Adapting Pretrained Language Models for Machine Translation

Pre-trained models are being made readily available by companies or research institutes so that the research community can reuse and repurpose them in other tasks. There are three common techniques for using these pretrained models: (1) continue training on the desired task, i.e. treat the pre-trained model as nothing more than a clever weight initialization, or most commonly known as fine-tuning, (2) train the output layer of the pre-trained model and keep the rest of the model fixed, (3) add small parts of trainable parameters, so called 'adapters' [1,2,3], throughout the network and train them, while keeping the rest of the network fixed.

This thesis focuses on the task of machine translation and tries to benefit from models pre-trained monolingually on a language modeling task. This difference in the task inevitably requires some adaptation of the model. The question is how large this adaptation should be and which parts of the model it should concern [4].

Minimally, the thesis will connect two language models, one to serve as the encoder in the source language and the other to serve as the decoder in the target language. The common component of attention between the encoder and decoder has to be initialized randomly and the decoder has to be adapted to produce the target sentence left to right but the rest of the network can remain fixed, reusing weights of the two language models.

Further experiments in the thesis will gradually allow more and more of the network to be fine-tuned for the translation task in question with the goal to examine which of the approaches delivers the best translation quality while considering the needed training time.

The experiments will also consider the aspect of the domain of the text, that is, the training data and the final test set will come from a slightly different domain. All the experiments can be limited to a single language pair, e.g. German to English translation.

Most of the evaluations will rely on automatic measures of translation quality such as BLEU, chrF and also more recent metrics like COMET. A very small manual evaluation of the best setups at the end is desirable.

References

1. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., Laroussilhe, Q.D., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-Efficient Transfer Learning for NLP. *ICML*.
2. Pfeiffer, J., Vulić, I., Gurevych, I., & Ruder, S. (2020, November). MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7654–7673. doi:10.18653/v1/2020.emnlp-main.617
3. Bapna, A., Arivazhagan, N., & Firat, O. (2019). Simple, Scalable Adaptation for Neural Machine Translation. *EMNLP*.
4. Winata, G.I., Wang, G., Xiong, C., & Hoi, S.C. (2021). Adapt-and-Adjust: Overcoming the Long-Tail Problem of Multilingual Speech Recognition. *ArXiv, abs/2012.01687*.