

CLIP-tology: Classifying Images of Unseen Classes with Zero-Shot Learning

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Nikash Sethi
Spring, 2021

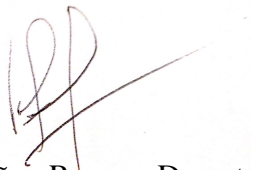
On my honor as a University Student, I have neither given nor received
unauthorized aid on this assignment as defined by the Honor Guidelines
for Thesis-Related Assignments

Signature
Nikash Sethi



Date: May 6, 2021

Approved
Vicente Ordoñez-Roman, Department of Computer Science



Date: May 6, 2021

The fields of artificial intelligence and computer vision have experienced revolutionary innovation in the past few years. Perhaps the most robust and applicable application within this field is zero-shot learning, where models are tasked with classifying categories of images they have never seen before. My technical work over the course of this semester focuses on improving the performance and usability of a state-of-the-art model for zero-shot learning called CLIP (Contrastive Language-Image Pretraining), recently released by OpenAI. In this paper, I explore in detail the success of different types of prompt engineering, which is used to provide auxiliary context to the model. Specifically, I explore the capabilities of CLIP by creating textual prompts that are augmented with external knowledge such as the use of hyponyms and entry-level categories. In this paper, I demonstrate that these external types of information lead to improved accuracy, which provides initial recommendations about how transfer-learning by prompt engineering might look in the future. I also explore the use case of CLIP in predicting an encoder from Wikipedia text samples of classes belonging to a dataset of birds, and using this encoder to learn attributes of classes of unseen images. Ultimately, my work advances a rapidly-progressing field within computer vision and points toward key areas of future innovation in zero-shot learning.

INTRODUCTION AND RELATED WORK

Computer vision technologies have achieved incredibly high accuracies in classifying known images. State of the art models today are excellent at identifying features of certain images and corresponding them with images from classes they have been trained on in the past. That said, this problem space does not solve many important real-world applications, as it is impossible to train models on images of every possible class. Zero-shot learning is a particularly interesting field within computer vision where models must classify categories of images that

they have not been trained upon at all (Wang et al., 2019, p. 1). Traditional computer vision relies on training models on large sets of images, to extract and represent features that they can later match with novel images. In zero-shot learning, though, test images may be from classes that models have never seen before, which presents a unique and open-ended problem that more closely encompasses real-world applications of computer vision.

Introducing Zero-Shot Learning

Zero-shot learning was first proposed as a technique for solving the problem of data scarcity (Li et al., 2019, 8690). This relatively novel field accounts for the lack of generalizability in classical computer vision classification tasks by no longer requiring models to be trained on images from each class. Instead, zero-shot models relying on attributes or other information that describes classes to which images can belong (Xian et al., 2019, p. 2253). Wang et al. (2019) find that zero-shot learning models are perfect for scenarios where target classes are large, rare, changing, or bottlenecked in amount of labeled data for training (p. 2).

The field of zero-shot learning has tremendous applications that extend to a variety of fields. Given that zero-shot learning is a more generalized and robust form of computer vision classification tasks, the problem spaces that zero-shot learning can be applied to mirror and expand on those already being solved with many current computer vision approaches.

Applications of zero-shot learning include computer vision in the form of videos, natural language processing, and more specific tasks like medical imaging (Wang et al., 2019, p. 27). Ultimately, zero-shot learning has the opportunity of revolutionizing a vast range of applications, which is why many researchers have started to focus in on innovating and progressing the new field.

Forms of Auxiliary Information for Zero-Shot Learning

For zero-shot learning to succeed in identifying images of unseen classes, auxiliary information is always necessary to convey properties that models would have learned had they been trained on these unseen classes (Wang et al., 2019, p. 4). Many current zero-shot learning approaches require structured lists of attributes that represent important features and qualities of image classes (Xian et al., 2019, p. 2253). With this information, models can learn to recognize and locate novel object instances with no prior training examples (Li et al., 2019, p. 8690). Structured lists of attributes are often compiled by datasets in the computer vision space that specifically look to solve instances of zero-shot learning problems. This is the case with the Animals with Attributes dataset that Xian et al. (2019) proposed, which has become a key benchmark dataset for much zero-shot learning innovation in the recent years (p. 2251).

However, very few large-scale datasets provide these attribute lists with their images, as such attribute lists are difficult to collect and standardize. Further, when classifying objects of many visually similar classes, as the case with datasets of birds or flowers, image instances across classes might have only subtle differences (Reed et al., 2016, p. 50). Using natural language textual descriptions to describe classes offers far more flexibility in conveying important features in classes without having to standardize attributes. These textual descriptions are much more common and attainable than lists of attributes, which is a necessity for scalable zero-shot learning innovation. According to Reed et al. (2016), these natural language texts are especially advantageous on datasets where classes have subtle differences (p. 50). In these scenarios, attribute lists would not be able to capture and emphasize slight visual differences between different categories of images. Instead, detailed textual descriptions can succeed in highlighting specific and detailed differences that heavily structured attribute lists could never be

designed to include. Reed et al. demonstrate that natural language encodes only the salient visual aspects for distinguishing categories without including attributes that are shared among many classes, which is advantageous to zero-shot learning models (p. 49).

Many zero-shot learning models work by learning semantic information across both seen images that the models have been trained on and auxiliary attribute information. Generally, instances of each class are often represented as a vector in a feature space, which allows objective comparisons between various images (Wang et al., 2019, p. 3). The dimensions of these engineered spaces are not designed by humans and are rather learned by machine learning models (Wang et al., 2019, p. 9). Rahman et al. (2018) propose an approach to embed vectors of image features and achieve high performance in a generalized zero-shot learning task where test images might be from seen or unseen classes (p. 5652). Li et al. (2019) take a similar approach to this embedding problem, and perform zero-shot learning with a language component that understands the meaning of textual descriptions (p. 8690). Once these embeddings are generated for each class, models can represent test images in the same embedding space and classify these test images according to the closest vector distance.

Generalized Zero-Shot Learning

In zero-shot learning tasks, images are trained on images from seen classes and tested on a strictly disjoint set of images from unseen classes. In other words, images at test time can only be classified to the set of unseen classes. A closely related field to this is generalized zero-shot learning, in which images can be assigned to either an unseen or seen class label with the highest compatibility score (Xian et al., 2019, p. 2253). This problem is much harder than classic zero-shot learning as models often struggle to treat seen and unseen classes equally, given that they have been previously trained on the seen classes. Generalized zero-shot learning struggles

because there is a strong bias toward seen classes, and so almost every test unseen instance is categorized into a seen class (Rahman et al., 2018, p. 5653). As a result, while generalized zero-shot learning accuracy scores are often far lower than their classic zero-shot learning counterparts, Xian et al. (2019) claim that the problem space offers a far more realistic version of computer vision (p. 2263). Recently, many researchers have been looking to develop powerful algorithms that perform and scale well in both classic and generalized zero-shot learning.

IMPROVING CLIP ACCURACY WITH PROMPT-ENGINEERING

My technical work focuses on providing context in the form of a strategy called prompt engineering to a newly released machine learning model. Contrastive Language-Image Pre-training (CLIP) is a state-of-the-art neural network model recently published by OpenAI that has achieved tremendous performance in the zero-shot learning problem space. The model learns visual concepts from natural language supervision and is pre-trained on image-text pairs sourced from across the internet (Radford et al., 2021, p. 1). While many similar models are trained on datasets of images belonging to classes that have attribute lists, CLIP's approach of using over a million image and caption pairs online offers a promising alternative that is far more robust than traditional methods (Radford et al., 2021, p. 1). This approach allows CLIP to learn from a diverse and massive training space that spans a wide range of categories and classification tasks, whereas most models are only successful on one or a few related datasets and therefore do not scale well to different tasks.

With this revolutionary approach of training on image and natural language text pairs, CLIP achieves state-of-the-art accuracy on multiple diverse benchmark tasks, making it competitive with many different models in optical character recognition, action recognition in videos, geo-localization, and many other fine-grained object classification tasks (Radford et al.,

2021, p. 1). This robustness in many different tasks makes CLIP applicable and effective to various use cases. The CLIP model jointly pre-trains an image encoder and a text encoder that takes textual captions correlated to each image (Radford et al., 2021, p. 2). These encoders are trained to match images with captions such that CLIP can learn to recognize how patterns within images correspond to textual text that describes these images. Figure 1 below overviews this contrastive pre-training process.

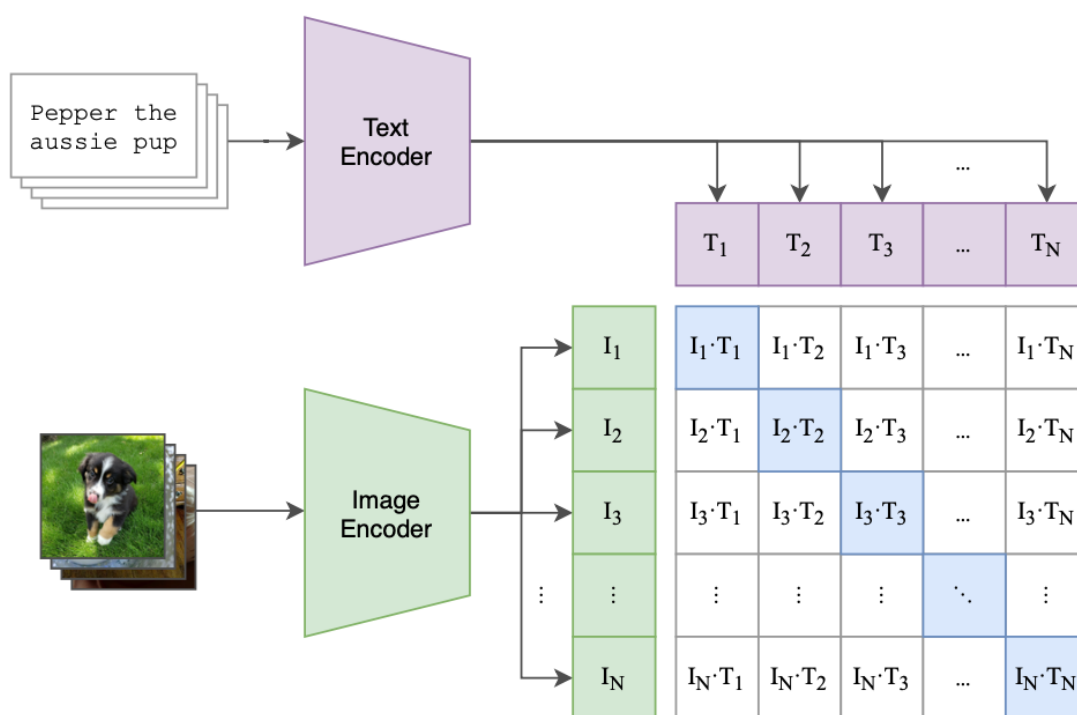


Figure 1: CLIP pre-training process. This image from Radford et al. (2021) demonstrates how CLIP trains a text encoder and image encoder to predict the correct image-text pairs of training examples (p. 2).

At test-time, CLIP is capable of serving as a zero-shot classifier. CLIP can achieve this by generating captions corresponding to each different class in a certain dataset, in the form “this is a photo of a [class name]”. Then, once CLIP sees a test image example, it can determine which class caption scores best against the image and classify that image accordingly. Figure 2

describes this process below. This approach is far more robust than most zero-shot classifiers because CLIP can be applied to a wide range of datasets and classify many different types of images by simply encoding captions of each class name in the dataset.

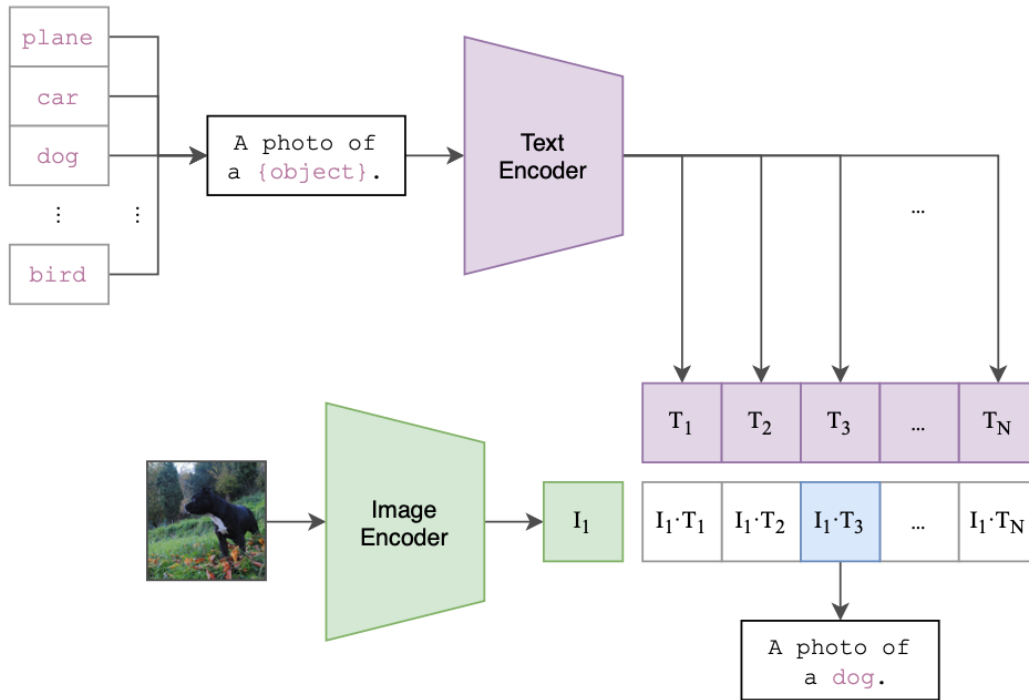


Figure 2: CLIP as a zero-shot classifier. This image from Radford et al. (2021) demonstrates how at test time, CLIP takes an image as input and encodes captions corresponding to each class name. Then, it uses pre-trained encoders to determine which caption corresponds best with the image, thereby classifying that image in a zero-shot manner. (p. 2).

My technical research seeks to improve this approach by changing the form and content of the captions that CLIP feeds into its text encoder for each class. Radford et al. (2021) define this sort of approach as prompt engineering, in which the templates of captions corresponding to each class are modified to either provide extra context to the model or cater the templates to specific datasets (p. 7). Another related approach that Radford et al. released initial findings for is ensembling, in which the scores of multiple different templates are averaged together when matching classes to images (p. 8). This is useful for when images at test time can vary greatly in

style. For instance, scores of captions like “a big photo of a [class name]” and “a cartoon [class name]” might be averaged together such that the model can know test images might take different forms and styles.

This past semester, I implemented various prompt engineering and ensembling methods for the ImageNet dataset. ImageNet is a database of hundreds of thousands of images of various classes and is a very popular benchmark dataset for many computer vision tasks (Stanford Vision Lab, 2016). The dataset is organized according to a hierarchy of words defined by the WordNet dataset and so the classes follow these hierarchies (Stanford Vision Lab, 2016).

As a first step for improving the performance of CLIP on ImageNet, I included hyponyms, which are names of parent classes according to the ImageNet hierarchy, in my prompts. With this improvement, my templates were now organized as “a photo of a [class name], which is a type of [hyponym name]”. After observing an increase in accuracy, I created an Amazon Mechanical Turk survey online to gather data on common human-generated category names to describe each class, and use these human-generated names instead of hyponym names. I then ensembled over a subset of templates predefined by CLIP to be able to recognize a diverse set of image types. Finally, I found that ensembling the top 5 human generated category names offered a strong improvement over simply using the best or most common category name. These results are defined in Table 1 on the following page.

CLIP Accuracies on ImageNet (%)		
	One Prompt Template	Ensembled Templates
No Extra Context	63.2	64.27
Standard ImageNet Hyponyms	63.65	64.46
MTurk Categories	63.41	64.51
Ensembled Top-5 Mturk Categories	64.96	65.81

Table 1: CLIP Accuracies. Radford et al. (2021) achieve an accuracy of 63.2% on CLIP. My various approaches to provide additional context in prompt templates and ensemble over results improves this accuracy.

As demonstrated in Table 1, averaging the scores of the top-5 human generated category names over an ensembled group of prompt templates achieves an accuracy that is 2.61% greater than that advertised in the Radford et al. (2021) paper where CLIP was initially released. While this improvement may seem marginal, my work offers a relatively low-effort increase on a state-of-the-art benchmark. Further, these improvements are an initial effort that point toward an opportunity to use prompt engineering to further increase the accuracy of CLIP. Ultimately, prompt engineering and ensembling are techniques that I recommend should be researched in depth for a variety of datasets beyond only ImageNet.

PREDICTING AN ENCODER WITH WIKIPEDIA DESCRIPTIONS

The second portion of my research semester involved using CLIP’s image encoder to predict a classifier for unseen classes using Wikipedia textual descriptions of classes. Specifically, I learned a classifier for the Caltech-UCSD Birds (CUB) dataset. This fine-grained dataset contains over eleven thousand images among two hundred classes of birds, and also defines certain attributes associated with these classes (Caltech-UCSD Birds 200, 2011). Due to the presence of these attributes, the CUB dataset is frequently used in research involving zero-shot learning as classifying birds presents a challenging task given how similar species of birds are.

I followed an approach similar to Ba et al. (2015) to predict a classifier. I first trained a simple logistic regression classifier for images belonging to seen classes in the CUB dataset. To do so, I encoded each image using CLIP’s pre-trained image encoder and used mean-pooling to reduce the dimension of each encoding. Training my logistic regression classifier ultimately generated a matrix of weights with each row corresponding to a class that the classifier had seen. Next, I trained a simple multi-layer perceptron (MLP) with one hidden layer on Wikipedia textual descriptions corresponding to each seen class. Specifically, I first calculated term-frequency inverse-document-frequency (tf-idf) features of each Wikipedia textual description to standardize the representations of each textual description. I trained these feature vectors to map to each row of my logistic regression classifier, in effect creating a mapping function from Wikipedia descriptions to logistic regression weights. I then used my trained MLP to predict and learn logistic regression weights of the remaining unseen classes by feeding in textual Wikipedia descriptions corresponding to classes the logistic regression model had not seen before. Once these weights were predicted, I appended the weights to my original logistic regression classifier, thereby learning a classifier for both seen and unseen classes without training on any seen images.

To gauge the performance of my learned classifier, I compute accuracy scores of seen and unseen classes separately. Further, I compute area under the curve scores for precision-recall as well as receiver operating characteristic scores. These measures are commonly used to analyze the performance of classifier models beyond accuracy scores. The specific results I achieve are listed in Table 2 below and compared to the results achieved by Ba et al. (2015), which was a motivating paper for this part of my technical research.

Learning a ZSL Classifier for CUB							
	Accuracy	ROC-AUC			PR-AUC		
	Unseen	Seen	Unseen	Weighted	Seen	Unseen	Weighted
Using CLIP	5%	0.972	0.513	0.88	0.924	0.033	0.746
Ba et al. (2015)	10%	0.98	0.85	0.953	0.37	0.13	0.31

Table 2: Learned Classifier Performance. The above table compares the performance of my CLIP implementation of a learned zero-shot classifier with the Ba et al. (2015) implementation.

As the above table demonstrates, the accuracy score of my CLIP implementation on unseen classes is worse than that of the Ba et al. (2015) paper. The weighted ROC-AUC score is also lower than what Ba et al. achieve, and the PR-AUC score is only higher because of there are far more seen classes which disguises the low unseen PR-AUC score. These low results may be due to the difficulty of mapping high dimensional vectors given a small training set. Since there was only one Wikipedia article per class, the number of samples that my multi-layer perceptron was trained on was the number of classes. Training multi-layer perceptrons on this few training examples can lead to the neural networks not learning and transferring patterns in the data. While the results demonstrated by my learned classifier do not meet or approach the state-of-the-art accuracies achieved by recent zero-shot learning models, the idea of learning a classifier from textual descriptions offers an interesting and novel approach to zero-shot learning that should be explored more in the future.

CONCLUSION

Through my technical research this semester, I had the opportunity of working with a cutting-edge publicly released machine learning model that is already being used for revolutionary applications in computer vision. My technical work involving CLIP for zero-shot learning explores opportunities for improvement on the robust model. The findings I discover in both prompt engineering techniques to provide context information and using CLIP to predict a

classifier for zero-shot learning are both novel approaches that present promising directions for further research. These opportunities for future work may include expanding my findings of prompt engineering and ensembling on a wider range of datasets and using different machine learning techniques and architectures to more effectively predict a classifier for zero-shot learning. Ultimately, my work seeks to advance the rapidly-progressing field of computer vision and artificial intelligence.

ACKNOWLEDGEMENTS

I would like to thank Professor Vicente Ordoñez-Roman of the Department of Computer Science for his mentorship with my technical research this semester. I would also like to thank Dr. Fuwen Tan, a PhD student studying under Professor Ordoñez-Roman for his guidance and ideas. Finally, I would like to thank Professor Richard D. Jacques for his help with my sociotechnical synthesis research.

REFERENCES

- Ba, J. L., Swersky, K., Fidler, S., & Salakhutdinov, R. (2015). Predicting deep zero-shot convolutional networks using textual descriptions. *2015 IEEE Conference on Computer Vision*, 4247–4255. doi:10.1109/ICCV.2015.483
- Caltech-UCSD Birds-200. (2011). *California Institute of Technology Vision*. Retrieved from <http://www.vision.caltech.edu/visipedia/CUB-200-2011.html>
- Li, Z., Yao, L., Zhang, X., Wang, X., Kanhere, S., & Zhang, H. (2019). Zero-shot object detection with textual descriptions. *AAAI Conference on Artificial Intelligence*, 33(1), 8690-8697. doi:10.1609/aaai.v33i01.33018690
- Paul, A., Krishnan, N. C., & Munjal, P. (2019). Semantically aligned bias reducing zero shot learning. *2019 IEEE Conference on Computer Vision and Pattern Recognition*, 7049-7058. doi:10.1109/CVPR.2019.00722
- Radford, A., Kimm, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Kreuger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *OpenAI*, 1-48.
- Rahman, S., Khan, S., & Porikli, F. (2018). A unified approach for conventional zero-shot, generalized zero-shot, and few-shot learning. *IEEE Transactions on Image Processing*, 27(11), 5652-5667. doi:10.1109/TIP.2018.2861573
- Reed, S., Ataka, Z., Lee, H., & Schiele, B. (2016). Learning deep representations of fine-grained visual descriptions. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 49-58. doi:10.1109/CVPR.2016.13
- Stanford Vision Lab. (2016). *ImageNet*. Retrieved from <http://www.image-net.org/>

Wang, W., Zheng, V., Yu, H., & Miao, C. (2019). A survey of zero-shot learning: settings methods, and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2), 1-37. doi:10.1145/3293318

Xian, Y., Lampert, C. H., Schiele, B., & Akata, Z. (2019). Zero-shot learning – a comprehensive evaluation of the good, the bad, and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9), 2251-2265. doi:10.1109/TPAMI.2018.2857768