Automated Close Captioning

Urmzd Mukhammadnaim [*1], Keelin Sekerka-Bajbus [†1], and Benjamin J. Macdonald [‡1]

[1]Faculty of Computer Science, Dalhousie University

February 17, 2022

Motivation

According to a meta-study done by Nucleus Research [], two-thirds of online transactions are abandoned by blind individuals due to the lack of accessibility. Furthemore, approximately 36 million are said to have some degree of visual impairment, with the number expected to triple by 2050 according to a statistic by the World Health Organization. With a spending power of almost half a trillion dollars a year, providing accessibility as a service to individuals with disabilities is a untapped market that is by large, not catered to. Furthermore, as the pandemic has forced human interactions to be in large part, remote, accessibility is no longer a privilege, but a basic human right. This project proposes a tool to begin enabling better accessibility integration on video-streaming platforms such as Netflix and YouTube, through the automated closed captioning of videos. More specifically, given a short video clip, we propose a model which can generate a single sentence describing the events in the input.

[*] urmzd@dal.ca, B00800045

[†] kl967083@dal.ca, B00739421

[‡] bn282348@dal.ca, B00803015

Data Collection and Processing

Collection

In this project, we propose utilizing the dataset designed for Deepmind Kinetic as a means of circumventing the collection of video clips sharing similar contexts and hence, reducing the complexity associated with finding an appropriate datset. The Deepmind Kinetic 700 dataset contains 650,000 video clips collected from YouTube categorized in one of 700 different movement taxa. In our case, only a selective set of videos will be used to train and validate our model. To be specific, 600 videos from 3 random classes will be selected, totalling to 1800 clips. The reason for the massive reduction in the dataset size is due to the additional labelling required for our specific use case. Deepmind Kenetic attempts to classify a video clip containing repetitive moments in one of 700 categories, whereas our model will attempt to generate a sentence describing the events occuring in the video. If the full dataset were to be labelled (650,000 clips), we would not be able to train the model due to the current time constraints. Additionally, our GPU resources are limited, and so, we do not have the capability to train models that require extensive video processing.

Processing

As mentioned earlier, the data retrieved from Deepmind Kinect's dataset requires extra processing to fit our use-case. As the our model attempts to generate sentences to describe the events in a video, each video must labelled with a sentence describing not only the target movement, but also any actions that occured before and after the occurrence of a particular movement. While certain movements are shared across videos in the same category, the context in which the video was captured may differ, making labelling non-trivial. Videos will be clipped and labelled with a short description of the events that occured in the video, at which point we will be able to begin training.

Subproblems To Solve

We anticipate multiple subproblems we will have to address. One includes the creation of a model to evaluate the similarity of two sentences for the purposes of quantifying the performance of our CNN-RNN model. With our previous experience utilizing word2vec, this will be a problem we can quickly address. Other subproblems include determining the appropriate architecture for our CNN-RNN model. Another would be later experimentation with hyper parameters for our model to optimize its performance. Should our model perform poorly, we may also wish to explore the normalization of our dataset. If we find that the quantity of training data is insufficient, data augmentation to increase this set would be a viable option to explore.

Next Steps

Over the next month, there are several tasks we must complete. By February 21, we expect to have finished the annotation of our video dataset with brief descriptions of the actions on screen. Once this is completed, the next step will be to work on the initial implementation. We would like this to include a simple model containing an embedding layer that is able to quantify the difference in similarity between two sentences. This will be critical in evaluating the accuracy of our generative model. Ideally, this will be completed the week of February 28. While the embedding model is being completed, another pair of teammates will continue working on the generative model. We anticipate to have completed a trainable iteration of this by the week of March 14, after which the focus will switch to experimenting with the construction of the model. Once we have determined the impact on accuracy of these varying hyperparameters, we will begin focusing on the writing of a paper expressing our findings, and the preparing of a project poster.

Division of Work

To ensure an even distribution of work between teammates, we will utilize issue tickets on git to assign to one another. We will have set meetings in which we will determine what tasks must be completed by the next time we gather, create the tickets and then assign them. Should one person complete an issue early, they will be able to assign themselves another ticket.

References