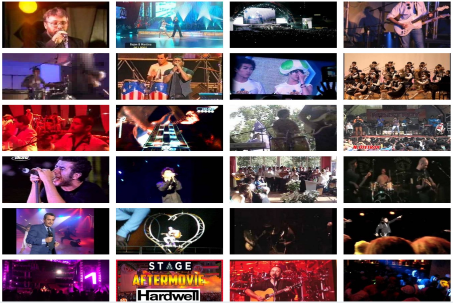
Predicting Video Tags Using Google's YouTube-8M Dataset

Tatyana Li

Michigan State University

Data

- **YouTube-8M V2**, largest multi-label classification dataset
- Kaggle competition
- ∼7 million YouTube videos (450,000 hrs of video)
- vocabulary of 4716 class labels (on average, 3.4 labels/per video)
- pre-extracted audio and visual features



Pre-processing

- train set ($\sim$5 mln.), validation set ($\sim$1 mln.), test set (700,650)
- visual frame-level features extracted using publicly available Inception network trained on ImageNet
- audio features extracted using VGG-inspired acousting model based on preliminary YouTube-8M version
- PCA (+ whitening) + quantization for dimensionality reduction
- total of 1024 video and 128 audio features

Metric

- Evaluation based on Average Global Precision (AGP), average precision across all predictions and all videos:

$$GAP = \sum_{i=1}^{N} p(i)\Delta r(i) \tag{0.1}$$

- where:
  N = number of final predictions for all videos
  p(i) is precision and
  Δr(i) is recall

Video-level models (Logistic Regression)

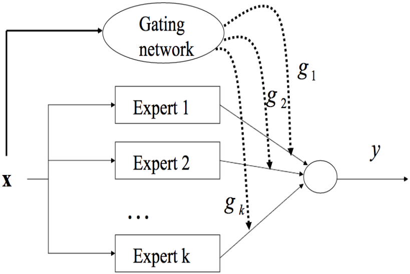
- Training on Google Cloud
- Independent binary logistic regression classifiers for each label, using L2-regularization (benchmark)
- Learn weights **w** by minimizing the total log-loss on the train set:

$$\lambda \|\mathbf{w}\|_{\mathbf{2}}^{\mathbf{2}} + \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathcal{L}(\mathbf{y}, \sigma(\mathbf{w}^{\mathbf{T}} \mathbf{x}_{\mathbf{i}})) \tag{0.2}$$

- **Average Global Precision (GAP) = 0.705**

Video-level models (Mixture of Experts)

- A mixture of logistic experts models (2 mixtures)
- Using L2-regularization for weights
- Gating network decides which experts to use, where $g_1 \ldots g_k$ are gating functions
- Used Softmax gating distribution



Marginal improvement: **Average Global Precision = 0.719**

Frame-level models (Logistic Regression) and RNN

- Independent one-versus-all logistic regression classifiers for each label, using L2-regularization
- **Low Global Average Precision (GAP) = 0.57**

- Recurrent Neural Network, LSTM architecture
- 2 stack LSTM layers, 1024 hidden units
- Long training process ∼11 hours
- **Improved Average Global Precision ∼0.80**

- To do ...

Thank you!