Google Cloud & YouTube-8M Video Understanding Challenge

- Starter Code -

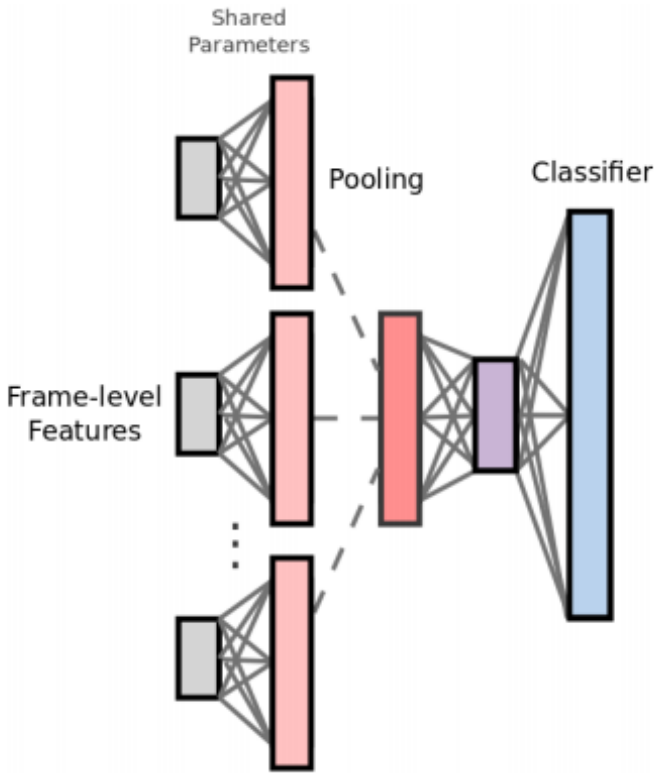
김성현

서울대학교 통계학과

2017년 5월 11일

Deep Bag of Frame (DBoF)

- 30개의 프레임, cluster size = 8192, hidden size = 1024.



Mixture of expert(Moe)

- $f_i(e|x)$: i번 모형의 entity e에 대한 예측값. 로지스틱 모형 사용.

$$f_i(e|x) = \sigma(\beta_{ei}^T x), \quad i = 1, ..., K$$

- $g_e(i|x)$: entity e에서 i번 모형에 대한 가중치.

$$g_e(i|x) = \frac{exp(w_{ei}^T x)}{\sum_{j=1}^{K+1} exp(w_{ej}^T x)}, \quad i = 1, ..., K+1$$

-

$$h(e|x) = \sum_{i=1}^{K} f_i(e|x) g_e(i|x)$$

train

- 모든 모형에 대해 일부분의 데이터를 사용해 작동하는지 시험.
- 적당한 batch size에서 구글 클라우드와 개인 컴퓨터에서 학습 가능.
- batch size와 gpu에 따라 학습 속도가 변함.
- 구글 클라우드에서는 tesla K80 8개까지 사용 가능.
- Logistic 모형과 DBOF + Logistic 모형은 어느 정도 학습이 완료.

train:Logistic

- Logistic 모형
- 8gpu 사용하여 50분 동안 5 epoch 학습.
- 4096개로 나누어진 validation 데이터파일 중 186개 평가.
  MAP: 0.299 | Avg_Hit@1: 0.780 | Avg_PERR: 0.633 |
  GAP: 0.693

Input Features	Modeling Approach	mAP	Hit@1	PERR
Frame-level, $\{\mathbf{x}^v_{1:F_v}\}$	Logistic + Average (4.1.1)	11.0	50.8	42.2
Frame-level, $\{\mathbf{x}^v_{1:F_v}\}$	Deep Bag of Frames (4.1.2)	26.9	62.7	55.1
Frame-level, $\{\mathbf{x}^v_{1:F_v}\}$	LSTM (4.1.3)	26.6	**64.5**	**57.3**
Video-level, μ	Hinge loss (4.3)	17.0	56.3	47.9
Video-level, μ	Logistic Regression (4.3)	28.1	60.5	53.0
Video-level, μ	Mixture-of-2-Experts (4.3)	29.6	62.3	54.9
Video-level, $[\mu; \sigma; \mathrm{Top}_5]$	Mixture-of-2-Experts (4.3)	**30.0**	63.3	55.8

train:DBOF

- DBOF + Logistic 모형
- 1gpu 사용하여 14시간 동안 3 epoch 학습.
- 4096개로 나누어진 validation 데이터파일 중 186개 평가.
  MAP: 0.350 | Avg_Hit@1: 0.813 | Avg_PERR: 0.670 |
  GAP: 0.741

Input Features	Modeling Approach	mAP	Hit@1	PERR
Frame-level, $\{\mathbf{x}^v_{1:F_v}\}$	Logistic + Average (4.1.1)	11.0	50.8	42.2
Frame-level, $\{\mathbf{x}^v_{1:F_v}\}$	Deep Bag of Frames (4.1.2)	26.9	62.7	55.1
Frame-level, $\{\mathbf{x}^v_{1:F_v}\}$	LSTM (4.1.3)	26.6	**64.5**	**57.3**
Video-level, μ	Hinge loss (4.3)	17.0	56.3	47.9
Video-level, μ	Logistic Regression (4.3)	28.1	60.5	53.0
Video-level, μ	Mixture-of-2-Experts (4.3)	29.6	62.3	54.9
Video-level, $[\mu; \sigma; \mathrm{Top}_5]$	Mixture-of-2-Experts (4.3)	**30.0**	63.3	55.8

train:submission

모형	GAP
(1st) WILLOW	0.84493
(benchmark) mean_rgb + mean_audio + Logistic	0.74711
(제출) rgb + DBOF + Logistic	0.71214
(제출) mean_rgb + Moe	0.70893
(benchmark) mean_rgb + Logistic	0.69360
(제출) mean_rgb + Logistic	0.69170

train:LSTM

- batch size 128, 1gpu일 때 초당 40개, 8gpu일 때, 초당 200개의 데이터 학습.
- 8gpu일 때, 1 epoch당 대략 7∼8시간 걸릴 것으로 예상됨.
- 개인 컴퓨터에서 cpu로 초당 0.5개의 데이터 학습.