Research Assessment: Final Product Part II

Kavan Mehta

Ms. Whitcomb

ISM 2

5 February 2023

Kavan Mehta

Ms. Whitcomb

ISM 2

5 February 2023

Research Assessment #17: Final Product Part II

Date: 5 February 2023

Subject: 3D Convolutional Neural Networks for Cross Audio-VIsual Matching Recognition

MLA citation(s):

Torfi, Amirsina, et al. "3D Convolutional Neural Networks for Cross Audio-Visual Matching Recognition." *ArXiv.org*, 13 Aug. 2017, https://arxiv.org/abs/1706.05739.

Assessment:

After speaking with my mentor Dr. Paschall from IDEXX Labs on my Final Product idea of utilizing both audio and video to recognize speech, I found out about multi-modal networks to use both speech and visual input features. He provided me with some great resources to explore multi-modal networks and their architecture that helps them combine both audio and video for data analysis. To understand a new type of approach to the problem of utilizing both audio and video effectively, I found a scholarly article, "3D Convolutional Neural Networks for Cross Audio-Visual Matching Recognition," which went over an effective approach better than the current state-of-the-art models using two non-identical 3D Convolutional Neural Networks (CNNs) that use audio and visual networks.

Through the journal article, I first reviewed the common association that my entire Final Product is based on: lip movements and the sounds of words are correlated with each other and hence, enable us to use both video and audio together to help speech recognition in environments

with distracting or background noise. I learned about the several complexities of this problem, such as "[recognizing] the part of the speech regardless of who is speaking" and in text-independent scenarios, "no prior information or restrictions are considered for the utterances" (Torfi et al. 1). These complexities are really essential to understand fully as they will help think about how I can make a more universal solution to the problem using a multi-modal network that can recognize a wide range of common vocabulary. This will also depend on if I make it dependent on the training data's restrictions of limited vocabulary, or try to fill in the gaps for certain phrases and vocabulary that could enhance the model. The research in the article primarily based its goal on creating "nonlinear mappings that learn a non-linear embedding space between the corresponding audio-video streams using a simple distance metric" to use the advantages of lip movements and speech audio (Torfi et al. 2). This goal requires the use of mathematical theory behind the deep learning model which I will have to understand as I move on in the future to implement an effective solution. Another interesting fact that will serve useful is that CNNs have been used before for solving this problem. There are two main features that we keep in mind: "Phonemes are the smallest distinguishable unit of an audio stream which are combined to create a spoken word, and a viseme is its corresponding visual equivalent" (Torfi et al. 2). This connects to how we can break up the entire video stream into many many individual components and use the data to analyze each one separately to train our model efficiently on words and phrases. I will require both spatial and temporal analysis of data as my machine learning model will need to keep track of the real data along with its time to keep the correlation between lip movements and speech audio usable for the project. The datasets and frameworks for pre-processing of the data in the presented research could significantly help guide me along with my mentor's resources on how I could align both audio and visual data together for the

multi-modal network to be effective in my Final Product. The model also utilizes an “input speech feature map, which is represented as an image cube, [corresponding[to the spectrogram” (Torfi et al. 3). The unique representation of the data is very similar to my Original Work project for speech recognition of simple phrases as I worked with spectrograms too! This should be a little more familiar and help me in completing my research and understanding of the algorithms and the approach of my Final Product.

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