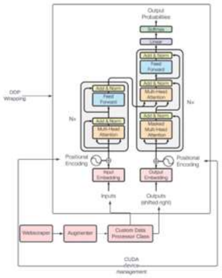
An S2S TLM for The University of Virginia’s College at Wise

Noah Sturgill

The University of Virginia’s College at Wise

Abstract

This research presents the design and implementation of a Transformer-based Sequence-to-Sequence (S2S) model tailored for an embedded chatbot for the University of Virginia’s College at Wise (UVA Wise). By leveraging PyTorch and CUDA for efficient computation, this study approaches the feasibility and challenges associated with implementing a niche Large Language Model. Tokenization for the model occurs using the cl100k_base tokenizer. The dataset required for training the model is procured through a carefully crafted web scraping mechanism using Scrapy, targeting relevant information from UVA Wise’s web resources. The raw data is then sanitized and augmented into a corpus usable in an S2S model.



[1] This diagram shows the architecture and processes involved for the model. The Distributed Data Parallel wrapping wraps all operations of the S2S Transformer. CUDA device management occurs within this wrapping, moving tensors at appropriate times. The Scrapy-based scraper efficiently gathers the data from relevant UVA Wise web resources. The scraped data is then augmented and processed through a custom data processor class; additionally, this class generates QA key pairs. Dynamic masking enables the model to handle batches efficiently based on sequence-length. The dynamic masking extends to the implementation of the encoder and decoder layers, allowing the model more flexibility in the adjustment of the attention mechanism. Similarly, the positional encoding scheme extends beyond the conventional PyTorch implementation, adapting to sequence lengths dynamically.

Objectives

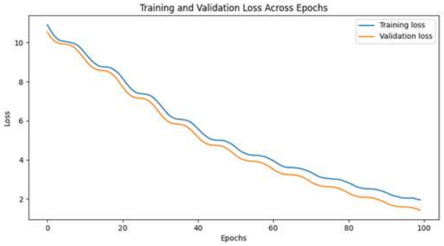
The primary objective of this research is to design and implement a Transformer-based Sequence-to-Sequence (S2S) model optimized for serving as an embedded chatbot for the University of Virginia’s College at Wise. The chatbot aims to provide immediate, accurate, and contextually relevant assistance to students by answering queries covering various college-related topics. To achieve this, the study sets forth the following specific objectives:

1. **Leverage Advanced NLP Techniques:** Advanced NLP Techniques shall be utilized using the Transformer architecture, renowned for its efficacy in handling sequence-to-sequence tasks. The implementation focuses on optimizing the model for efficiency and scalability by employing PyTorch as the learning framework and CUDA for GPU acceleration.
2. **Develop a Robust Dataset:** Employ web scraping technologies, specifically Scrapy, to collect and construct a comprehensive and relevant dataset from UVA Wise’s web resources.
3. **Customize the S2S Model for UVA Wise:** Tailor the Transformer model to cater specifically to the needs of UVA Wise students. Such tailoring requires ensuring the model can handle the variability and specificity of student queries while providing accurate and contextually relevant responses.

Preliminary Results

Initial training sessions showcased promising loss progression, supporting the model’s potential to effectively learn from the constructed data. Specifically, the implementation showed a gradual decrease in the loss function across epochs, suggesting the model was progressively increasing its ability to predict the target sequence given input sequences. However, it is crucial to note that these findings are preliminary and are derived from a modest dataset extracted from UVA Wise web resources.

Given the model’s architecture and the complexity of natural language, evaluation on solely loss metrics does not provide the complete picture. As such, plans are in place to incorporate BLEU and ROUGE scores as part of future evaluation metrics.



Conclusions

The preliminary results are encouraging, given the favorable loss progression of the S2S model. Looking forward, the project is poised for several developmental trajectories. The enhancement of the dataset through expanded web scraping efforts will remain a critical focus. A richer dataset will contribute to the refinement of the model’s performance and its capability to handle a more diverse array of queries. Another avenue for future development lies in the optimization of the model architecture and training process. The computational limitations of development require future training to be conducted on a cluster for optimal model performance. The initial training of the model was conducted on a singular NVDA 3070 Ti.

The generation method will require fine-tuning to ensure that queries contain an appropriate tone for addressing student queries. An outside dataset, such as the DREAM dataset, is being considered as a means of objective measurement of the model; additionally, alternative datasets can provide metrics of comparison against other models on defined datasets. The ongoing development of the model will continue to aim at furthering the usage of Large Language Models as tools within the Education infrastructure.

References

[1] Anon. Language modeling with nn.transformer and torchtext. Retrieved March 31, 2024 from https://pytorch.org/tutorials/beginner/transformer_tutorial.html

Acknowledgements/Contact Information

Noah Sturgill – nds6g@uvawise.edu
Developed using: PyTorch CUDA

RESEARCH POSTER PRESENTATION DESIGN © 2019
www.PosterPresentations.com