# Semantic Video Classification Challenge

**Xiulong Liu**
Department of Electrical and Computer Engineering
University of Washington
xl1995@uw.edu

**Tianyi Zhang**
Department of Electrical and Computer Engineering
University of Washington
zhang506@uw.edu

## Abstract

In this project, we explore deep neural networks for semantic video classification. The general pipeline is to use extracted frame level features to predict the multiple labels that each video is associated with. Logistic model is our baseline approach, and we explore the LSTM model and the state of the art NetVLAD model to give some insights on semantic video classification.

## 1 Introduction

Semantic video classification is providing semantic tags for videos, where each video can be associated with multiple categories. For example, a video can be categorized as sport, and further to be basketball. In our project, we aim at predicting the probability of each video belongs to each category and choosing the top 5 as our labels for the result. Our data comes from YouTube8M dataset, including 3862 total categories with 6.1million videos. And we treat the problem as multiple binary classification problem.

The challenge of the task lies in the fact that each video has only a few positive labels, whereas all the other categories are not. And videos are comprised of consecutive frames, where not only image features for each frame are significant for predicting labels, but also the temporal dependencies between each frame plays important roles as well. And it is well known that Recurrent Neural Network(including all its variants) is good at modeling sequential data like speeches, videos and texts. Therefore, we want to see how well it performs on semantic video classification task. And further, we want to see how state of the art model performs compared to RNN and give some insights on why they could achieve even better performance on this specific task.

## 2 Method

### 2.1 Dataset

Youtube8M dataset contains over 6 million videos with 3862 labels. The average number of label is 3 per video. Two types of features are available, frame-level and video-level. All features are obtained from a GoogLeNet pre-trained on ImageNet. For our semantic video classification need, we uses the frame-level feature. To obtain frame-level feature, the original video is down sampled into 300 frames. For each frame, a 1024-dimension feature is generated. Along with video id and label, these

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

three types of content are packed into a tfrecord file. Due to the massive among of data, we only used 10% of the total 6 million video to train our network.

### 2.2 logistic

As a baseline model, we implement a simple logistic model to compare with the performance of other models. We implement a fully connected network with one hidden layer. The each cell in output layer uses the sigmoid function as activation function. Each label is treated as a binary classification problem and sigmoid cross entropy is used as loss function.

### 2.3 LSTM

We uses LSTM model to treat the input data as a sequence to obtain the feature in temporal feature. LSTM is derived from RNN model. RNN contain multiple connected cells. Each cell can have its own input and output. The weight of each cell is passed to the next cell to preserve the features in sequence input. LSTM is designed for the network to remember the relatively important input in each cell and filter out the input that does not contribute much to the output. LSTM uses sigmoid functions to act as gates for the information to pass selectively. All gates in a LSTM cell control the state of this cell. The output of each cell is determined by the state. We use two layers of stacked LSTM, each layer contains 512 cells. For a 300-frame input feature, each cell takes a frame as input and output labels. Only the last output is considered result label and output in other cells are used for loss calculation and back propagation. We use sigmoid for each individual label to output a probability. The loss function is implemented is sigmoid cross entropy to estimate the loss across 3862 labels.

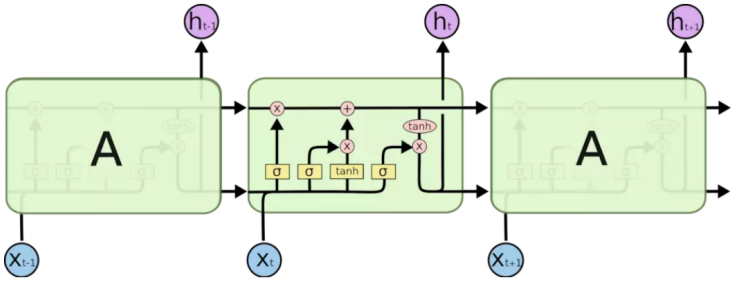


Figure 1. LSTM structure.

### 2.4 NetVLAD

The idea of NetVLAD comes from the traditional computer vision descriptor VLAD. The VLAD descriptor stands for "vector of local aggregated descriptor". Generally, we have a vector representation of each local region in an image, i.e, HOG descriptor. However, simply concatenating all local feature vectors within an image to form a global descriptor is not always a good idea becuase the dimension of global feature would be very high in this case. Furthermore, the relationship between each local descriptor cannot be addressed by simply concatenating descriptors one by one. To solve this problem, VLAD first uses kmeans clustering to provide cluster representation of features in local descriptor space. And then each local descriptor is subtracted from the cluster center with which the descriptor is associated with. In this case, each local descriptor forms a vector that points from its cluster center to itself. Aggregating all these vectors within a cluster will give a compact representation in exact the same dimension as each local descriptor, and concatenating all cluster representation gives us the vlad vector. Formally, the vlad descriptor per cluster could be formulated as:

$$V(j,k) = \sum_{i=1}^{N} a_k(x_i)(x_i(j) - c_k(j)) \tag{1}$$

where $V(j,k)$ denotes the $j^{th}$ dimension of cluster k and $c_k$ is the the center of cluster k run by k-means on entire local descriptors on image. NetVLAD borrows the idea from VLAD and turns it into a network layer that could gather all local descriptors into a single vector. However, the cluster assignment cannot be hard assignment because it would not be back-propagated using gradient descent otherwise. What happens inside the NetVLAD layer is that the cluster assignment is soft by means of calculating the probability of the cluster to which the descriptor belongs to using softmax. The overall struture of NetVLAD is shown in Figure2.

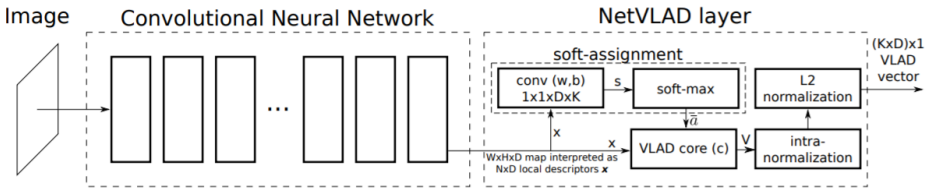


Figure 2. NetVLAD structure.

And the right part of the diagram is the core of VLAD layer. In this setting, the cluster center and the softmax units are learnable parameters. Number of clusters is hyper-parameter. The layer is formulated as below.

$$V(j,k) = \sum_{i=1}^{N} \frac{e^{w_k^T x_i + b_k}}{\sum_{k'} e^{w_{k'} x_i + b_{k'}}} (x_i(j) - c_k(j)) \tag{2}$$

### 2.5 State of the art model

The champion model of the YouTube-8M Video Understanding Challenge on Kaggle is called Gated NetVlad [4]. Gated NetVLAD is a model which includes both enhancement in feature extraction and gating. A context gating unit is introduced in this model and it replaces LSTM to extract the context information. Moreover, Gated NetVLAD uses ensemble learning to overcome the downside of single feature extraction model. Multiple models are used to further process the feature and mix the result together before gating. The way of gating introduced in Gated NetVLAD is call context gate. The formula is shown in (3), where $\sigma$ is sigmoid function, W and b are trainable weights and $\circ$ is element-wise multiplication. This gate is able to directly re-weight the input X to preserve the specific meaning (in this case, the score of labels) in the layer [4]. Then a Mixture of the expert model with context gating used for classification.

$$Y = \sigma(WX + b) \circ X \tag{3}$$

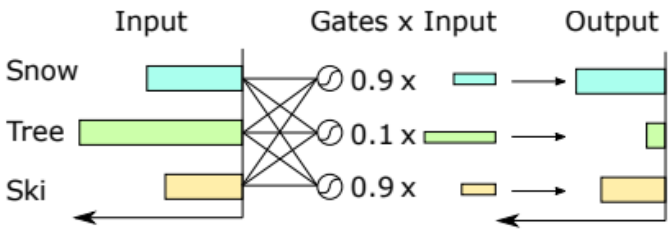


Figure 3. Context gate re-weight.

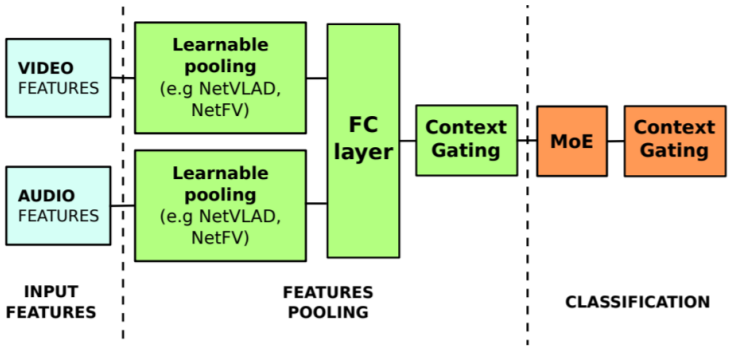


Figure 4. Gated NetVLAD model structure.

## 3 Result

We use 600000 videos to train our LSTM model and 10000 videos to test. Since the dataset is massive, we only train our model for 5 epochs with the learning rate of 0.001, batch size of 128, 0.05 of learning rate decay every 400000 examples and Adam optimizer. The total time used for training 5 epoch is over 12 hours. We uses three different indices to evaluate the performance of the model[2]. GAP is the average precision on the top 20 labels per example. Hit@1 is the probability of the top 1 label being the ground truth label. PERR is the annotation precision when the same number of labels per video are retrieved as there are in the ground-truth[3]. The improvement fro m logistic model to LSTM model is significant. NetVLAD shows even better performance than LSTM. Since LSTM model introduces the concept of gate and NetVLAD enhances the method of feature extraction, the result is very likely to be better to include both models. As the state of the art model, Gated NetVLAD's performance is 5.8% better than NetVLAD and 9.6% better than LSTM.

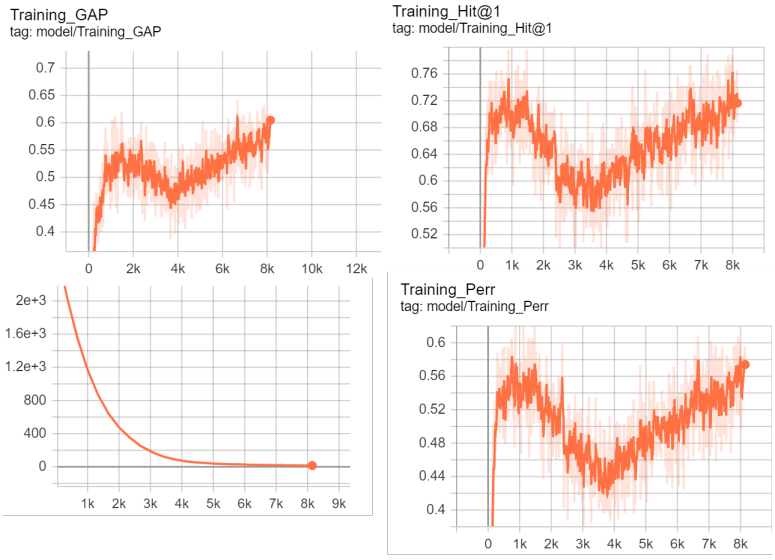


Figure 5. Result of logistic model.

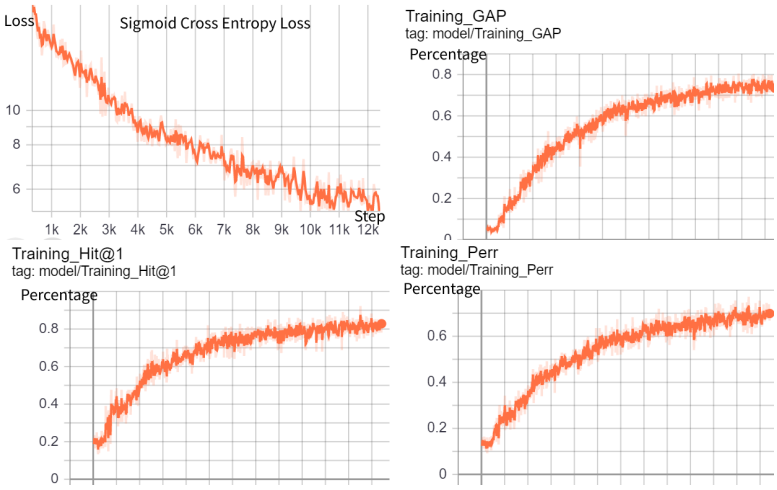


Figure 6. Result of LSTM model.

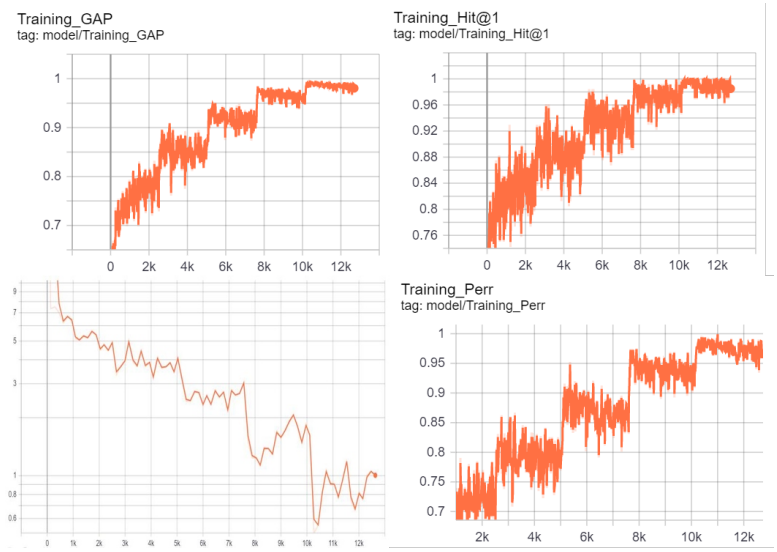


Figure 7. Result of NetVLAD model.

Table 1. Performance of models on GAP, PERR and Hit@1.

| Model | GAP | PERR | HIT@1 |
|---|---|---|---|
| Logistic | 57.7 | 55.7 | 71 |
| LSTM | 73.6 | 68.3 | 81 |
| NetVLAD | 77.4 | 72.6 | 83.9 |
| Gated NetVLAD | 83.2[4] | NA | NA |

## 4 Conclusion

LSTM model and logistic model both achieve relatively good result. LSTM model overperforms logistic model by 16% in GAP on test dataset. On one hand, the logistic model is hard to retrieve temporal information, while its result is still good. This may indicate that the feature in every frame is more important than the temporal information. On the other hand, the increase of performance for LSTM model indicate that the temporal information has certain contribution to the result. The idea of model enhancement is implemented by modifying the feature extraction part or using gating to keep features that are more meaningful to the video. With both methods, the Gated NetVLAD achieves the state of the art.

## References

[1] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[2] Abu-El-Haija, Sami, et al. "Youtube-8m: A large-scale video classification benchmark." arXiv preprint arXiv:1609.08675(2016).

[3] Kim, Hyun Sik, and Ryan Wong. "Google Cloud and YouTube-8M Video Understanding Challenge.“

[4] Miech, Antoine, Ivan Laptev, and Josef Sivic. "Learnable pooling with context gating for video classification." arXiv preprint arXiv:1706.06905 (2017).

[5] Arandjelovic, Relja, et al. "NetVLAD: CNN architecture for weakly supervised place recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.