YouTube-8M Video Understanding

Kun Huang
khuang47@stanford.edu

INTRODUCTION

MOTIVATION:

- Video understanding is a challenging task for numerous applications and research.
- This project addresses the problem of multi-label video classification and temporal localizations for user-generated videos.

INPUTS:

- YouTube-8M frame-level features dataset and segment-rated dataset.

APPROACH:

- Video-level
  - visual and audio features aggregation with NetVLAD.
  - Mixture-of-Experts for final classification
- Segment-level
  - Transfer learning based on video-level model.
  - Context-ignore and context-aware combined model.

RESULTS:

Video-level model achieves 85% global average precision. Segment-level model achieves 82% mean average precision.

DATA

- YouTube-8M dataset released by Google. Millions of YouTube videos, with machine-generated annotations from a diverse vocabulary of 3,800+ visual entities
- YouTube-8M Segments Dataset:which includes human verified labels at the 5-second segment level

METHODS

$$V(j,k) = \sum_{i=1}^{N} a_k(\mathbf{x}_i)\,(x_i(j) - c_k(j)) \;\Rightarrow\; V(j,k) = \sum_{i=1}^{N} \frac{e^{\mathbf{w}_k^T \mathbf{x}_i + b_k}}{\sum_{k'} e^{\mathbf{w}_{k'}^T \mathbf{x}_i + b_{k'}}}\,(x_i(j) - c_k(j))$$

EQN 1: Vector of Locally Aggregated Descriptors (VLAD)

EQN 1: NetVLAD, with VLAD integrated with supervised learning

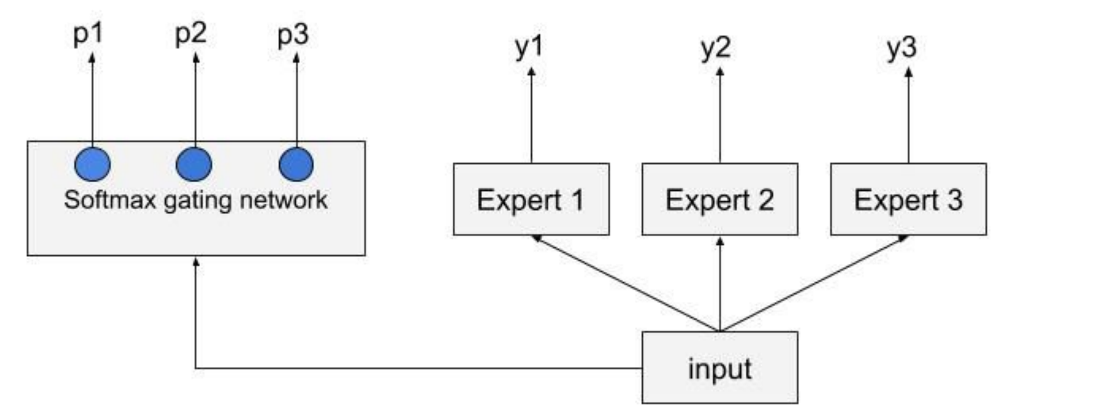

Figure 1: Mixture-of-Experts: experts specifies in different regimes, manager determines the relevance of experts.

$$p(c_k) = \sum_{j=1}^{E} p(c_k|e_j)p(e_j)$$

EQN 1: MoE formula

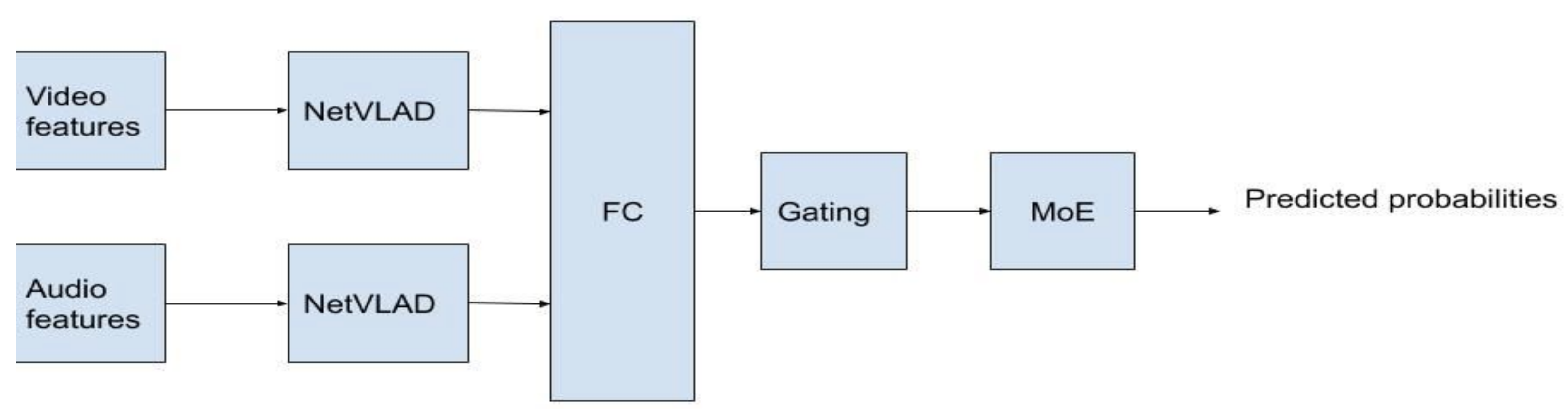

Figure 2: Video-level model. NetVLAD layer for features aggregation, MoE for the final classification

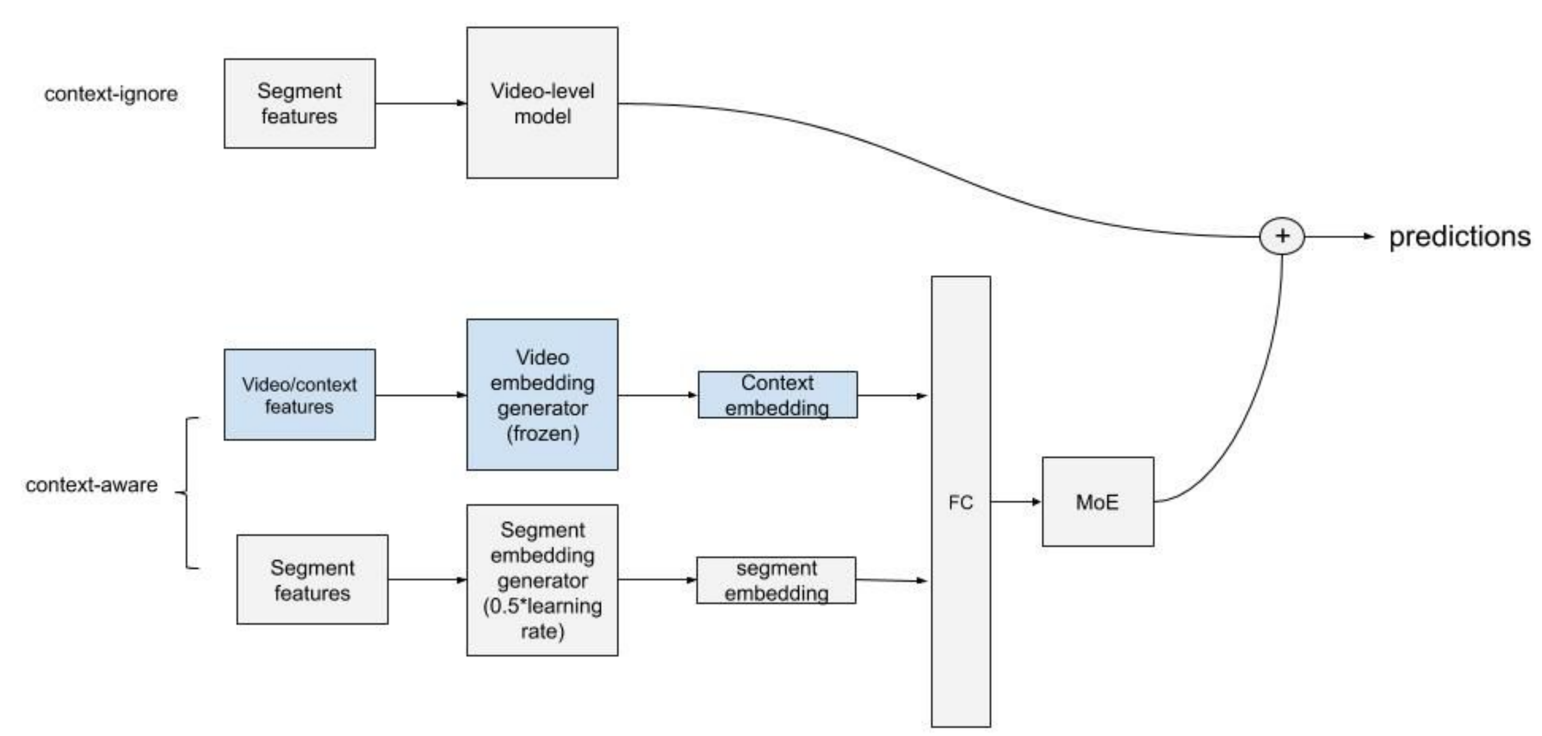

Figure 4: Temporal localization architecture: context-ignore model is video-level model fine-tuned on segment dataset. Context-aware model encodes entire video and each segment with video-level model, then delivers the embeddings to fully connected classifier.

RESULTS

Video-level classification: Global average precision of 85%. Temporal localization: Mean average precision of 82%

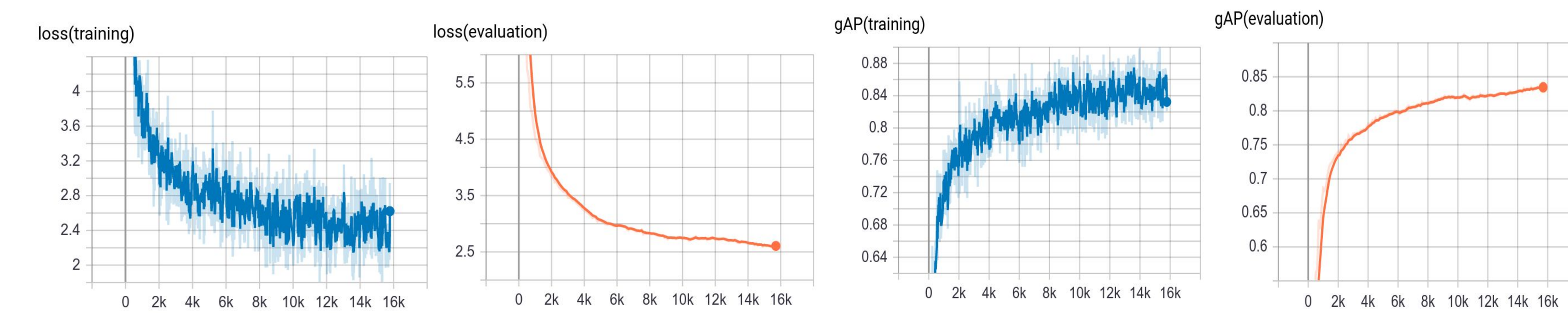

Figure 5: Training and evaluation process of video-level classification

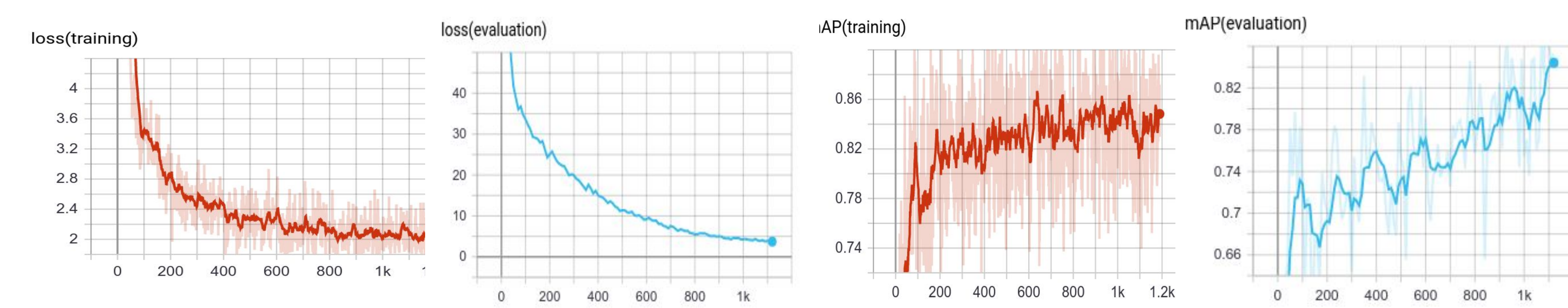

Figure 6: Training and evaluation process of temporal localization.

CONCLUSION

SUMMARY:

- Classifier with NetVLAD aggregation and Mixture-of-Experts achieves gAP of 85% in large-scale video classification.
- Transfer learning with context-aware and context-ignore combined model achieves mAP of 82% in temporal localization

FUTURE WORK:

- Incorporate temporal features in video classification as the current algorithm is focused on static features. E.g. combine NetVLAD, RNN and MoE.
- Reduce model size. Current video-level model is 3.72G, and segment-level model is 10G.