

# Automated Close Captioning

Urmzd Mukhammadnaim<sup>1</sup>  
B00800045  
urmzd@dal.ca

Keelin Sekerka-Bajbus<sup>1</sup>  
B00739421  
k1967083@dal.ca

Benjamin J. Macdonald<sup>1</sup>  
B00803015  
bn282348@dal.ca

<sup>1</sup>Faculty of Computer Science, Dalhousie University

## Problem Statement

According to a report done by Nucleus Research [nuc, 2019], two-thirds of online transactions are abandoned by blind individuals due to the lack of accessibility. Furthermore, approximately 36 million people 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 [wor, 2021]. With a spending power of around half a trillion dollars a year [Yin et al., 2018], 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 several potential approaches, including 3D ResNets, RNN-CNN hybrids and BERT models to generate a single sentence describing the events in the provided input.

## 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 context. The DeepMind Kinetic 700 dataset contains 650,000 video clips collected from YouTube categorized in one of 700 different movement taxa [Smaira et al., 2020]. For the sake of time, only a selective set of videos will be used

to train and validate our model. To be specific, 600 videos will be sampled from 3 arbitrary classes, yielding  $600 \times 3 = 1800$  clips. The reason for the large reduction in the sample size is due to the additional labelling required for our use case. While DeepMind Kinetic attempts to classify a video clip containing repetitive moments in one of 700 categories, our model will be attempting to generate a sentence summarizing the events of a video. If the full DeepMind Kinetic dataset were to be used, we would not have enough time to train the model. 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 Kinetic’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 be labelled with a sentence describing not only the target movement, but also any actions occurring before and after the execution of a particular movement. While action definitions are shared between siblings in the same category, the context in which the video was captured may differ, making labelling non-trivial. To ensure enough time remains to train the model, preprocessing will consist only of trimming the videos and labelling the shortened video with a short description of the events within them.

## Approaches

In the preliminary analysis of the stated problem, a few different approaches were found. Before describing some of the approaches, it should be clarified what subproblems require solving. The first problem is event classification, given a video, what key moments were encountered? The second problem is language modelling, given the key moments, what sentence best encapsulates the different events? For this project to succeed, both of these questions must be answered. Down below, we describe some relevant works and potential extensions to better fit our use case.

### I3D

[Smaira et al., 2020] details the usage of an RNN-CNN hybrid architecture (I3D) to determine the most prevalent action in a video (DeepMind Kinetic 700). To accomplish our goal, the model can be extended to classify events beyond movement patterns. The classified events can then be fed through an embedding model to allow for a greater vocabulary, at which point, a language model can be used to generate a sentence that best encapsulates the events that occurred.

## 3D ResNet

On the same dataset mentioned above, [?] was able to use a 3D ResNet for marginally better performance. The same classification was done with a different architecture. As a result, the process mentioned above can be, by large, carried over.

## BERT

VideoBERT is an extension of the BERT architecture that additionally utilizes video tokens [?]. The model found great success in the completion of masked sentences given a video clip. However, similar to the above mentioned models, the dataset used was extremely large. However, the model was still able to perform within reason on training sets with 5000 video clips, making this an interest model to explore the impact of data augmentation, regularization and normalization on performance. Furthermore, as our task is to generate a sentence, it may be worth exploring if part of speech tags can be applied to prevent the need for prefilling sentences.

## 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, other 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

- [nuc, 2019] (2019). The internet is unavailable.
- [wor, 2021] (2021). Vision impairment and blindness.
- [Smaira et al., 2020] Smaira, L., Carreira, J., Noland, E., Clancy, E., Wu, A., and Zisserman, A. (2020). A short note on the kinetics-700-2020 human action dataset. *CoRR*, abs/2010.10864.
- [Yin et al., 2018] Yin, M., Smith, D.-M., Overton, C., and Shaewitz, D. (2018). A hidden market: The purchasing power of working-age adults with disabilities.