The YouTube-8M Kaggle Competition:
Challenges and Methods

Haosheng Zou*, Kun Xu*, Jialian Li, Jun Zhu
Presented by: Yinpeng Dong
All from Tsinghua University
2017.7.26
Contents

- Introduction & Definition
- Challenges
- Our Methods & Results
- Other Methods
Introduction

- GAP evaluation

\[ GAP = \frac{20N}{\sum_{i=1}^{M} \frac{p(i)}{i} \cdot \frac{1}{M}} \]

- GAP = 1.0 ⇔

  ![Diagram](3.4 labels / video on average)

- Low confidence predictions should be suppressed enough (3.4 labels / video on average).
Problem Definition

- We focus on exploiting frame-level features.
- 4716 binary classification tasks.

- Input: $\{v_1, v_2, ..., v_T\}, \{a_1, a_2, ..., a_T\}$
- Output: Probability of labelling $e_1, e_2, ..., e_{4716}$.

- Rough model:
- Frame understanding block: fixed-length descriptor $x_{video}$
- Classifiers block: 4716 binary classifications
Challenges

1. Dataset Scale
2. Noisy Labels
3. Lack of Supervision
4. Temporal Dependencies
5. Multi-modal Learning
6. Multiple Labels
7. In-class Imbalance
Challenges (cont.)

1. Dataset Scale:
   - 5M (or 6M) training videos, 225 frames / video, 1024 (+128) dimension features / frame.
   - Disk I/O in each mini-batch.
   - Validation takes several (~10) hours.
   - Downsample; smaller validation set; …

2. Noisy Labels:
   - Rule-based annotated labels, not crowdsourcing
   - 14.5% recall w.r.t. crowdsourcing, positive→negative
   - Negative dominates; learning the annotation system
   - Ensemble; more randomness; …
Challenges (cont.)

3. Lack of Supervision:
   - No information about each frame.
   - Only video-level supervision for the whole model.
   - Attention; auto-encoders; …

4. Temporal Dependencies:
   - Features haven’t yet taken into account.
   - Humans can still understand videos at 1 fps.
   - RNNs; clustering-based models (e.g. VLAD); …
Challenges (cont.)

5. Multi-modal Learning:
   - “every label in the dataset should be distinguishable using visual information alone”
   - Audio features do help.
- Different fusion techniques.

6. Multiple Labels:
   - Uniquely extracted $x_{video}$ should be incredibly descriptive for 4716 binary classification tasks.
   - Labels all usually present or not in groups. Implicit correlation from a shared frame understanding block may not be sufficient.
Challenges (cont.)

7. In-class Imbalance:
   - 5M training videos
     - > 500K positive: 3 labels
     - > 100K positive: < 400 labels
     - Hundreds of positive: ~ 1000 labels
   - Imbalance ratio $\frac{100K}{5M} = \frac{1}{50}$ for 90% binary classification
   - Loss manipulation; specific techniques; ...

Our Methods, High-Level

- Random cropping: Take 1 frame every 5 frames
  - Rougher temporal dependencies
  - Only the start index is randomized

- Multi-Crop Ensemble:
  - One model, varying the start index
  - Uniformly averaging

- Early Stopping:
  - Fix 5 epochs of training at most
  - Train directly on training and validation sets.
Our Methods, Model

- Prototype: stacked LSTM (1024-1024) + LR / 2MoE
  - 4716 binary classifiers
  - Stacked (Bi)LSTM (Late Fusion) (Layer Normalization) etc.
- Layer Normalization
- Late Fusion
Our Methods (cont.)

- **Attention**

- **Bidirectional LSTM**
Our Results

| Model                              | Public   | Private  |
|------------------------------------|----------|----------|
| baseline (on Kaggle)               | 0.74711  | 0.74714  |
| prototype (full, visual only)      | 0.78105  | 0.78143  |
| prototype (full)                   | 0.80224  | 0.80207  |
| prototype (crop)                   | 0.80204  | 0.80190  |
| BiLSTM+LR+LN                       | 0.80761  | 0.80736  |
| BiLSTM+MoE                         | 0.81055  | 0.81067  |
| BiLSTM+MoE+attention               | 0.81232  | 0.81227  |
| BiLSTM+MoE (full)                  | 0.81401  | 0.81399  |
| ENSEMBLE (16)                      | 0.83477  | 0.83470  |
| ENSEMBLE (36)                      | **0.83670** | **0.83662** |
Other Methods

- Separating Tasks
  - Different frame understanding block, thus different video descriptor for each meta-task
  - 25 verticals as meta-tasks, too slow (15 examples / s)

- Loss Manipulation
  - Ignore negative labels when predicted confidence < 0.15

- Unsupervised Representation Learning
  - Using visual to reconstruct both visual and audio features
Conclusion

1. Dataset Scale
2. Noisy Labels
3. Lack of Supervision
4. Temporal Dependencies
5. Multi-modal Learning
6. Multiple Labels
7. In-class Imbalance
Thank you!
Q & A