Capstone Proposal

August 7, 2019

1 Machine Learning Engineer Nanodegree

1.1 Capstone Proposal

Jason Hall August 7, 2019

1.2 Proposal

1.2.1 Domain Background

Technologies that are capable of identifying actions and events in video will be utilized for a variety of purposes.

Use cases include: * automatic alerts from military, commercial, and retail video surveillance systems * auto-labeling of videos for search engines and indexing * a number of other consumer related products where an action could be taken upon certain recognized actions/events

One of the leading efforts on this front is the MIT-IBM Watson AI Lab project named "Moments in Time". The project can be found here: http://moments.csail.mit.edu/

1.2.2 Problem Statement

Action and event recognition technology is still in its infancy. In 2018 the MIT-IBM Watson AI Lab held a contest to produce AI that could recognize actions and events in 3 second video clips found in their "Moments in Time Dataset". The dataset includes over 800k training videos, 33k validation videos, and has 339 classes. The metric used to rank submissions is top-k accuracy (using the average of k=1 and k=5): Where k=1, the model must predict the correct label. Where k=5, the correct label must be in the top 5 highest probability estimates the model produced. The winner of the contest produced a score of 0.5291. I will try to improve on that score. The model I produce can then be used outside of the dataset with other videos and for different uses.

1.2.3 Datasets and Inputs

The dataset I will use for this project is the "Moments in Time Dataset" provided by MIT-IBM Watson AI Lab. The link to the download can be found here: http://moments.csail.mit.edu/#download

The dataset includes over 800k training videos, 33k validation videos, 67k test videos and has 339 classes. Each video is a 3 second clip. The videos are pre-processed to be 256x256 and 30 fps. This is excellent for training a model. To extend the trained model for other data, only light pre=processing will be required.

1.2.4 Solution Statement

The solution to this problem will be to achieve a high averaged top-1 and top-5 accuracy. Currently the best a team has done is to achieve 0.5291 on the test dataset.

1.2.5 Benchmark Model

The benchmark for this problem is the DEEP-HRI model produced by Hikvision from the MIT-IBM Watson AI Lab contest. Their averaged top-1 and top-5 accuracy was 0.5291. My model will be rated against theirs, using the same metric. The results of the contest and their score can be found here: http://moments.csail.mit.edu/results2018.html

1.2.6 Evaluation Metrics

The evaluation metric used is the average of top-1 accuracy and top-5 accuracy. Top-1 accuracy is simply the accuracy of the model in predicting the ground truth class label. The top-1 accuracy score is the percentage of time the model has the ground truth class label in its five highest probability estimates. Said differently, if the ground truth class label is in the model's top five guesses, the model is deemed to have accurately predicted that example for the top-5 accuracy score.

1.2.7 Project Design

For this project I will explore using convolutional neural networks (CNN), as they are the leading algorithm for image and video recognition. I will first create a simple stacked CNN architecture. I will then benchmark that with the training and validation datasets. Then I will begin testing if data augmentation influences the model accuracy, specifically I will check to see if RGB color channels are necessary, or if simple black and white (one channel) can be used. If black and white can be used, it will speed up the model significantly. I will then experiment with more complicated CNN designs. Specifically I will feed the input into multiple stacked CNN models with different size filters. This should theoretically allow the model to capture features of different resolutions. Finally, I will experiment with feeding the features of the CNN models into LSTM networks to capture the temporal sequence found in the videos.