Unsupervised Multitask NLP Models

Introduction

The field of natural language processing (NLP) is primarily composed of systems that have been trained with a specific function in mind. A wide range of NLP tasks exist, including question answering, translation, summarization, reading comprehension, sentiment analysis, and others. Lacking in the NLP field are good generalized systems that can accomplish multiple tasks relating to NLP, much like a human can.

The goal of OpenAI's Generative Pre-trained Transformer (GPT) artificial intelligence system is to achieve this generalized ability. GPT is modeled using deep neural networks to accomplish language modeling, among other tasks. This review focuses on the system in this series called GPT-2, as outlined in Radford et al., 2019[1]. It also compares its architecture to another generalized model developed by Google, called BERT[2] (Bidirectional Encoder Representations from Transformers). Finally, the performance of GPT-2 on several NLP tasks is summarized.

Generalized NLP Models

Several systems have been created for generalized NLP modeling. A single task system produces an output given an input. A multitask system does the same but has a range of tasks that can be performed. Central to a generalized framework is the principle of task transfer, or transfer learning—that the principles and strategies for accomplishing one task can also apply to another task. This transferring ability allows generalization to exist.

One way the effectiveness of generalized models can be determined is by measuring zero-shot performance for different NLP tasks. Zero-shot learning involves correctly predicting the class of items belonging to a class that was not identified during training.

GPT-2 Architecture

The first step to create the GPT-2 architecture is to train on a dataset. The dataset used was created specifically for GPT-2. All outbound links from the social media platform Reddit, on posts with at least 3 karma (indicating positive feedback from Reddit users), were followed and the resulting pages were scraped to create the training dataset. This approach allowed a huge amount of data to be collected, while ensuring that a high percentage of the data is intelligible (as Reddit users would not give positive feedback for spam links or nonsensical pages), which solves a major issue of other large datasets. The resulting dataset is called WebText. After de-duplication and other cleaning, the training dataset contains 8 million documents, or 40GB of text.

[1] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. Language Models are Unsupervised Multitask Learners. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[2] Jacob Devlin, Ming_wei Chang, Kenton Lee, Kristin Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805v2. https://doi.org/10.48550/arXiv1810.04805.

The next step is to model the input using a version of byte pair encoding (BPE)[3], where most words are encoded as single tokens, while rare words are encoded as a sequence of tokens. This serves to combine the empirical benefits of word-level modeling with the generality of character-level modeling.

Finally, GPT-2 uses a transformer-based language-modeling architecture.[4] A transformer is a form of deep learning that employs the concept of self-attention, which determines what parts of the input data are most important for a task. Transformers use a semi-supervised approach in their architecture, specifically unsupervised pretraining[5] and supervised fine-tuning.[6] GPT-2 infers a task from the input received. For example, from the input "translate apple to French," GPT-2 would infer the task is translation.

The final GPT-2 model contains over 1.5 billion parameters but is still shown to underfit the WebText dataset. The model has been further expanded and improved in GPT-3.

GPT-2 vs BERT

GPT-2 and BERT are both generalized NLP models that were primarily trained for language modeling. GPT-2 and BERT differ in that GPT-2 uses only the decoder block of the transformer architecture and BERT uses only the encoder block. In language modeling, GPT-2 only predicts a word based on preceding words. BERT predicts a word based on words that come before, as well as after, in a sentence. This presents a risk of the model predicting something based on pure memorization rather than actual learning. To prevent this, BERT asks the model to learn the word in a sentence after masking out 15% of words in the sentence. Where GPT-2 focuses on predicting the next word, naturally creating sentences along the way, BERT employs a sentence-to-sentence architecture, looking at the likelihood of one sentence following another.

GPT-2 Performance

First, we will look at GPT-2 performance on the primary task for which it was trained—language modeling. Then, we will look at its generalized performance on other tasks compared to general and task-specific architectures.

For language modeling, GPT-2 improved accuracy on predictions for the LAMBADA and two Children's Book Test datasets over a BERT-based model from 0.5923 to 0.6324, 0.857 to 0.933, and 0.823 to 0.8905, respectively. Additionally, it improved perplexity on the LAMBADA dataset from 99.8 to 8.63. Overall, GPT-2 performed better than BERT-based models on 7 out of 8 test datasets in a zero-shot setting.

[3] Rico Sennrich, Barry Haddow, Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. arXiv:1508.07909v5. https://doi.org/10.48550/arXiv.1508.07909

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention isall you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[5] Ilya Sutskever, Rafal Jozefowicz, Karol Gregor, Danilo Rezende, Tim Lillicrap, Oriol Vinyals. Towards Principled Unsupervised Learning. arXiv:1511.06440v2. https://doi.org/10.48550/arXiv.1511.06440.

[6] Dai, A. M. and Le, Q. V. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems*, pp. 3079-3087, 2015.

For reading comprehension, GPT-2 obtained an F1 score of 55 on the development set. This is a great improvement over the performance of 3 out of 4 baseline systems used for reading comprehension, but its performance was still eclipsed by the BERT-based system, which is approaching a human performance of 89 F1. Still, this is an impressive result.

GPT-2 performed better than some baseline systems for translation, summarization, and question answering tasks, but was far eclipsed by the best approaches in these categories.

Conclusion

GPT-2 provides an approach for multiple NLP tasks, such as question answering, summarization, translation, and others without being specifically trained to perform these tasks. Although GPT-2 is not usable for many of these tasks, as it has considerable shortcomings in accuracy/perplexity for them, it has substantially improved performance compared to some baseline models and logged especially impressive results for its primary task of language modeling, as well as for reading comprehension. These accomplishments demonstrate that significant progress continues to be made in developing NLP systems.