Examiner's Report on Dissertation of

NTU Ph.D. Candidate Andrew KOH

Part 1: Review of dissertation

Summary review and recommendation

This dissertation addresses several applications in the area of cross modal (audio and text) machine learning, namely automated audio captioning (AAC) and language based audio retrieval (LBAR). Several novel methods that show improved performance are developed. Most of the new techniques are inspired by methods applied in somewhat analogous earning applications such as automated video captioning, but very significant insight and modification was required in order to adapt them to audio-textual scenarios. Considerable insight into state of the art deep neural network based machine learning was required to conceive and successfully implement this work, demonstrating that the candidate has a thorough and up to date understanding of the research literature of the field, and sufficient insight to alter and apply them in creative ways. The demonstrated improvements are meaningful, and the new techniques introduced here provide a base for future research. I thus find this work to be an innovative and substantive contribution that warrants award of a doctoral degree.

Review by Section

The Abstract clearly and fairly represents the work presented in the dissertation. Acronyms in the abstract such as AAC should be spelled out the first time they are used.

Chapter 2 provides a cogent summary of the fields of deep neural network based masks, and particularly to AAC and automated audio captioning, and language based audio retrieval. The explanations are clear, The document judiciously selects the level of detail, and balances well between broader exposition of the field and a more detailed focus on those areas most relevant to the topics of the thesis. The chapter also provides higher level insight enabling the reader to understand and appreciate the insight behind the author's extension of these concepts to the rather different domains of audio and text for AAC and LBAR. Well done!

On page 25, the statement "Similarly, the Audio Captioning Transformer [75] also pretrains the encoder on audio tagging before training on audio tagging." seems inconsistent or at least confusing and should be clarified or corrected.

On page 35, the following sentence copied from the text lacks a verb, and its intent is unclear: "In image retrieval, different methods such as better feature extraction using large pretrained models [98, 99], improved global and local feature alignment using variants of attention mechanism [100–103]."

The primary contribution in Chapter 3 Is the RLSSR addition to existing architectures to improve the consistency between the audio and textual internal representations, in the hope of improving the final caption consistency. This idea is a novel approach for embedding this conceptual insight into a practically implementable system in a manner allowing initialization with pretrained models. The results in the tables on page 64 do indeed show some modest improvement, when starting with pre-trained models.

The minimal benefits when starting from scratch surprise me, because it seems like it ought to provide more freedom to match the internal representations. The thesis suggests that the training sample size is insufficient in the cold start situation to provide improvement, but sufficient to provide some benefit with pre-training. This explanation is possible, but not compelling, because one might expect more relative improvement. While likely beyond the scope of this dissertation, more thorough understanding of this behavior might give clues as to how to make this approach work better, and might be a good direction for future investigation. I might begin by examining the performance when the test set equals the training set, because benefits should be expected even with a small corpus if the concept adds substantially more capability to the network.

On page 64, the best performing algorithms are not always highlighted correctly in the tables. This must be fixed.

Chapter 4 seeks to improve language based audio retrieval. As in Chapter 3, the author posits that improving coherence or similarity of the text and acoustic model embeddings will lead to significant improvements in this task. Another key consideration is that very large training datasets and pre-trained models for each of these two tasks already exist, whereas no comparable training corpus for language based audio retrieval exists nor is likely to emerge soon, so making use of these pretrained models and thus implicitly their huge labeled training datasets seems likely to benefit this task both in cost and performance. The author proposes to unify these embeddings by in effect fusing them into a single common embedding by the use of a converged tied model, which is a subnetwork that produces a common set of inputs shared by both the audio and textual components of the LBAR. And once again, a contrastive intermediate-layer term is included in the overall training metric to specifically drive the tied block subnetwork component to better consistency between the audio and textual joint embedding. The intuition behind this approach seems sound and innovative. It offers several advantages in terms of practical implementation, such as the ability to easily make use of large pre-trained models, and to avoid retraining those if so desired.

This new system with converged tied layers works better than the baseline. The performance is boosted greatly by using transformer layers in the converged tied model. The use of pretrained models appears to be essential, because the model completely fails to converge when training from scratch, at least for the small-by-comparison LBAR training datasets available. Slight improvements are sometimes observed when the pre-trained models are allowed to adapt further. Use of the contrastive supplement in the performance metric is required to improve over the baseline.

The discovery of this particular fortuitous combination of techniques providing a large leap in performance is gratifying, but makes one wonder exactly why this is the case, and whether still better combinations exist. While probably beyond the scope of this dissertation, further investigation into the nature of these results might yield valuable insights to guide future

research. Since the transformer networks can learn context, it would be worth taking a closer look at the retrieval results to ask whether the transformer has learned to recognize a few specific contexts (for example to have paired a few sets of audio and textual keywords and their features), or whether it has contributed to more general, across the board improvement in the alignment of the audio and textual features. I would suggest looking at whether the retrieval results are “clumpy”, in that it works very well for some subsets and fails completely on most others, or whether there’s a more general partial improvement over the majority of samples. The answer might suggest different ways forward with respect to future research.

Chapter 5 borrows the concept of curriculum learning from other machine learning applications and applies it to automated captioning. The curriculum learning premise is to order the training samples by increasing difficulty, training with easier subsets in the early epochs, and gradually increasing the difficulty with each epoch. This is largely an empirical study in which the candidate conjectures that “stopwords” (more generic grammatical elements such as articles and prepositions) increase the difficulty of the captioning task, and that removing such elements leaves an easier task akin to keyword matching that is inherently simpler for the network to learn. The author accordingly introduces an increasing percentage of stopwords across epochs to make the curriculum more challenging. To test this conjecture, an alternative, word frequency based curriculum is introduced in which the early epochs are dominated by stopwords, thus providing a contrast against which to compare the proposed keyword-first curriculum.

The empirical results bear out the author’s hypothesis and that the stopwords-last curriculum training performs better than a baseline system without graduated training, and that the word-frequency-based curriculum including stop words shows only modest gain over the baseline. While this appears to bear out the author’s hypothesis at words with syntactical rather than descriptive meaning particularly challenge the learning process, let me suggest that an easy opportunity may be overlooked here. The frequency based training will include some very common keywords along with the stopwords in the early epochs, and the stopwords may be confounding an otherwise beneficial strategy of common keywords first. A hybrid keyword-frequency-first/stop-words-later curriculum might capture the benefits of both of these curricula, and would explain the positive but lesser gains of the undifferentiated frequency based curriculum over the baseline. This should be an easy extension to test that might produce significantly better results for minimal effort.

The highlighting of best results for each metric in Table 5.3 is inconsistent. This should be fixed.

The summary of the thesis contributions in Chapter 6 is appropriate.

The suggestions for future research demonstrate a mature understanding of the field of the thesis and are sound and appropriate.

The last sentence of Section 6.1.2 states “This model obtained a R1 score of 0.11, a 266% improvement over the baseline score of 0.11.” The obvious error should be corrected.

These thesis document contains many small grammatical and typographical errors. It should be proofread once again before publication.