Reply to Examiner No. 1

Name of Student: Andrew Koh Jin Jie

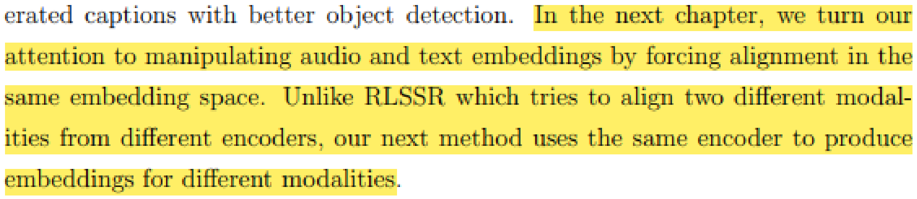
Degree: Doctor of Philosophy

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

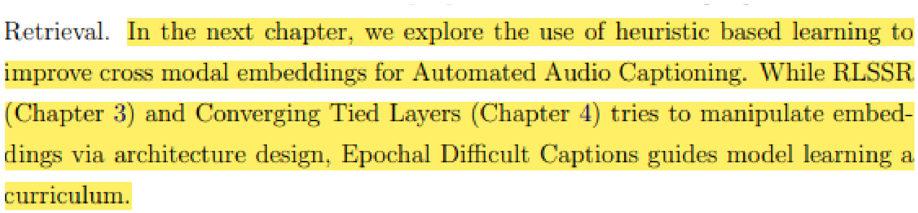
General comments

1. **Examiner's comment:** The main issue is that the connection between each section is not clearly explained. Although all sections focus on audio captioning and retrieval, each section seems independent of the other. A more general should be summarized to illustrate the contribution of the research.

   Response: Thank you for this comment. We expanded the summary of each research chapter and have indicated how the current chapter relates to the next chapter.



erated captions with better object detection. In the next chapter, we turn our attention to manipulating audio and text embeddings by forcing alignment in the same embedding space. Unlike RLSSR which tries to align two different modalities from different encoders, our next method uses the same encoder to produce embeddings for different modalities.



Retrieval. In the next chapter, we explore the use of heuristic based learning to improve cross modal embeddings for Automated Audio Captioning. While RLSSR (Chapter 3) and Converging Tied Layers (Chapter 4) tries to manipulate embeddings via architecture design, Epochal Difficult Captions guides model learning a curriculum.

2. **Examiner's comment:** Some of the expressions could be modified to make the thesis more integrated. For example, in multiple places, "papers" should be revised as "work" or "chapter".

   Response: Thank you for this comment. We have revised the thesis and replaced mentions of "papers" with "work" where appropriate

05 November, 2023

__________________________ __________________________

Signature of Student Date