Dr hab. inż. Maciej Piasecki, prof. PWr.
Katedra Sztucznej Inteligencji
Politechnika Wrocławska
maciej.piasecki@pwr.edu.pl

25th March 2025

Reviewer's opinion
on Ph.D. dissertation authored by

Sebastian Jaszczur

entitled:

Automating Competency Handling in Ontology Development Process

1. Problem and its impact

What is, in your opinion, the most important problem discussed in the dissertation?
The presented dissertation has a surprising hybrid character, i.e. it is based on a series of 3 research papers, all published in two prestigious scientific conference (namely: Conference on Neural Information Processing System (NeurIPS) 2020 and 2024, and International Conference on Machine Learning (ICML)), but next it has been expanded with two unpublished works that do not fulfil formal requirements to be included in a thematic series of publications presented as a PhD thesis. These two texts were uploaded to arXiv – a commonly used preprint repository, but are not peer-reviewed works published in proceedings, journal or book. It is worth to notice that these two unpublished works have already been cited more than 50 times, but this does not change their status.
The title of the Sec. 1.8 "List of Publications" is very misleading and wrong, as the list includes 3 publications and 2 unpublished works (sic!) – uploaded into the *arXiv* repository and without clear information about their current status. It is not a simple a editing error, as the reviewer searched intensively for the publication details and could not find anything but *arXiv* and rejection information on Open Review system. Moreover, the author is not correct declaring "My thesis comprises of five papers", as, from the point of view of PhD procedure, it comprises of only three published scientific papers. The other two unpublished works should be treated as additional expansion added to the thesis, and, e.g., only mentioned in the introductory chapter. That is why I described the presented thesis as a kind of 'hybrid'. It would be much clearer and much better to present a regular dissertation citing and based on all different works of the author, instead of the present form sitting in the middle between a series of publication and a regular thesis.

Because the thesis is based *de facto* on a series of publication, it is not surprising that its goal is quite general and summarises the goals of the three individual papers in the series in some way:

> "This PhD thesis examines the potential of improving Large Language Models (LLMs) with Conditional Computation. [...] This PhD thesis examines the potential of improving Large Language Models (LLMs) with Conditional Computation."

Still, the fourth work (not a part of the series) does not fit to this goal, as "Structured Packing in LLM Training" is not related to Conditional Computation. In addition, in the fifth work (also belonging to the series) experiments were performed on relatively small models, but with Conditional Computation.
The main problem with the goal is that the term: "Large Language Models" has not been defined in the entire thesis and all three papers. It is true that this is common practice in literature, but here this term

has been intentionally used in the goal description, so it should be defined. The presented experiments are mostly dealing with models of the size below 1 billion parameters, that are rather not considered LLMs (especially recently) and also were not in the moment of writing the papers comprising the thesis.

Concerning the three published research papers, that should be considered as the basis of the dissertation from the formal point of view (because the other two are not published), their goal has not been characterised but fits into the general goal. These three papers explore Mixture of Experts to make the transformer architecture sparse and improve computational cost – first of all inference efficiency, and Scaling Laws, expanded to the Mixture of Experts applications.
The main assumption formulated is as follows:

> "Feed-forward layers contain the majority of Transformer parameters and consume the majority of the Transformer's computational budget, measured in terms of FLOPs. Consequently, replacements of the feed-forward layer, like Mixture of Experts, are the primary focus of this work."

Is it a scientific one?
Both aspects, i.e. conditional computation architectures of deep neural networks and scaling laws, especially in terms of predicting computational budget required to achieving the assumed performance, are scientific questions, even if studied in relation to language models of smaller scale.

Does it have a practical meaning?
Both issues should have direct influence on practical applications. This has not been convincingly presented in the thesis, at least explored deeply enough, as only some parts of the experiments were related to practical NLP tasks, and even those were limited to well-known and commonly used benchmarks that are over-explored and do not necessarily tell the truth of practical applications.
A positive exception is Terraformer proposed in the first paper and its application to summarisation of longer articles (with a selection of examples in the appendix).

2. Contribution

What is the main, original contribution of the dissertation?
This question should be approached from two perspectives. The first is formal, but very important for the PhD procedure, namely, what is the personal contribution of the author to the different works, and to series as a whole. The second is around the question of what the main original contribution of the presented work to the development of the given scientific domain is? The domain that can be identified as Natural Language Processing.

Concerning the first perspective, and the three papers comprising the proper series (as being really publications in contrast to the 4th and 5th unpublished works), the author declared he was responsible for the creative aspects of 1s, the aspects of design, coordination and experimental verification in 2nd, and the idea, design and "supervision" in 3rd. Thus, we may conclude that the author had very significant responsibility in the relation to the core innovative aspects of these three papers.
What may worry, especially the reviewer, is that the procedure adopted in the faculty does not require confirmation from the co-authors (confirmed by the letters exchanged), while there are PhD students among them and they may also refer to the same papers during their PhD procedures.
It is worth to record, that the reviewer has not received confirmation of the contribution of Mr Jaszczur to the presented three papers signed by the co-authors.

The potential problem is already visible in the case of the fifth work which pop-ups in the web under the name of the other co-author as being responsible for it. However, it must be immediately emphasised that these two unpublished works, namely fourth and fifth, should not be treated as proper elements of the series presented as a thesis. This leads to the question: what was the reason for including them into the presented series? Especially as the role of Mr. Jaszczur was not very significant in them. One can guess that this was done to increase the length of the presented series, as three works are a series, but not very long. Nevertheless, the first three papers have been published in the world top rank conferences, present interesting novel solutions and consistently explore closely related topics. They are also obviously related to the thesis goal.

Coming to the second perspective — contribution to the domain — the three papers and the fifth work too, are exploring ways of improving the efficiency of neural language models by Conditional Computation and deeper understanding of relations between the network structure, training data and its performance, as well as taking a novel perspective on the expected performance within the assumed computation budget. Both issues are very important for the domain, even more important now, as rapidly growing cost of applications of Large Language Models causes more attention given to their efficiency. In closer look, in the first paper the authors propose introduction of „sparse variants" of the different layers of the Transformer architecture, i.e. a variant of the Mixture of Experts scheme in which „a full weight matrix" is trained and „then only activate specific parts of it for each input token during decoding". An innovative extension of Mixture of Experts was next appropriately experimentally evaluated. Very significant speed up in decoding was observed in comparison with the state of the art, which is especially important for applications.
It is worth noticing that the authors systematically explored the idea not only replacing the Feed Forward (linear) layers, but also different components of the attention heads (but without bigger success).
Moreover, a complete, novel expanded structure of a transformer was proposed: Terraformer — "a Transformer model that uses reversible layers for memory efficiency and sparse attention to handle long sequences.". The authors showed that it is not only faster in text generation, but also it is a good basis for a challenging task of summarising longer scientific articles – one of a few examples of practical applications in the thesis.
In the second paper, the authors study the relationship between the number of parameters, training data size (training tokens) and the sparse transformer structure vs its performance. They propose expansion of the scaling laws to sparse transformer networks and introduce a novel hyperparameter of granularity. It allows for studying a range of Mixture of Experts architecture from the point of view of efficiency and "the optimal adjustment of the size of experts". A novel "parametric scaling law for predicting the final loss value L based on granularity G, the total number of non-embedding parameters N, and the number of training tokens D." was introduced. It facilitates estimating the performance relations between a transformer and Mixture of Experts model in relation to the aforementioned parameters, e.g. for more informed design decisions and "computing-optimal settings of Mixture of Experts models", but also guiding future research.

The third paper explores further the idea of Sparse Transformers targeting the problem of their discontinuity related to the way in which subgroups of parameters – "experts" – are selected. A new "Mixture of Tokens, a novel continuous Transformer architecture closely related to sparse Mixture of Experts" is proposed and thoroughly studied. It introduces an innovative and intriguing idea of a combined representation of tokens delivered to the experts that makes the architecture better integrated with the training process and matching its continuous character. It is a novel generalisation of the Mixture of Experts architecture that increases its flexibility and improves performance, e.g.

> “Our best MoT models not only achieve a 3×increase in training speed over dense Transformer models in language pretraining but also match the performance of state-of-the-art MoE architectures.”

Its major limitation is larger memory complexity, as the authors noticed, and this is especially problematic in the case of inference not performed in batches, like in the case of standard use of Large Language Models.

The fourth work (unpublished) deals with an important challenge for language models, i.e. effective processing of large input context. The presented work concentrates on one specific aspect of the problem, i.e. better utilisation of “training data, keeping other components, such as the architecture and training objectives, unchanged.” As a result, it is loosely connected to the other works of the thesis but brings observations that relatively simple amends can result in practical gains. A simple scheme of composing longer context input data from available collections, called Structured Packing for Long Context (SPLiCe) was proposed. Its core is simply building training examples by linking text fragments that are semantically related. The proposed method can explore structures of the data collections but also use some similarity measures. The authors showed good results with using a so simple means like BM25 information retrieval technique. While the whole approach brings promising results (e.g. knowledge transfer between domains of program codes and text), the scope of experimental evaluation is quite limited to mainly BM25, single link data organisation and smaller language models of 2B–6B tokens. As in the case of most works, less attention was paid to detailed analysis application tasks in which the expanded models were tested. It has not been presented how the proposed training data generation can be connected to the sparse transformers that are the core of the thesis.

Finally, the fifth work (also unpublished) shows that the studied Mixture of Experts can be used for successful expansion of the Mamba deep neural network and especially language models based on it. The authors provided “empirical validation of our hypothesis that interleaving Mamba with MoE can improve the performance of a model” and they did it in a very careful and convincing manner, e.g. performing “extensive ablations on the model architecture, number of experts, and other design choices”. They also tested the introduction of sparse layers into Mamba blocks and parallel placement of Mixture of Expert layers but without significant success. Thus, the fifth work repeats to a very large extent the experimental scheme known from the papers one and two. However, the Mixture of Tokens architecture introduced in the third paper was not used, without any explanation.

So, the expansion pattern seems to work on yet another deep network different, but similar to Transformer. However, the key question is if Mamba is worth attention and efforts in its scaling? Concerning its potential, it is a language model equivalent to BERT, that can be hardly considered a Large Language Model (especially according to contemporary understanding), and so being outside the declared scope of the work series of this thesis.

3. Correctness

Can we trust what is claimed in the dissertation?

All papers in the series are well planned and written in a systematic way. All the decisions and steps are very well described. This is especially visible in the first and second paper. Concerning the latter, all the derivation of the proposed scaling law for Mixture of Experts are carefully motivated and introduced step by step. The formulated claims are validated by experiments performed and analysed in a proper way. All the experiments were done on textual datasets of large, practical scale. The only limitation was that the authors focused mostly on different forms of text generation perplexity, and experiments with application of the language models in different text processing tasks received much less attention, especially from the point of view of the in-depth analysis of the obtained results. Most of such experiments were done on well-known benchmark datasets, and, e.g., error analysis and comparison are shallow. Yet another question mark is a limited scale of the models trained and evaluated in comparison to "Large Language Models" announced to in the thesis' goal, at least according to the present understanding of this class of models. However, in some experiments the estimations made suggest transferability of the results onto large scale models, too.

Are the arguments correct? Indicate the flaws you have noticed, if any.
Most of the presented works are focused on improvement of inference efficiency and perplexity of the language models. Vast majority of experiments show that the solutions proposed by the authors express better performance, or at least constant positive tendency in comparison to the state-of-the-art models. However, behind this generally positive picture one can notice some less clear aspects. They may also explain relatively modest up-take and influence of these solutions proposed.

In the first papers it is argued that accuracy of Scaling Transformers and the proposed "Terraformer" model are similar, but we may observe that the differences are quite small, and not certainly significant. Moreover, they are not analysed in depth.
As it was mentioned, in most cases the author concentrates on perplexity which is helpful but only an indirect signal of the future performance of a model while fine-tuned to the different tasks. For example, Terraformer was evaluated only in one task for which Rouge metrics were reported but not analysed (in addition, Rouge metrics by themselves also provide only shallow, technical characteristics of the errors). Concerning the distance between the model generative power and the future applications, e.g., approaches to knowledge editing in Large Language Models suggest that it is the Feed Forwards layers where the factual knowledge is stored, while these layers are targeted in the techniques proposed in the thesis.
In the second paper, the evaluation is quite shallow, only loss is used and reported, but in fact it is hardly specified in the paper.
In the third paper, most of experiments were conducted "on two model scales: a 77M Medium model and a 162M Base model", and no model larger than 1B parameters was trained.

The fourth and fifth works (that do not belong to the paper series, so we put them on side), also include some issues. In the fourth work, it is quite surprising that the BM25 ranking measure results in the best performance. This effect has not been studied there deeply enough. However, it was not also compared to several other possible measures and, e.g., bi- and cross-encoder architectures.
In the fifth work, experiments were also performed on very small models 25M and 100M (this may be caused by the characteristics of the Mamba model). Moreover, initially, only perplexity is defined as the evaluation measure, while later accuracy is mentioned, but without a proper definition.
Structure and presentation of ideas

All five works have not only been copied into the thesis but interlinked in a way helpful for the reader. Their introductions and conclusions have been slightly rewritten to reveal the links among the works, and to emphasises the narration set up in the introductory first section. This structure distinguishes very positively the thesis from many similar ones based on series of publications. However, as it was already noted, the thesis has a hybrid form, and the last two works are not publications and should not be presented as being a part of the series proper. It would be much more honest and fairer in relation to the reviewer and procedure to present this additional works only within the introductory section as a kind of further work. An even better option would just to write a regular dissertation and expand it with all other relevant research results that had not been included in the three publications forming the series. The introductory section presents the structure of the thesis and main research tasks undertaken in very good way and is well balanced between delivering an overview, defining the research directions, and keeping it relatively short.

4. Knowledge of the candidate

What are the chapters of the dissertation (or sections in chapters) that resemble a tutorial and thus confirm a general knowledge of the candidate in the discipline of ***Information and Communication Technology****. What areas of that discipline are covered by those chapters/sections? What do you think about quality of those chapters/sections?*

The first section presents deeper understanding of the author of the domain deep neural networks and language models. Related works in the three publications of the series show also good orientation of the candidate in the state of the art. However, the papers are quite narrowly focused and do not go deeper into more generally understood Natural Language Processing, that is especially visible in the limited scope of evaluation experiments. Without no doubts, the author has good knowledge in Machine Learning and many technical aspects of Computer Science, e.g. open-source codes have been published for all works discussed in the thesis.

What is your opinion on the list of references? What is the degree of its completeness?

The list of references is a good background for the discussed issues and was representative on the moment of writing the three publications of the series.

5. Other remarks

Additional detailed comments:

p 6: "oblicznia warunkowe" – a typo

p 12: "Feed-forward layers contain the majority of Transformer parameters and consume the majority of the Transformer's computational budget" — lack for source for the claim

p 15: Eq. 1.3 — the parameters are not precisely named and explained in text

p 22: "three out of these five publications" — not all of them are really publications, so only two of them are a proper part of the publication series being a basis for this dissertation

p 24: "P2" is repeated affecting the numbering further on

p 39: if active parameters are counted in a way excluding the routing parameters, does not it bias the complexity analysis?

p 41: (3.3) — at least h is not explained

p 47: Table 3.3 — not clear if these results are reported on the basis of analytical estimations or experiments/simulations

p 57: "the efficiency of Tokens" — a typo

p 57: "Our main result is a substantial speed-up of MoT models compared to dense Transformers (Figure 4.6) and results comparable to sparse MoEs (Figure 4.5)." — what is the reason as there is no routing and all experts always process mixtures of tokens?

p 67 : "relatively short, compared to pre-training, fine-tuning." — fine-tuning is misleading, continually pre-training, continuation of pre-training?, on the same page "we continue training on ..."

p 67: "50/50 mixture of RedPajama," — missing citation, 50/50 is unclear, lacks explanation p 69: “which are text classification” —quite general characteristics

p 71: “SPLiCe models have improved question answering capabilities.” too hasty or too far going, the differences are very small, and the general level is low

p 71: “SPLiCe improves the two-shot performance.” — unclear, no information both in the table caption and in the text

P 71: Sec. „Lost in the middle” — the selected task seems

p 73: “use ’standard’, as” — editing, wrong opening quotation mark

P 82: “smaller, □25M” — mysterious, unexplained square symbol

p 82: “We optimize and report cross-entropy loss, which equals log perplexity per token.” — but this type of limited evaluation does not show applicability of the proposed solution, especially when combined with small size of the models

p 82: „a capacity factor of 1.0” — unclear

p 83: Table 6.2 — what was the reason for using the different number of experts in Mamba and the transformer referred to?

p 89: “Transformer-MoE consistently achieves higher accuracy than MoE-Mamba” — how was this measured?

p 94: “we would advocate for more research to be presented in the form of the scaling laws” — why is this so important? Scaling laws are useful estimations, meticulously tuned in experiments. They provide some predictions, but as the author noticed they provide a very partial picture.

6. Conclusion

Taking into account what I have presented above and the requirements imposed by Article 187 of the *Act of 20 July 2018 - The Law on Higher Education and Science* (with amendments)[1], my answers to the following three basic criteria – questions:

- *Does the dissertation present an original solution to a scientific problem?*
- *Does the candidate possess general theoretical knowledge and understanding of the discipline of Computer Science?*
- *Does the dissertation support the claim that the candidate is able to conduct scientific work?*

are positive and I recommend to proceed with further steps of the PhD procedure and public defence of the thesis of Mr Sebastian Jaszczur.

However, I would like to emphasise that, as

- the presented series of three papers, in which Mr Jaszczur had a leading creative role, but being a member of the team, is relatively short,
- as well as that this series was artificially expanded with two unpublished work,

the formal status of the thesis is not completely clear. I would like to discuss this issue during the defence.

Maciej Piasecki

Signature

[1] http://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20190000276