Project Proposal: Toward General NLP - Lifelong Learning of Cross-Domain Language Classification Tasks

Authors
Jiaye Zhu / jiayezhu@usc.edu
Runqi Pang / runqipan@usc.edu
Hanxiang Liu / liuhanxi@usc.edu
Rongcan Fang / rongcanf@usc.edu
Haozhe Zhang / haozhe@usc.edu

1 Project Domain & Goals

Natural Language Processing (NLP) has been one of the primary focuses in the field of Artificial Intelligence. With the thrive of Attention and Transformer architectures (Vaswani et al., 2017) in recent years, the NLP research has been empowered by more powerful language modeling and understanding tools for better and in-depth applications. Current pre-trained language model (PLM) like BERT (Devlin et al., 2018) is trained with a large amount of computational resource and are powerful tools for embedding the natural language. However, such language models need to be *finetuned* for downstream tasks such as classification, summarization, and question-answering. The fine-tuned model is usually task-specific, meaning it can only perform well on that task. If a fine-tuned model is trained again on another task, it will be more likely to forget the old task, this phenomenon is referred to as *catastrophic forgetting* (Lange et al., 2019). The common practice is to fine-tune PLM for each downstream task offline. This method will end up in storing multiple model weights in order to perform multiple tasks. It still cannot process online datastream and will lead to significant memory overhead.

We propose to look into the mitigation of catastrophic forgetting in learning multiple cross-domain language classification tasks, with the constraint of limited computational resource and memory. This domain, namely **Continual Learning**, has been widely examined in Computer Vision and Robotics, but relatively less in NLP (Biesialska et al., 2020). Our goals for this project are:

1. Examine and reproduce recent related works in the Continual Learning of NLP.
2. Compare and contrast these methods for their pros and cons.
3. Compose a new cross-domain language classification dataset and perform evaluation.
4. Introduce new components into existing frameworks for CL specifically in NLP.

In a more advanced stage of this project, we seek to explore more efficient components that can be added to the Transformer model or the overall framework to increase the performance compared to existing works. For example, we can add additional layers to the PLMs to reduce forgetting by balancing the *task recency bias* like in (Wu et al., 2019). From data availability perspective, we can increase the size of supporting data by utilizing a generative model like Variational Autoencoder (VAE) (Kingma and Welling, 2013). We will test these ideas with comparison to the baseline methods.

2 Related Works

Pre-trained Language Model (PLM) is widely applied in current NLP research and applications. It allows us to conduct experiments in state-of-the-art tools without spending numerous time and resources in training from scratch. The Huggingface platform (Wolf et al., 2020) provides various types PLMs of different sizes. Considering the project's limited time and resources, we use a small and fast PLM called *DistilBERT* (Sanh et al., 2019) for experiments.

Continual Learning in NLP. According to (Lange et al., 2019), CL methods can be divided into replay, regularization-based and parameter-isolation methods. Some simple methods like Experience Replay (Rolnick et al., 2018) and Elastic Weight Consolidation (Kirkpatrick et al., 2017) are strong baselines in Continual Learning. They are developed for solving general Continual Learning problems while we wish to focus on the field of NLP specifically. (Wu et al., 2022) proposed a comparative study paradigm to investigate Continual Learning in PLMs. We will follow their approach and establish baselines on replay and regularization-based methods for more datasets and extending the evaluation to cross-domain classification.

3 Datasets

We will begin with a standard dataset for incremental language classification:

- ArXiv Papers (Clement et al., 2019). A dataset of scholar papers for multi-class classification of categories.

In addition, we will use these datasets to form a task stream across multiple domains to evaluate the model's capability in a more ”general” case.

- Food.com Recipe & Review Data. Classify the ratings (1-5) given the review and recipe texts.
- Amazon Reviews. We use the ”Video Games” sub-category and classify the ratings (1-5) given the reviews.
- Clothing Fit Data. Classify the ratings (1-10) given the review of clothing.
- IMDB (Maas et al., 2011). The Large Movie Review Dataset of binary classification for sentiment analysis.
- GLUE Benchmark (Wang et al., 2018). The General Language Understanding Evaluation benchmark contains nine tasks (eight of them are suitable for our problem).

The task stream will be constituted of 12 tasks with 39 classes in total. The first three datasets can be found at `https://cseweb.ucsd.edu//~jmcauley/datasets.html`. These datasets are already cleaned and stored in structured files; therefore minimal pre-processing is needed.

4 Technical Challenges

As addressed in (Biesialska et al., 2020) and (Wu et al., 2022), there are limited works looking into Continual Learning for Natural Language Processing. It's hard and goes beyond the course syllabus as this field is one of the cutting-edge research topics in NLP. We choose to perform compare & contrast study in established baselines on new datasets to give another view on CL in NLP.

Since we are using PLMs, the first challenge will be setting up Continual Learning scenarios on the datasets, which requires extensive coding and debugging to build a data foundation for our evaluations. Second, we need to implement the common CL approaches on top of the Transformer models, which requires us to have a basic understanding of its inner process, and modification to PLMs may be needed. Third, we ought to perform hyper-parameter tuning for a fair comparison of different methods. We will mitigate the challenges in evaluation by following the process and metrics discussed in (Wu et al., 2022).

In the exploring stage, the main challenge will be identifying which of the new ideas can effectively improve performance. To that, we need to gain a deeper understanding of Transformer, Attention, and possibly generative structures like VAE. We also need to design and perform ablation studies to experimentally verify the effect of our ideas.

5 Application fields

To the best of our knowledge, current Natural Language Processing applications associated with Deep Learning are still heavily relying on the fine-tuning of every single task on Pre-trained Language Models. Although the PLMs have excessive model capacity to achieve good performance on many downstream tasks, the fine-tuning strategy only exhibits PLMs' capability in width and not in depth. Also, with strict constraints on memory and computational resource, fine-tuning for each task is less viable and an incremental learning scheme is required.

The potential application fields of our project are widely-across many domains. For example, in a natural language-based recommendation system, we might need to update the model as the situation changes, and an NLP continual learning framework can just do incremental learning instead of re-finetune from the PLM. Other types of downstream tasks like dialogue systems can also benefit from Continual Learning, as the language and dialogue are constantly evolving and we need the PLM to adapt to changes incrementally.

References

[Biesialska et al.2020] Magdalena Biesialska, Katarzyna Biesialska, and Marta R Costa-Jussa. 2020. Continual lifelong learning in natural language processing: A survey. *arXiv preprint arXiv:2012.09823*.

[Clement et al.2019] Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. 2019. On the use of arxiv as a dataset.

[Devlin et al.2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

[Kingma and Welling2013] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes.

[Kirkpatrick et al.2017] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

[Lange et al.2019] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory G. Slabaugh, and Tinne Tuytelaars. 2019. Continual learning: A comparative study on how to defy forgetting in classification tasks. *CoRR*, abs/1909.08383.

[Maas et al.2011] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June. Association for Computational Linguistics.

[Rolnick et al.2018] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Greg Wayne. 2018. Experience replay for continual learning. *CoRR*, abs/1811.11682.

[Sanh et al.2019] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

[Vaswani et al.2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

[Wang et al.2018] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *CoRR*, abs/1804.07461.

[Wolf et al.2020] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

[Wu et al.2019] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. 2019. Large scale incremental learning. *CoRR*, abs/1905.13260.

[Wu et al.2022] Tongtong Wu, Massimo Caccia, Zhuang Li, Yuan-Fang Li, Guilin Qi, and Gholamreza Haffari. 2022. Pretrained language model in continual learning: A comparative study. In *International Conference on Learning Representations*.