Exploring LSTMs for Video Analysis

Michael N. Lombardo, Michael.Lombardo@uoit.net
Faculty of Science, University of Ontario Institute of Technology
Supervisor: Faisal Z. Qureshi

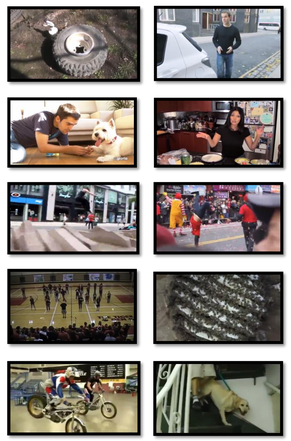
vclab.ca

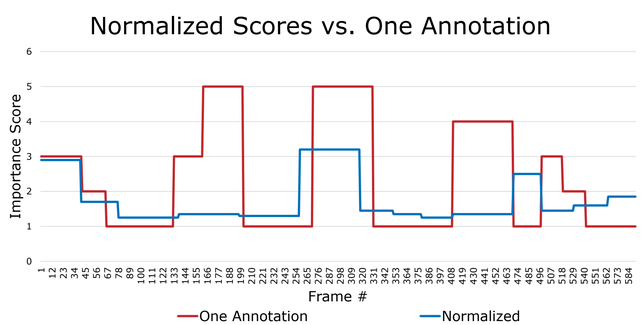
Introduction

- Explore the use of recurrent neural networks to capture spatio-temporal structure of a video with a view to identify "interesting" video segments.
- Use data containing hand-labeled videos to train a long short-term memory network for predicting 1) video segment boundaries and 2) video frame interestingness scores.

Dataset: TVSum50

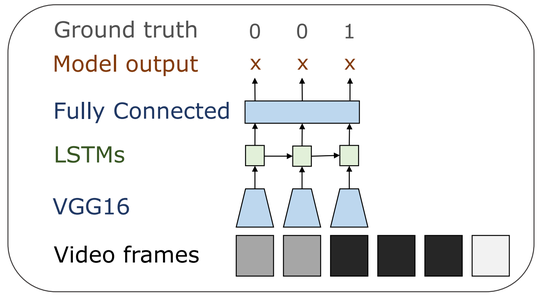
- 50 videos (collected from YouTube)
- 10 categories, 5 videos per category
  - Changing vehicle tire
  - Getting vehicle unstuck
  - Grooming an animal
  - Making sandwich
  - Parkour
  - Parade
  - Flash mob gathering
  - Bee keeping
  - Attempting bike tricks
  - Dog show
- Each video is annotated by 20 users
- Users are asked to assign a value between 1 and 5 to each frame





1. Boundary Detection

- **Problem**: classify whether or not a boundary occurs at current frame given the information in the previous two frames and the current frame
- Video length = 9 frames
- LSTM steps = 3
- LSTM hidden state = 256

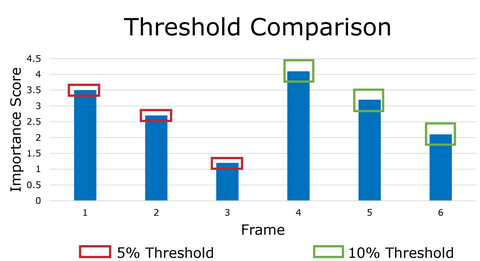


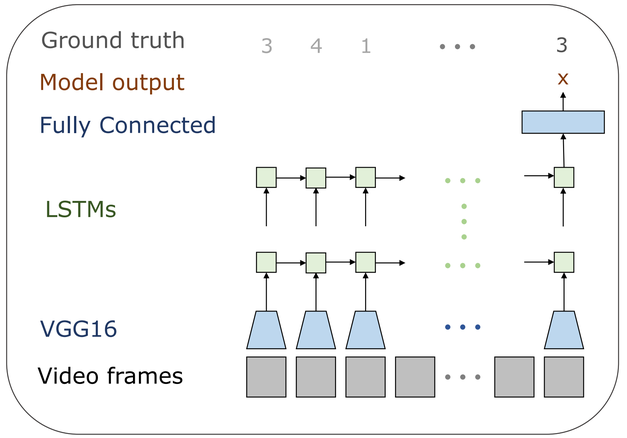
Results

Epoch	Accuracy	Epoch	Accuracy
0	12.85%	100	99.79%
10	93.40%	150	99.80%
25	97.22%	200	99.86%
50	99.73%	250	99.88%

2. Computing Frame Interestingness Scores

- **Problem**: use regression to assign a score between 1 and 5 to each frame
- Total number of frames = 352,000
  - Subsampling: every 5th frame was selected
- LSTM steps:
  - 16
  - 32
  - 64
- Architectures (layer/hidden states)
  - 1/256
  - 1/512
  - 2/256
  - 2/512





Results

Architecture	Average Accuracy
A – 5%	32.20%
A – 10%	75.70%
B – 10%	64.97%
C – 10%	52.58%
D – 10%	75.40%

Experiments used a batch size of 256, sequence length of 16, and various thresholds of 5% & 10%.

Sequence Length	Prediction Threshold	Average Accuracy
16	5%	32.20%
16	10%	75.70%
32	5%	30.26%

Sequence Length	Prediction Threshold	Average Accuracy
32	10%	46.03%
64	5%	30.41%
64	10%	53.58%

Experiments used a batch size of 256, each training/testing split included testing on different categories.

Acknowledgements

A special thank you to the University of Ontario Institute of Technology for allowing me to complete my Honour's Thesis, and the supervision of Faisal Qureshi.

Future Directions

- Explore other neural networks to add more layers to the Classification process for the importance scores of frames.
- Determine the representativeness, and uniformity of each frame in a given video.