Kaggle Competition
Google Cloud & YouTube-8M Video Understanding Challenge
5th place solution

Deep Learning Methods for Efficient Large Scale Video Labeling

M. Pękalski, X. Pan and M. Skalic

CVPR'17 Workshop on YouTube-8M Large-Scale Video Understanding
Honolulu, HI
July 26, 2017

Agenda

1. Team *You8M*
2. Models
3. Data Augmentation and Feature Enginnering
4. Training Methods
5. Key to Success

Team *You8M*

Marcin Pękalski: Master’s degree in Mathematics and Master’s degree in International Economics. Currently a data scientist at Kambi, a B2B sportsbook provider. *Stockholm, Sweden*.

Miha Skalic: Holds Master's degree in Biotechnology. Now working towards a PhD in Biomedicine at University Pompeu Fabra . *Barcelona, Spain*.

Xingguo E. Pan: Trained as a physicist. After getting a PhD, he went into the financial industry. *Chicago, USA*.

Frame level Models

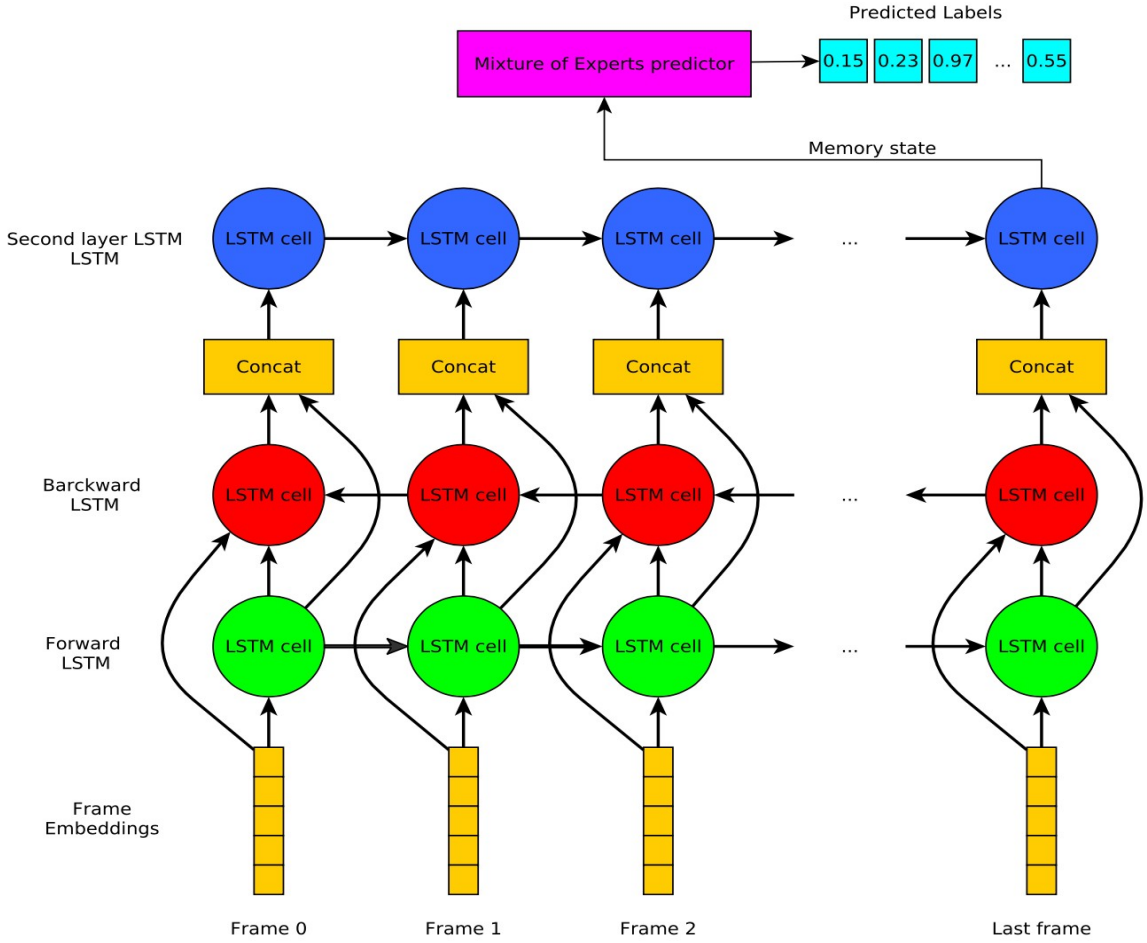
Bi-directional LSTM

- Dynamic length RNN
- Two models running in oposite directions
- MoE with two experts applied to last layer
- 6 epochs took 3 days

Bi-directional GRU

- Similar structure as LSTM
- Layer sizes 625x2, 1250
- Trained with 5 folds

Bi-directional LSTM model



MoNN3Lw

3 FC layers:

- 2305x8
- 2305x1
- 2305x3

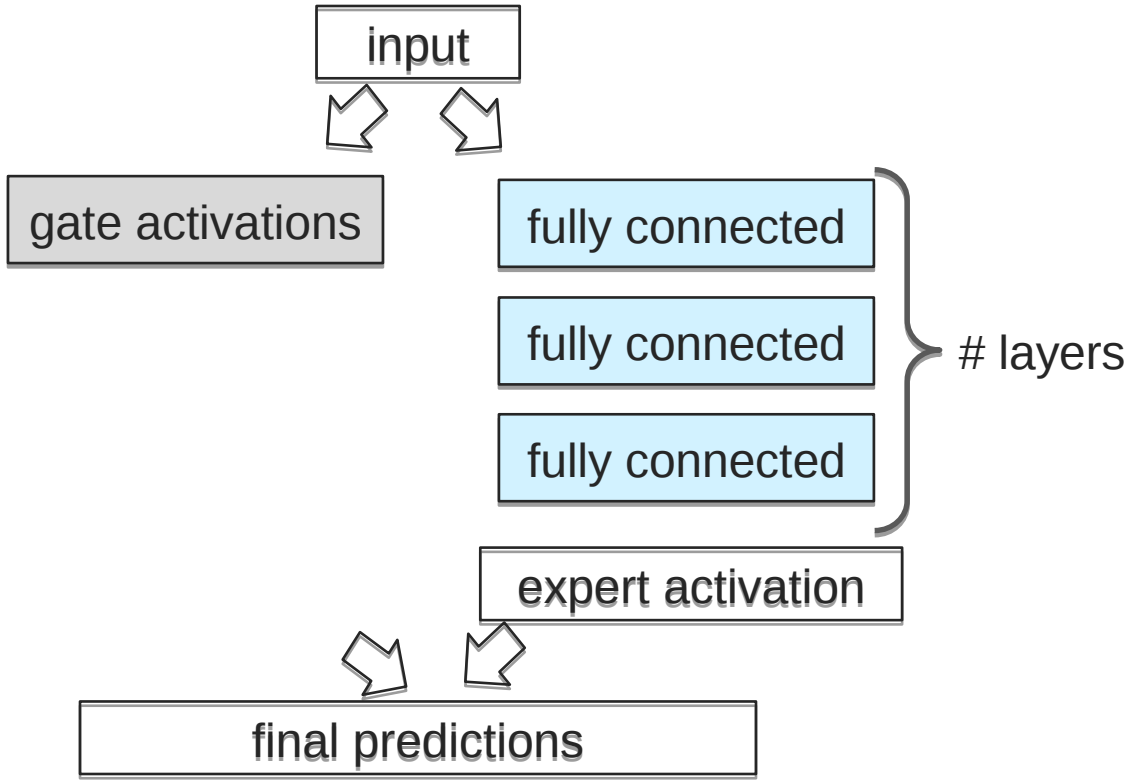
3 experts

Features

- mean_rgb/audio
- std_rgb/audio
- num_frames

5 epochs took 9 hours on GeForce GTX 1080ti

MoNN: Mixture of neural network experts



Models

	GAP single checkpoint single fold	GAP models of same family, checkpoints and folds ensemble
MoNN	0.823	0.835
LSTM	0.820	0.835
GRU	0.812	0.834

GAP All models ensemble: **0.8419** *Best score 0.8424*

- All models are implemented with Tensorflow.
- GAP scores are from private leaderboard.

Model Correlation

Models	MoNN	LSTM	GRU
MoNN	1.0	0.96	0.96
LSTM		1.0	0.98
GRU			1.0

Data augmentation
&
Feature
engineering

Video splitting:

- Split one video into two halves
- training samples: 6.3 million => 18.9 million

Video Level features:

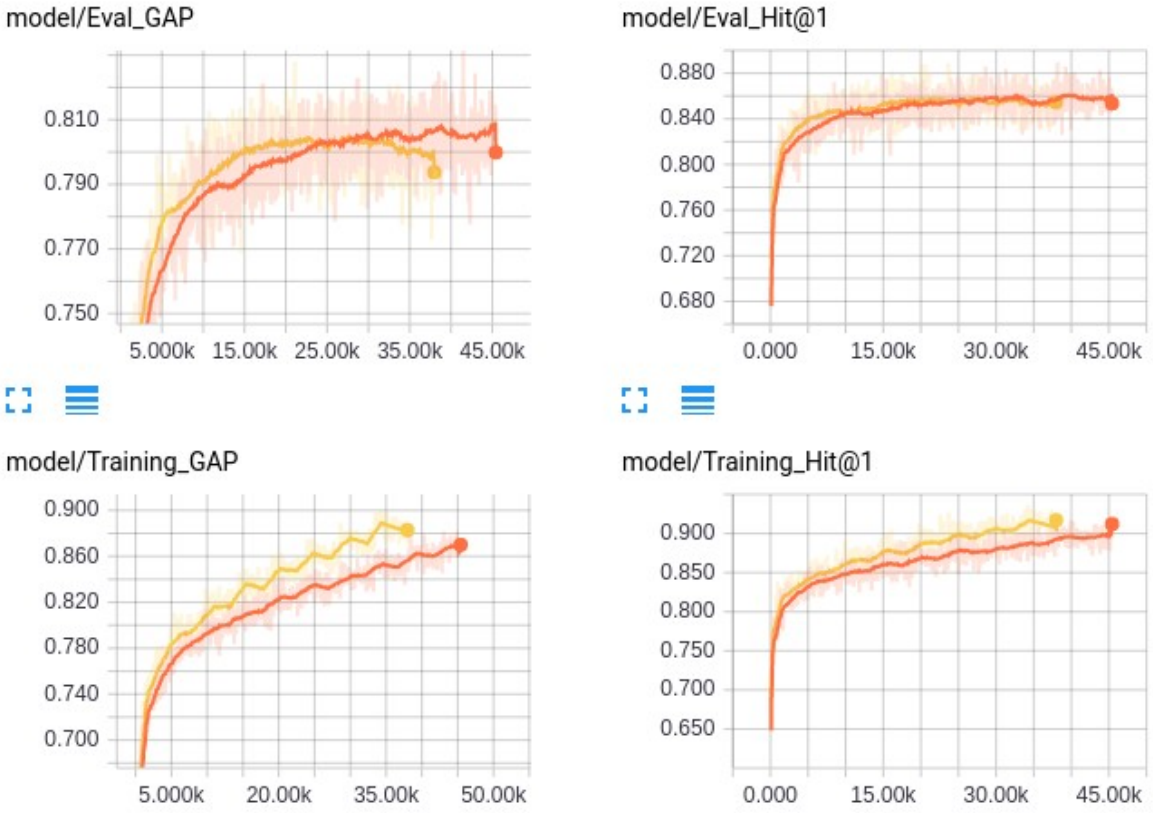
- mean-rgb/audio
- std-rgb/audio
- 3rd moments of rgb/audio
- num_frames
- Moments of entire video
- top/bottom 5 per feature dimension

Training

Being able to see the out-of-sample GAP greatly facilitated our management of the training process and of model and feature selection.

Example:
Both yellow and red models can well fit the train data, but yellow model peaked around 20 K steps on test data.

Monitor out-of-sample performance while training



Training

Training Methods that we experimented with and that helped to deepen our understanding of the data and models.

Dropout

Truncated Labels

Batch, Global and No Normalization

Exponential Moving Average

Training on Folds

Boosting Network

Key to Success

We tried a lot of **different architectures** (wide and narrow, deep and shallow, dropouts, ema, boosting, etc.).

Combined video and frame level models.

Tried **different ensemble weighting** techniques, utilizing individual model's GAP score and the correlation between models.

Generated **different features**: pre-assembled them into data for efficient training process.

Data augmentation.

Training on **folds of data** and averaging the results over folds and checkpoints.

Resources & Acknowledgements

arXiv Paper:

- https://arxiv.org/abs/1706.04572

Source Code:

- https://github.com/mpekalski/Y8M

Acknowledgements:

- The authors would like to thank the Computational Bio-physics group at University Pompeu Fabra for letting us use their GPU computational resources. We would also like to thank Jose Jimenez for valuable discussions and feedback.