Large-Scale Video Classification with Cloud Computing

EECS 351, Winter 2017, UM Ann Arbor

Tyler Kohan: tkohan@umich.edu

Murat Turkeli: mturkeli@umich.edu

Yunsheng Bai: yba@umich.edu

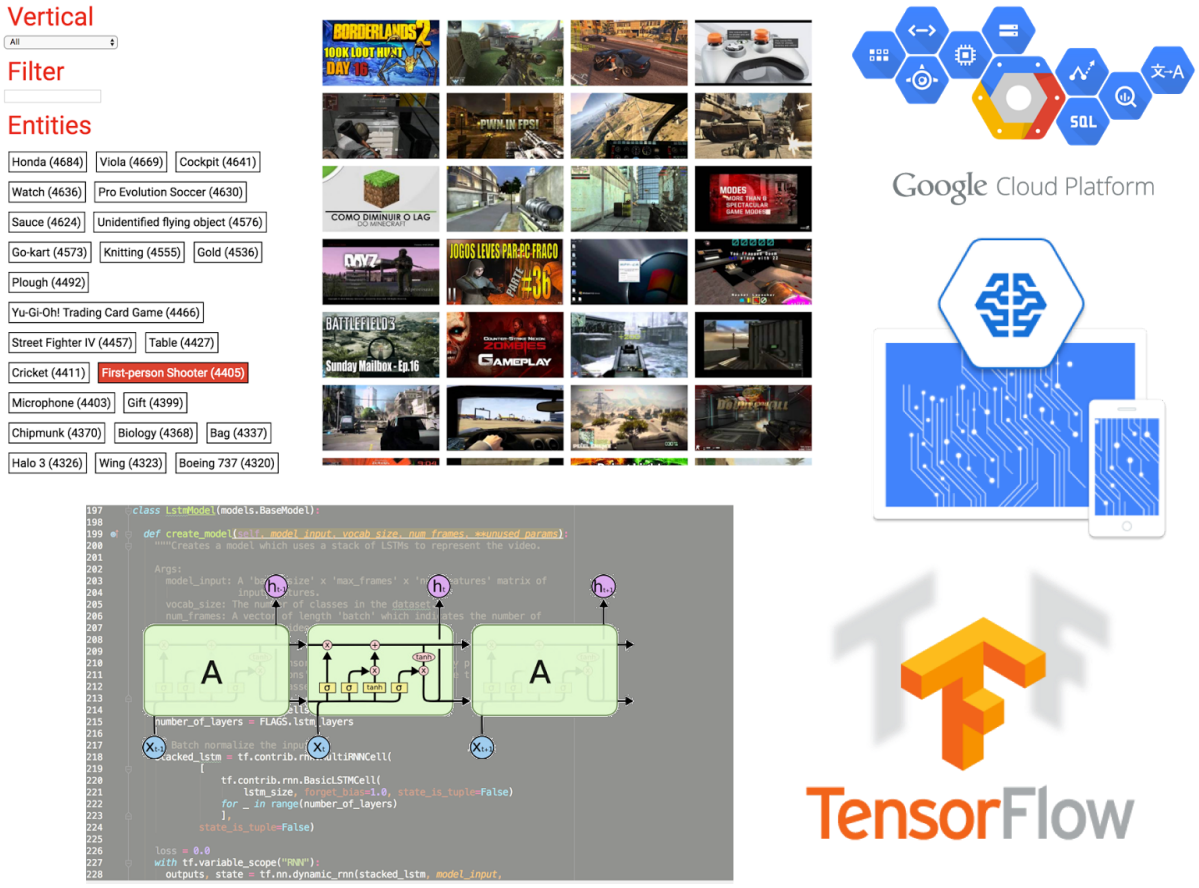

Figure 1: Overview. This project involves a large-scale video dataset, deep learning with TensorFlow, Google Cloud Platform, etc.

1. Problem Statement

According to Kaggle, video captures a cross-section of our society. And major advances in analyzing and understanding video have the potential to touch all aspects of life from learning and communication to entertainment and play. In this project, we are participating in a public Kaggle competition sponsored by Google. Google is inviting the Kaggle community to join efforts to accelerate research in large-scale video understanding, while giving participants early access to the Google Cloud Machine Learning (Cloud ML) beta platform.

In this competition, we will develop classification algorithms which accurately assign video-level labels using the new and improved YouTube-8M V2 dataset. By taking part, we will not only explore **machine learning** using **large video dataset** trained on the **Google Cloud Platform**, but also try to state-of-the-art benchmarks.

2. Dataset

We use the YouTube-8M dataset (figure 2). According to Google's research paper, YouTube-8M contains more than **8 million videos**—over 500,000 hours of video—from 4,800 classes. It also comes with **precomputed state-of-the-art audio-visual features** from every second of video (3.2B feature vectors in total). The total size of the frame-level features is 1.71 Terabytes. The total size of the video-level features is 31 Gigabytes. According to Google's official website, this makes it possible to get started on this dataset by training a baseline video model in less than a day on a single machine! We will use Python with TensorFlow to develop the models, and Google Cloud Platform to train the models. See more details below.

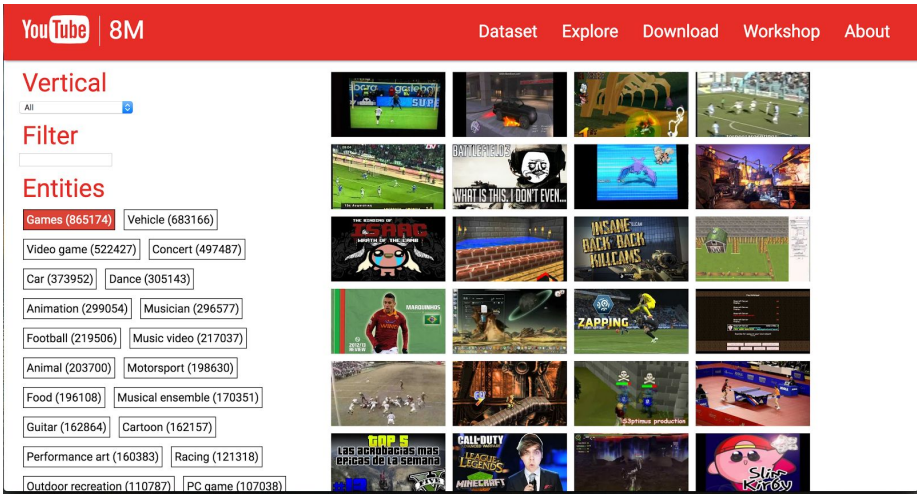

Figure 2: YouTube-8M V2 dataset has 4716 classes. Reference: https://research.google.com/youtube8m/explore.html

The significance of YT-8M should be stressed. Today, one of the greatest obstacles to rapid improvements in video understanding research has been the lack of large-scale, labeled datasets open to the public. For example, the availability of large, labeled datasets such as ImageNet has enabled continued breakthroughs in machine learning and machine perception. Figure 3 illustrates the scale of YouTube-8M, compared to existing image and video datasets. To that end, Google's recent release of the YouTube-8M (YT-8M) dataset represents a significant step in this direction. Making this resource open to everyone from students and industry professionals is expected to kickstart innovation in areas such as representation learning and video modeling architectures.

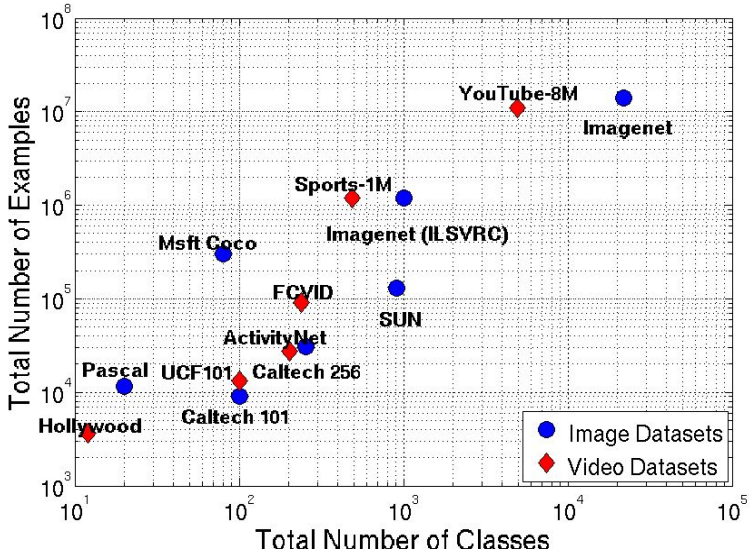

Figure 3: The progression of datasets for image and video understanding tasks. Large datasets have played a key role for advances in both areas. Reference: Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., & Vijayanarasimhan, S. (2016). Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*.

3. Methods

We will develop several machine learning models using TensorFlow and train them on Google Cloud Platform.

Machine learning models: Our Github repo is a fork of a starter repo which already implements the Logistic Model, Mixture of Expert Model, Deep Bag of Frame (DBof) Model, and Long Short-Term Memory (LSTM) Model using TensorFlow. We will try combinations of these models as well as new ones including Discrete Fourier Transform (DFT), Decision Tree Model, Random Forest Model, Support Vector

Machine (SVM) Model, Convolution Neural Network (CNN) Model, etc. Figure 4 shows the architecture of the DBof Model.

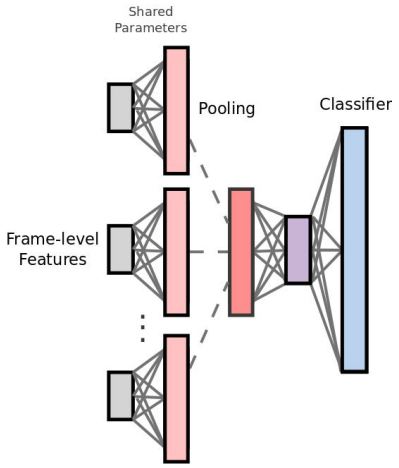

Figure 4: The network architecture of the DBoF approach. Reference: Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., & Vijayanarasimhan, S. (2016). Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*.

Training: Since the dataset is too large to store on our local computer, we will use Google Cloud Platform (with GPU) to train our models. Figure 5 shows the screenshots of working with Google Cloud Platform. According to the Kaggle policy, a free trial account includes $300 in credits, and participants who expend their $300 free trial credits may be eligible to earn additional Google Cloud credits. We will start using free credits. We will use the gradient descent algorithm on the Google Cloud Platform to train our models.

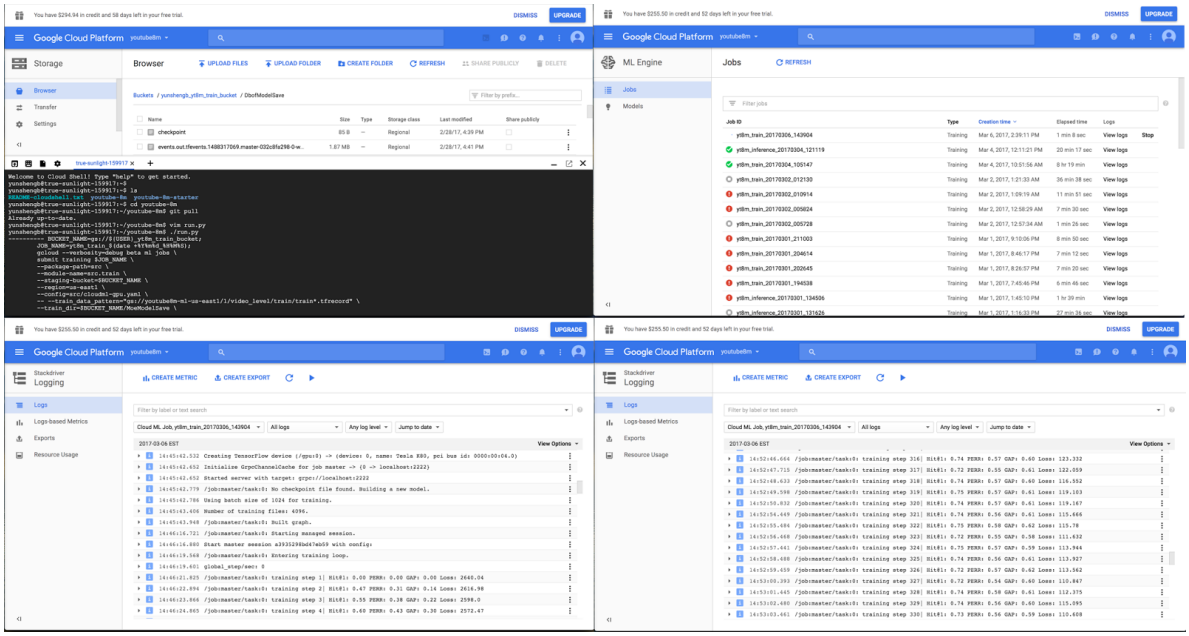

Figure 5: Google Cloud Platform in use.

4. Quantitative Measures

According to Kaggle, submissions are evaluated using the Global Average Precision (GAP) at k, where k=20. For each video, we will submit a list of predicted labels and their corresponding confidence scores. The evaluation takes the predicted labels that have the highest k confidence scores for each video, then treats each prediction and the confidence score as an individual data point in a long list of global predictions, to compute the Average Precision across all of the predictions and all the videos.

If a submission has N predictions (label/confidence pairs) sorted by its confidence score, then the Global Average Precision is computed as:

$$GAP = \sum_{i=1}^{N} p(i)\Delta r(i)$$

where N is the number of final predictions (if there are 20 predictions for each video, then N = 20 * #Videos), p(i) is the precision, and r(i) is the recall. Precision is the fraction of correct labels that we predict over the number of labels that we predict. Recall is the fraction of correct labels that we predict over the number of ground truth labels.

5. Links

Dataset: ***https://research.google.com/youtube8m/***
Paper: ***https://arxiv.org/pdf/1609.08675.pdf***
Kaggle: ***https://www.kaggle.com/c/youtube8m***
Github: ***https://github.com/yunshengb/youtube-8m***