Lionel Quintanilla, Andrew De La Rosa, Pouya Tavakoli, Brian Tran, Matthew Zaldana

Professor Maryam Qudrat

ENGR 361

30 November 2021

A Machine Learning Approach to Closed Captioning

For our research hypothesis, we want to create an application that uses machine learning techniques to produce closed captions on any video player with accuracy that exceeds current methods. We plan to achieve this by exploring two specific aims: first, by devising a machine learning technique that can create captions with a reasonable degree of accuracy, and secondly by creating a user-facing application that can take in video inputs and display captions by overlaying of these videos.

The issue of improving the accuracy of closed captioning is imperative because it intersects both legal and moral considerations. The main necessity for accurate captions comes from the fact that closed captions are legally required in the United States and other countries. However, deaf and hard of hearing people who rely on captions to understand video media are being underserved because current methods of captioning struggle with accuracy and readability. This is caused partly because conventional captioning methods take time to write and can cost significant sums. Still, current methods for creating captions automatically meant to solve this are not particularly accurate or widely available. While there is research being done to improve the current methods, we want to take a novel new approach and use machine learning to bypass the issues with the current methods.

To research our hypothesis, we will conduct two separate experiments based on our specific aims. For the first experiment, we will create a list of three to six machine learning-based algorithms for creating closed captions from images and videos. We will then set up each algorithm and the required artificial intelligence on a dedicated computer and have it caption a series of dataset containing images and videos. We will repeat this process for each machine learning technique so that their results can be compared. To analyze the results, we will quantify the results of each technique by taking the captions they generated from the datasets and running them through several machine learning tests for natural language to determine the accuracy of the captions. The accuracy scores will then be averaged for each technique to compare them.

For the second experiment, we will gather a mixed focus group of deaf/hard of hearing volunteers and average volunteers will test a user-facing closed captioning application we will develop. Each volunteer will be given a set of instructions on how to use the application and computer with the application load. The focus group will then use the set of instructions to test the usability and performance of the most important functionalities of the application. In particular, the questionnaire will seek to quantify how intuitive the application is to use, how interoperable it is with current video players, and how readable the display captions. The questions will be scored on a scale of 1-5, so that average scores can be generated from the volunteers.

In terms of experimental outcomes and impact they will have on our hypothesis, the machine learning technique we choose should provide immediate improvement in accuracy over existing methods. Since our research has shown that a score of roughly 20 to 30 on natural language tests provides for reasonable accuracy with automatically generated closed captions, the technique we choose should fall within these parameters. The benefit of the technique falling within this range is that it would demonstrate the captions are accurate not just at the word level, but also at the sentence level. This should improve not just the readability of the captions, but also help the grammar and vocabulary, which are the biggest barriers to making captions easy to understand. For the closed captioning application, the optimal results would be the focus group providing an average score of 4 on the questionnaire. In order to be practical for daily use, the application should be easier to use, or at least as easy to use, than existing closed captioning tools for web-based video players. Since the optimal goal of the application is to be able to work will all types of multimedia content, a high average score would prove that the application could see use as a general-purpose captioning tool, rather than being confined to the Internet.