Audio Captioning and Retrieval with Improved Cross-Modal Objectives

The thesis includes five works related to automated audio captioning and language-based audio retrieval.

Chapter 1 introduces the overall objectives of the thesis, the motivations, and the major contributions.

Chapter 2 introduces the two tasks, related works, the datasets, and evaluation metrics.

Chapter 3 presents the first work. The first work focuses on the automated audio captioning problem and proposes a self-supervised reconstruction latent space similarity regularization method to improve it. The self-supervised training is performed by recreating the audio embeddings from the text prompt.

Chapter 4 introduces the second work. The second work targets the language-based audio retrieval problem. It proposes a method called Converging Tied Layers to align the audio and text representations using transformer layers. It shows that using such a layer helps to improve retrieval performance.

Chapter 5 unveils the third work. The third work proposes to use curriculum learning to improve automated audio captioning. This work observes that using curriculum training allows the method to achieve a 5.5% increase in performance.

Besides the three works mentioned above, the thesis also includes two additional works to improve word sense disambiguation via transfer learning and audio tagging.

Chapter 6 concludes the thesis and discusses future works.

This thesis is generally well structured, with each chapter serving a specific purpose or addressing a given issue.

At the same time, this thesis could be further improved by addressing the following issues:

1. 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.
2. 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”.

Overall, the thesis has made original and substantial contributions to the research topic. I give my recommendation of acceptance of the thesis for the award of the degree.

Questions:

At the same time, this thesis could be further improved by addressing the following issues:

1. 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.
2. 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”.