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 account for dependencies between tokens of the input sequence. Additionally, in a single step during inference, the decoder returns the top most likely tokens rather than only a single token and then selects the best sequence given this set of tokens from the previous step given again a set of tokens. This process is generally known as beam search. Furthermore, the decoder uses the output of the encoder and may thus only start once the encoder is entirely 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. [Vaswani et al., 2017, Hochreiter and Schmidhuber, 1997, Chung et al., 2014]

A great body of research to improve the performance of recurrent 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.

Attention is lorem ipsum dolor sit amet Praesent ac lorem non dolor tempus consequat in id ipsum. Donec rutrum elit eleifend nibh porttitor, ut tempus nunc ultricies. In lobortis congue ex ac placerat. Integer tempus arcu nec tincidunt vulputate. Donec efficitur placerat ante ut rhoncus. Quisque et nisi ac massa volutpat commodo. Integer ornare, nunc ac pretium pretium, arcu libero eleifend sapien, ut laoreet leo neque id erat. These models are called transformer networks.

In their most recent work, [Vaswani et al., 2017] propose a novel model architecture that they call the Transformer, which A very novel type of model architecture for neural machine translation was introduced by [Vaswani et al., 2017]. In their work, [Vaswani et al., 2017]

Transformer networks improve on recent state-of-the-art. Transformer network without any recurrent nature outperforms previous state-of-the-art. Highly parallelizable (in comparison) yet still computationally expensive.

2 Proposed Approach

In their work, [Kim and Rush, 2016] have demonstrated that applying knowledge distillation to a NMT model allows to generate a significantly simplified and thus much faster version of the original model

with comparable accuracy. The simplified student network appears to perform with comparable accuracy even if beam search is omitted entirely during inference.

We propose to apply knowledge distillation as described by [Hinton et al., 2015] and [Kim and Rush, 2016] to the NMT model that [Press and Smith, 2018] present and evaluate its effectiveness on the performance during inference and potential consequences on the accuracy of the model in the framework that [Press and Smith, 2018] used to evaluate their work.

3 Task Description

3.1 First Group of Tasks

3.2 Second Group of Tasks

3.3 Evaluation

3.4 Literature Review

3.4.1 Sub-subtitle

4 General Thesis Guideline

5 Tutor

Mathias Müller

6 Signatures

Lorem ipsum

References

[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*.

[Hinton et al., 2015] Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[Kim and Rush, 2016] Kim, Y. and Rush, A. M. (2016). Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*.

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

[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.