University of Zurich[UZH]

Department of Computational Linguistics

University of Zurich
Department of Computational Linguistics
Binzmühlestr. 14
CH–8050 Zürich
Phone +41 44 635 43 31
Fax +41 44 635 68 09
www.cl.uzh.ch

UZH, Department of Computational Linguistics, Binzmühlestr. 14, CH–8050 Zurich

Rico Sennrich
SNSF Professor
Phone +41 44 635 71 31
sennrich@cl.uzh.ch

Zurich, February 26, 2023

Thesis Review: Dušan Variš

This thesis by Dušan Variš, titled “Learning Capabilities of the Transformer Neural Networks” forms a substantial body of work that analyses the learning capabilities of Transformers with different approaches, using a mixture of controlled synthetic tasks and machine translation task.

Overall, the thesis is very well crafted. The research questions are addressed in a set of carefully designed empirical studies, and result in a number of novel findings, such as demonstrating the limited generalisation capability of standard Transformers towards outputs of different length than those seen during training, or producing words composed of a higher number of subwords. Without doubt, the thesis demonstrates Dušan Variš’ ability for creative scientific work.

The thesis begins with providing background on multi-task learning and neural sequence-to-sequence modeling. These sections adequately prepare the reader for Dušan Variš’ own contributions, and I especially commend the breadth of literature considered in the chapter on multi-task learning, and also in the intial part of chapter 4 on generalization. Going beyond the literature in natural language processing and machine learning, the chapters also tie the generalisation abilities of deep learning methods to those of humans, drawing on literature from cognitive science and psychology.

Chapter 4.1 makes a convincing case that the generalisation abilities of Transformer models are limited when models are provided with very homogeneous training data, specifically data where the end of sentence occurs within a narrow range of values at training time. While such a setting is a bit extreme and artificial, demonstrating this lack of generalisation is important in the context of previous work that tried to draw conclusions about long-distance capabilities of different neural architectures where a similar mismatch existed between training and test length of sentences. Based on the discussion in the beginning of chapter 4, one thing I expected were experiments that vary the Transformer capacity to empirically test the relationship between overparametrization and generalisation, but this omission is forgivable given that other research questions were answered. Similarly, the thesis demonstrates in 4.3 a lack of generalisation when it comes to the ability to form new words from subwords, if there is a mismatch in their length between training and test time.

Chapter 4.2 extends experiments about generalisation capabilities by filtering the training data to be more dissimilar in terms of word distributions with respect to the validation/test set. Two similarity metrics are considered, TF-IDF and a language model. While the experiments do not confirm the hypothesis that there is a lack of generalisation, I wonder how results were affected by the choice of filter. As discussed in the thesis, language model scores were not normalized by length, resulting in a removal of short sentences first, which probably more effectively introduces a length bias rather than affecting distributional similiarity.

Chapter 5 presents a study of Elastic Weight Consolidation (EWC) as a regularizer during fine-tuning of pre-trained models. The thesis applies and tests EWC empirically on a new task, machine translation, but also introduces variants of the algorithm to normalise the Fisher information matrix, and stabilizing the contribution of the regularizer. While both the modifications of the algorithm and the experimental setup are convincing, the evaluation shows that the method underperforms some baselines. For language model regularization, one finding is that regularizing the model to stay close to the left-to-right language modeling mode may be detrimental for performance of the bidirectional encoder. In the multilingual MT experiments, different choices of EWC regularization offer different trade-offs between performance on the initial and fine-tuned dataset. I understand that computational resources are limited, but one question left open here is the effect of the hyper-parameter λ, which is constant across experiments, since EWC normalization and stabilization may affect the strength of regularization with constant λ.

Chapter 6 presents a modular Transformer based on a controller that controls the activation of Transformer components on the token or sequence level. Differential training is enabled via a Gumbel-sigmoid function, and an auxiliary training objective is added to control the “budget” (proportion of active modules). One notable difference to some previous work is that previous work has combined modularization with an increase in the number of modules/parameters, effectively reducing the parameter bottleneck, whereas in this work, the modularization is done without such an increase, merely allowing the model to learn a higher degree of specialisation via modules. It is thus remarkable that the method does achieve substantial improvements in multilingual translation.

In summary, the contributions of the thesis have importance in the field of machine translation, where his results on generalisation will inform the choice of training data and further efforts in developing more generalisable architectures, and his contributions in multi-task learning with EWC regularisation and modularisation provide valuable insights for future work. Given the generality of the Transformer, the sequence-to-sequence model under investigation, the insights also bear relevance to natural language processing and machine learning in general. In conclusion, I reconfirm that the thesis forms a

substantial contribution to the scientific knowledge, and is worthy of the award of a Doctoral degree.

Yours sincerely,

Rico Sennrich
SNSF Professor