Aggregating Frame-level Features for Large-Scale Video Classification

in the Google Cloud & YouTube-8M Video Understanding Challenge

Shaoxiang Chen[1], Xi Wang[1], Yongyi Tang[2], Xinpeng Chen[3], Zuxuan Wu[1], Yu-Gang Jiang[1]
[1]Fudan University [2]Sun Yat-Sen University [3]Wuhan University

Google Cloud & YouTube-8M Video Understanding Challenge

- Video multi-label classification
  - 4,716 classes
  - 1.8 classes per video
- Large amount of data (used in the challenge)

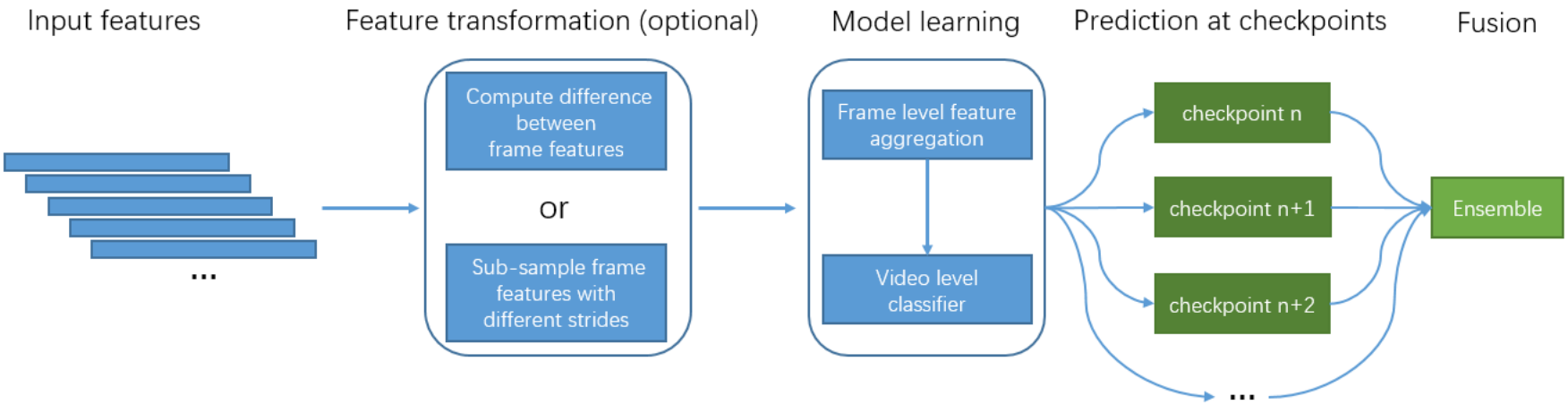
Partition	Number of Samples
Train	4,906,660 (70%)
Validate	1,401,828 (20%)
Test	700,640 (10%)
Total	7,009,128

- Audio and rgb features extracted from CNNs are provided in both frame and video level.

Summary

- Our models are: RNN variants, NetVLAD and DBoF.
- By default we used the MoE (Mixture of Experts) as video level classifier.
- Our solution is implemented in TensorFlow based on the starter code.
- It takes 3-5 days to train our frame-level models on a single GPU.
- Achieved 0.84198 GAP on the public 50% test data and 4th place in the challenge.
- Paper: https://arxiv.org/pdf/1707.00803.pdf
- Code & Documentation: https://github.com/forwchen/yt8m

Model learning overview



Models & Training

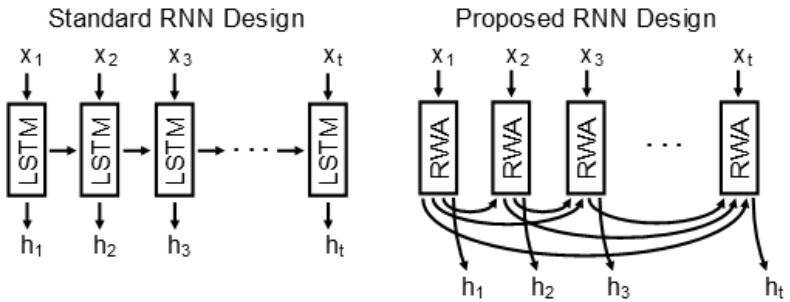
Model	Variations
LSTM	-
LSTM	Layer normalization & Recurrent dropout
RNN	Residual connections
GRU	-
GRU	Bi-directional
GRU	Recurrent dropout
GRU	Feature transformation
RWA	-
NetVLAD	-
DBoF	-
MoE	-

Training settings:

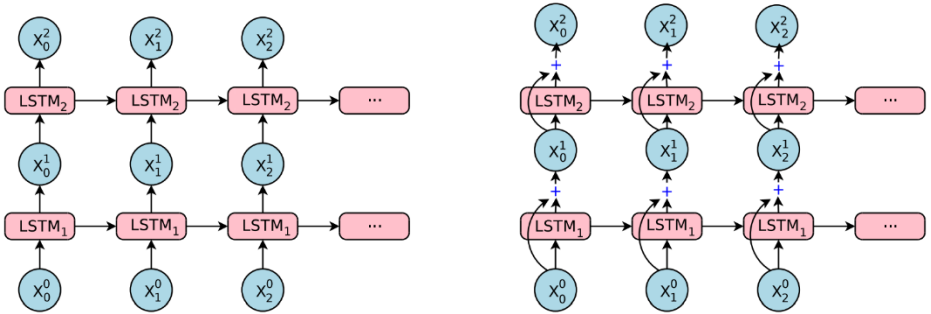
- Learning rate: 0.001, decays every epoch
- Batch size: 256 (RNNs), 1024(NetVLAD)
- Adam optimizer

DBoF & MoE: https://github.com/google/youtube-8m
Recurrent Weighted Average: https://github.com/jostmey/rwa
RNN with residual connections: https://github.com/NickShahML/tensorflow_with_latest_papers

Models details



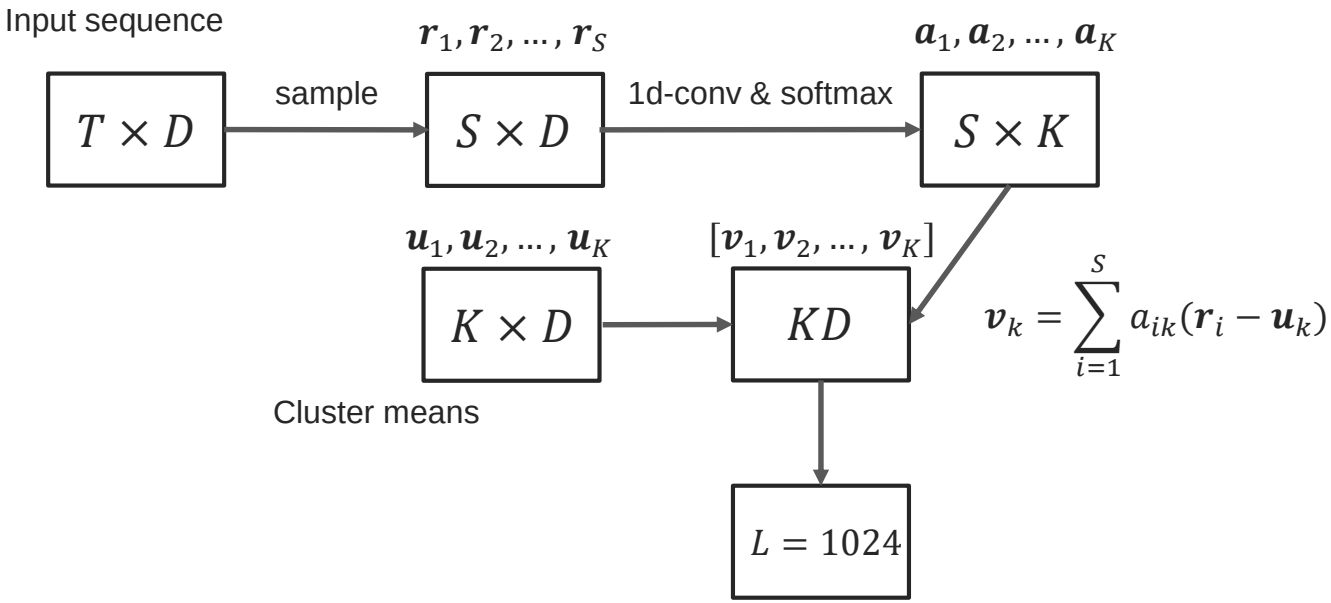
Structure of RWA (Recurrent Weighted Average) from [1]



Structure RNN with residual connections from [2]

[1] Ostmeyer, Jared, and Lindsay Cowell. "Machine Learning on Sequential Data Using a Recurrent Weighted Average." *arXiv 2017.*
[2] Wu, Yonghui, et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." *arXiv 2016.*

Models details



Arandjelovic, Relja, et al. "NetVLAD: CNN architecture for weakly supervised place recognition.“ CVPR 2016

Results (Single model)

Model	GAP@20
NetVLAD	0.79175
LSTM	0.80907
GRU	0.80688
RWA	0.79622
RNN-Residual	0.81039
GRU-Dropout	0.81118
LSTM-Layernorm	0.80390
GRU-Bidirectional	0.80665
GRU-feature-trans	0.78644

Results (Ensemble)

Ensembles	GAP@20
NetVLAD	0.80895
LSTM	0.81571
GRU	0.81786
RWA	0.81007
RNN-Residual	0.81510
GRU-Dropout	0.82523
Ensemble 1	0.83996
Ensemble 2	0.83481
Ensemble 3 (searching)	0.83581
Ensemble 4	**0.84198**

Fusion weights:

- Empirical, based on valid/test set performance
- Searching / learning over small split of valid set

Model ensembles: fusion of 3-5 checkpoint predictions
Ensemble 1: fusion of 9 model ensembles
Ensemble 2: fusion of another ~20 models with equal weights
Ensemble 3: fusion of the same models as 2, weights are obtained with searching
Ensemble 4: fusion Ensemble 1, 3 and others.

Q&A