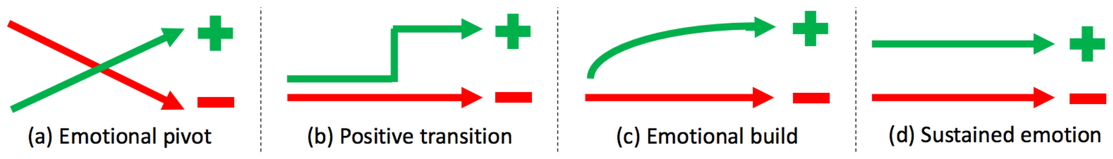
Story Understanding in Video Advertisements

Keren Ye, Kyle Buettner, Adriana Kovashka
Department of Computer Science, University of Pittsburgh

BMVC 2018 Newcastle upon Tyne

Introduction

- Creative narrative techniques are used in advertisements, which inspire our project.
  Freytag's pyramid:
  exposition, rising action, ***climax****, denouement*
  Four archetypes of dramatic structure (*Young 2008*):



- We crowdsource climax annotations on 1,149 videos.
- We develop unsupervised and supervised methods to predict the climax.
- We build a ***sentiment*** prediction model based on the predicted climax and other semantically meaningful features.

Dataset

- We base our work on the PITT video ads dataset (***Hussain et al., CVPR 2017***).
  We use both the topic and sentiment annotations.

Topic	***17,345***	***Sentiment***	***17,345***	...

- We gather additional climax annotations for 1,149 videos.
  Q: When does the climax occur?

  A: 1 min 20 secs
  We choose workers with at least 98% approval rate.
  Each video is annotated by four workers.
  We end up with 1,149 videos.

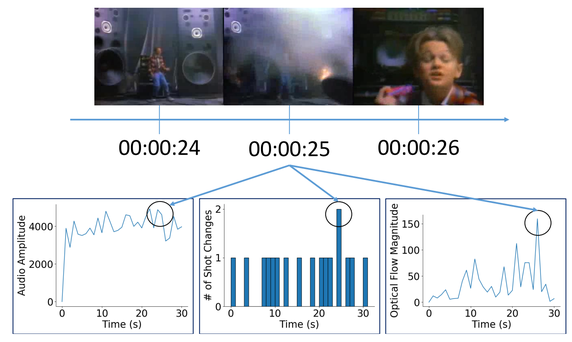
Method

- ***Feature indicators*** (processed one frame per second)

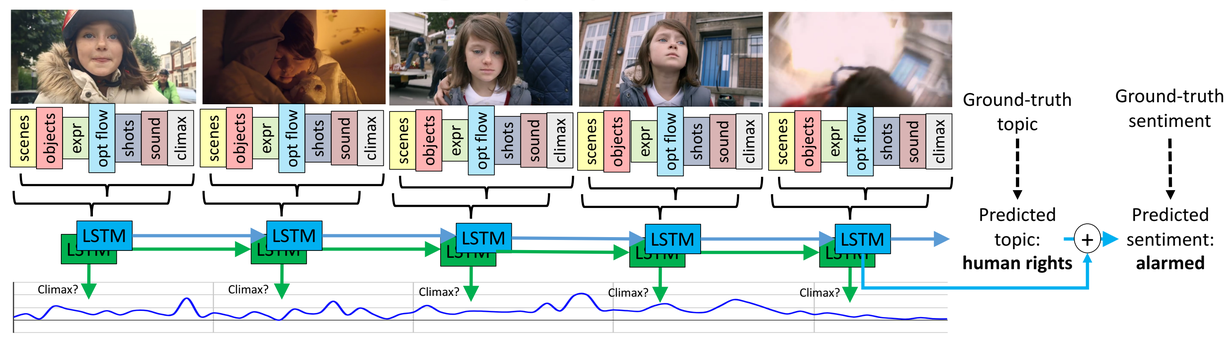
sound | shots | opt flow | scenes | objects | expr | topic

✓ Audio amplitude ($\boldsymbol{a}^k$), max amplitude of audio in the k-th frame.
✓ Shot boundary indicator ($\boldsymbol{b}^k$), 0 or 1 depending on whether a shot boundary occurs in the frame
✓ Optical flow magnitude ($\boldsymbol{o}^k$), a scalar value denoting average optical flow magnitude in the k-th frame (using the implementation of *Ranjan et al., 2017*)
✓ Type of place / scene ($\boldsymbol{pl}^k$), a 365-way scene distribution predicted by Place365 model (*Zhou et al., 2017*)
✓ Object ($\boldsymbol{ob}^k$), a 80-D vector capturing the presence of COCO objects, aggregated by max-pooling of the detection result (using *Huang et al., 2017*)
✓ Facial expressions ($\boldsymbol{fa}^k$), 8 facial expressions and valence-arousal values predicted by a deep model trained on AffectNet data (*Mollahosseini et al., 2017*). We use the OpenFace (*Amos et al., 2016*) to detect faces in the frame.
✓ Topic, a 38-D video level topic distribution distilled using topic annotation of PITT video ads dataset

- ***Unsupervised climax prediction***: we directly use audio amplitude, shot boundary frequency, and optical flow magnitude to predict climax.



- ***Supervised climax prediction***: we predict climax using an LSTM (64 hidden units) that outputs 0/1 for each frame.
- ***Supervised sentiment prediction***: we use an LSTM (64 hidden units) to encode the frames.
  ✓ We also add the predicted climax (1D) as extra information.
  ✓ Multi-task learning: topic prediction (38D)



Evaluation

- ***Climax prediction***
  ✓ Accuracy: we treat the prediction as correct if ground-truth climax is close (within 0, 1, 2 sec)

	top-1 prediction			top-3 prediction		
Method	w/in 0 s	w/in 1 s	w/in 2 s	w/in 0 s	w/in 1 s	w/in 2 s
heuristic-guess	0.031	0.083	0.121	0.122	0.299	0.430
shot boundary (unsup)	0.068	0.179	0.265	*0.221*	**0.457**	**0.588**
optical flow (unsup)	0.064	0.152	0.220	0.163	0.380	0.513
audio (unsup)	**0.077**	0.171	0.255	0.178	0.403	0.534
LSTM, ResNet only	0.071	*0.206*	**0.290**	0.190	0.400	0.523
LSTM, all feats (Ours)	**0.077**	**0.209**	*0.287*	**0.226**	*0.439*	*0.546*

- ***Sentiment prediction***
  ✓ Mean Average Precision (mAP): average precision over evenly spaced recall levels
  ✓ Accuracy@1: fraction of correct top-1 predictions
  ✓ Agree with k: assign a ground-truth label to a video only if at least k annotators agree on the sentiment

	Agree with 1		Agree with 2		Agree with 3	
Method	mAP	acc@1	mAP	acc@1	mAP	acc@1
Hussain et al.	0.283	0.664	0.135	0.435	0.075	**0.243**
Our model	**0.313**	**0.712**	**0.160**	**0.449**	**0.094**	0.241

- ***Ablation studies***
  ✓ We show improvement over the baseline that uses only the CNN feature
  ✓ Each experiment uses the CNN feature combined with a specific feature indicator

Improvement	creative	educated	alarmed	fashionable	emotional	angry
sound	96.0%	110.5%	106.4%	122.7%	85.7%	165.6%
shots	94.5%	99.8%	108.2%	140.9%	84.8%	61.4%
opt flow	99.5%	106.3%	113.1%	120.5%	77.8%	112.9%
scenes	83.2%	122.0%	121.8%	140.4%	80.7%	84.6%
objects	94.1%	107.3%	114.3%	127.2%	102.1%	67.2%
expr	92.2%	104.6%	112.8%	119.5%	129.4%	87.7%
topic	121.3%	110.9%	110.9%	158.8%	77.2%	148.5%

Acknowledgement

NSF
Grant Nr 1566270

Google
Faculty Research Award

NVIDIA hardware grant