Research Assessment: Final Product Part III

Kavan Mehta

Ms. Whitcomb

ISM 2

12 February 2023

Kavan Mehta

Ms. Whitcomb

ISM 2

12 February 2023

Research Assessment #18: Final Product Part III

Date: 12 February 2023

Subject: Transformer-based audio-visual speech recognition

MLA citation(s):

Serdyuk, Dmitriy, Otavio Braga, and Olivier Siohan. "Transformer-based video front-ends for audio-visual speech recognition." *arXiv preprint arXiv:2201.10439* (2022).

Assessment:

After learning about two different approaches, 3D Convolutional Neural Networks and Transfer Learning with pre-trained models, I spoke to my mentor Dr. Paschall from IDEXX Labs on my Final Product. We decided to explore other approaches and datasets in order to get a more holistic idea of how we could implement an audio-visual model to optimize speech recognition in environments with noise. Thus, to understand a new type of approach to the problem of utilizing both audio and video effectively, I found a scholarly article from Google's AI Research, "Transformer-based video front-ends for audio-visual speech recognition," which went over a newer approach to using Transformer architecture for the visual part of the multi-modal neural network.

Through the journal article, I first reviewed the usual approaches as described by Google researchers. The use of a 3D Convolutional Neural Network is very traditional in the industry and demonstrating the utilization of Transformer architecture is extremely new in the field. As I

know already through my Original Work projects, the research paper emphasized that “a self-attention based transformer was proposed for a variety of sequential tasks” and audio-visual speech recognition is definitely one of these tasks (Serdyuk et al. 1). This made the argument of researchers very convincing that if transformers have enhanced accuracies compared to traditional networks like RNNs and CNNs in sequential tasks, why can’t they be used in audio-visual networks too? The article then emphasizes some complexities about alignment between video and audio as that was challenging for the new ViT (Visual Transformer) that was used for the experiment. Since they were testing the efficiency of the ViT compared to the traditional 3D CNN for this application, they performed experiments with various datasets and kept certain aspects constant like the A/V ASR Model Architecture. They used videos that were “23 to 30 frames per second” and “[cropped] the videos centering near the mouth region” with sizes of “128 x 128” (Serdyuk et al. 2). I think that these various parameters could be very helpful to me when I am creating my own model as these are standard industry/research parameters that yield more optimized results. I was also able to learn more about the transformer architecture as it holds a “14-layer transformer encoder with 512 hidden dimensions, 8 attention heads, and the relative positional embedding” with a “conformer encoder with 17 layers” (Serdyuk et al. 2). These specifications were very interesting as the transformer combines both visual and audio inputs into different nodes combining them later to create the final output of the speech that is being recognized. Also, the use of positional embeddings is a concept that I am familiar with due to my experience with natural language processing and projects for my Original Work in ISM 1 and 2. The datasets that the researchers used were also unique as they used the data from Google’s platforms like Youtube with professional transcriptions such as “YTDEV18 set” and they used “LRS3-TED” for evaluation of their training (Serdyuk et al. 3).

Furthermore, through the various experiments of the researchers at Google, they demonstrated that the “ViT outperforms the convolutional baseline in certain settings” (Serdyuk et al. 4). Hence, I think that as I discuss various options for my approach with my mentor, Transformers could be an extremely valuable asset to have the best accuracy using both video and audio for optimized speech recognition in environments with both babble and overlapping noise.

I aspire to continue to learn about other methods of implementation and professional audio-visual datasets I could use to utilize both audio and video to solve the optimization problem in speech recognition. I hope to continue working with my mentor through discussions and create my Final Product!