Reply to Examiner No. 2

Name of Student: Andrew Koh Jin Jie

Degree: Doctor of Philosophy

Thesis Title: Audio Captioning and Retrieval with improved Cross-Modal Objectives

General comments

1. **Examiner's comment:** 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.

   Response: Thank you for this comment. We have revised and clarified the sentence.

   2.1. Automated Audio Captioning

   captioning task. Similarly, the Audio Captioning Transformer [75] also pretrains the encoder on audio tagging before training on audio captioning.

2. **Examiner's comment:** 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].

   Response: Thank you for this comment. We have revised and fixed the sentence structure.

   Image retrieval and audio retrieval models are trained and evaluated in a similar fashion using a ranking loss and the COCO caption process [16]. In image retrieval, different methods such as better feature extraction using large pretrained models [98, 99], and improved global and local feature alignment using variants of attention mechanism [100–103] are used to improve retrieval performance. Transfer learning by pretraining models using supplementary pre-training tasks like the masked modelling task [104–106] is also a popular approach. However, these methods have not been tried on audio retrieval as it is a relatively new task.

3. **Examiner's comment:** On page 64, the best performing algorithms are not always highlighted correctly in the tables. This must be fixed.

   Response: Thank you for this comment. The table has been revised to highlight the best performing algorithm.

Model	$BLEU_1$	$BLEU_2$	$BLEU_3$	$BLEU_4$	$ROUGE_L$	METEOR	CIDEr	SPICE	SPIDEr
Base (with transformer encoder)	0.516	0.330	0.219	0.142	0.348	0.152	0.319	0.102	0.210
Base (with transformer encoder) + PANN	**0.542**	**0.363**	**0.248**	**0.163**	**0.362**	**0.161**	**0.369**	**0.108**	**0.242**
Base sans transformer enc	0.523	0.337	0.227	0.151	0.353	0.153	0.332	0.102	0.211
Base with PANN sans transformer enc	0.538	0.350	0.235	0.153	0.362	0.158	0.348	0.106	0.227

TABLE 3.3: Ablation experiments to determine usefulness of the transformer encoder. Base refers to our default model, consisting of the convolutional encoder, transformer encoder and transformer decoder. PANN refers to loading the pretrained weights for transfer learning.

Model	$BLEU_1$	$BLEU_2$	$BLEU_3$	$BLEU_4$	$ROUGE_L$	METEOR	CIDEr	SPICE	SPIDEr
Base	0.516	0.330	0.219	0.142	0.348	0.152	0.319	0.102	0.210
Base + L2 loss	0.518	0.335	0.228	0.152	0.352	0.151	0.326	0.102	0.214
Base + L1 loss	0.515	0.338	0.234	0.159	0.351	0.151	0.325	0.098	0.212
Base + PANN	0.542	0.363	0.248	0.163	0.362	0.161	0.369	0.108	0.242
Base + PANN + L2 loss	**0.552**	**0.370**	0.251	0.166	0.369	0.163	0.375	**0.112**	0.240
Base + PANN + L1 loss	0.551	0.369	**0.252**	**0.168**	**0.373**	**0.165**	**0.38**	0.111	**0.246**

TABLE 3.4: Results of using the RLSSR module. L1 and L2 loss refers to the distance metric used to optimize the RLSSR module. PANN refers to applying transfer learning using the pretrained weights from the pretrained audio neural network.

Model	$BLEU_1$	$BLEU_2$	$BLEU_3$	$BLEU_4$	$ROUGE_L$	METEOR	CIDEr	SPICE	SPIDEr
Baseline [83]	0.389	0.136	0.055	0.015	0.262	0.084	0.074	0.033	0.054
Fine-tune PreCNN Transformer [66]	0.534	0.343	0.230	0.151	0.356	0.160	0.346	0.108	0.227
AT-CNN10 [79]	**0.556**	0.363	0.242	0.159	0.368	**0.169**	0.377	**0.115**	-
Base + PANN + L1 loss	0.551	**0.369**	**0.252**	**0.168**	**0.373**	0.165	**0.380**	0.111	**0.246**

TABLE 3.5: Comparison with other state of the art. Our model consistently beats the previous state of the art on the $BLEU_n$, $ROUGE_L$, and CIDEr scores.

4. **Examiner's comment:** The highlighting of best results for each metric in Table 5.3 is inconsistent. This should be fixed.

   Response: Thank you for this comment. The table has been revised to highlight the scores of the best performing method.

Model	$BLEU_1$	$BLEU_2$	$BLEU_3$	$BLEU_4$	$ROUGE_L$	METEOR	CIDEr	SPICE	SPIDEr
System 1 - cross entropy	**0.559**	0.358	0.237	0.153	0.169	0.374	0.382	**0.116**	0.249
System 1 - cross entropy + EDC stopwords	0.558	**0.362**	**0.242**	**0.159**	**0.170**	**0.375**	**0.391**	0.115	**0.253**
System 1 - scst	0.641	**0.417**	**0.277**	**0.174**	0.182	**0.407**	0.432	0.124	0.278
System 1 - scst + EDC stopwords	**0.642**	0.409	0.272	0.172	0.182	0.402	**0.444**	0.124	**0.284**
System 2 - cross entropy	0.553	0.367	0.248	0.160	0.162	0.372	0.359	0.111	0.235
System 2 - cross entropy + EDC stopwords	**0.558**	**0.376**	**0.258**	**0.172**	**0.167**	**0.376**	**0.381**	**0.115**	**0.248**

TABLE 5.3: Comparison of performance of systems trained on Epochal Difficult Captions (EDC) stopwords curriculum against their counterpart

5. **Examiner's comment:** 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.

   Response: Thank you for this comment. We have revised and fixed the number (over the baseline score of 0.03)

The best model described in this study uses a CNN10 audio embedding, RoBERTa base text embedding, and a transformer encoder layer with 4 layers and 96 dimensions. This model obtained a R_1 score of 0.11, a 267% improvement over the baseline score of 0.03.

6. **Examiner's comment:** These thesis document contains many small grammatical and typographical errors.

   Response: Thank you for this comment. We have proofread the thesis and fixed grammatical and typographical errors.

05 November, 2023

_________________________ ____________________________

Signature of Student Date