YouTube-8M:
A Large-Scale Video Classification

서울대학교 통계학과
김 용대, 김 사라
2017.05.04

YouTube-8M Dataset

- Frame-level data (total size of 1.71 TB)
  Each video has
  a. “video_id”: unique id for the video,
  b. “label”: list of labels of that video,
  c. Each frame has “rgb”: float array of length 1024,
  d. Each frame has “audio”: float array of length 128.

  A video v is given by a sequence of frame-level features (rgb) $x^v_{1:F_v}$,
  where x^v_j is the feature of the j-th frame from video v.

YouTube-8M Dataset

- Video-level data (total size of 31GB)
  Each video has
  a. “video_id”
  b. “labels”
  c. “mean_rgb”:float array of length 1028
  d. “mean_audio”: float array of length 128

Evaluation Metric

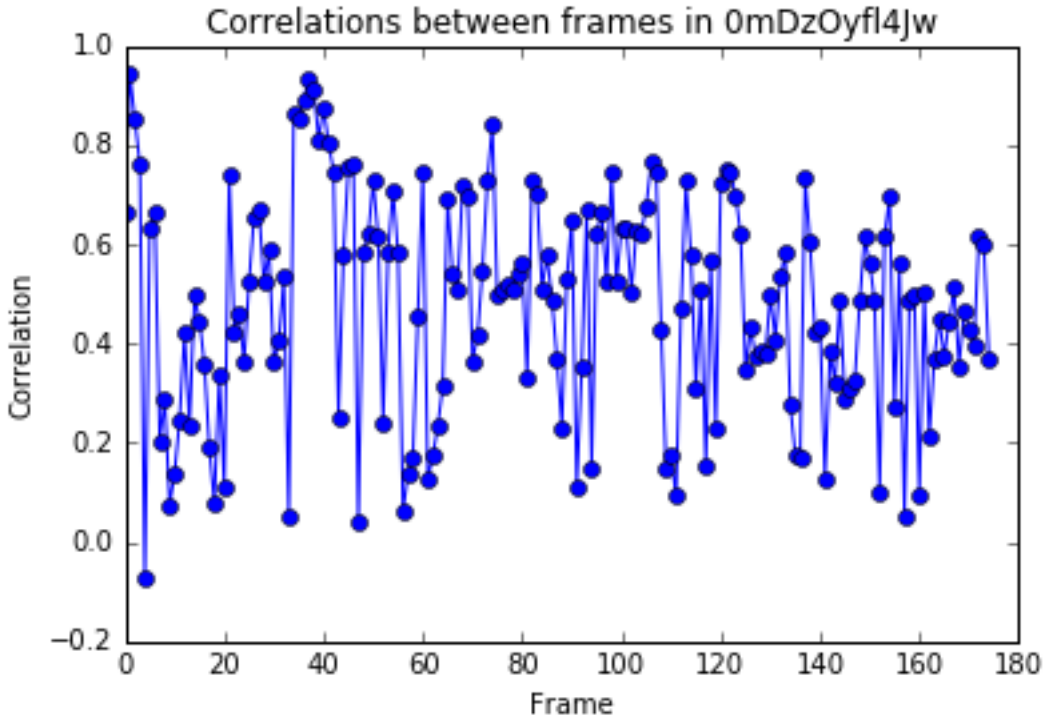
- **Global Average Precision (GAP) at k**

  For each video, we take the predicted labels that have the highest k confidence scores. Then we treats each prediction and score as an individual data point to have $k \times \#Videos$ predictions.
  If $k \times \#Videos$ predictions sorted by its confidence score, we compute the precision, recall, and GAP:

$$GAP = \sum_{i}^{k \times \#Videos} p(i)\Delta r(i).$$

Correlation plot

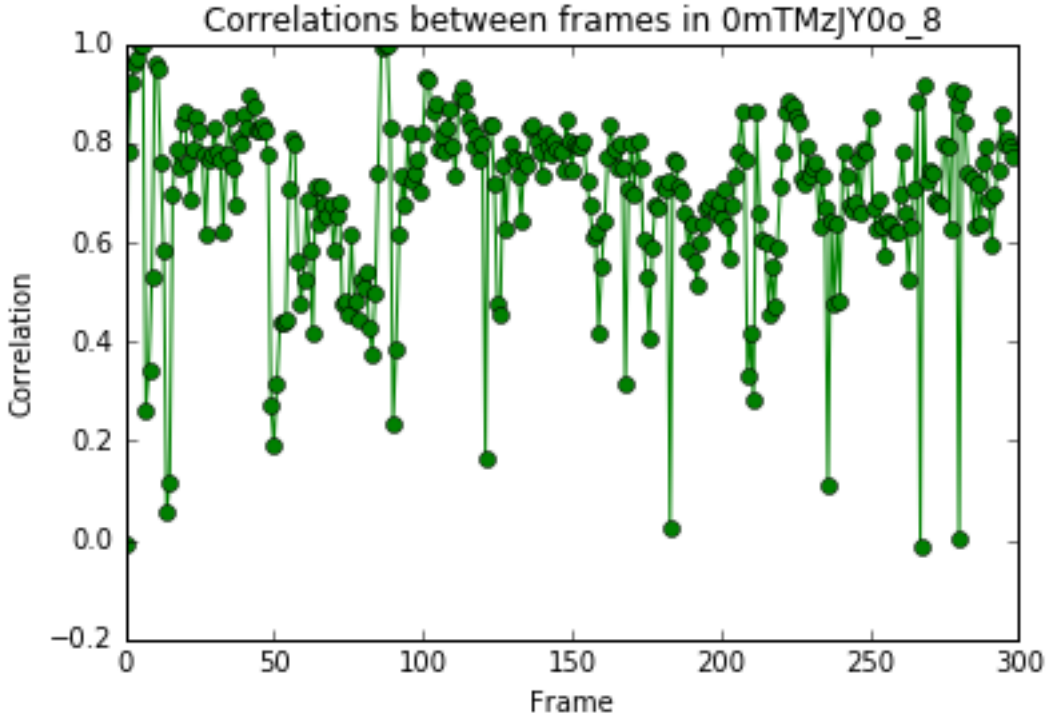
- Frame-level data 의 크기를 줄이기 위해 scene detection 을 사용.
- 한 비디오의 frame-level data 에서 연속된 두 개의 프레임 feature vector의 correlation 을 구하여 scene detection 을 진행.



Video id: 0mDzOyfl4Jw
176초 길이의 비디오 파일내의 correlation plot.
여기서 correlation 값이 0.3이하인 개수는 38개.

Correlation plot

- Frame-level data 의 크기를 줄이기 위해 scene detection 을 사용.
- 한 비디오의 frame-level data 에서 연속된 두 개의 프레임 feature vector의 correlation 을 구하여 scene detection 을 진행.



Video id: 0mTMzJY0o_8
300초 길이의 비디오 파일내의 correlation plot.
여기서 correlation 값이 0.3이하인 개수는 13개.

Correlation plot

- (추가 논의 필요)
  이전과 같이 일정한 기준 이하의 correlation 값을 가지는 부분에서 scene이 변하였다고 생각할 수 있기 때문에, 변하는 부분들을 잡아내서 구간 내의 feature vector 들을 각 성분 마다 평균을 취한 mean feature vectors 을 사용하여 RNN 을 training.

Multi-label classification

- 현재 진행 부분.

  1. Video label data 에서 class 을 선택:
     총 4716개의 class 들의 빈도수를 세어, 빈도 수가 높은 상위 100개의 class 을 선택.

  2. Video-level data 선택:
     1번에서 선택된 100개의 class 들 중 한 개라도 포함된 video 및 video-level data 을 이용하여 모형 적합.

  3. class 수가 4716개인 분류문제를 class 수가 100개인 분류문제로 변환.
     여기서 1번에서 선택된 class 에 포함되지 않는 class은 사용하지 않음.

Multi-label classification

- 현재 진행 부분.

4. 사용된 모형:
   a. Logistic model (100)
   b. 2 Fully connected layers (512, 256) + Dropout(0.5)
      + logistic (100)

   - optimizer: sgd
   - hidden layer activation: relu
   - output layer activation: sigmoid
   - loss: binary_crossentropy
   - metrics: top_k_categorical_accuracy

Multi-label classification

- 결과

Model	Hit 1 ratio	GAP
Logit	0.6906	0.7791
Deep logit	0.6690	0.7792

Discussion

1. Frame-level 의 scene detection 을 이용하여 RNN 을 이용하여 training.

2. Label 사이의 dependency 가 있는 경우에서 Multi-label classification 접근 방법 (ex. Mixture of experts) 및 loss function 의 결정.

3. 빈도 수가 낮은 class 들의 처리.