for Optics and Photonics, 2009.

[32] M. Mackiewicz, J. Berens, and M. Fisher, "Wireless capsule endoscopy color video segmentation," *IEEE Transactions on Medical Imaging*, vol. 27, no. 12, pp. 1769–1781, 2008.

[33] Y. Chen and J. Lee, "A review of machine-vision-based analysis of wireless capsule endoscopy video," *Diagnostic and therapeutic endoscopy*, vol. 2012, 2012.

[34] P. Sivakumar and B. M. Kumar, "A novel method to detect bleeding frame and region in wireless capsule endoscopy video," *Cluster Computing*, vol. 22, no. 5, pp. 12219–12225, 2019.

[35] S. Iobagiu, L. Ciobanu, and O. Pascu, "Colon capsule endoscopy: a new method of investigating the large bowel," *Journal of Gastrointestinal and Liver Diseases*, vol. 17, no. 3, pp. 347–352, 2008.

[36] P. W. Mewes, P. Rennert, A. L. Juloski, A. Lalande, E. Angelopoulou, R. Kuth, and J. Hornegger, "Semantic and topological classification of images in magnetically guided capsule endoscopy," in *Medical Imaging 2012: Computer-Aided Diagnosis*, vol. 8315, p. 83151A, International Society for Optics and Photonics, 2012.

[37] V. Kodogiannis and J. N. Lygouras, "Neuro-fuzzy classification system for wireless-capsule endoscopic images," *International Journal of Electrical, Computer, and Systems Engineering*, vol. 2, no. 1, pp. 55–63, 2008.

[38] A. Van Gossum, M. Munoz-Navas, I. Fernandez-Urien, C. Carretero, G. Gay, M. Delvaux, M. G. Lapalus, T. Ponchon, H. Neuhaus, M. Philipper, *et al.*, "Capsule endoscopy versus colonoscopy for the detection of polyps and cancer," *New England Journal of Medicine*, vol. 361, no. 3, pp. 264–270, 2009.

[39] A. V. Mamonov, I. N. Figueiredo, P. N. Figueiredo, and Y.-H. R. Tsai, "Automated polyp detection in colon capsule endoscopy," *IEEE transactions on medical imaging*, vol. 33, no. 7, pp. 1488–1502, 2014.

[40] Y. Yuan, B. Li, and M. Q.-H. Meng, "Improved bag of feature for automatic polyp detection in wireless capsule endoscopy images," *IEEE Transactions on automation science and engineering*, vol. 13, no. 2, pp. 529–535, 2015.

[41] A. Tsuboi, S. Oka, K. Aoyama, H. Saito, T. Aoki, A. Yamada, T. Matsuda, M. Fujishiro, S. Ishihara, M. Nakahori, *et al.*, "Artificial intelligence using a convolutional neural network for automatic detection of small-bowel angioectasia in capsule endoscopy images," *Digestive Endoscopy*, vol. 32, no. 3, pp. 382–390, 2020.

[42] K. Pogorelov, O. Ostroukhova, M. Jeppsson, H. Espeland, C. Griwodz, T. de Lange, D. Johansen, M. Riegler, and P. Halvorsen, "Deep learning and hand-crafted feature based approaches for polyp detection in medical videos," in *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 381–386, IEEE, 2018.

[43] F. Deeba, S. K. Mohammed, F. M. Bui, and K. A. Wahid, "A saliency-based unsupervised method for angiectasia detection in endoscopic video frames," *Journal of Medical and Biological Engineering*, vol. 38, no. 2, pp. 325–335, 2018.

[44] Y. Yuan, J. Wang, B. Li, and M. Q.-H. Meng, "Saliency based ulcer detection for wireless capsule endoscopy diagnosis," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 2046–2057, 2015.

[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[46] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, 2014.

[47] A. F. Smeaton, P. Over, and A. R. Doherty, "Video shot boundary detection: Seven years of trecvid activity," *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 411–418, 2010.

[48] A. Hanjalic, "Shot-boundary detection: unraveled and resolved?," *IEEE transactions on circuits and systems for video technology*, vol. 12, no. 2, pp. 90–105, 2002.

[49] R. W. Lienhart, "Comparison of automatic shot boundary detection algorithms," in *Storage and retrieval for image and video databases VII*, vol. 3656, pp. 290–301, International Society for Optics and Photonics, 1998.

[50] J. S. Boreczky and L. A. Rowe, "Comparison of video shot boundary detection techniques," *Journal of Electronic Imaging*, vol. 5, no. 2, pp. 122–128, 1996.

[51] J. Mas and G. Fernandez, "Video shot boundary detection based on color histogram.," in *TRECVID*, 2003.

[52] P. P. Mohanta, S. K. Saha, and B. Chanda, "A model-based shot boundary detection technique using frame transition parameters," *IEEE Transactions on multimedia*, vol. 14, no. 1, pp. 223–233, 2011.

[53] Y. Gong and X. Liu, "Video summarization using singular value decomposition," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, vol. 2, pp. 174–180, IEEE, 2000.

[54] W.-S. Chu, Y. Song, and A. Jaimes, "Video co-summarization: Video summarization by visual co-occurrence," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3584–3592, 2015.

[55] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 907–919, 2005.

[56] M. Vasilakakis, D. K. Iakovidis, E. Spyrou, and A. Koulaouzidis, "Weakly-supervised lesion detection in video capsule endoscopy based on a bag-of-colour features model," in *International workshop on computer-assisted and robotic endoscopy*, pp. 96–103, Springer, 2016.

[57] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," *arXiv*, Dec 2016.

[58] Z. Shou, D. Wang, and S.-F. Chang, "Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs," *arXiv*, Jan 2016.

[59] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, "Temporal Action Detection with Structured Segment Networks," *arXiv*, Apr 2017.

[60] Nguyen, Phuc, Liu, Ting, Prasad, Gautam, Han, and Bohyung, *Weakly Supervised Action Localization by Sparse Temporal Pooling Network*. Chichester, England, UK: Wiley, Dec 2017.

[61] G. Gkioxari and J. Malik, "Finding Action Tubes," *arXiv*, Nov 2014.

[62] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei, "End-to-end Learning of Action Detection from Frame Glimpses in Videos," *arXiv*, Nov 2015.

[63] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S.-F. Chang, "CDC: Convolutional-De-Convolutional Networks for Precise Temporal Action Localization in Untrimmed Videos," *arXiv*, Mar 2017.

[64] T. Lin, X. Zhao, and Z. Shou, "Single Shot Temporal Action Detection," *arXiv*, Oct 2017.

[65] J. Gao, Z. Yang, and R. Nevatia, "Cascaded Boundary Regression for Temporal Action Detection," *arXiv*, May 2017.

[66] L. Wang, Y. Qiao, X. Tang, and L. Van Gool, "Actionness Estimation Using Hybrid Fully Convolutional Networks," *arXiv*, Apr 2016.

[67] Z. Shou, H. Gao, L. Zhang, K. Miyazawa, and S.-F. Chang, "Autoloc: Weakly-supervised temporal action localization in untrimmed videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[68] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition," *arXiv*, Aug 2016.

[69] L. Wang, Y. Xiong, D. Lin, and L. Van Gool, "UntrimmedNets for Weakly Supervised Action Recognition and Detection," *arXiv*, Mar 2017.

[70] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National science review*, vol. 5, no. 1, pp. 44–53, 2018.

[71] M. Rochan and Y. Wang, "Weakly supervised localization of novel objects using appearance transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[72] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1049–1058, 2016.

[73] M. Shi, H. Caesar, and V. Ferrari, "Weakly Supervised Object Localization Using Things and Stuff Transfer," *arXiv*, Mar 2017.

[74] A. Khoreva, R. Benenson, E. Ilg, T. Brox, and B. Schiele, "Lucid data dreaming for object tracking," in *The DAVIS challenge on video object segmentation*, 2017.

[75] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille, "Weakly- and Semi-Supervised Learning of a DCNN for Semantic Image Segmentation," *arXiv*, Feb 2015.

[76] Z. Shen, J. Li, Z. Su, M. Li, Y. Chen, Y.-G. Jiang, and X. Xue, "Weakly supervised dense video captioning," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1916–1924, 2017.

[77] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual translation embedding network for visual relation detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5532–5540, 2017.

[78] K. Soomro, H. Idrees, and M. Shah, "Action Localization in Videos Through Context Walk," 2015. [Online; accessed 22. Jul. 2021].

[79] S. Ma, L. Sigal, and S. Sclaroff, "Learning Activity Progression in LSTMs for Activity Detection and Early Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1942–1950, IEEE, Jun 2016.

[80] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. C. Niebles, "SST: Single-Stream Temporal Action Proposals," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6373–6382, IEEE, Jul 2017.

[81] V. Escorcia, F. C. Heilbron, J. C. Niebles, and B. Ghanem, "DAPs: Deep Action Proposals for Action Understanding," in *Computer Vision – ECCV 2016*, pp. 768–784, Cham, Switzerland: Springer, Sep 2016.

[82] K. K. Singh and Y. J. Lee, "Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-supervised Object and Action Localization," *arXiv*, Apr 2017.

[83] C. Sun, S. Shetty, R. Sukthankar, and R. Nevatia, "Temporal Localization of Fine-Grained Actions in Videos by Domain Transfer from Web Images," *arXiv*, Apr 2015.

[84] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free? - weakly-supervised learning with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[85] W. Bae, J. Noh, and G. Kim, "Rethinking class activation mapping for weakly supervised object localization," in *European Conference on Computer Vision*, pp. 618–634, Springer, 2020.

[86] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, pp. 31–71, Jan 1997.

[87] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," 2016.

[88] Z. Jie, Y. Wei, X. Jin, J. Feng, and W. Liu, "Deep Self-Taught Learning for Weakly Supervised Object Localization," *arXiv*, Apr 2017.

[89] V. Kantorov, M. Oquab, M. Cho, and I. Laptev, "ContextLocNet: Context-Aware Deep Network Models for Weakly Supervised Localization," *arXiv*, Sep 2016.

[90] P. Tang, X. Wang, X. Bai, and W. Liu, "Multiple Instance Detection Network with Online Instance Classifier Refinement," *arXiv*, Apr 2017.

[91] R. Girshick, "Fast R-CNN," *arXiv*, Apr 2015.

[92] Y. Xu, C. Zhang, Z. Cheng, J. Xie, Y. Niu, S. Pu, and F. Wu, "Segregated Temporal Assembly Recurrent Networks for Weakly Supervised Multiple Action Detection," *arXiv*, Nov 2018.

[93] N. Vasconcelos and A. Lippman, "Statistical models of video structure for content analysis and characterization," *IEEE Trans. Image Process.*, vol. 9, pp. 3–19, Jan 2000.

[94] C. Loukas, "Video content analysis of surgical procedures," *Surg. Endosc.*, vol. 32, pp. 553–568, Feb 2018.

[95] S. E. de Avila, A. da_Luz Jr, A. d. A. Araújo, and M. Cord, "Vsumm: An approach for automatic video summarization and quantitative evaluation," in *2008 XXI Brazilian Symposium on Computer Graphics and Image Processing*, pp. 103–110, IEEE, 2008.

[96] S. Cai, W. Zuo, L. S. Davis, and L. Zhang, "Weakly-supervised video summarization using variational encoder-decoder and web prior," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 184–200, 2018.

[97] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial lstm networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 202–211, 2017.

[98] T. Han, W. Xie, and A. Zisserman, "Video representation learning by dense predictive coding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

[99] J. Zhu, Z. Zhu, and W. Zou, "End-to-end video-level representation learning for action recognition," in *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 645–650, IEEE, 2018.

[100] F. Jiang, J. Yuan, S. A. Tsaftaris, and A. K. Katsaggelos, "Anomalous video event detection using spatiotemporal context," *Computer Vision and Image Understanding*, vol. 115, no. 3, pp. 323–333, 2011.

[101] D. Brezeale and D. J. Cook, "Automatic video classification: A survey of the literature," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 3, pp. 416–430, 2008.

[102] C. G. Snoek and M. Worring, *Concept-based video retrieval*. Now Publishers Inc, 2009.

[103] N. Dimitrova, H.-J. Zhang, B. Shahraray, I. Sezan, T. Huang, and A. Zakhor, "Applications of video-content analysis and retrieval," *IEEE multimedia*, vol. 9, no. 3, pp. 42–55, 2002.

[104] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pp. 321–330, 2006.

[105] G. Geisler and G. Marchionini, "The open video project: research-oriented digital video repository," in *Proceedings of the fifth ACM conference on Digital libraries*, pp. 258–259, 2000.

[106] H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah, "The thumos challenge on action recognition for videos "in the wild"," *Computer Vision and Image Understanding*, vol. 155, pp. 1–23, 2017.

[107] S.-J. Wang, S. Wu, X. Qian, J. Li, and X. Fu, "A main directional maximal difference analysis for spotting facial movements from long-term videos," *Neurocomputing*, vol. 230, pp. 382–389, 2017.

[108] G. Medioni, I. Cohen, F. Bremond, S. Hongeng, and R. Nevatia, "Event detection and analysis from video streams," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, pp. 873–889, Aug 2001.

[109] J. Luiten, I. E. Zulfikar, and B. Leibe, "Unovost: Unsupervised offline video object segmentation and tracking," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2000–2009, 2020.

[110] M. J. Halvey and M. T. Keane, "Analysis of online video search and sharing," in *Proceedings of the eighteenth conference on Hypertext and hypermedia*, pp. 217–226, 2007.

[111] R. Killick, P. Fearnhead, and I. A. Eckley, "Optimal detection of changepoints with a linear computational cost," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1590–1598, 2012.

[112] J. Chen and A. K. Gupta, *Parametric Statistical Change Point Analysis*. Basel, Switzerland: Birkhäuser, 2012.

[113] Z. Harchaoui, E. Moulines, and F. R. Bach, "Kernel change-point analysis," in *Advances in neural information processing systems*, pp. 609–616, 2009.

[114] M. Bosc, F. Heitz, J.-P. Armspach, I. Namer, D. Gounot, and L. Rumbach, "Automatic change detection in multimodal serial mri: application to multiple sclerosis lesion evolution," *NeuroImage*, vol. 20, no. 2, pp. 643–656, 2003.

[115] R. Malladi, G. P. Kalamangalam, and B. Aazhang, "Online bayesian change point detection algorithms for segmentation of epileptic activity," in *2013 Asilomar Conference on Signals, Systems and Computers*, pp. 1833–1837, IEEE, 2013.

[116] M. Staudacher, S. Telser, A. Amann, H. Hinterhuber, and M. Ritsch-Marte, "A new method for change-point detection developed for on-line analysis of the heart beat variability during sleep," *Physica A: Statistical Mechanics and its Applications*, vol. 349, no. 3-4, pp. 582–596, 2005.

[117] P. Yang, G. Dumont, and J. M. Ansermino, "Adaptive change detection in heart rate trend monitoring in anesthetized children," *IEEE transactions on biomedical engineering*, vol. 53, no. 11, pp. 2211–2219, 2006.

[118] J.-F. Ducré-Robitaille, L. A. Vincent, and G. Boulet, "Comparison of techniques for detection of discontinuities in temperature series," *International Journal of Climatology: A Journal of the Royal Meteorological Society*, vol. 23, no. 9, pp. 1087–1101, 2003.

[119] J. Reeves, J. Chen, X. L. Wang, R. Lund, and Q. Q. Lu, "A review and comparison of changepoint detection techniques for climate data," *Journal of applied meteorology and climatology*, vol. 46, no. 6, pp. 900–915, 2007.

[120] D. Rybach, C. Gollan, R. Schluter, and H. Ney, "Audio segmentation for speech recognition using segment features," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4197–4200, IEEE, 2009.

[121] M. F. R. Chowdhury, S.-A. Selouani, and D. O'Shaughnessy, "Bayesian on-line spectral change point detection: a soft computing approach for on-line asr," *International Journal of Speech Technology*, vol. 15, no. 1, pp. 5–23, 2012.

[122] S. Chib, "Estimation and comparison of multiple change-point models," *Journal of econometrics*, vol. 86, no. 2, pp. 221–241, 1998.

[123] H. Cho and P. Fryzlewicz, "Multiple-change-point detection for high dimensional time series via sparsified binary segmentation," *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pp. 475–507, 2015.

[124] I. Cleland, M. Han, C. Nugent, H. Lee, S. McClean, S. Zhang, and S. Lee, "Evaluation of prompted annotation of activity data recorded from a smart phone," *Sensors*, vol. 14, no. 9, pp. 15861–15879, 2014.

[125] A. Wald, *Sequential analysis*. Courier Corporation, 2004.

[126] S. Aminikhanghahi and D. J. Cook, "A survey of methods for time series change point detection," *Knowledge and information systems*, vol. 51, no. 2, pp. 339–367, 2017.

[127] H. V. Poor and O. Hadjiliadis, *Quickest detection*. Cambridge University Press, 2008.

[128] A. Tartakovsky, I. Nikiforov, and M. Basseville, *Sequential analysis: Hypothesis testing and changepoint detection*. CRC Press, 2014.

[129] F. D. la Torre Frade, J. Campoy, Z. Ambadar, and J. F. Cohn, "Temporal segmentation of facial behavior," in *Proceedings of (ICCV) International Conference on Computer Vision*, October 2007.

[130] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282–289, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., Jun 2001.

[131] M. Csörgő and L. Horváth, "20 nonparametric methods for changepoint problems," *Handbook of statistics*, vol. 7, pp. 403–425, 1988.

[132] D. Gong, G. Medioni, S. Zhu, and X. Zhao, "Kernelized temporal cut for online temporal segmentation and recognition," in *European Conference on Computer Vision*, pp. 229–243, Springer, 2012.

[133] R. P. Adams and D. J. MacKay, "Bayesian online changepoint detection," *arXiv preprint arXiv:0710.3742*, 2007.

[134] A. Lung-Yut-Fong, C. Lévy-Leduc, and O. Cappé, "Homogeneity and change-point detection tests for multivariate data using rank statistics," *Journal de la Société Française de Statistique*, vol. 156, no. 4, pp. 133–162, 2015.

[135] F. Desobry, M. Davy, and C. Doncarli, "An online kernel change detection algorithm," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 2961–2974, 2005.

[136] M. J. Rasch, *Analysis of neural signals: Interdependence, information coding, and relation to network models*. PhD thesis, Graz University of Technology Graz, Austria, 2008.

[137] C. Truong, L. Oudre, and N. Vayatis, "Selective review of offline change point detection methods," *arXiv*, Jan 2018.

[138] C. Rohrbeck, "Detection of changes in variance using binary segmentation and optimal partitioning," 2013. [Online; accessed 17. Jul. 2021].

[139] J. Chen, V. Kellokumpu, G. Zhao, and M. Pietikäinen, "Rlbp: Robust local binary pattern.," in *BMVC*, 2013.

[140] T. Lindeberg, "Scale invariant feature transform," *DIVA*, vol. 7, no. 5, p. 10491, 2012.

[141] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," in *Computer Vision – ECCV 2006*, pp. 404–417, Berlin, Germany: Springer, May 2006.

[142] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *MIR '07: Proceedings of the international workshop on Workshop on multimedia information retrieval*, pp. 197–206, New York, NY, USA: Association for Computing Machinery, Sep 2007.

[143] M. Chen, X. Shi, Y. Zhang, D. Wu, and M. Guizani, "Deep features learning for medical image analysis with convolutional autoencoder neural network," *IEEE Transactions on Big Data*, 2017.

[144] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[145] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv*, Sep 2014.

[146] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," *arXiv*, Apr 2014.

[147] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv*, Dec 2015.

[148] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," *arXiv*, Feb 2016.

[149] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *arXiv*, Aug 2016.

[150] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for MobileNetV3," *arXiv*, May 2019.

[151] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[152] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[153] Feb 2009. [Online; accessed 3. Aug. 2021].

[154] A. Ng *et al.*, "Sparse autoencoder," *CS294A Lecture notes*, vol. 72, no. 2011, pp. 1–19, 2011.

[155] M. Tschannen, O. Bachem, and M. Lucic, "Recent advances in autoencoder-based representation learning," *arXiv preprint arXiv:1812.05069*, 2018.

[156] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[157] E. H. Spriggs, F. De La Torre, and M. Hebert, "Temporal segmentation and activity classification from first-person sensing," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 17–24, IEEE, 2009.

[158] T. Grundy, R. Killick, and G. Mihaylov, "High-dimensional changepoint detection via a geometrically inspired mapping," *Statist. Comput.*, vol. 30, pp. 1155–1166, Jul 2020.

[159] G. D. Wambui, Gichuhi. A. Waititu, and A. Wanjoya, "The Power of the Pruned Exact Linear Time(PELT) Test in Multiple Changepoint Detection," *undefined*, 2015.

[160] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *European conference on computer vision*, pp. 540–555, Springer, 2014.

[161] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[162] S. Zhang, H. Tong, J. Xu, and R. Maciejewski, "Graph convolutional networks: a comprehensive review," *Comput. Soc. Netw.*, vol. 6, pp. 1–23, Dec 2019.

[163] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive Representation Learning on Large Graphs," *arXiv*, Jun 2017.

[164] F. Wu, A. Souza, T. Zhang, C. Fifty, T. Yu, and K. Weinberger, "Simplifying Graph Convolutional Networks," in *International Conference on Machine Learning*, pp. 6861–6871, PMLR, May 2019.

[165] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *arXiv preprint arXiv:1402.1128*, 2014.

[166] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," *arXiv preprint arXiv:1409.2329*, 2014.

[167] W. W. Lo, S. Layeghy, M. Sarhan, M. Gallagher, and M. Portmann, "E-graphsage: A graph neural network based intrusion detection system," *arXiv preprint arXiv:2103.16329*, 2021.

[168] J. Liu, G. P. Ong, and X. Chen, "Graphsage-based traffic speed forecasting for segment network with sparse data," *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[169] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The Graph Neural Network Model," *IEEE Trans. Neural Networks*, vol. 20, pp. 61–80, Dec 2008.

[170] M. Bianchini, M. Maggini, L. Sarti, and F. Scarselli, "Recursive neural networks learn to localize faces," *Pattern Recognit. Lett.*, vol. 26, pp. 1885–1895, Sep 2005.

[171] R. Zeng, W. Huang, M. Tan, Y. Rong, P. Zhao, J. Huang, and C. Gan, "Graph Convolutional Networks for Temporal Action Localization," 2019. [Online; accessed 11. Jul. 2021].

[172] S. Yan, Y. Xiong, and D. Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition," Apr 2018. [Online; accessed 11. Jul. 2021].

[173] P. Ghosh, Y. Yao, L. Davis, and A. Divakaran, "Stacked Spatio-Temporal Graph Convolutional Networks for Action Segmentation," 2020. [Online; accessed 11. Jul. 2021].

[174] D. Huang, P. Chen, R. Zeng, Q. Du, M. Tan, and C. Gan, "Location-Aware Graph Convolutional Networks for Video Question Answering," *AAAI*, vol. 34, pp. 11021–11028, Apr 2020.

[175] F. Manessi, A. Rozza, and M. Manzo, "Dynamic graph convolutional networks," *Pattern Recognit.*, vol. 97, p. 107000, Jan 2020.

[176] L. Yao, C. Mao, and Y. Luo, "Graph Convolutional Networks for Text Classification," *AAAI*, vol. 33, pp. 7370–7377, Jul 2019.

[177] J. Gao, T. Zhang, and C. Xu, "Graph Convolutional Tracking," 2019. [Online; accessed 11. Jul. 2021].

[178] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei, "End-to-end learning of action detection from frame glimpses in videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2678–2687, 2016.

[179] B. Ghanem, J. C. Niebles, C. Snoek, F. C. Heilbron, H. Alwassel, R. Khrisna, V. Escorcia, K. Hata, and S. Buch, "Activitynet challenge 2017 summary," *arXiv preprint arXiv:1710.08011*, 2017.

[180] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 961–970, 2015.

[181] Y. Chen, L. Tao, X. Wang, and T. Yamasaki, "Weakly supervised video summarization by hierarchical reinforcement learning," in *Proceedings of the ACM Multimedia Asia*, pp. 1–6, 2019.

[182] Z. Li and L. Yang, "Weakly supervised deep reinforcement learning for video summarization with semantically meaningful reward," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3239–3247, 2021.

[183] M. Hagenbuchner, A. Sperduti, and A. C. Tsoi, "A self-organizing map for adaptive processing of structured data," *IEEE Trans. Neural Networks*, vol. 14, pp. 491–505, May 2003.

[184] A. Sperduti and A. Starita, "Supervised neural networks for the classification of structures," *IEEE Trans. Neural Networks*, vol. 8, pp. 714–735, May 1997.

[185] A. C. Tsoi, G. Morini, F. Scarselli, M. Hagenbuchner, and M. Maggini, "Adaptive ranking of web pages," *ResearchGate*, pp. 356–365, Jan 2003.

[186] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, pp. 604–632, Sep 1999.

[187] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[188] P. Nguyen, T. Liu, G. Prasad, and B. Han, "Weakly supervised action localization by sparse temporal pooling network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6752–6761, 2018.

[189] S. Paul, S. Roy, and A. K. Roy-Chowdhury, "W-TALC: Weakly-supervised Temporal Activity Localization and Classification," *arXiv*, Jul 2018.

[190] M. Rashid, H. Kjellstrom, and Y. J. Lee, "Action graphs: Weakly-supervised action localization with graph convolution networks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[191] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329–353, 2018.

[192] Z.-H. Zhou, "Multi-instance learning: A survey," *Department of Computer Science & Technology, Nanjing University, Tech. Rep*, vol. 1, 2004.

[193] J. Foulds and E. Frank, "A review of multi-instance learning assumptions," *The knowledge engineering review*, vol. 25, no. 1, pp. 1–25, 2010.

[194] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," *Advances in neural information processing systems*, pp. 570–576, 1998.

[195] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3d action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1647–1656, 2017.

[196] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), pp. 8024–8035, Curran Associates, Inc., 2019.

[197] B. Gong, W.-L. Chao, K. Grauman, and F. Sha, "Diverse sequential subset selection for supervised video summarization," *Advances in neural information processing systems*, vol. 27, pp. 2069–2077, 2014.

[198] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *European conference on computer vision*, pp. 766–782, Springer, 2016.

[199] M. Rabbani and P. W. Jones, *Digital image compression techniques*, vol. 7. SPIE press, 1991.

[200] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.