## External Examiner Report

Name: Mohammed Hasanuzzaman

I report here on the Ph.D. thesis titled "Multimodal Summarization", submitted by Mateusz Krubiński at the Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic. The submitted document contains 156 pages, including the acknowledgements and references sections, and is written in English. It is divided into five main chapters, starting with a background study (Chapter 1), followed by a formal introduction to the task of Multimodal Summarization (Chapter 2). The following two chapters focus on appropriate evaluation strategy (Chapter 3) and exploring datasets (Chapter 4), respectively. Experimental results are presented in Chapter 5, followed by the conclusion and future work.

## Summary Comments

The work presented in the thesis is of acceptable quality, and the student has demonstrated the considerable depth of knowledge gained throughout his PhD journey. The focus on Multimodal Summarization with Multimodal Output (MSMO) is both timely and well-articulated. The research leverages advancements in Machine Learning and Natural Language Processing to address this issue, focusing on developing an end-to-end system capable of producing multimodal summaries and creating resources to foster future

research in this direction. The structure of the thesis is coherent, showing a clear progression from background literature to approach, experiments, and conclusions.

## General Feedback/ Correction Recommendations

1. The dataset used in this thesis is primarily from the news domain. News significantly influences society and carries a profound political dimension. Therefore, computational models used for processing news articles must be rigorously evaluated not only for their technical performance but also through the lenses of ethics, fairness, and transparency. I recommend considering ethical implications to ensure that the work presented contributes responsibly to society and mitigates potential risks.

2. Create a smoother transition by summarising the previous chapter's achievements and linking them to the current chapter's goals.

3. Key terms such as 'Region Proposal Network (RPN)' and 'Faster R-CNN' are mentioned without sufficient explanation. Provide brief explanations when first introducing these terms.

## External Examiner Report

Name: Mohammed Hasanuzzaman

4. Provide a candid discussion of the limitations of the current approach, such as potential biases and the scope of generalizability of the results. Suggest ways to mitigate these limitations, such as incorporating more diverse datasets or refining algorithms.

5. I appreciate that the student highlighted the significant contribution of the MTEQA metric. I agree that MT and summarisation share similarities, but the exact nature of their evaluation differs. The metric is not closely integrated with the central theme of the thesis. Specifically, there is a lack of documentation on how this metric can be applied to MSMO or related applications.

6. The 'Background', 'Multimodal Summarization' (formal task definition), and 'Quality Evaluation' chapters occupy a significant portion of the thesis (pages 5-39), primarily focusing on existing studies. These chapters are somewhat lengthy and detailed. A more concise summary of the key points could improve readability and provide a smoother transition into subsequent sections, better highlighting the student's own contributions.

7. **Introduction**

- Emphasise the potential impact of your research on the scientific community and society at large.

- The research goals, research questions, and objectives are not clearly articulated. Include well-defined research goals, questions, and objectives. Clearly explain the importance of each element and how they address specific gaps in the literature.

8. **Quality Evaluation**

- Use consistent terminology throughout the chapter. Terms like "system output," "model output," and "hypothesis" should be used consistently.

- COMES: What are the inherent challenges in adapting an MT metric directly for summarisation? How does COMES handle the different natures of MT and summarisation data? What are the specific cases where COMES underperforms?

- While the proposed metric is well-motivated, a detailed discussion on why it is preferable over traditional metrics like ROUGE or BERTScore in specific scenarios would enhance its significance.

- The performance of the proposed metric needs to show more success in multilingual settings. Discuss how differences in language syntax, morphological complexity, domain-specific vocabulary, and contextual nuances might impact its performance.

## 9. Experiments

- The reasoning behind the assumption that a single, unified MSMO approach can handle the complexities of different modalities better than modular architectures that handle each modality separately is not very clear, apart from what is prevalent in the literature. It is also not evident from the results that the Transformer-based unified model performs better consistently. Elaboration in this regard would be great.
- If possible, conduct more extensive human validation of the datasets.
- The validation of visual outputs is not consistent with the proposed metric, which is primarily designed to evaluate text output. A more robust strategy that incorporates human judgment comprehensively would be beneficial.

## 10.  Conclusion

- Include a clear and concise summary of the main findings and emphasise the technical contribution to the field.

- Create a section where each research question is restated, and the corresponding findings are summarised.

- Clearly identify the challenges encountered during the research and discuss their potential impact on the findings.

- It is crucial to connect the findings with practical applications and their potential influence on society at large.

11.    I would recommend adopting a more cautious approach when making claims like the following: "For example, while machine learning models can classify an image as either a cat or a dog with greater accuracy than humans, or generate translations that annotators consider superior to those created by humans…". This statement might hold true for specific domains, datasets, and contexts, but I do not believe it is universally applicable across all verticals.

12.    Do not use we/our work/method/approach/solution; it is your PhD, your work.

13.    Break down complex sentences into shorter, clearer ones.

14.     Footnote markers follow punctuation marks.

15.     Never use 'get' or any of its variants. Use 'obtain', 'derive', 'is' …

16.     Always define the complete form of an abbreviation/acronym before using it (e.g. LSTM, TF-IDF, etc.)

17.     Never start a sentence with 'And', 'But' or 'Also'. W.r.t. the latter, use 'In addition', 'Furthermore.'

18.     Through proofreading would enhance the quality of the Thesis.

## External Examiner Report
### Name: Mohammed Hasanuzzaman

Mohammed Hasanuzzaman

Lecturer/Assistant Professor

Queen's University Belfast, UK

&

Funded Investigator, ADAPT Centre- A World-Leading SFI Research Centre

Chercheur Associé, GREYC UMR CNRS 6072 Research Centre, France

Room 01.006, 21 Stranmillis Road, EEECS, Queen's University Belfast, Belfast, BT7 1NN, UK.

Website: https://mohammedhasanuzzaman.github.io