Examination statement about dissertation manuscript “*Audio Captioning and Retrieval with improved Cross Modal Objectives*” of **Andrew Koh**

The doctoral dissertation deals with machine learning methods for automatic audio captioning, which aims to produce textual descriptions of general audio signals, and language-based text retrieval, which aims to retrieve relevant audio samples from a databased, based on an input textual query. The topic is very timely, since modern machine learning methods have enabled learning models that can learn cross-modal information, and there are datasets for learning such models.

The scientific contributions of the thesis include presenting methods that guide training of a captioning system by using components and loss terms aim to reconstruct audio features from generated captions, presenting methods that share layers between audio and text branches of a deep neural network in their similarity calculation, and using curriculum learning where simplified captions are presented in early epochs to help the training procedure. The methods are evaluated using an established benchmark dataset used in the field, and at least some of the proposed methods are shown to improve the performance.

The topics addressed in the dissertation are challenging and timely, the methods are suitably chosen, and they enable addressing the topics appropriately. The results of the thesis have been also presented previously in five peer reviewed conference publications. The candidate is the first author three of them, and his role is clearly stated and sufficient to justify the inclusion of all the publications in the dissertation.

The thesis is in general quite well written. However, in many parts of the text, important details are missing. This should be addressed before the manuscript is of a suitable quality for a doctoral dissertation. There are also some minor inaccuracies, and minor problems with the language. The overall organization of the thesis could also be improved – there seems to be some repeated information, which is perhaps because the text has been compiled from several peer-reviewed publications. In the bibliography, publication information is missing from a large number of publications. I am giving a list of detailed observations related to the above issues in the appendix of the evaluation statement.

The experimental setups used in the thesis are mostly suitable to address the research objectives. The research is based on state-of-the-art datasets and methods and addressing limitations in them. In some parts of the thesis work the experimental setup does not evaluate the targeted phenomena. For example in Chapter 4, the developed methods are compared to a baseline which is based on a different model architecture (CNN vs CRNN) and is trained with different data – a setup which does not allow evaluating the effectiveness of the proposed converging tied layers, since the differences in the performance can be resulting of different data or model architecture. Ideally the experiments should be designed so that one can evaluate the effect of one factor (e.g., converging tied layers) independent of others, and if such evaluation cannot be done, the manuscript should at least clearly state that obtaining better results than a baseline cannot be used to draw conclusions about the effectiveness of one component in the proposed method. In the current version of the manuscript a reader may get misunderstanding about the conclusions, which should be avoided. I am giving detailed comments about these issues also in the appendix.

In general the thesis does not include much critical analysis or discussion of the results. A mature doctoral thesis is expected to include discussion about limitations of the work done, including the factors that I am pointing out in the review.

Overall, the thesis addresses important and timely scientific topics, but it will need revisions related to the issues I have raised before it will meet the standards of a doctoral thesis.

APPENDIX: Detailed list of issues in the manuscript

Page xiii (Abstract): When referring to automatic captioning and text-based retrieval, the thesis describes that “Both tasks require a model that is not only able to comprehend the acoustic events occurring within an audio clip, but also able to translate that information into natural language”. However, it is not clear why text-based audio retrieval would require translating information into natural language. My understanding is that retrieval could be done without that, by mapping audio and text into the same domain where the relevance score is calculated. However, if retrieval requires translation information in audio to natural language, it should be clarified where exactly such information is needed.

Page 2: The thesis describes “language-based audio retrieval involves not only detecting and comprehending acoustic sound events but also translating and aligning these events with natural language”. Similarly to the above issue, it is not clear where translation to natural language is needed.

standardized high/low -> impulse/background

Page 15: The thesis refers to “low level events” and “high level events” but it is not clear what makes an event low or high level, and what is the difference between these two categories.

Page 15: The thesis writes that “Impulse events are often short lived and and almost always overlap with background events”. It is not fully clear what ”short lived” means. Does it refer to short duration? Mentioning that impulse events overlap with background events seems weird, since I do not see how the impulsiveness of an event would be related to the background. added clarification that model has to be aware of both concurrent events

Page 15: The thesis refers to “a high time frame”. I assume this refers to the number of time frames, but the expression that is used is not clear. rephrase: over a longer period of time

Page 20: It is not clear if “pretrained audio encoder” refers to one or multiple encoders. In either case, the language should be corrected (either by adding “a” or “the” or changing to plural”). changed to plural

Page 20 refers to LSTMs, GRUs, and RNNs as alternative encoder components. However, LSTMs and RNNs are specific types of RNNs, so the description does not give very accurate information about their relationship. reparagraphed to indicate RNN is a broader category

Page 25 has an expression “pretrains the encoder on audio tagging before training on audio tagging” where I assume the latter “tagging” should be “captioning”. fixed

Page 27: scientific text should not use contractions such as “isn’t”. fixed

Page 32: text “SPIDEr [90] (portmanteau of SPICE and CIDEr) is a linear combination of SPICE and CIDEr, and was optimized by using a policy gradient [92] method” gives the impression that the metric was optimized using the policy gradient method. However, reference [90] seems to be about optimizing a machine learning model (given a fixed metric), so the text is misleading. unsure what this is referring to as the reference is about using PG to optimize SPICE and CIDEr

Page 40: variables $\mathbf{f}_{pi}$ and $\mathbf{f}_{ni}$ use inconsistent formatting; in some expressions, i, n, and p are superscripts of f, but in some expressions they are not. In general, the mathematical notations should be consistent. fixed

Page 40: Eq. (2.2) defining the triplet loss uses dot products as similarity measures. However, this definition of the triples loss would lead to minimizing the similarity of positive samples and maximizing the similarity of dissimilar ones, so there is a sign error in the formula. included another eqn to update use of dot product

Page 41: It is described that a contrastive loss is used in the thesis, and that contrastive loss has not previously been used in language based audio retrieval. However, the triplet loss which has been extensively used on the topic is also a contrastive loss, so this description is misleading. In general, the relationship between the triplet loss and the other contrastive loss that is used should be made more clear to understand what value the other contrastive loss adds to the triplet loss. added clarification, described differences

Page 53: In section 3.2.1, the vanishing gradient problem is described to originate from the use of attention mechanism. This is misleading, since the attention as such does not cause the vanishing gradient problem. The problem originates more from the autoregressive sequence models where all the previous time steps are used as an input in the decoder. fixed

added clarification in figure 3.2 caption and added dimension sizes to figure 3.3

Pages 55-60: the dimensionality of data representation at different processing steps is not given, and specifically there is no information what is the dimensionality of the data that is inputted to the transformer. This makes it difficult to understand the functioning and role of the transformer encoder properly. Based on the illustration of figure 3.2, the output of the CNN encoders is a 527-length vector, but in this case there would be no need for any sequence model like a transformer, which is confusing. Since on page 55 it is described that testing transformer encoder fully in audio captioning is one of the goals of this chapter, the processing steps and data in each of them should be described in sufficient detail to allow understanding how the processing is exactly done.

Page 60: There is two incomplete sentences “which uses them to generate capt” and “To assess the contribution of the transformer encoder to the overall performance of the audio captioning system.” fixed

Page 60: It is described that the input of the decoder includes a sequence of word tokens. However, it is not clear what word tokens these are. I assume they are the previously outputted words, but this is not explained properly. In general, the functioning of the decoder should be explained at sufficient detail to allow reproducing the results. clarified: previously decoded words

Page 61: The text describes that “The RLSSR module takes as input the output text embedding (T) from the last layer of the decoder”. However, it is not clear if this is the sequence of embeddings or an embedding of a specific time step. clarification: of all decoded time steps

Page 61 refers to ”average max pooling” and ”average global pooling” operations, but it has not been defined how these exactly operate (which operations are done over what dimensions). added numbers and clarification

Pages 63 and 77 describe the hyperparameters used in the experiments. It has not been explained what criteria has been used to choose their values. Particularly it should be explained whether validation or test data is used for choosing or optimizing the hyperparameter values. added clarification: ie no hp tuning, used similar values from previous work

Page 62 describes the procedures for calculating the evaluation metrics. This is redundant with the explanation in Section 2.1.5. shortened

Page 64, Table 3.3 included method "Base – transformer enc". It is not fully clear whether this is a model with or without the transformer encoder. replaced - with 'sans'

Page 65: "huber" should be capitalized, typo "simliarity"

fixed

Page 66: "euclidean" should be capitalized

Page 66: The best developed method is described to outperform the best related model. It is not fully clear how the reference methods are chosen. For audio captioning there exists several models that produce clearly better results than the ones presented in the thesis. There is no need that the proposed method outperform those, but it will be good to discuss how the proposed methods compare to the state of the art, and what are the factors affecting the reasons why there are methods which performance is clearly better. Table 3.5 that presents the results of the methods refers to these other models as "state of the art". However, there existing several methods that give significantly better results than the models presented in the Table, so it should be clarified in what terms the chosen methods are chosen state of the art. added clarification (at point of research), models from dcase or peer- reviewed conference

Page 72: "Acoustic information" should be lowercase. fixed

Page 72: The expression "minimally close in distance" is confusing. It can be understood as the closeness being minimal (e.g., things are distant), even though I think the point is that the distance is minimal. This could be rephrased to avoid ambiguity. fixed

Page 73: There is no information about the caseline CRNN model architecture.

Pages 73-76: the proposed method consists of CNN layers to produce audio embeddings. There is no information about the dimensionality of the embeddings, which makes it difficult to understand the role of the studied transformer or tied layers. included in section 4.5

Page 76: The text refers to Section 2.3 for the contrastive loss, but Section 2.3 does not contain information about the loss. It will be good to include a formula for the contrastive loss used, to understand how it differs from the triplet ranking loss that is used as a part of the total loss. The triplet ranking loss is also a contrastive loss, so terming the supplementary contrastive only as a contrastive loss is misleading, and unclear how these two contrastive losses complement each other. updated reference
added paragraph on complementary

Page 77: The model with converging tied layers and contrastive loss is described to perform significantly better than the baseline. However, the encoders of these two models

rebuttal: point of transfer learning is to leverage pretrained embeddings which will be naturally pre-trained on different sets of training data. however, the finetuning of the tied layers is all done using the same data (clotho).
added clarification that tied layers are always trainable

are trained with different data (including the pre-training), so conclusions about the relative performance of different model architectures cannot be drawn. Ideally the models should be compared in a setting where the same training data is used for all the models, so that one can evaluate the effect of the models. Ideally additional experiments should be done with same training data, or if this is not feasible the text should be revised to make it clear that the results cannot be used to draw any conclusions about the effect of the converging tied layers.

Page 78: Table 4.3 reports the results of models trained without the supplementary contrastive loss. However, the CRNN results in the table seem to be the same as the CRNN results in Table 4.2, where the supplementary contrastive loss is presumably included. This is not fully clear since this is not explicitly described, and it is not clear whether the CRNN model in Table 4.3 uses the supplementary contrastive loss or not. It would be good to discuss reasons why the lack of supplementary contrastive make the developed models fail. The caption refers to results "without contrastive loss", but presumably these still include the triplet ranking loss which is also a contrastive loss, so the description is misleading.

this has been described in earlier section that the referred contrastive loss is the CLIP loss. also added clarification that the CRNN is the baseline model, and added a paragraph on how the contrastive loss complements

Page 80: The summary describes that "Converging Tied layers are proposed to increase efficiency and to merge the two disjoint audio and text encoder". However, currently the results do not demonstrate any increase in the efficiency caused by the converging tied layers, since the tested models are (pre-)trained with different data.

same as rebuttal

Page 84: in formula (5.2), there is D in both left- and right-hand side of the equality. I assume there should be D' in the right-hand side.

fixed

Pages 88-89 describe the procedures for calculating the evaluation metrics. This is redundant with the explanation in Section 2.1.5.

fixed

Page 94: "Keyword" should be lowercase

fixed

Page 99: The conclusions describe that "Converging Tied Layers significantly improved the performance compared to the baseline". However, currently the results do not demonstrate any increase in the efficiency caused by the converging tied layers, since the tested models are (pre-)trained with different data. This should be clarified to avoid any false conclusions about the usefulness of the converging tied layers.

Page 102: "One" should be lowercase.

fixed

Bibliography:

done

-The publication information is missing from a large number of publications. For example from the first two pages of the bibliography (109-110), this includes references 5, 16, 17, 18. I am not listing all the publications missing the publication information, but proper publication information (e.g., journal or conference name, issue, year) should be added to all the references
-There is a large number arxiv publications. In case of those, one should find out if the arxiv paper is published in a peer-reviewed conference or journal, and instead of arxiv refer to the peer-reviewed version.

-Many bibliography items contain capitalization errors (e.g., bert, chatgpt)