Evaluating Transformers as Embedding layers

Daniel Gedon
Department of Information Technology
Division of Systems and Control
Uppsala University
daniel.gedon@it.uu.se

Abstract

In this course project we study the use of pretrained transformers as embedding layers for NLP tasks. We compare embeddings on word level, character level and combined word-character level for the task of word sense disambiguation (WSD). We pretrain the transformers on a large corpus on data and then finetune it for the WSD task. As a comparison method we use standard embedding layers. The complete code is self written for pretraining and finetuning and is freely available.

1 Introduction and Background

For many NLP tasks, the use of large pretrained language language models is the standard method nowadays. Attention based models as pretrained model have pushed the state of the art from its first use in the NLP area by Bahdanau et al. [2016], which adds a RNN on top of a self-attention mechanism for a machine translation task. Vaswani et al. [2017] presented a new model structure, the transformer, which is only based on attention and does not use an additional RNN or CNN. This model is pretrained in a self-supervised fashion on a large corpus of freely available and unlabeled language data.

Two main pretrained model categories can be distinguished for our purposes. **Autoregressive models** make use of the decoder in the transformer architecture. These models generate a language model by predicting the next token in a sequence with information only from previous tokens. Well known examples are GPT (Radford et al. [2018]), GPT-2 (Radford et al. [2019]) and the most recent GPT-3. The second category are **autoencoding models**. These models rely on the encoder in the transformer architecture. The model is not limited to previous tokens but can look at all tokens. The inputs are corrupted to have a denoising autoencoder, similar to Vincent et al. [2010]. The most famous example is BERT (Devlin et al. [2019]) and its further improvements like RoBERTa (Liu et al. [2019]) or DistilBERT (Sanh et al. [2020]). In practice, a certain percentage of tokens in the input is corrupted by masking them out. The model can use information before and after that token in order to predict the masked tokens and therefore learn a language representation.

In this course project we use transformers on different text abstraction levels (i.e. word level, character level and a combination of both) as pretrained models and compare their performance on a common, simple NLP task. The goal is not to push the limits in terms of overall performance but to compare the different levels of abstraction in the transformer. Our code is freely available [1].

2 Methods

2.1 Transformer as Embedding

We make use of a BERT-like architecture and learning objective.

[1] https://github.com/dgedon/nlp_transformer_embeddings

1st WASP course on Natural Language Processing (WASP NLP 2020), Linköping/Gothenburg, Sweden.

2.2 Comparison Method

3 Experiments

3.1 NLP Task

3.2 Datasets

3.3 Results

4 Conclusion

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473 [cs, stat]*, May 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*, July 2019.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. page 12, 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. page 24, 2019.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]*, February 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv:1706.03762 [cs]*, June 2017.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. page 38, 2010.