University of Zurich^UZH

Institute of Computational Linguistics

DRAFT: Eager Machine Translation

Master Thesis Description

Prof. Martin Volk, Ph.D.

1 Introduction

Current Neural Machine Translation (NMT) models employ a so-called encoder-decoder architecture. That is, NMT models commonly include two distinct modules, an encoder and a decoder, such that the encoder translates the input sequence into a list of vectors and the decoder translates this list of vectors into the output sequence, one symbol at a time.

The encoder and decoder modules typically are modelled as recurrent neural networks (RNNs), a special type of neural network that uses the output of the previous step in the following step to provide the model with information about previous steps of the input sequence. The decoder module uses the output of the encoder and may thus only start once the encoder is finished.

This leads to an inherently sequential nature in the model architecture that causes training and inference to be computationally very expensive and time consuming. [Chung et al., 2014, Sutskever et al., 2014, Wu et al., 2016, Cho et al., 2014, Bahdanau et al., 2014, Luong et al., 2015]

A great body of research to improve the performance and accuracy of neural language models and encoder-decoder architectures has since been conducted in the form of (add some references here, like factorization tricks, conditional computation and what not). Research with similar objectives include the application of attention mechanisms in addition to the encoder and decoder modules [Bahdanau et al., 2014, Luong et al., 2015, Vaswani et al., 2017].

Attention circumvents the issue that traditional neural language models need to represent all relevant information of the source sentence in a single fixed-length vector. An attention mechanism computes a weighted sum of all encoder states in every decoder step and therewith effectively learns to translate and align jointly. These models were shown to be particularly effective on longer sequences [Bahdanau et al., 2014, Luong et al., 2015].

In their most recent work, [Vaswani et al., 2017] propose a novel architecture that they call the Transformer which is based entirely on attention mechanisms. Their model employs the traditional encoder-decoder architecture but does not use any recurrent networks.

A big drawback of the transformer architecture is that computing attention is particularly expensive during inference. TODO: elaborate on drawbacks and thus motivation for alternative approaches (You may not need Attention).

[Press and Smith, 2018] propose a model that combines the traditional encoder and decoder steps and eagerly returns a translation after every token of the input sequence. Particularly, [Press and Smith, 2018] investigate the potential of neural machine translation models that do not use attention mechanisms. Instead, they propose a special pre-processing step that aligns dependencies in the input and target sequences by adding a special ε-token. TODO: elaborate on eager and show example

That is, in cases in which the translation of a given token of the input sequence follows subsequently in the output sequence, the model learns to return the ε-token and thus delay further translation for another time step. The ε-token can be considered a way for the model to express that it requires more information before making a conclusive decision. That is, the ε-token, effectively, can be considered a *wait*-token.

[Press and Smith, 2018] quantitatively show that their model performs on par with traditional models as introduced by [Bahdanau et al., 2014]. However, they do not qualitatively or empirically assess the behavior of their model with regards to the ε-tokens. They report that during their experiments translation quality improved if they padded the input sentence with ε-tokens just before the EOS-token but do not elaborate on this further. It may thus be that the ε-token is not helping the model to learn to wait in the case of non-monotonic translations but enables the model to exploit the increased sentence length by simply waiting initially and acquiring multiple tokens of the source sentence before starting to or proceeding to eagerly return translations.

TODO: show examples of exploit and without exploit

In the context of waiting and pondering, [Graves, 2016] have introduced an approach that allows recurrent neural networks to learn the number of steps to take between receiving an input and returning an output. TODO: elaborate: XYZ have shown that increased depth leads to more performant networks. On the other hand, deeper networks require more resources during training and inference and more prone to overfit.

[Graves, 2016] extend the traditional architecture with a state transition model that allows the network to perform a variable number of computation steps for every input it receives. The model learns the number of steps to perform for a given input and therewith learns for which input tokens it requires a deeper computation, it learns for which tokens to wait and ponder, so to speak.

2 Proposed Approach

We propose to qualitatively assess the effect of adding ε-tokens in the eager translation model that [Press and Smith, 2018] introduce. Particularly, we investigate if the model learns to wait in cases of non-monotonic translations or if the model simply exploits the increased sequence length. For this we design and construct an artificial task in which we displace dependent tokens by a suitable margin and then evaluate the performance of the model on these examples. Finding and constructing this task will be part of our research.

We further propose to reproduce the model as outlined by [Press and Smith, 2018] and use it as the baseline for our experiments. We then want to extend the model that [Press and Smith, 2018] propose with Adaptive Computation Time as introduced by [Graves, 2016] and compare the behavior of the two approaches qualitatively. That is, for more complex parts of sequences, we expect a model with ACT to perform a larger number of computation steps than for simpler tokens and argue that this should reflect in an increased number of iterations for long term dependencies, i.e. non-monotonic translations.

3 Task Description

3.1 Do ε-tokens allow a model to learn to wait?

- Reproduce model and results from [Press and Smith, 2018]
- Design artificial task to assess quality of ε-tokens in [Press and Smith, 2018]
- Train and evaluate model with artificial task
- Qualitatively assess behavior of the model trained on artificial task

3.2 Can ACT be used to learn to wait and ponder in cases of non-monotonic translations?

- Adapt source code from [Press and Smith, 2018] and add ACT
- Train and evaluate model with artificial task
- Qualitatively assess behavior of the model trained on artificial task
- Compare behavior of the new model with the baseline model

3.3 Evaluation

3.4 Literature Review

Summarze and provide an overview of relevant work in the field.

4 General Thesis Guideline

The typical rules of academic work apply and must be followed. At the end of the thesis, a final report has to be written and the work has to be presented. The report should be clearly organized and follow the usual academic report structure.

5 Tutor

Mathias Müller

6 Signatures

Lorem ipsum

References

[Bahdanau et al., 2014] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473.*

[Cho et al., 2014] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078.*

[Chung et al., 2014] Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555.*

[Graves, 2016] Graves, A. (2016). Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983.*

[Luong et al., 2015] Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025.*

[Press and Smith, 2018] Press, O. and Smith, N. A. (2018). You may not need attention. *arXiv preprint arXiv:1810.13409*.

[Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

[Wu et al., 2016] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.