Automatic Labelling of Videos using Deep Neural Network

CSE 847 Project Proposal

Tarang Chugh
chughtar@msu.edu

Rahul Dey
deyrahul@msu.edu

1. PROBLEM DESCRIPTION

In the last decade, the world wide web has witnessed a massive explosion of multimedia data due to the rise of online media-sharing services such as Youtube, Facebook, etc. While many studies have tackled the problem of analyzing and understanding static images [1], improvements in video understanding research has been lagging due to unavailability of large-scale labelled dataset. However, analysis and understanding of video content on these video-sharing services provide insights to make the interface more engaging and customized to individual needs. Relevant search results, improved video recommendations, and content filtering are some of the many desired properties, which require the understanding of video content. In this project, we aim to analyze and understand video content to accurately assign labels. Google's recent release of the Youtube-8M dataset accompanied [2] with the launch of the Kaggle competition[1] to benchmark the performance on this task, has motivated us to take up this arduous task. We plan to take a deep learning based approach to design the classifier. We will also make use of the Google Cloud Machine Learning beta platform[2] available to the participants of this competition.

2. SURVEY

Convolutional Neural Networks (CNN) have emerged as a powerful tool for image based automation tasks such as recognition, segmentation, enhancement, encoding etc. It has become the preferred choice for such tasks as compared to other machine learning algorithms such as SVM, k-NN, Multilayer Neural Networks, etc. because of their unique ability to extract useful features from two dimensional inputs. However, when it comes to spatio-temporal inputs such as videos, CNN in itself is often found to be performing poorly. In our survey, we came across several approaches adopted by researchers to perform tasks such as video classification and labelling. We will discuss some of the approaches here briefly.

In case of videos, traditional CNN architectures suffer due to three main issues. The first issue relates to the large amount of computational resources required to train a CNN for classifying videos. For instance, videos of frame size 178×178, takes weeks to train a CNN using powerful GPUs. Performance of a deep neural network relies on the abundance of data. Thus, in case of videos, the size of data becomes even more, because of their three dimensional (two spatial and one temporal) nature. The second main issue is that CNN doesn't have any inbuilt mechanism to learn temporal patterns in the input data. And lastly, videos are usually of varying time duration. Even if all the input videos are resized to a particular resolution, different durations make them harder to be fed to a predesigned fixed network.

Several researchers have taken different approaches to overcome some of these issues. Karpathy et al. [3] suggested an approach where the input video is divided into two separate streams of data, a *context* stream that learns low resolution features in the frames, and a *fovea* stream that that learns high resolution features from the middle portion of the video. Thus, the effective resolution can be reduced substantially. This takes advantage of the camera bias since usually the object of interest is usually in the center of the frame. Another approach relies to on treating an entire slice of video having T number of frames and feeding them into a modified first convolutional layer by extending them to be of size $rows \times columns \times 3 \times T$. When successive slices are fed into the CNN, the first layer can find out the difference in them, thus characterizing the global motions in the video. Similar attempts have been made, such as in [4], to extend CNN to multimedia input by treating space and time as equivalent dimensions of the input and performing convolutions in both time and space.

In recent years, Recurrent Neural Networks have come up as a promising tool in learning sequential and temporal data. They have been applied successfully in many speech and text based tasks such as recognition, translation, sentiment estimation, synthesis etc. Recently, Venugopalan et al. [5] modeled frames of videos using pre-trained CNN and sequence of words using a pre-trained RNN on images associated with sentence captions to translate videos to natural language. We plan to combine the power of CNN to learn spatial data and of RNN to learn temporal data in our goal of understanding videos. The final goal of the project is to be able to recognize objects and activities in videos and label the videos accordingly which can be of tremendous use for video sharing websites in performing auto-labeling and tagging, user-recommendations, content filtering, retrieval and sorting, etc.

3. PROJECT TIMELINE

The first milestone for us is to prepare ourselves for this challenging task. It is crucial to perform a comprehensive literature review to understand the existing methodology and algorithms, and understand their strengths and weak-

[1] https://www.kaggle.com/c/youtube8m
[2] https://cloud.google.com/ml/

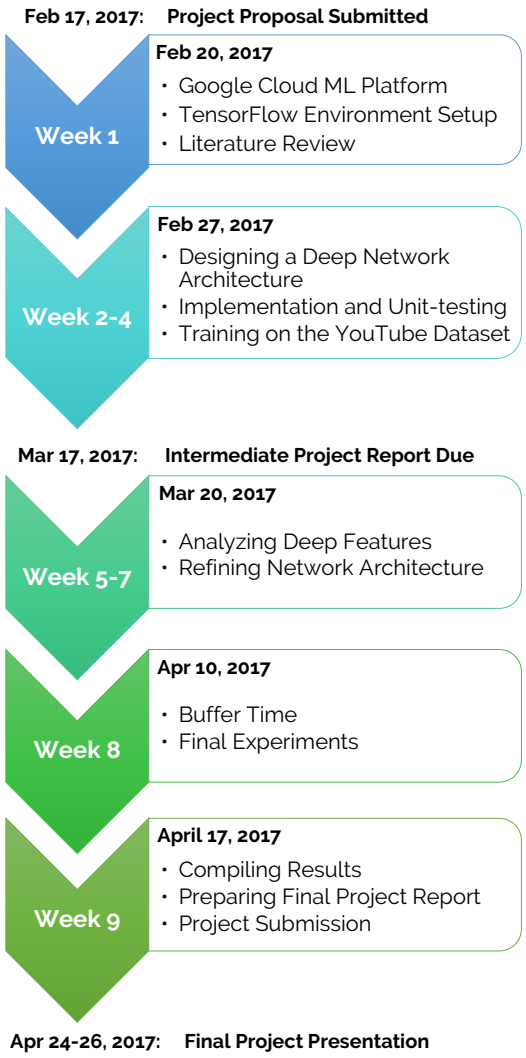

Figure 1: Timeline presenting the targeted milestones in the project.

nesses. We will primarily focus on the studies based on deep learning. We will also get acquainted with the Google Cloud Machine Learning platform and setup the required environment to use TensorFlow. We have allotted 1 full week for this task, which also includes a buffer time if we face any challenges in getting used to the new environment.

The next milestone for us it to implement our understanding. This will include design, implementation and training of a deep neural network architecture. We also plan to put sufficient time in getting acquainted with the YouTube dataset.

Once we obtain a preliminary set of results, we will spend time to understand the limitations of our architecture and refining it. This will include parameter and architecture tuning using cross-validation. We also plan to analyze the deep features learnt by the network, to improve our understanding and take steps towards improving the overall performance.

Lastly, we will run the final set of experiments and compile the results to prepare for our final project report and presentation. Figure 1 shows the overview of our project timeline.

4. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[3] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[4] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013.

[5] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014.