DS-GA 1012 Natural Language Understanding Proposal

Continual Learning with BERT using Meta-Learning

Ant Ngo [an3056], William Huang [wh629], Christine Shen [zs1534], Abha Sahay [as13492]

1 Motivation:

Yogatama et al. defines general linguistic intelligence as the ability to reuse previously acquired knowledge to adapt quickly to new tasks. Ideally, successful natural language understanding (NLU) models can achieve such human-like intelligence and perform well by learning from a continual stream of tasks. Large pre-trained models, like BERT (Devlin et al., 2018), based on Transformer architecture (Vaswani et al., 2017) have achieved state-of-the-art performance on a wide array of NLU tasks. However, Yogatama et al. (2019) found that these models fail to meet this definition of general linguistic intelligence and suffer from catastrophic forgetting in a continual learning setting. The goal of this project is to utilize meta-learning methodologies such as OML (Online aware Meta-learning) (Javed and White, 2019) to further train BERT in an attempt to learn representations that are more robust to catastrophic forgetting.

2 Methodology:

For our baseline, we will replicate results from Yogatama et al. (2019) of training BERT on the curriculum of SQuAD (Rajpurkar et al., 2016) $\rightarrow$ TriviaQA (Joshi et al., 2017). During the fine-tuning of TriviaQA, we will monitor the performance of SQuAD to track the model's catastrophic forgetting. This is depicted in the bottom stream of Figure 1.

For our experiment, we will further train BERT using OML to learn representations that are robust to catastrophic forgetting. We refer to this new pre-trained model as BERT-M. We will use several training datasets (Trischler et al., 2016; Dunn et al., 2017; Yang et al., 2018; Kwiatkowski et al., 2019) from MRQA 2019[1]. We exclude SQuAD and TriviaQA from meta-learning to avoid overfitting. For our OML procedure, we update the prediction learning network (PLN) using the task specific objective in the inner loop of the algorithm. We then update the weights of the representation learning network, BERT, in the outer loop using the masked language model objective similar to BERT's original pre-training procedure. Once we've learned BERT-M, we train the model on the same continual learning curriculum as BERT as described in the top stream of Figure 1. By comparing the two results, we hope to show our hypothesis that further pre-training BERT with meta-learning will lessen the catastrophic forgetting of previously learned tasks.

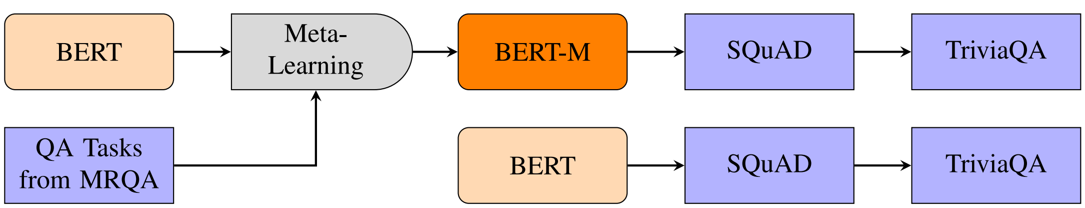

Figure 1: Top stream depicts the use of meta-learning while the bottom stream depicts the baseline for comparison.

3 Data and Tools:

We will utilize the code[2] shared by Javed and White (2019) as a starting point and further modify it using PyTorch (Paszke et al., 2019) and HuggingFace's Transformers library (Wolf et al., 2019) to make it appropriate for BERT and NLU tasks. Further, we will use the QA datasets from the MRQA Github[3].

4 Collaboration Statement:

All team members contributed equally to writing the proposal.

[1] https://mrqa.github.io/shared
[2] https://github.com/Khurramjaved96/mrcl
[3] https://github.com/mrqa/MRQA-Shared-Task-2019#datasets

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. **Bert: Pre-training of deep bidirectional transformers for language understanding**.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. **Searchqa: A new qa dataset augmented with context from a search engine**.

Khurram Javed and Martha White. 2019. **Meta-learning representations for continual learning**.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. **Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension**.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. **Natural questions: A benchmark for question answering research**. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. **Pytorch: An imperative style, high-performance deep learning library**.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. **Squad: 100,000+ questions for machine comprehension of text**.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. **Newsqa: A machine comprehension dataset**.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. **Attention is all you need**.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. **Huggingface's transformers: State-of-the-art natural language processing**.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. **Hotpotqa: A dataset for diverse, explainable multi-hop question answering**.

Dani Yogatama, Cyprien Masson D Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, and Phil Blunsom. 2019. **Learning and evaluating general linguistic intelligence**.