Predicting Video Tags Using Google's YouTube-8M Dataset (tentative title)

Tatyana Li
litatyan@msu.edu

1. PROBLEM DESCRIPTION

On September 2016, Google announced a release of the large labeled dataset, YouTube-8M, for video understanding research. The dataset contains about 8 million YouTube videos, amounting to more than 500K hours of video content. Earlier this month, **Google hosted a Google Cloud and YouTube-8M Video Understanding Challenge** on a Kaggle platform, providing an updated version of the dataset with new labels and newly added audio features. This project will use the above dataset to implement and compare classification algorithms to predict video labels based on the video and frame-level features provided.

2. RELATED WORK

[Note: Due to the fact that Google announced the above competition two days ago, the survey of the related work presented below is limited, and will be expanded further in future reports.]

There has been significant progress in the area of video understanding. YFCC-100M datasets [1] with 800K videos and metadata, containing video titles, descriptions and tags; ActivityNet [2], which is a large-scale video benchmark with several thousand videos and 200 human activity classes, and can be used to compare algorithms for human activity understanding. As of today, the Sports-1M [3] has been considered one of the largest video datasets, containing around 1M video instances. However, there has been no video datasets, comparable in scale and diversity to YouTube-8M. The updated YouTube-8M dataset contains over 8M videos and 1.9B video frames and provides the richest resource for research in video understanding and representation learning.

3. PRELIMINARY PLAN

What will be done between now and intermediate report due date (March 17, 2017).
Due to the large size of the data files (31 GB of video-level data, about 1.7 TB of frame-level data), they are hosted on Google Cloud, along with the training and validation sets ground truth labels. I will set up a Google Cloud account to retrieve the training and test files, as well as set up the required environment using cloud shell by installing prerequisite packages and dependencies and the latest version of TensorFlow. Next, I will work on a more comprehensive survey of the existing research related to video understanding and the machine learning algorithms used to predict the key labels of a video. Along with literature review, I will work on exploring the data and finalize data-preprocessing and feature engineering. Along with the release of the dataset, Google Research team also released a technical report[4], where the authors used Support Vector Machine, Logistic Regression and Mixture of Experts to classify video labels. The classifiers reported by the authors could be a good set of models to start with and the reported classification results will be used as benchmark baseline results. Ultimately, the purpose of this project is to identify the classification model that gives the best label predictions based on the test data provided by Google, which is not publicly available.

What will be done between March 17, 2017 and Final Report due date (April 28, 2017).
Within this timeframe, I will work on implementation of the classification algorithms, with intermediate submissions of the predictions on Kaggle. That way, I would be able to evaluate the prediction accucacy of the model on the test set, which is not publicly available. At this time, all other required sections of the report would also be expanded, refined and completed by the project due date.

4. REFERENCES

[1] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning, ICML*, pages 448–456, 2015.

[2] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. ctivitynet: A large-scale video benchmark for human activity understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, 2015.

[3] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1725–1732, Columbus, Ohio, USA, 2014.

[4] Abu-El-Haija Sami and et al. Youtube-8m: A large-scale video classification benchmark. arxiv preprint arxiv:1502.07209. 2016.