PhD Thesis Report

Thesis Title: Learning Capabilities of the Transformer Neural Network
Thesis Author: Mgr. Dušan Variš
Reviewer: Mgr. Ondřej Dušek, Ph.D.

Thesis Contents Summary

Dušan Variš's PhD thesis is analyzing and experimenting with neural networks based on the transformer architecture, focusing especially on their ability to learn in complex settings. The transformer architecture, originally developed for machine translation (MT), is arguably one of the most important neural architectures nowadays, as it underlies most neural natural language processing (NLP) systems and sequence transformation tasks in general. This includes pretrained large language models (LMs), which are at the forefront of current research in artificial intelligence (AI). The author specifically looks at the transformer's data overfitting, incremental/continual learning and multi-task learning properties. He experimentally evaluates his research questions on synthetic string-editing tasks as well as multilingual machine translation.

Although the author's published works over the course of his PhD are much more extensive and relatively broad (centering around MT, but including other NLP tasks such as tagging or image captioning), the research presented in the thesis is mainly based on two published short conference papers (Chapters 4 and 5). The experiments are substantially extended in both cases. In addition, the thesis offers another independent set of as of yet unpublished experiments (Chapter 6). The experimental contributions involve specifically:

- *Transformer overfitting to sequence lengths and vocabulary* in Chapter 4: This includes three sets of experiments related to transformer overfitting, the first of which was published at EMNLP 2021. (1) Experiments with target sequence length show severe overfitting to training data sequence lengths in transformers and thus their limited generalization capability. (2) Overfitting to particular vocabulary is not present despite original expectations. (3) On the other hand, transformers are sensitive to the maximum number of subwords per word in the training data, which again limits their generalization abilities.

- *Incremental learning and the use of elastic weight consolidation (EWC)* in Chapter 5: In an initial experiment (published at ACL-SRW 2019), the author extends the general EWC setting to the task of MT and shows it as a possible regularization option for an MT model with unsupervised pretraining; the performance, however, does not beat a backtranslation scenario. Based on EWC problems found in the first experiments and following related work, the author proposes modifications to the EWC formulation (normalization and stabilization), reducing the method's sensitivity to hyperparameter choice. Further experiments with a continual learning scenario in multilingual MT (high-resource languages to low-resource ones) show that EWC presents a tradeoff between learning new language pairs perfectly and forgetting old ones.

- *Modular extensions to the transformer architecture* in Chapter 6: The author proposes novel extensions to the transformer architecture: a controller module decides on switching on or off a set of modules working in parallel on each transformer layer. Hard decisions on module state are smoothed via Gumbel-sigmoid or straight-through estimation. The results on MT show that the modularized transformer can get modest performance gains while reducing the overall number of active model parameters at a given time, but it does not show any specialization of the individual models (e.g., with respect to language pair).

The experimental chapters are accompanied by the necessary theoretical and formal chapters as follows:

- Chapter 1 presents a brief introduction and motivating background, introduces the main research questions of the thesis and presents the author's contributions. It also provides a short overview over the following chapters and comments on the publications underlying the thesis.

- Chapter 2 represents a relatively brief general overview of the problems of multi-task learning and incremental learning in neural networks.

- Chapter 3 gives an overview of neural sequence modeling, with a particular focus on the transformer architecture and its training.

- Chapter 7 adds a short conclusion after the experimental chapters, summarizing responses to the research questions and sketching potential future work.

Overall Evaluation

The extent of research work carried out here fulfills the usual requirements for a PhD thesis. The author demonstrated his ability to conduct novel and independent research. This work expands our understanding of the transformer architecture and provides a solid basis for further research in the area.

What I really liked about the thesis are the research questions, the reasoning behind the evaluation and the use of synthetic tasks as well as practical MT experiments. I believe that the experiments in the thesis are novel and interesting; their overall structure also very nicely demonstrates the author's deep understanding of the architecture and machine learning theory in general. I liked that the author is pushing the transformer architecture and testing its limits in non-conventional settings. This does not go with the general mainstream flow of today's NLP research and asks about deeper questions regarding the architecture, essentially bringing in more machine learning theory than usual in our field, but still staying deeply rooted in NLP. I really appreciate that the thesis also includes experiments that do not confirm the author's initial hypotheses (as is the case with potential transformer vocabulary overfitting in Section 4.2).

The results on transformer length overfitting (on both overall sequence lengths and subwords per word; Sections 4.1 and 4.3) are really interesting and should inspire more further works on overcoming these problems. The experiments with EWC in Chapter 5 present an interesting trade-off option between forgetting a previous task and learning a new task perfectly. This may be useful where storage is scarce. The transformer modularization described in Chapter 6 reaches slight performance gains while using fewer parameters at any given time, which makes for a promising regularization option, even if it does not yield task-specialized modules as originally expected.

I am generally happy with the way the experiments were conducted, but I do believe that more attention should have been paid to general randomness of neural network experiments. The results presented seem to be based on single randomly initialized runs only, and they do not include any statistical significance tests. This reduces their overall explanatory power. In addition, hyperparameters are mostly just listed, but their choice is not very well explained. In one case, they are admittedly left unoptimized (Section 6.3). I understand that computational resources are at the core of the issue here, but I do believe that multiple random seeds would have been easily computed at least for the synthetic string editing tasks, and that simple statistical significance tests such as bootstrap resampling can be performed relatively easily in all experiments. In addition, I believe that the hyperparameters could have been optimized in a scaled-down setting but on the same task for Section 6.3.

As a side note, I was not so keen on the discussions of general AI and parallels of transformer neural network learning with human cognition. I believe the inspiration by human cognition in artificial neural networks is very loose at best, and discussing these parallels does not help to understand the neural architectures very much. Furthermore, it may have the undesired side effect of over-hyping the actual state of the art, which is currently nowhere near general AI. However, this is purely my own personal taste and I acknowledge that it may be controversial.

I regard the thesis text as generally well written, mostly self-contained, albeit a little rough around the edges. The language shows occasional traces of edits (such as missing or duplicate words, or non-matching syntactic constructions). The theoretical chapters, especially Chapter 2, could go into more detail and mention approaches a little closer to the author's own work (even though these closely related works are referenced from the experimental chapters). Furthermore, a few related works are commented on but not really explained (see details below). I am also not sure about the order of Chapters 2 and 3 – I believe swapping them would make the text flow better.

A similar problem with the text being a little rough is the absence of chapter content description or any introductory commentary or guidance to the reader in most cases. Most chapters delve straight into the contents without providing much background or context; it is OK and understandable, but not very reader-friendly. The thesis offers a very good English presentation at a near-native level, which is only hampered by the non-standard occasional punctuation or determiner use (with no significant effect on readability).

Recommendation

Based on the evaluation above, my recommendation is that the thesis be **approved** for a PhD. My reservations are relatively minor and do not challenge the thesis as a whole.

Detailed Comments

I am including some detailed comments as suggestions to the author, especially for the case he decides to publish the currently unpublished parts of the thesis. They all revolve around minor issues.

Regarding Contents:

Chapter 1

- Most of the research question introduction feels more like a detailed summary; it is OK but I was not expecting to get the answers already.
- Some statements in the research question background paragraphs would be better backed by citations.
- There are a few unclear references ("this behavior", "this belief" on pg. 4).
- I am not sure how the quadratic attention cost impedes the transformer's "effectiveness"? Do you mean efficiency here?
- On top of pg. 5, I believe you mean the dissimilarity of vocabulary distributions?
- Some references (e.g., module budget, adversarial datasets) are not yet clear at this point in the text.
- The chapters overview is missing Chapter 6.

Chapter 2

- The references to Bapna and Firat (2019) and Zhang et al. (2021a) are not explained.
- The introduction of multilingual MT is a little strange on pg. 16-18 and should be better commented (potentially in its own subsection).

Chapter 3

- The reference to Bogoychev et al. (2020) is unclear.
- Positional embeddings (Gehring et al., 2017) are not explained.
- It is not clear why the improvements to positional encoding described on pg. 28 did not catch on.
- Random sampling is not actually explained (only commented on) in Sect. 3.2.2.
- The final paragraph of Chapter 3 would better fit Chapter 4 introduction.

Chapter 4

- The Type I error on pg. 40 is not well explained.
- The averages in Table 4.1 do not seem to match the above values for the 0-10 and 16-20 columns.
- I am not sure what you mean by "some form of overfitting can be desired" (pg. 54).
- Is the number of training examples in Figure 4.12 measured in thousands, or really as individual examples?
- The last paragraph on pg. 56 is hard to follow.
- The reference to "subword length" in Section 4.3 is ambiguous – this could mean the number of characters per subword, and can only be understood by inferring from context. Note that "tokens" typically refer to subwords, not words, in neural sequence processing.

Chapter 5

- The methods for avoiding CF (end of pg. 66) should have been described in more detail, either here or in Chapter 2.
- A similar problem applies to the introduction to Section 5.1: The works cited there should have been explained, at least briefly.
- You never mention explicitly in Section 5.1 (apart from the title) that what you are doing here is actually describing EWC.
- The regularization schemes mentioned at the start of Section 5.3 would better be explained in more detail.
- I am not sure why you introduce the incremental learning taxonomy on pg. 77. I do not see it explaining anything.
- In Section 5.3, you should mention explicitly that you are working with synthetic data again. Also, mentioning MT at the end of the introductory text is a little surprising, given that you did not explain that task at all in the preceding text.
- I am not sure Figure 5.6 should use lines to connect the different values of λ.
- You mention a "multi-task setting" at the end of Section 5.3.2, but I do not believe this is what you are doing in Section 5.3.3 – I believe it is still incremental learning?
- Having at least a small-scale manual analysis for the MT experiments in Section 5.3, like you do in Chapter 6, would have been preferable.
- It is not clear which "bilingual baseline" you mean on pg. 86.
- How different are your experiments from the experiments by Thompson et al. mentioned on pg. 87?

Chapter 6

- Note that Chomskyan hypotheses about language acquisition are not undisputed; I would consider the hypothesis you mention on pg. 89 also an "oversimplification".
- The argument about the lack of control mechanism and the mixture-of-experts on pg. 90 is not entirely clear, you may want to point out in what way it is different from multi-head attention.
- It is not entirely clear how exactly your architecture shown in Figures 6.1 and 6.2 relates to the deep averaging network architecture.
- It would have been better to also include a verbal description of Equation 6.11.
- The string operations in Section 6.2 would be better explained with some examples.
- You should comment in the main text of the thesis on the number of different modules used.
- I did not entirely understand your argument on encoder-related module selection on pg. 98.
- Regarding your remark that individual MT language pairs are "not as distinguished" on pg. 102, I think you could potentially argue that these are distinguished more clearly than for the string edits, due to using different scripts/vocabulary.
- Please do not use the word "significant" unless you actually run statistical significance tests (pg. 103).
- I would say that many-to-many MT models are not "more complex", the architecture is the same – just the data/task is more complex (pg. 104).
- The captions for Figures 6.9 and 6.10 are somewhat confusing; you should point out that the different values in the same layers are for the individual modules.
- I am not sure what you want to say by the sentence starting "The result of the multilingual evaluation [..]" on pg. 111.

Regarding Writing: These are all very minor problems, but most of them are repeated multiple times throughout the text; I am only referring to one or two examples per problem type here:

- The article use is not always standard, the most frequent problem is the overuse of definite article (e.g., "the deep learning [..] algorithms are able to reach the solutions [..] with low generalization error" on pg. 38). In this particular case, it would be better to remove the definite articles (abstract nouns in plural) and add an indefinite in the last phrase (adjectival modifier).
- The punctuation follows the Czech standard more than the English convention in some cases (e.g., "for a scenario, when [..]" on pg. 71, "The trade-off [..] suggests, that [..]" on pg. 87).
- There are several spurious abbreviations – defined but never or rarely used (e.g., DL, CI, PI, MEE); "GPU" is defined after use.
- Some pronominal references are not entirely clear and would have been clearer if the respective noun was repeated (e.g., on pg. 38 "They" referring to Belkin et al. or "they" referring to "transformers" at the start of Section 4.1).
- Some references to other parts of the text are very vague (e.g., Footnote 3 on pg. 28 or the end of Chapter 4 introduction without explicit links).
- There are a few instances of incorrect modal use (e.g., "some form of overfitting can be desired" on pg. 54, "may" would be more appropriate).
- The bibliography has a few duplicate entries (e.g., Cho et al., 2014; Masson D'Autume et al., 2019). The Conneau & Lample entry is both duplicate and referencing a whole proceedings volume, although the intention was probably to reference a particular paper.

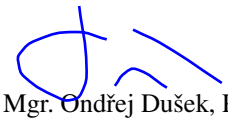
Questions

After reading the thesis, I have a few questions for the author, some of which could be discussed during the defense:

- Do you have any ideas about solutions to the IID problem in evaluation described on pg. 39?
- Can you explain why the performance on long strings is zero in Table 4.1 whereas it is reduced but non-zero on short strings?
- In Section 4.2, I believe the various "newstest" datasets are quite different, covering different topics from different years? Why do you not use the same set for filtering the training examples?
- How do you set the λ parameter in Section 5.2.2, and how come it is so different from the value range used in subsequent experiments? Is it solely caused by the normalization?
- Why did you choose top 4 best checkpoints (and not, say, 3 or 5) in Section 5.2.2?
- Did you omit applying masked language modeling in Section 5.2.2 simply due to lack of resources, or would there be another problem?
- For the synthetic experiments in Chapters 5 and 6, you use transformer models of depth 1 only. That feels like a surprising choice, why did you make it and how would using two or more transformer layers influence the results?
- Regarding the MT experiments in Section 5.3, is having all the corpora available for tokenization realistic? How would you work in a situation where you start with the high-resource languages only?
- While I agree with the statement that "the output of the multi-head attention block is always a combination of all attention heads", I do believe that input-conditioned attention weights can influence the size of the individual contributions (up to making some of them insignificant) – am I missing something here?
- What was the reason for choosing different string operations in Section 6.2 to what you used in Chapters 4 and 5? Did you also try using *copy* and *reverse* instead?

- Would it make sense to measure the module selection entropy per layer, not just per module?
- What happens when no modules are selected in a certain layer? I believe that some information must still flow to the next layer?
- How big was the overlap between annotators on the German translations in Section 6.3 and how often did they agree on the ratings? Is the high total number of ratings in Table 6.3 (443, much higher than the stated 163 sentences) due to using each rating individually on the overlapping sentences? How come that the different models do not have the same number of sentences rated (in both Tables 6.3 and 6.4)?

Prague, 5 March 2023



Mgr. Ondřej Dušek, Ph.D.
Institute of Formal and Applied Linguistics
Charles University, Faculty of Mathematics and Physics
V Holešovičkách 747/2, 18000 Praha 8