Large Scale video classification using YouTube-8M

Emma An (anran), Anqi Ji (anqi), Edward Ng (edjng) — *Stanford University*

Motivation

- The amount and themes of video we deal with are immense
  On YouTube: **300** h uploaded/min, **5** B watched/day
- YouTube-8M dataset: 7 M video URLs, 3.2 B features, 4716 labels
- Goal: achieve higher accuracy rate using audio and visual features
- Input: frame-level visual and audio features (1 frame/sec)
- Output: multiple labels summarizing the key topics of the video

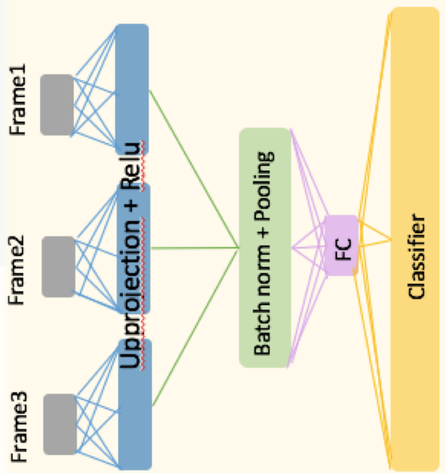
Methods

Traditional video classification limitation

- 1. Only deal with visual features
- 2. Long runtime due to redundant temporal info
- 3. Dataset is topic specific

We are using a combination of frame-level audio and visual features and explore different ways to achieve a balance in them on the YouTube-8M, largest dataset ever.

Deep Bag of Frames (DBoF)

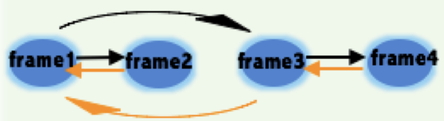


Gated Recurrent Unit (GRU)

Adaptive shortcut connections between different frames

$$f(h_{t-1}, x_t) = u_t \odot h_t + (1 - u_t) \odot h_{t-1}$$

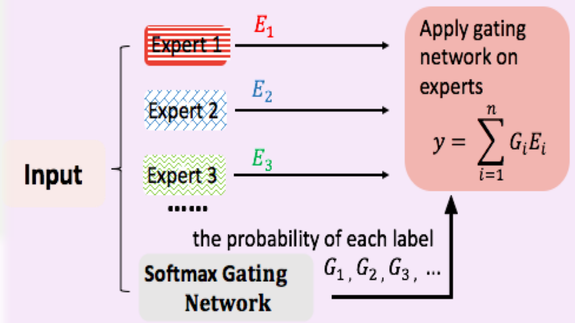
Candidate Update $h_t = \tanh(W[x_t] + U(r_t \odot h_{t-1}) + b)$
Reset gate $r_t = \sigma(W_r[x_t] + U_r h_{t-1} + b_r)$
Update gate $u_t = \sigma(W_u[x_t] + U_u h_{t-1} + b_u)$



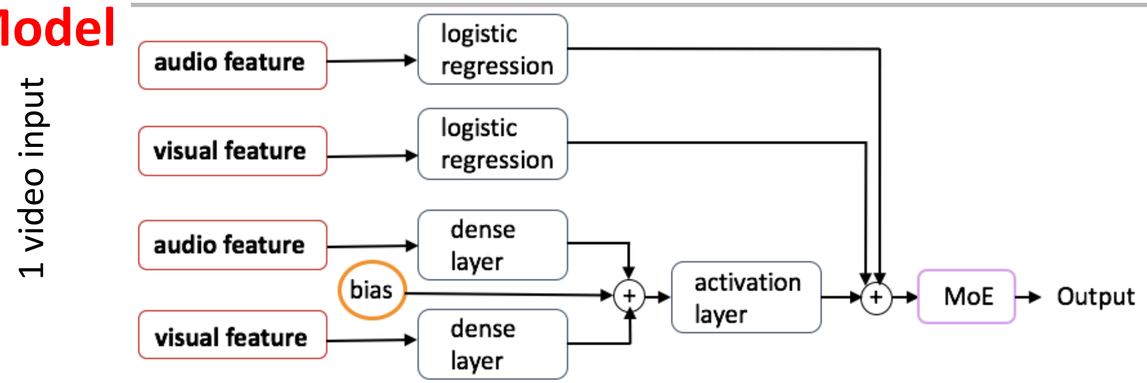
Execution:
1. Select a readable subset r
2. Read the subset r⊙ h
3. Select a writable subset u
4. Update the subset h

Model of Experts (MoE)

Each expert is a CNN and it outputs a label



Model



- We trained our model using Tensorflow on Google cloud using GPU
- Average training runtime is 1 to 3 hours for frame-level models

Results

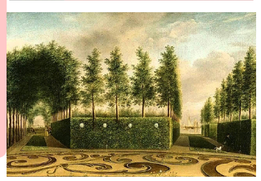
Good cases

Vehicle, Volkswagen passenger cars, car, Volkswagen, tire
vehicle
car
sedan
tire
wheel

lip gloss, lipstick
cosmetics
lipstick
lip gloss
fashion
rouge

Bad cases

ketchup, food, sauce, tomato
food
games
video games
animal
newscaster



harpsichord
games
vehicle
video game
dance
car

Metric

Hit@k: the fraction of test samples that contained at least one of the ground truth labels in the top k predictions
PERR (precision at equal recall rate): precision of scoring labels among all ground-truth labels
MAP: mean average precision
GAP (global average precision): official metric for Kaggle Challenge: area under the precision/recall curve

Result

	Avg_Hit@1	Avg_PERR	MAP	GAP
logistic (video-only)	0.788	0.646	0.646	0.707
logistic (audio-only)	0.565	0.431	0.089	0.429
Dense (audio + video)	0.836	0.703	0.387	0.775
MoE (audio + video)	*0.84*	*0.709*	*0.415*	*0.782*
LSTM (video-only)	0.645	0.573	0.266	N/A

Conclusion

1. Video-labels maintain reasonable level of performance compared to frame-level models
2. Classifying user-generated content is noisy, both in labeling and input
3. Multi-modal models greatly improves classification