Question Answering by Transformer-XL with Auto Parameter Tuning

Stanford CS224N Default Project

Chi Wang
Department of Computer Science
Stanford University
chiwang@stanford.edu

February 16, 2021

1 Key Information to include

- External collaborators (if you have any): N/A
- Mentor (custom project only):
- Sharing project: N/A

2 Research paper summary (max 2 pages)

Title	Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context
Venue	Association for Computational Linguistics (ACL)
Year	2019
URL	https://arxiv.org/abs/1901.02860

Table 1: Paper bibliographical information [1].

Background. Transformers have a potential of learning longer-term dependency, but are limited by a fixed-length context in the setting of language modeling. There are two critical limitations of using a fixed length context. First, the largest possible dependency length is upper bounded by the segment length; Second, simply chunking a sequence into fixed-length segments will lead to the context fragmentation problem.

This TransformerXL paper proposed a new architecture to handle the long term dependencies and address the limitation of fixed length context, it obtained strong results not only in long text articles but also short-term dependency.

Summary of contributions. The main technical contributions from TransformerXL include introducing the notion of recurrence in a purely selfattentive model and deriving a novel positional encoding scheme. These two techniques form a complete set of solutions, as any one of them alone does not address the issue of fixed-length contexts. Transformer-XL is the first self-attention model that achieves substantially better results than RNNs on both character-level and word-level language modeling.

Limitations and discussion. Transformer-XL obtains strong perplexity results, models longer-term dependency than RNNs and Transformer, achieves substantial speedup during evaluation, and is able to generate coherent text articles. The author looks forward so see Transformer-XL beeb applied to the fields of text generation, unsupervised feature learning, image and speech modeling.

There is one thing worth notice from the paper is TranformerXL model only gain small advantage on small datasets and short sequences compared to state-of-the-art LSTM models, it requires some fine tuning and proper regularization. This is what we need to pay attention when training with SQUAD dataset.

Why this paper? To a question and answering problem, it's very critical to have the model works well for both long and short dependency, since we might want to consider the entire context at once for calculation attention. TransformerXL is the suggested paper in our final project handout, and it's also the most cited paper from Google scholar on this topic.

Wider research context. TransformerXL is a language model, it tries to address the limitation of using fixed-length context so it can perform well on long dependency context. With its capability of understanding long sequences, we could apply it on question answering, source code understanding [2] and build generalized pretraining model for language understanding tasks [3].

3 Project description (1-2 pages)

Goal. In this project, I want to achieve two goals:
First, Implement a question answering system from scratch with QANet architecture[4] and apply the TransformerXL ideas by modifying the attention layer.

Second, Try out how automatic hyperparameter tuning could improve the model performance when doing fine tuning. At the fine-tuning phases, I will expose model parameters as variables with a search range to Katib[5] (an automatic hyperparameter tuning system), and apply Hyperband[6] and Bayesian

Optimization[7] search algorithm to find the best hyper parameters. I hope it will improve model performance at least in a small margin.

If I have extra time, I want to build another question answering model with pretrained BERT model and compare the result.

Task. This project is about to build a question answering system to perform well on SQUAD 2.0 dataset, the detail is defined at the project handout.

Data. SQUAD 2.0 dataset, the detail is defined at course project handout, "The SQuAD Data" section.

Methods. In this project, I plan to re-implement QANet[4], a model based off of the self-attention mechanisms in Transformer, other than closely mirrors the QANet, I will replace its self-attention layers by a Transformer-XL, a new type of attention mechanism specifically designed to handle long-term dependencies.

For the fine-tuning, I will change the model code to expose different optimization options as parameters with a search range, such as regularization options, model size and number of layers, optimization algorithm options, so it can work with automatic hyperparameter tuning tools (Katib[5]) to find the most suitable hyperparameters.

Baselines. The baseline model is BiDAF[8] which is already provided in the project starter code, what I will do is to add character embedding.

Evaluation. For the evaluation metric, I will stick to the default evaluation that consists of the Exact Match (EM) score and the F1 score (see details in project handout) on the provided test set. During training, we also monitor the evolution of negative log-likelihood (NLL) to stop training when the model starts to show signs of overfitting.

References

[1] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context, 2019.

[2] Thomas Dowdell and Hongyu Zhang. Language modelling for source code with transformer-xl, 2020.

[3] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2020.

[4] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. Qanet: Combining local convolution with global self-attention for reading comprehension, 2018.

[5] Katib: Automatic hyperparameter tuning.

[6] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 18(1):6765–6816, 2017.

[7] Eric Brochu, Vlad M Cora, and Nando De Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

[8] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*, 2016.