Automated Close Captioning

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

[1]Faculty of Computer Science, Dalhousie University

February 16, 2022

Motivation

According to a meta-study done by Nucleus Research [cite], 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. However, 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.

Procssing

As mentioned earlier, the data collected requires extra processing. 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 repetive movement, but also any sequential 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 be different, making labelling non-trivial.

Next Steps