Video Understanding: Project Progress Report

Sanmathi Kamath, Pranjali Kokare, Alekhya Munagala

Problem Statement: Review

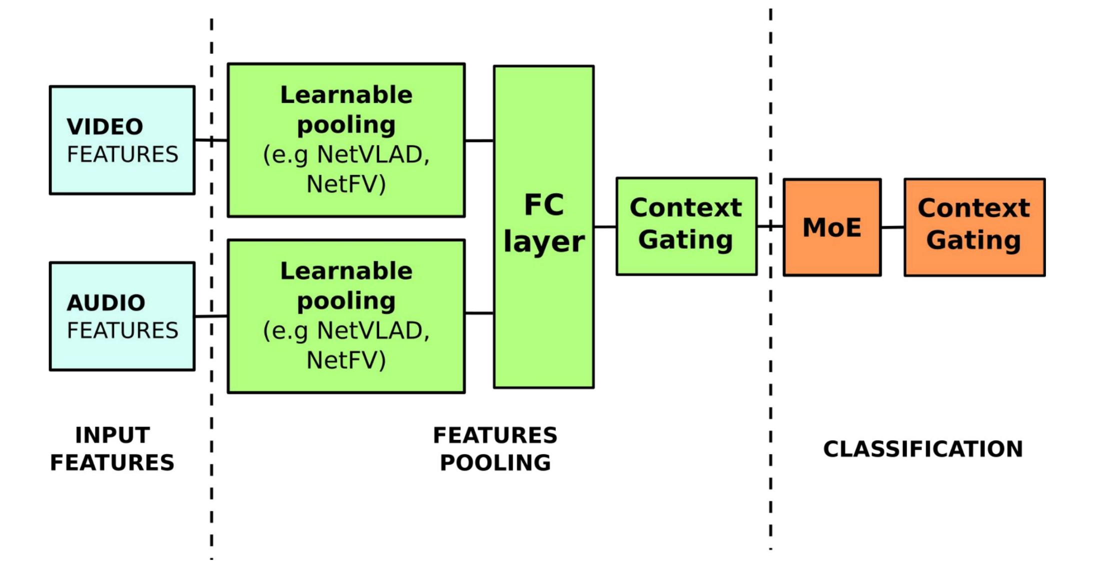
- Goal: Given a video, we would like to predict multiple labels/categories for each video.
- Understanding videos can be used for:
  - Automatic Tagging of new videos
  - Easy video retrieval/ search
  - Video Recommendation

Celebration, Birthday, Blowing Candles, Cake

Literature Review

Learnable pooling with Context Gating for video classification

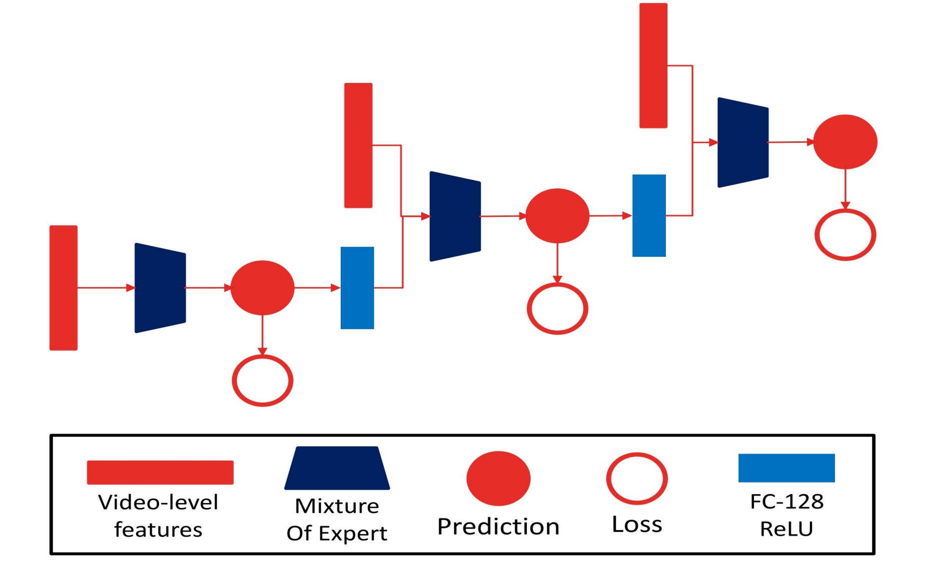
- Existing pooling techniques are simple but suboptimal
- Investigates learnable pooling techniques
- Context Gating: Non linear interdependencies between features and output label space
- Mixture of Experts



Literature Review

The Monkeytyping Solution to the YouTube-8M Video Understanding Challenge

- Proposes a deep learning architecture which uses input features and previous predicted labels
- Attention Weighted Stacking
  - Uses attention network to generate weights for LSTM pooling



Literature Review

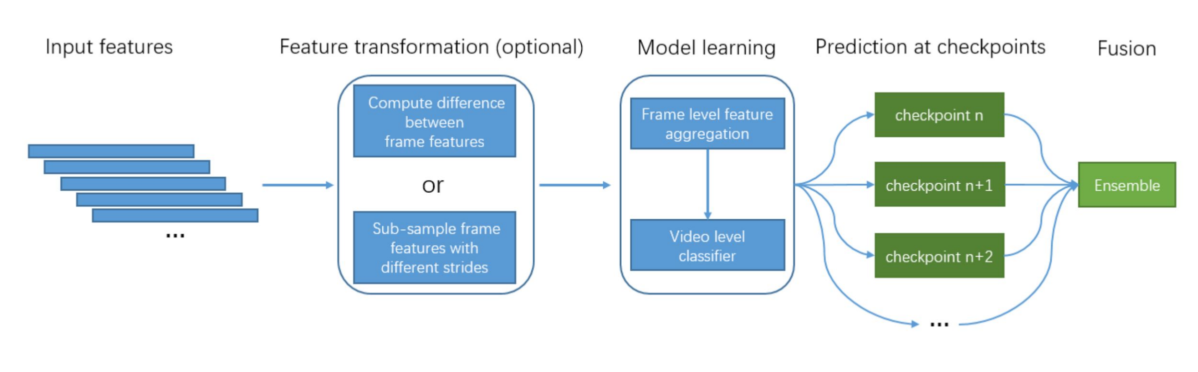
Temporal Modeling Approaches for Large-scale Youtube-8M Video Understanding

- Two-stream sequence model
  - Trains two bidirectional LSTM and GRU models for audio and video features separately and then runs them through attention layers and concatenates them,
- Fast-forward sequence model
  - The fast-forward connection takes the outputs of previous fast-forward and recurrent layer as input, and uses a fully-connected layer to embed them,
- Temporal residual neural networks
  - Temporal convolution neural networks are utilized to transform the original frame-level features into a more discriminative feature sequence, which can be further fed to LSTM.

Literature Review

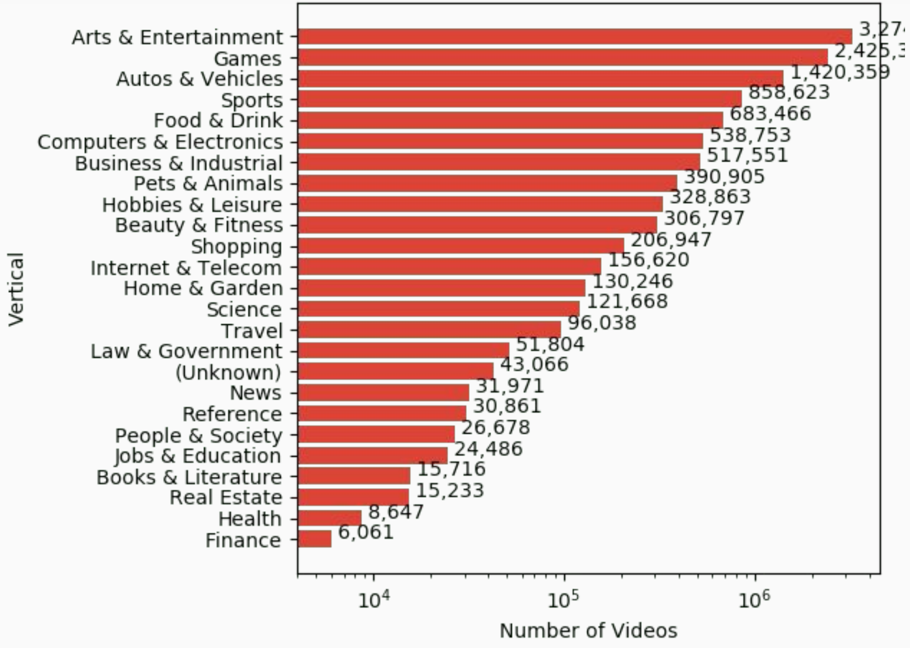
Aggregating Frame-level Features for Large-Scale Video Classification

- Feature transformation:
  - Leverage Temporal Information
  - Adjacent Frame Difference
- Predictions from multiple model checkpoints fused, output of multiple such models fused.



Understanding the Data

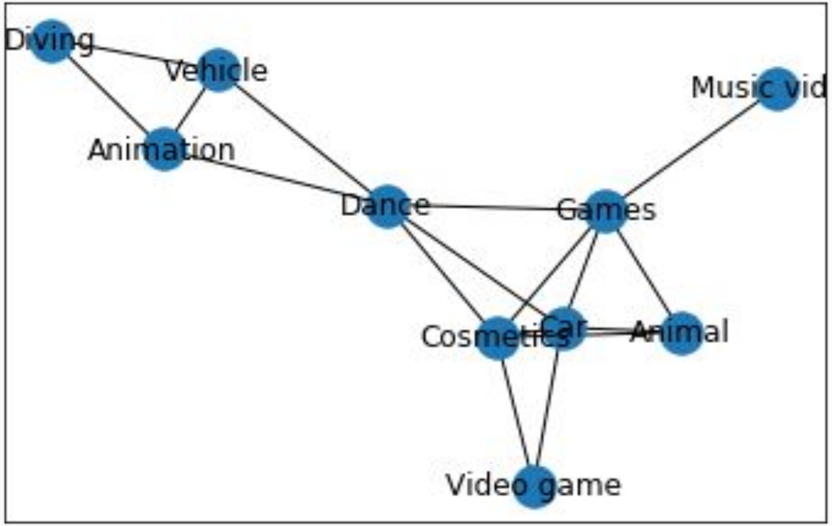
- 6.1 Million Videos
- 2.6 Billion Audio/Visual features
- 3682 classes
- Average 3 labels / video
- 3844 shards



Understanding the Data

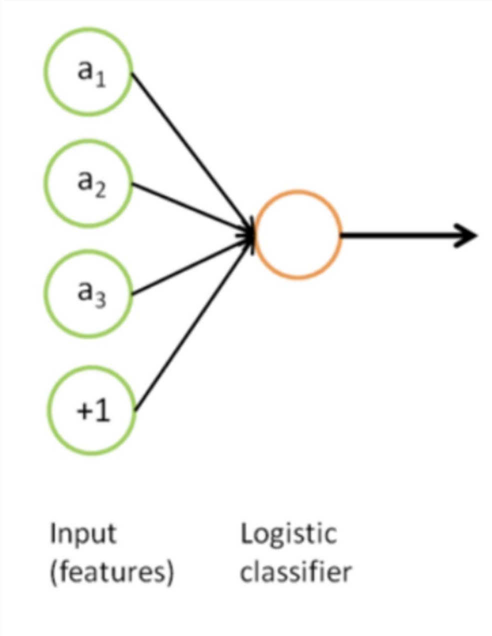
Features:

- Contains Youtube video ID and labels.
- 1024 RGB features
  - Mean of all frames for Video-level data
  - Calculated every second, upto 300 seconds for frame-level data
- 128 audio features
- Provided as tensorflow.Record



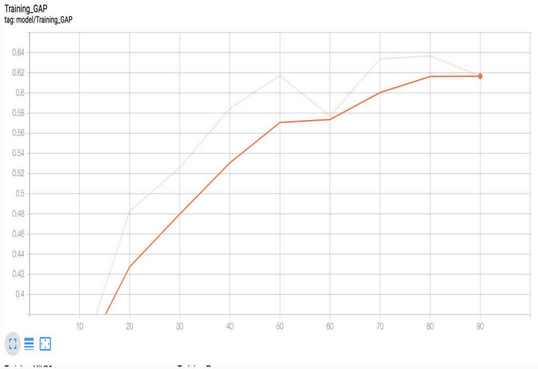
Frame Level Model

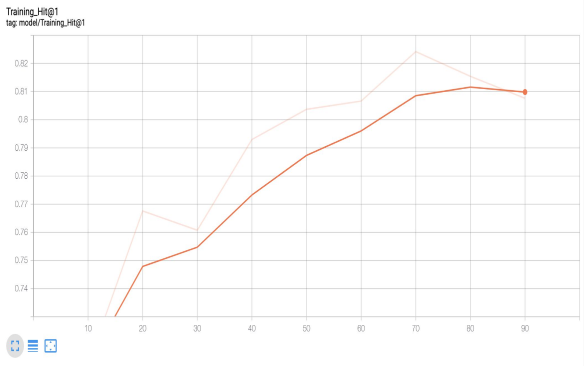
- Each video is sampled and scores are generated for each frame.
- One-vs-all classifier for each of the labels
- For inference, the scores of each frame are aggregated to predict top k labels for the video
- Average pooling to reduce the effect of outliers
- Fully-connected model with sigmoid non-linearity(Logistic model) and Average pooling

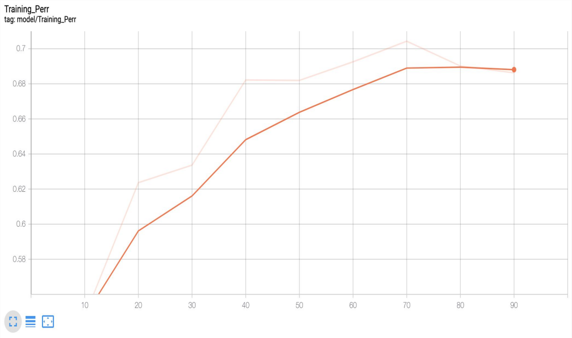


Training Process

We monitored our training process using the following metrics:







Global average Precision(GAP)

Hit@1

Precision@Equal Recall Rate

Dissection of AlexNet

- Used the Network Dissection tool (http://netdissect.csail.mit.edu/)
- Concepts defined over Broden dataset divided into following categories:
    - scene
    - object
    - part
    - material
    - texture
    - color
- Activations generated from each unit is compared against every concept.
- Intersection over Union (IoU) score computes confidence with which the concept has been detected.

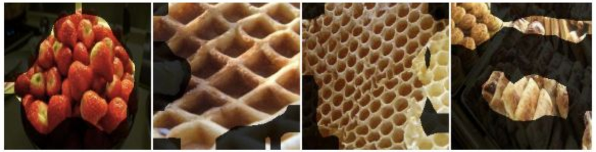
Results of Dissection

dog

unit 161 (object) IoU 0.10

waffled

unit 20 (texture) IoU 0.12



red

unit 57 (color) IoU 0.08

food

unit 47 (material) IoU 0.04

mountain snowy

unit 87 (scene) IoU 0.08

hair

unit 140 (part) IoU 0.06

Work Progress

- ➢ 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 tradeoffs between model architecture and performance.

Next Steps

- Tuning parameters using evaluations metrics
- Implementing transfer learning with AlexNet architecture for Frame Level Model
- Creating a framework for video level classification
- Network dissection of trained final models

References

- David Bau , Bolei Zhou , Aditya Khosla, Aude Oliva, and Antonio Torralba, “Network Dissection: Quantifying Interpretability of Deep Visual Representations”, arXiv:1704.05796
- Antoine Miech, Ivan Laptev, Josef Sivic, "Learnable pooling with context gating for video classification", arXiv:1706.06905
- He-Da Wang, Teng Zhang, "The Monkeytyping Solution to the YouTube-8M Video Understanding Challenge", arXiv:1706.05150
- Fu Li, Chuang Gan, Xiao Liu, Yunlong Bian, Xiang Long, Yandong Li, Zhichao Li, Jie Zhou, Shilei Wen, "Temporal Modeling Approaches for Large-scale Youtube-8M Video Understanding", arXiv:1707.04555
- Shaoxiang Chen, Xi Wang, Yongyi Tang, Xinpeng Chen, Zuxuan Wu, Yu-Gang Jiang, "Aggregating Frame-level Features for Large-Scale Video Classification", arXiv:1707.00803

Thank you!