Google Cloud and Youtube-8M Video Understanding Challenge

Hyun Sik Kim (hsik@stanford.edu) Ryan Wong (rawong@stanford.edu)

Stanford ENGINEERING
Electrical Engineering

Overview of project

Our project tackled the Youtube-8M challenge – the multi-label classification of videos in the Youtube-8M dataset.

Youtube-8M dataset

- 7 million videos with 4,716 different labels (avg. of 3.4 labels / video)
- Frame-level features extracted from the Inception network
- Video level features are a simple mean across frames

Total no. of videos	Over 7 million (70% train / 20% validation / 10% test)
Total no. of labels	4,716 (avg. of 3.4 per video)
Original video length	120-500 seconds
No. of encoded frames	Up to 360 frames / sec per video
Visual features	1,024 dimensional (8-bit each)
Audio features	128 dimensional (8-bit each)

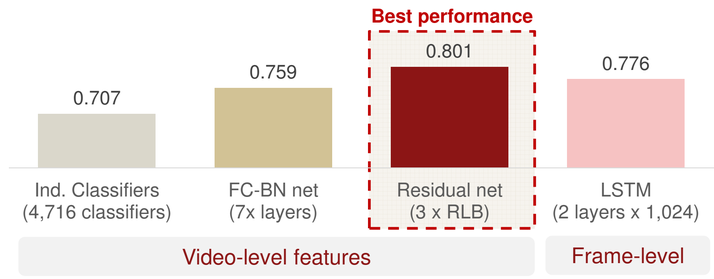
Summary of results

Performance on validation set

Model		Hit@1(1)	PERR(1)	mAP(1)
Independent classifiers (w/out audio)		0.789	0.646	0.376
Mixture of experts (MoE) (w/out / with audio)		0.728 / 0.772	0.562 / 0.611	0.110 / 0.125
FC-BN network	2 x layers	0.792 / 0.826	0.646 / 0.687	0.244 / 0.283
(w/out / with audio)	5 x layers	0.756 / 0.844	0.595 / 0.712	0.111 / 0.346
	7 x layers	0.772 / 0.807	0.613 / 0.653	0.125 / 0.144
Residual network (with audio)	**3 x RLB(2)**	**0.853**	**0.725**	**0.399**
LSTM (w/out audio)	2 layers x 1,024	0.841	0.708	-

(1) [x] / [y] in table entries where [x] denotes performance without audio features (just visual) and [y] denotes performance with visual+audio.
(2) Residual learning block (RLB). See architecture diagram below.

Global Average Precision (GAP) on test set

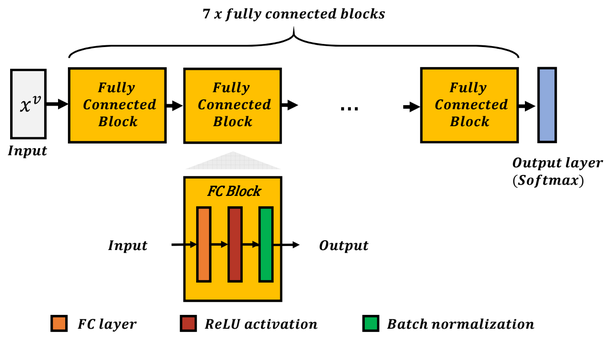


1. **Residual networks exhibited the best performance** – easiest to optimize and more robust to hyperparameters
2. **Temporal cues are difficult to learn** – well-designed video-level feature models can outperform frame-level ones (e.g. LSTM)
3. **Audio features materially improved performance**
4. **More sophisticated video-level aggregation of features might improve performance** – strong results based on simple mean of frames

FC-BN network (video-level features)

Multi-layer feed-forward network comprising of repeating fully-connected (FC) with ReLU and batch-normalization (BN) layers.

7 x layers network

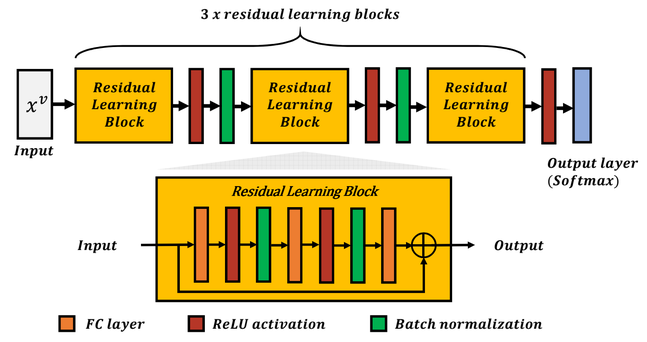


- ▲ Relatively easy to train
- ▲ Audio features materially improved performance
- ▼ Inability to learn temporal relationships between frames

Residual network (video-level features)

Multi-layer feed-forward network comprising of residual learning blocks that have FC with ReLU and BN layers.

3 x Residual Learning Block (RLB) network

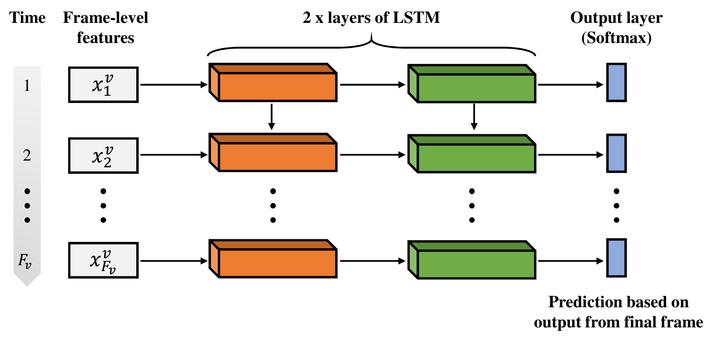


- ▲ Best GAP performance of 0.80
- ▲ Easier to optimize and more robust to hyperparameters
- ▼ Inability to learn temporal relationships between frames

LSTM (frame-level features)

Multi-layer LSTM network based on frame-level features.

2 layers x 1,024 units per LSTM cell



- ▲ Can learn temporal relationships and label dependencies
- ▼ Most computationally expensive to train
- ▼ Temporal cues difficult to learn – outperformed by residual network