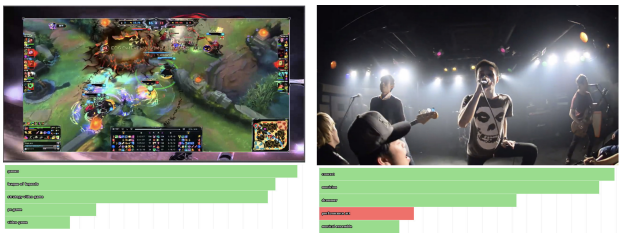
Video Understanding

Sanmathi Kamath, Pranjali Kokare, Alekhya Munagala

***Abstract*— With an increase in generation and consumption of videos, efficient retrieval of videos has become a challenging task. By providing multiple labels/categories to each video, we hope to develop a better video understanding, as a step towards solving this problem.**

I. Problem Statement

Today, videos have become an important means of information. Statistics show that around 300 hours of videos are uploaded to YouTube every minute and nearly 5 billion videos are watched every day. This has resulted in huge amounts of unorganized video data. Classifying and assigning labels/categories (popularly known as tags) to videos and improving video understanding can lead to better video search and discovery, improved video recommendations and pave way for efficient video retrieval.



(a) Games, PC games, League of Legends, Video Game, Strategy Video Games

(b) Concert, Drummer, Musician, Musical Ensemble, Performance Art

Fig. 1. Example labels for a video frame

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 [1]. The YouTube-8M challenge [2] provides a great source of data and inspires us to explore the area of Video Understanding. Given a video, the challenge is to provide multiple labels/categories that could be used to describe the video. Using this dataset, we aim to create a compact model for large-scale video understanding to tag each video with multiple labels.

II. Problem Importance

While images are a powerful source of information, temporal continuity of events in videos help in delivering information more intuitively. For classifying a video, we need a model that can efficiently capture this temporal information. Developing such models come with a cost of computational time. Hence, finding a compact model with balance between speed and accuracy is essential.

Tagging a video manually is time consuming and subjective. Humans have different perceptions which might result in noisy labels. Also, video datasets are usually limited to a specific genre. Most well studied video datasets, such as Sports-1M[5], ActivityNet[6], UCF-101[7] are all confined to a certain theme of videos. These challenges show the need for developing a generic and accurate model for video understanding.

III. Proposed Approach

A. Dataset

YouTube-8M is the largest multi-label video classification dataset, composed of 8 million videos—500K hours of video—annotated with a vocabulary of 4800 visual entities. Each video is between 100 to 500 seconds long.

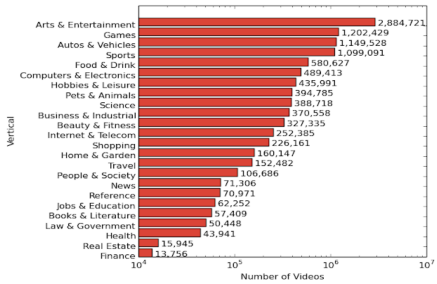


Fig. 2. Distribution of top 25 most labelled visual entities in the YouTube-8M Dataset

B. Our Approach

We propose to tackle the task of video understanding in two ways.

First, we wish to research and implement a frame-level classifier. The video will be sampled, and each frame will be classified into multiple labels. The scores will be combined using average pooling to predict top k labels for the video. This leverages the fact that adjacent frames are highly correlated. Frame-level classifier + average pooling will provide a good baseline model.

Second, we wish to research and implement a video-level classifier, which includes the temporal information. This will extract video-level features to predict the k labels with highest confidence scores for each video.

Our focus in the second step would be to research and implement a Deep Neural Network Architecture, since it would have the capabilities to capture more complex features. The next step would be to analyze our implemented models to understand how weights are being learnt and what features are being extracted by these models. Eventually, we would like to understand if the features extracted by the deep architectures can be encoded and utilized for solving our goal via lightweight architectures. At the same time, we would like to evaluate the trade-offs of using a lightweight architecture model using complex features vs using a deep neural network architecture for the same task.

IV. Validation of Implementation

Validating our implementation is heavily related to evaluating the performance of our models. Therefore, for this section, we will focus on answering how we will evaluate our code. The implemented model will generate a vector of tags for each given video. The correctness of these tags will be evaluated based on the evaluation metrics described in the following section.

V. Performance Evaluation

A few evaluation metrics that can be used for evaluating the performance of this technique are given below. As there are more than one correct labels for a given video, basic accuracy calculation is not sufficient to measure the performance of our model.

One of the evaluation metrics that are used for evaluating these types of problems is Hit@k. This is the fraction of test samples that contain at least one of the ground truth labels in the top k predictions.

Another evaluation metric is Global Average Precision (GAP). 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.[1]

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

where N is the number of predictions, $p(i)$ is precision and $r(i)$ is recall.

VI. Resources

The resources that we will use are :

- Software: PyTorch
- Hardware/Cloud: Google Cloud, Google Colab, GPU Cluster
- Dataset: YouTube-8M Dataset [2]

VII. Goals

The goals we aim for this project are as follows:

- 75%
  - Data Analysis and Pre-Processing
  - Review and Analyze existing model architectures
  - Implementing Frame-level Model
- 100%
  - Researching and Implementing Video-level model
  - Analysis and Evaluation of the implemented methods
- 125 %
  - Analyzing the weights learnt by deep neural network to explore the idea of using traditional machine learning models
  - Evaluating the trade-offs between model architecture and performance

References

[1] https://www.kaggle.com/c/youtube8m-2018/overview

[2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, Sudheendra Vijayanarasimhan, "YouTube-8M: A Large-Scale Video Classification Benchmark", arXiv:1609.08675

[3] Duarte, K., Rawat, Y., Shah, M. (2018). "VideoCapsuleNet: A simplified network for action detection. In Advances in Neural Information Processing Systems" (pp. 7610-7619)

[4] Lin, R., Xiao, J., Fan, J. (2018). "NextVLAD: An efficient neural network to aggregate frame-level features for large-scale video classification." In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 0-0).

[5] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In CVPR, 2014.

[6] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In CVPR, pages 961–970. IEEE Computer Society, 2015.

[7] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. CoRR, abs/1212.0402, 2012.