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

Tatyana Li
litatyan@msu.edu

1. PROBLEM STATEMENT

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. 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, however, publicly available.

2. RELATED WORK

Until very recently, the research on image recognition had been applied to small labeled image datasets, e.g. Caltech 101/256 [1], PASCAL [2]. Due to a rapid increase in computational power, there has been a significant advancement in image understanding research employing much larger-scale datasets, e.g. ImageNet [3] and SUN [4]. As an example, Krizhevsky et al. [5] performed training of a large, deep convolutional neural network to classify a subset of 1.2M images from the ImageNet database into 1000 different classes. The deep neural network contained 60M parameters and 650,000 neurons, with 5 convolutional layers. The researchers showed that using purely supervised learning, the deep convolutional neural network is capable of producing high classification accuracy. The authors report top-1 and top-5 error rates of 37.5% and 17%, which was significantly lower, compared to the previously established benchmark. Sermanet et al. [6] used convolutional neural nets for digit classification of house numbers from SVHN classification dataset, containing 32x32 images, with three color channels. The authors report that with L4 pooling, they achieved a state-of-the-art performance, with classification accuracy close to 95%.

Although, relatively less work has been done on large-scale video classification, compared to the image domain, still, significant progress has been made in the area of video understanding research. The scale of the datasets progressed from smaller labeled video datasets Hollywood 2 [7], Weizmann [8], containing only several thousands of video clips to much larger-scale datasets, such as YFCC-100M dataset [9] with 800K videos and metadata, containing video titles, descriptions and tags; ActivityNet [10], 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 [11] has been considered one of the largest video datasets, containing around 1M video instances. To date, 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. The dataset comes with "pre-computed state-of-the-art features for 1.9 billion video frames", which "levels the playing field" for researhers to utilize the dataset on a scale never seen before.

3. DATA

[P.S. To be expanded further.] YouTube-8M dataset represents a benchmark in video understanding, with the key purpose to identify the main topics associated with a video. The dataset contains YouTube videos from a diverse range of categories from gaming and entertainment to sports, foods, health, science and education. It sets this dataset apart from its' predecessors, where, for the latter, categories were limited to a single domain, such as, for example, sports or action. The Google researchers used a video annotation system [12] to produce topic annotations, and then obtain videos for a certain topic. Overall, the YouTube-8M dataset contains 4800 classes with close to 8.3M videos. A given video may have more than one class label, with an average of 1.8 labels per video. The train set, released by Google, contains 5,786,881 video examples, validation set contains 1,652,167 videos, and test set contains 825,602 videos. The dataset comes with standard frame-level features. However, there's still a lot of room for experimenting with video-level representation learning approaches.

4. MODELS

[P.S. To be expanded further.] In a technical report, published by Google, the researchers used video-level features to train independent binary logistic classifiers for each label, batch SVMs and a mixture of experts, first introduced by Jacobs and Jordan [**?**]. To overcome the challenge of scale which this dataset poses, the authors, instead of performing batch optimizations, which would have been impossible, used online learning algorithms, such as Adagrad to perform model updates, using small mini-batch examples.

5. PRELIMINARY PLAN

What has been done.
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 and ground truth labels. I 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. I worked on a survey of the existing research related to video understanding and the machine learning algorithms used to predict the key labels of a video. As I progress through the project, it's possible, that I would encounter other relevant research papers, that will subsequently be added to the literature review section of the paper. Google created a starter code in a designated github repository, that I ran in Cloud ML as a starting point to learn how to train, evaluate, and create predictions. As mentioned earlier, the classifiers reported by the authors of the technical report, could be a good set of models to start with. Following this logic, as a starting point, I decided to train simple binary logistic classifiers, in order to obtain first predictions and establish a benchmark to further improvement. However, due to an increasingly large size of the train data, the models had been taking an exceedingly long amount of time to converge. Currently, I'm creating a smaller set of train data to perform initial training and obtain predictions on a single machine. Afterwards, the models will be traing on a Google Cloud ML platform.

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.

6. REFERENCES

[1] G. Griffin, A.Holub, and P.Perona. Caltech-256 object category dataset. technical report 7694, california institute of technology,. 2007.

[2] M. Everingham, V.V. Cool, C.K.I. Williams, J.Winn, and A.Zisserman. The pascal visual object classes (voc) challenge,. 2009.

[3] J.Deng, W.Dong, R.Socher, L.Jia Li, K.li, and L.Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*, 2009.

[4] J.Xiao, K.A.Ehinger, J.Hays, A.Torralba, A.Olivia, and J.Xiao. Sun database: Exploring a large collection of scence categories. 2013.

[5] A. Krizhevsky, I.Sutskever, and G.E.Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems, NIPS*, pages 1097–1105, 2012.

[6] P. Sermanet, S.Chintala, and Y.LeCun. Convolutional neural networks applied to house numbers digit classification. In *ICPR*, 2012.

[7] I.Laptev, M.Marszalek, C.Schmidt, and B.Rozenfeld. Learning realistic human actions from movies. In *Conference on Computer Vision and Pattern Recognition, IEEE*, 2008.

[8] M.Blank, L.Gorelick, E.Shechtman, M.Irani, and R.Basri. Actions as space-time shapes. In *Proceedings of the International Conference on Computer Vision, ICCV*, 2005.

[9] 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.

[10] 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.

[11] 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.

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