Thesis Project Portfolio

Cognitively Inspired Energy-Based World Models

(Technical Report)

Investigating Privacy and Security Risks of Using ChatGPT for Schoolwork

(STS Research Paper)

An Undergraduate Thesis

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

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

Ganesh Nanduru

Spring, 2024

Department of Computer Science

Table of Contents

Executive Summary

Introduction

We developed the generative machine learning solution described in our technical report, Modeling through Internal Predictions, to explore an improvement on typical autoregressive machine learning that better represents human cognition. As neural networks are inspired by the structure and function of the human brain, we wanted to design our model to be more accurate to the latest findings in cognitive science. For our STS report, we were motivated to switch our focus to the data privacy of ChatGPT for academic purposes because we felt it bridged the applicability of our technical research to our academic environment. We understood that generative artificial intelligence is a quickly developing field that is gaining popularity in applications outside data science - to emphasize the importance of our work in generative AI in an academic context, we explored how ChatGPT, a leading model in GenAI, poses unique opportunities and risks to academic institutions across various disciplines and school districts.

Technical Report Summary

One of the predominant methods for the pre-training of capable foundation models across fields is prediction in the output space of the next element of an autoregressive sequence. In Natural Language Processing, this takes the form of Large Language Models (LLMs) predicting the next token; in computer vision, this takes the form of autoregressive transformers predicting the next frame/token/pixel.

However, we argue this differs from human cognition in three respects. First, when humans make predictions about the future, these shape internal cognitive processes, a characteristic that is not possible with traditional autoregressive models that make predictions in the output space. Second, humans naturally evaluate the strength of predictions regarding future states—a capability that is not possible with traditionally trained autoregressive models.

Third, on the basis of the previous capability of determining when predictions are "good enough," humans naturally leverage a dynamic amount of time when making predictions about the future. We believe all these capabilities to be essential in the success of humans at high-level decision making and reasoning tasks. Therefore, to circumvent limitations of traditional autoregressive pre-training, in this work we present Modeling through Internal Predictions (MIP), an approach that involves training an Energy-Based Model (EBM) to predict the compatibility of a given context and a predicted future state. In doing so, MIP enables models to form predictions of future sequences internally, allowing these predictions to actively influence and shape the model's internal state, much like how predictions guide human cognitive processes. Furthermore, we unveil a distinct EBM training methodology, Regressive Energy Based Modeling, which leverages ground truth energy labels to turn EBM training into a regression task. This novel approach underlies the development of MIP—enabling stable and rapid training. While we've applied MIP to computer vision, we believe MIP has broad applicability as a generalizable foundation model building approach and discuss its implementation in several fields. Our results indicate that MIP provides an exciting avenue towards the training of foundation models with improved prediction capabilities. One such avenue we extend our study into is natural language processing. We apply Regressive Energy Based Modeling to a transformer-based large language model and train on The Pile, an 825 GiB open-source language modeling dataset. We assess our model using The Pile's bits-per-byte (BPB) benchmark.

STS Research Paper Summary

ChatGPT, an artificial intelligence chatbot, is used by many students for its question answering and essay-style generative capabilities. While ChatGPT is helpful for boosting

academic productivity, it poses new concerns of data privacy and safety for its users. To investigate the research question of, “How safe is it for students to use ChatGPT for school assignments?”, This paper utilizes actor-network theory (ANT) to construct a detailed network of corporations, academic institutions, cultural concepts, and technical innovations encompassing the chatbot. This paper analyzes this network to discover actors compromising the data security of using ChatGPT. Namely, the cultural actor: students’ trust in ChatGPT’s data privacy and the organizational actors: vulnerable corporations that have access to ChatGPT’s data, are identified as key concerns for students using the chatbot for their schoolwork. In applying ANT for a data security investigation, this research provides a new application of STS frameworks in the digital world. Most importantly, this research aims to provide students, teachers, and academic institutions within the engineering discipline with a comprehensive understanding of the risks of students using ChatGPT for schoolwork.

Reflection

By working on the STS report and Capstone project simultaneously, my group was able to expand the scope of my problem definition to better understand the importance and consequences of the solution proposed in the technical report. Learning sociotechnical systems analysis skills during our STS research greatly improved the introduction section of our Capstone project, as it allowed us to identify what specifically our contributions to the field were and what impact our solution would have in the broader area of machine learning research. In particular, we found value in the connection between machine learning and human cognition to expand the effectiveness and explainability of our solution. If we were to redo the Capstone project without also working on the STS report, we would lose some of the important

connections between our solution and cognitive studies and related machine learning research that were integral to communicating the purpose and impact of our technical solution. Finally, writing the STS paper developed our exploratory research and academic writing skills, which contributed to the overall flow and base knowledge reflected in the final technical report.