Master Thesis Review

Charles University, Faculty of Mathematics and Physics

Thesis author	David Burian
Thesis title	Document embedding using Transformers
Submitted	2024
Program	Computer Science **Specialization** Artificial Intelligence

Review author	Dušan Variš **Role** reviewer
Position	Institute of Formal and Applied Linguistics

Review text:

The thesis studies the effects of embedding distillation from various pretrained embedding architectures (Transformer, Paragraph Vector) into a single sparse-attention Transformer model. The main aim of the proposed approach is to improve a document-level text embedding which requires capturing very wide textual context. The author observes two qualities of the existing models: ability to incorporate wide context without structure in a bag-of-words fashion (Paragraph Vector) and an ability to capture context with some of its underlying structure (Transformer SBERT model). The author leverages these varying abilities by optimizing a student model with a combined loss function to partially reflect text embeddings of both of these model architectures and demonstrates that it can lead to improvements on some of the evaluated document-level classification tasks and the studied retrieval tasks.

The author provides good overview of the studied problem and a summary of the related work. They also describe the similarities and differences of their method to the previous approaches. In Chapter 4 (Experiments), they propose the loss functions for training student models using the SBERT and the Paragraph Vector teacher. They provide a detailed analysis on the individual contribution of each approach and how they are affected by a different choice of hyperparameters. They use the similarity-based loss to distill knowledge from SBERT (structural) and a Canonical Correlation Analysis (CCA) for distillation from Paragraph Vector. The loss function assignment seems to be based on the author's assumptions about the models - what I found missing from the experiments is the comparison how well the loss functions proposed for one teacher model can be applied to the other teacher model and vice versa.

For better clarity, they introduce a normalized accuracy metric in Chapter 4 to measure model performance over the whole set of proposed tasks. The fine-grained look at the individual per-task performance was missing, same as the description of the individual task datasets. It was, however, later properly introduced in Chapter 5. Such late introduction can lead to the confusion of the

reader and should be avoided.

Besides the direct comparison of the proposed improvements during development in Chapter 4, I also appreciated the positive/negative document clustering analysis in Figure 4.4 demonstrating achieving the desired behavior of the structurally trained student model. Similar analysis of the desired behavior for contextually trained student was missing (e.g. how uncorrelated the resulting features are as they become less distant from the teacher embeddings).

Some figures in Chapter 4 were not completely clear to read due to description of the used symbols being only mentioned in the text itself. It should be included in the figure caption for better readability. Also, I was not able to find the meaning of the empty circle symbol present in some of the figures. Highlighting (or further color coding) the most important setups (further used in the follow up experiments) in some of the Chapter 4 figures would also help with readability but it is not as crucial. Further visual cues connecting repeating setups between the figures would also help. Lastly, I think that the thesis would also benefit from replacing the 4.6, 4.7 and 4.8 with a visual illustration of said projection networks or at least adding them for the reader's sake. I also suggest avoiding the use of the ArXiv citations if the cited papers have been published in a peer-reviewed journal or conference.

Chapter 5 provides a final evaluation of the best proposed variants of students and compares them with the teachers and the baseline Longformer model. The evaluation contains detailed look at the performance with respect to the individual tasks. The author also shows that while the student models can work better in a settings where only a small amount of data is available for the downstream task tuning, they start falling behind the SBERT model when the amount of available data increases. Even though the final results might seem slightly underwhelming (on classification tasks), the experiments were well designed and nicely wrapped up the study of the proposed approach. The final analysis could however be improved by a more in-detail study of the per-dataset performance, perhaps by providing a small scale manual evaluation.

To sum up, I found that the thesis motivation of combining models with different attributes for knowledge distillation is sound and the author studied this problem sufficiently with the good amount of experimental support considering the hardware available to the student. Some parts of the thesis could be improved by providing additional analysis of the models, for example theembeddings of the CCA-trained students or the analysis of the studied datasets with regards to the less expected results, such as very good performance of Longformer on imbd and PV on arxiv dataset; some level of manual inspection of the results would be helpful. The thesis presentation style still has some room for improvement but I still find it acceptable for a master-level student.

Some questions for the author:

- Has the author any suggestion how to quantify the suggested embedding model qualities (structural, contextual)?

- Have you considered swapping the proposed loss functions between the teacher models? And if not, what was your reasoning behind the decision?
- How do you deal with difference between the model dimensions when computing the structural loss? (It was not clear whether the SBERT teacher and Longformer student have same final embedding size and the links to the sources suggest otherwise)
- You claim that in your experiment settings the structural teacher is more important than the contextual one, leading to better results with the former. Have you considered that SBERT might simply be a more powerful model? Did you consider a setup where the performance difference of the two teachers would be reversed?
- It is not clear why you chose DM100 as the best variant of the DM model for the experiments (Figure 4.5 shows the DM1024 as the best performing one).
- Have you considered evaluating the model performance only on the long documents (longer than the maximum SBERT lentgh) to show the weaknesses of SBERT?

I recommend the thesis for defense.

I suggest to not consider the thesis for the annual award.

In Prague, 31. 5. 2024

Signature: