Problem 0 - The Researcher's Challenge

Victor Armegioiu

October 26, 2018

Abstract

This text addresses the first challenge proposed by AI-MAS for the selection process. We will briefly outline a prototype model, the problems we might encounter on the way of developing this project and provide a few key insights concerning a possible roadmap.

1 Introduction and prototype

Hand gestures are ubiquitous in our day to day lives. We often use them in order to efficiently convey information that can bypass the need for other types of communication, in certain circumstances. In collaboration with Facebook's Research Laboratory, we seek to make use of hand gestures as to optimize the user experience in terms of interaction with Facebook's interface.

This particular task can essentially be formulated as a Computer Vision problem. Given a dataset of videos of people performing hand gestures, the model should be able to map the videos to labeled hand gestures from a finite subset of such gestures. Using these labeled hand gestures, Facebook could easily infer the intended actions of the users, delivering them the content they require.

With this in mind, we can use a simple CNN-RNN/LSTM architecture for the actual model. We'd use $3D$ filters in order to account for the temporal dimensions, and RNNs for efficient sequence prediction.

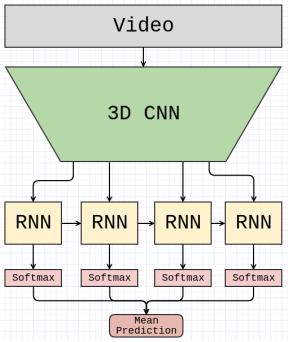

Figure 1: Model architecture visualization.

2 Challenges and development pipeline

This section aims to summarize a few of the challenges we will meet on the way while developing our model. Having a dataset of high quality is a must for any machine learning model.

This particular dataset would be tough to build as it requires involvement of hundreds of human subjects. One other challenge that may arise during the construction of the dataset would be the variance of the gestures. Specifically, we have to be able to account for the fact that two different gestures may map to the same meaning, in different cultures. Fortunately this one has an easy fix, we could simply provide a map of names to video examples that displays what should a particular gesture mean.

Another challenge that will most definitely surface would be on the side of the end-users, since they may choose to not give the app access to a video stream of themselves, hence rendering the concept useless. This particular problem is especially tough and would need to be carefully addressed by Facebook by gauging the users' interest in this potential feature.

In terms of technical issues, we need to consider several factors. First of all, the capacity of the model to recognize gestures from streams of low video quality. This aspect is two-fold, as the algorithm would work optimally for a specific frame rate that's highly dependent on the end-user's internet connection; another concern arises regarding the overall quality of the phones' cameras. We also need to take into account the computation effort of the model and the impact it would have on the user experience.

The actual development process would be roughly comprised of the following steps:

- Gauging the percentage of interested users
- Creating and curating the dataset
- First implementation of the aforementioned prototype model
- Test and refine the prototype
- Present a demo of the prototype at a dedicated keynote conference