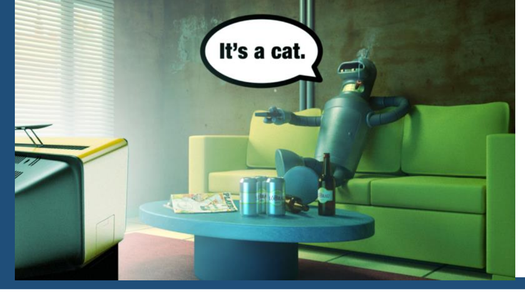
YouTube Video Classification

Alexandre Gauthier and Haiyu Lu



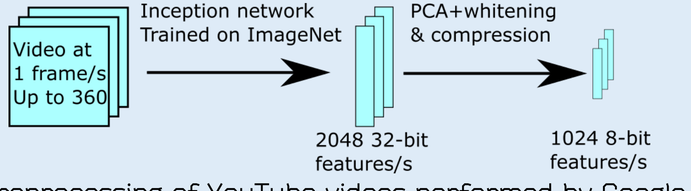
Background

- Classify YouTube videos into categories
- Categories can be used in video search
- Large dataset provided by Google
- **Difficulties:**
  Large dataset
  Lots of data in each video
  Combining spatial and temporal data

YouTube-8M[1] Dataset

- 8,264,650 videos
  70% training; 20% validation; 10% test

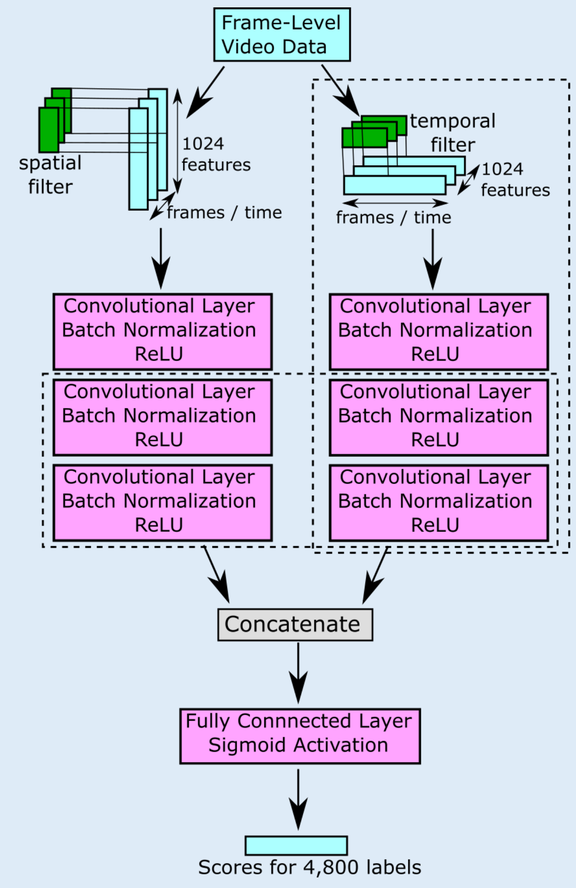
- **Features**: Individual frames preprocessed using network pretrained on ImageNet
  1 frame per second analyzed
  Visual data: 1024 features/frame
  Audio data: 128 features/frame
  Video-level data averaged over all frames



Preprocessing of YouTube videos performed by Google

- **Labels**: 4,800 classes
  Each video has an average of 1.8 labels
  Generated from human raters and automated curation

agau@stanford.edu, hylu@stanford.edu



Algorithm

- **Early-Fusion CNN[2]:**
  Convolves over frame-level features
  Filter depth = # of frames

- **Spatial-temporal CNN:**
  Two CNNs in parallel
  One convolves over features (Spatial)
  One convolves over frames (temporal)
  Concatenate outputs at end

Accuracy Metrics

- **Mean Average Precision (mAP):** Average area under precision-recall curve.
  $\text{precision} = \frac{tp}{tp+fp}$ $\text{recall} = \frac{tp}{tp+fn}$
  Higher mAP score is better
- **Hit@k**: Fraction of the test samples that contain at least one of the correct labels in the top k predictions
- **Precision at equal recall rate**: Predict the same number of labels per video as there are in the validation data
  PERR = the fraction you get correct

Results

	Modeling Approach	mAP	Hit@1	PERR
Google's models	Logistic	11.0	50.8	42.2
	Deep Bag of Frames	26.9	62.7	55.1
	LSTM	26.6	64.5	57.3
Our models	**Early-Fusion CNN**	**27.6**	**77.0**	**62.2**
	Spatial-temporal CNN	**29.7**	**77.4**	**63.1**

Convolutional models perform better
Spatial-temporal CNN performs best
Takes both feature and frame information into consideration

References

[1] arXiv:1609.08675v1 (Main Google paper)
[2] Karpathy, Andrej, et al. "Large-scale video classification with convolutional neural networks." *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2014.
Title Image: @joshuamaule and @surlyrightclick