Bachelor Thesis Review

Faculty of Mathematics and Physics, Charles University

Thesis author	Samuel Michalik
Thesis title	Deep Learning and Visualization of Models for Image Captioning and Multimodal Translation
Year submitted	2020
Study program	Computer Science
Study branch	IOI
Review author	Mgr. Rudolf Rosa, Ph.D. Reviewer
Department	Institute of Formal and Applied Linguistics

Overall

	good	OK	poor	insufficient
Assignment difficulty	X			
Assignment fulfilled			X	
Total size *... text and code, overall workload*		X		

The work introduces a useful tool for visualisation of attention in models for image captioning and multimodal machine translation. The tool is nicely done, solves a current task for which there is a lack of tools, and has a potential of being used by machine learning researchers in practice.

The work also experiments with several attention models described in existing literature, trying to compare them and interpret them using the developed tool.

According to the assignment, the main goal of the thesis are the experiments and their analysis, while developing the visualisation tool is only an intermediate goal, needed to perform the analyses. However, in the actual thesis, the focus largely shifted to the visualisation tool. In particular, the analysis of the performed experiments is extremely brief and problematic in many aspects. It does a good job at demonstrating the applicability of the developed tool for a real research problem, but otherwise leaves much to be desired.

On the other hand, it is worth noting that even experienced researchers often struggle with correctly analyzing and interpreting their results. The main goal of the thesis thus might have been way to demanding for a bachelor student with little experience in the field.

In any case, the thesis required the student to study various advanced concepts well beyond the bachelor studies and to learn to use several tools and frameworks, which I believe the student mastered reasonably well.

Thesis Text

	good	OK	poor	insufficient
Form *... language, typography, references*		X	X	
Structure *... context, goals, analysis, design, evaluation, level of detail*		X		
Problem analysis			X	
Developer documentation	X			
User Documentation	X			

There is a number of typos and other minor errors, but the text is reasonably well written and understandable. Some figures and tables are not referenced from the text and sometimes not even explicitly commented on. The design of the list boxes (e.g. page 8) is disturbing.

I particularly like the high-level introduction to machine learning, which shows a very good understanding of machine learning by the student. Both of the documentation are nicely written, guiding the user as well as the developer through the application with an adequate level of detail.

The weakest point of the text is the analysis of experiment results in 3.7.3. Even though analyzing the trained models is one of the main goals of the thesis according to the assignment, it consists of less than one page of text and is simply very bad, not really saying much and even so being wrong or misleading in a lot of what it does say. The analysis includes several pages of tables and figures, but these are only commented on very briefly and often not even explicitly refered to, leaving most of the analysis and interpretation work for the reader to do themselves. It also lacks a clear story: it is not very clear what the actual purpose of the analysis is.

The quantitative analysis tries to jointly analyze results of image captioning and multimodal machine translation when clearly the results are quite different on these two tasks and should be analyzed and interpreted separately. It also fails to comment on the striking massively low performance of general attention in multimodal MT. Moreover, it severely misinterprets the validation loss and validation BLEU curves:

- For image captioning, the loss keeps rising from the start, not "after only a few epochs".
- It does not seem that "BLEU score on the validation set continued to improve, as training proceeded, in most experiments"; this seems true for Model 0 and Model 1 in multimodal MT, but not for the other models or the other task, where it keeps improving for a few epochs and then stabilizes or starts decreasing.
- This whole issue of models seemingly getting worse and worse through training is just explained as overfitting, but is not further commented on, as of whether this has any implications towards the validity of other results or what should be done about it – to me, there seems to be something quite wrong with the training process, which makes me wonder whether any of the other results are meaningful in any way.

The qualitative analysis states some facts, but the context of the findings is missing. Are we trying to make the model more interpretable? Is a more interpretable model better? Are we interested in the relation between model performance and its interpretability?

Thesis Code

	good	OK	poor	insufficient
Design *... architecture, algorithms, data structures, used technologies*	X			
Implementation *... naming conventions, formatting, comments, testing*	X			
Stability		X		

The frontend is written in JavaScript and the backend in Python. The implementation uses adequate modern tools and technologies. The code is well structured, clean, commented, and accompanied by tests. The application is rich in functionality, is quite versatile, and goes beyond what was required by the assignment.

Overall grade Very Good

Award level thesis No

Date

Signature