Predicting Media Interestingness

Deep Learning for Multimedia Processing

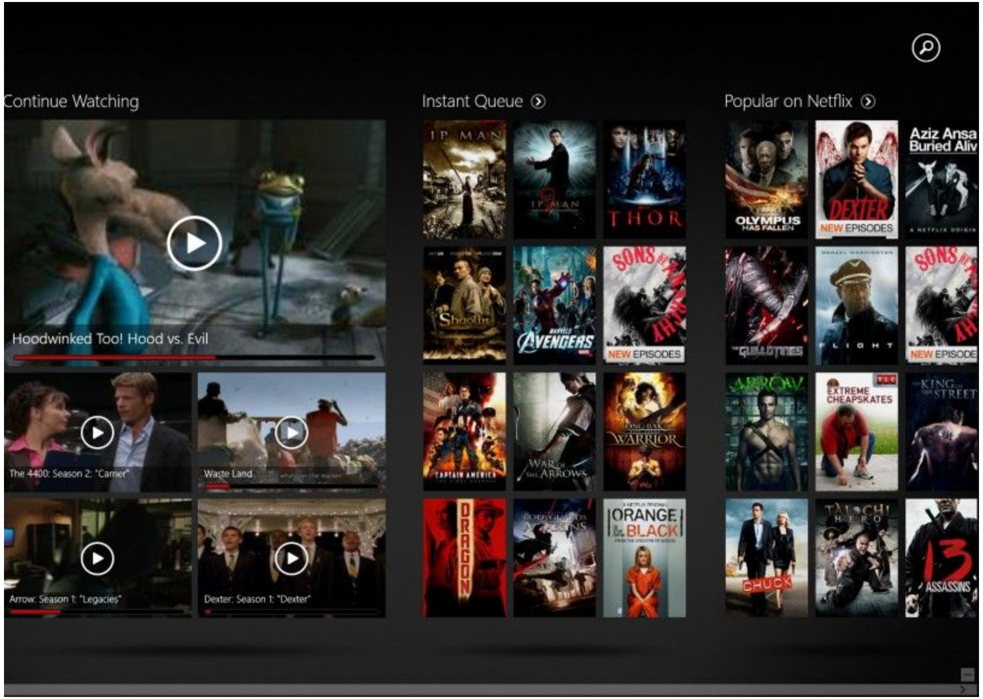
Lluc Cardoner

Outline

- **Motivation**
- Predicting image interestingness
- Results
- Predicting video interestingness
- Results

Motivation

MediaEval Benchmark



What is interesting?

What is interesting?

Not interesting

Interesting

Problem definition

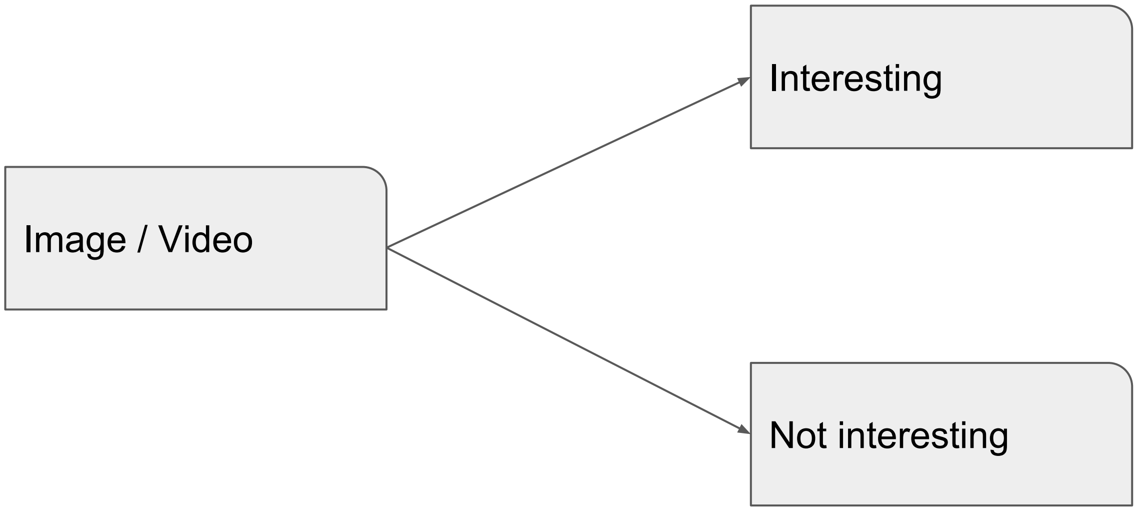


Image / Video

Interesting

Not interesting

MediaEval conclusions 2016

Features

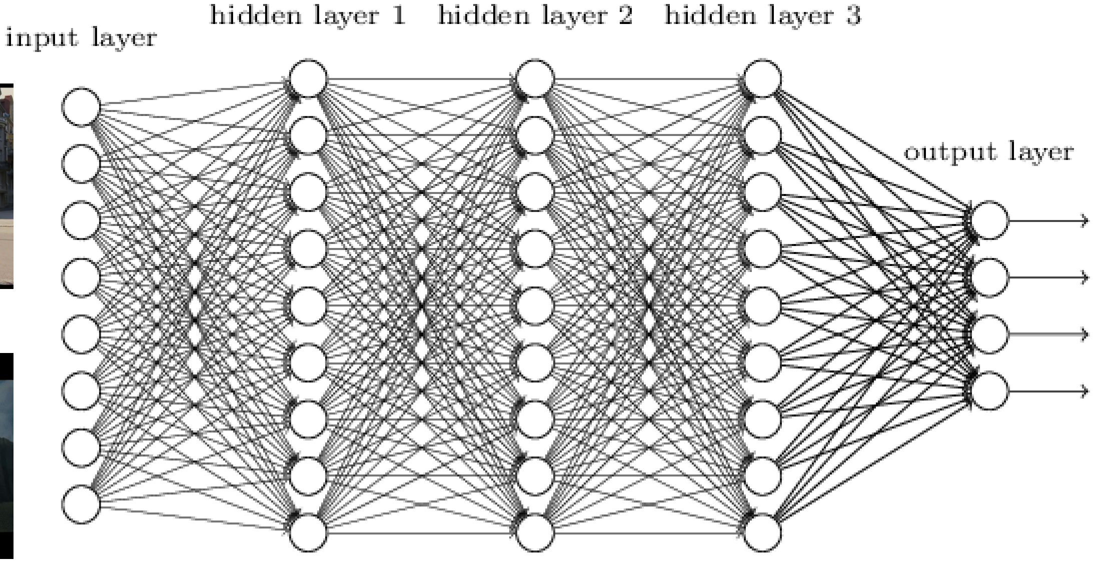
- Image: CNN features
- Video: Multi-modal (visual + audio)

Models

- SVM mostly used
- Few end-to-end deep learning architectures
- Video: time dependencies

Demarty, Claire-Helène, et al. "Predicting Interestingness of Visual Content." *Visual Content Indexing and Retrieval with Psycho-Visual Models* (2017).

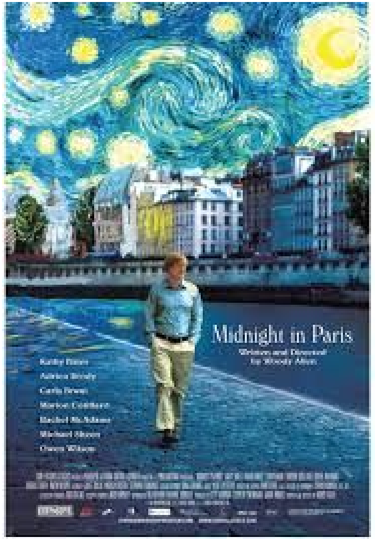
End-to-end deep learning approach



Dataset 2016: Data

- 52 movie trailers - development
- 26 movie trailers - testing

Total: 13 GB

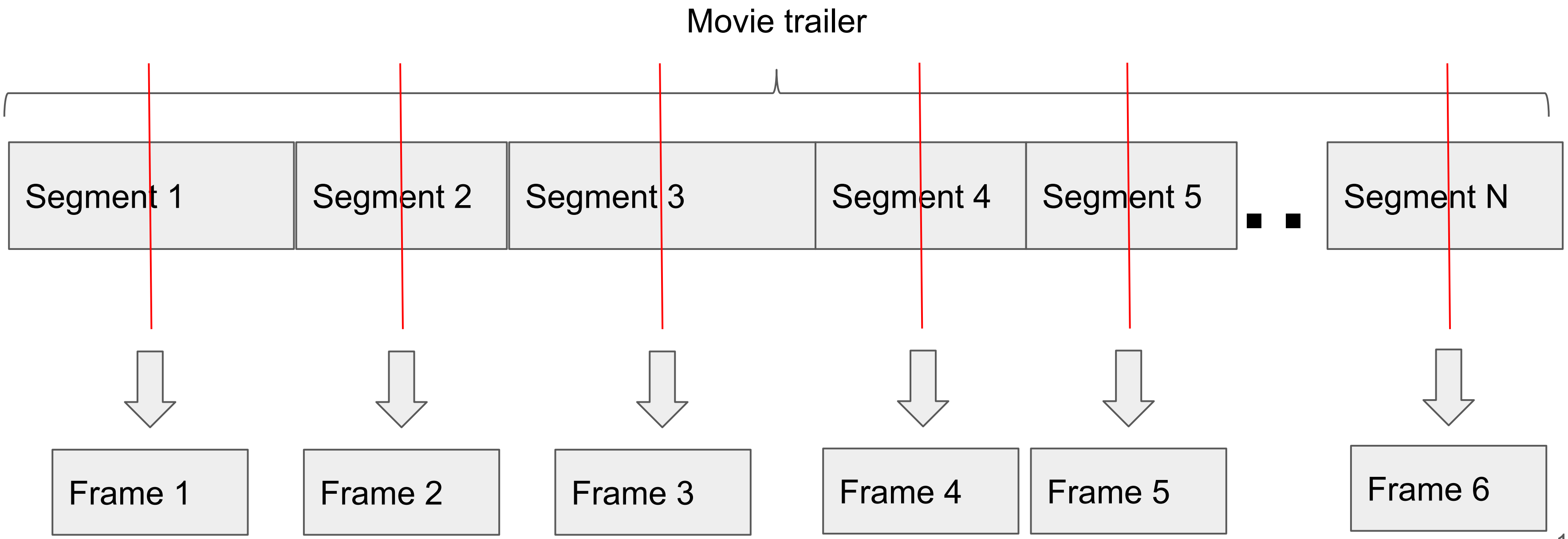


Outline

- Motivation
- **Predicting image interestingness**
- Results
- Predicting video interestingness
- Results

Dataset 2016: Frames

- 52 movie trailers - development
- 26 movie trailers - testing



Dataset: Ground truth

- Classification: 2 classes
  - 0 - not interesting
  - 1 - interesting

- Confidence values
  - Between 0 and 1

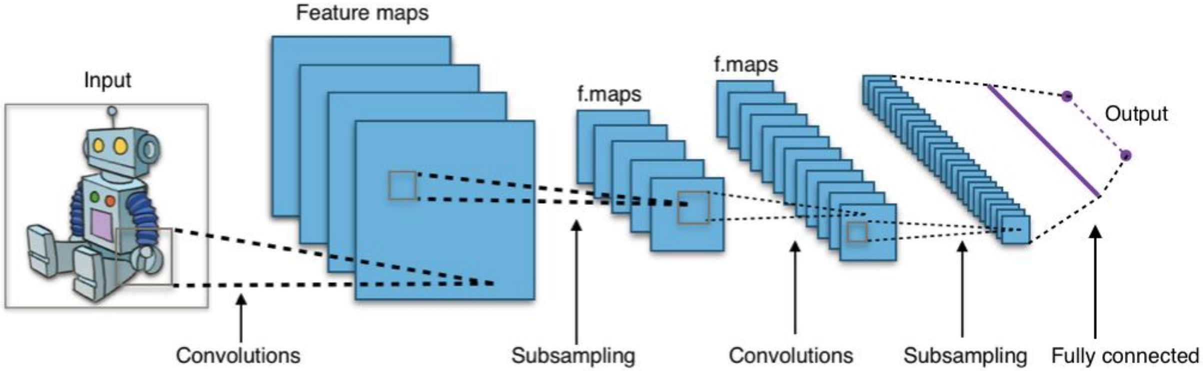
- Rank of the frame or segment in the video

Interesting: 1.0 → 1

Not Interesting: 0.026 → 0

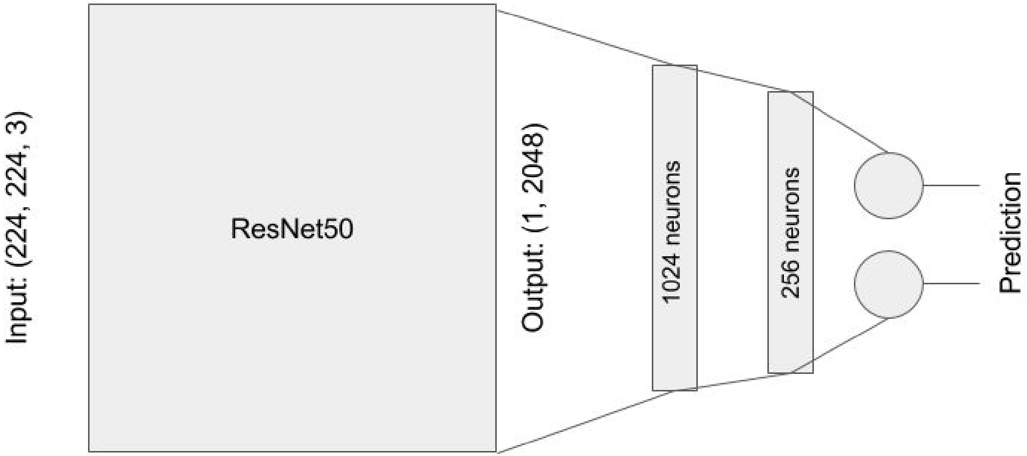
Predicting image interestingness

- ResNet50
  - Transfer learning
  - Fine tuning



He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

Adding layers

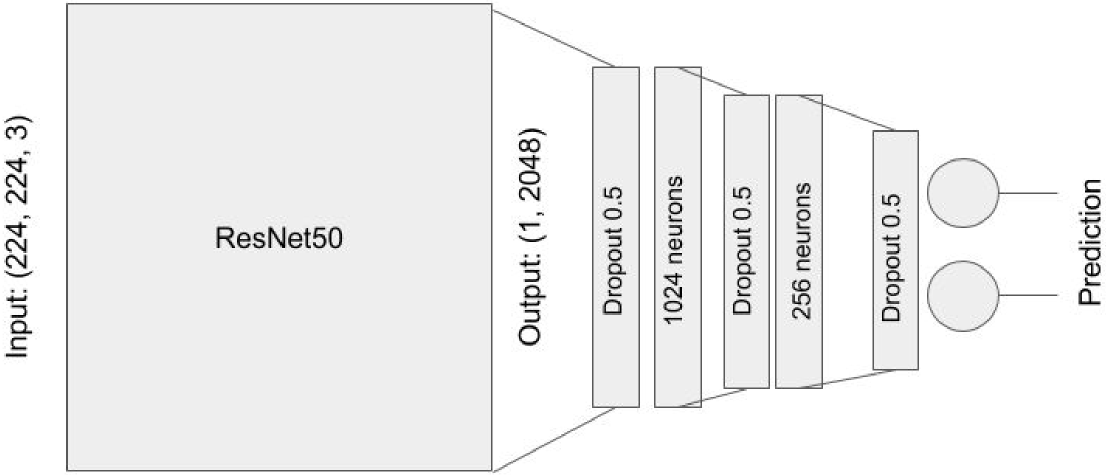


Problem: overfitting

Data augmentation

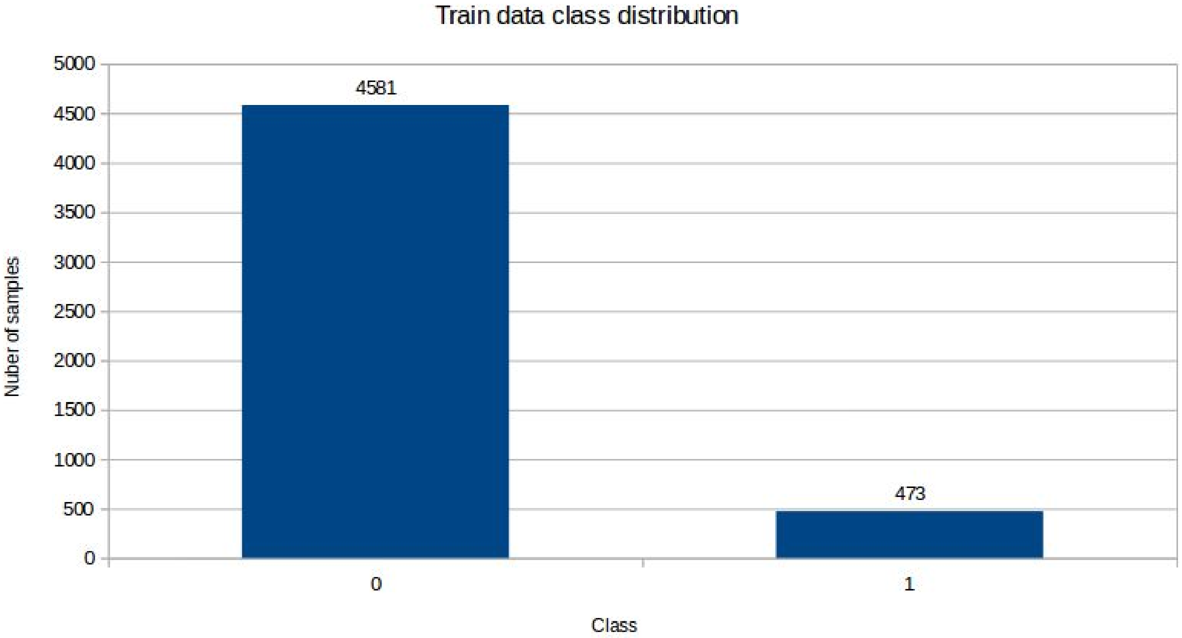
- Image Data Generator
    - Horizontal flip
    - Shuffling

Dropout

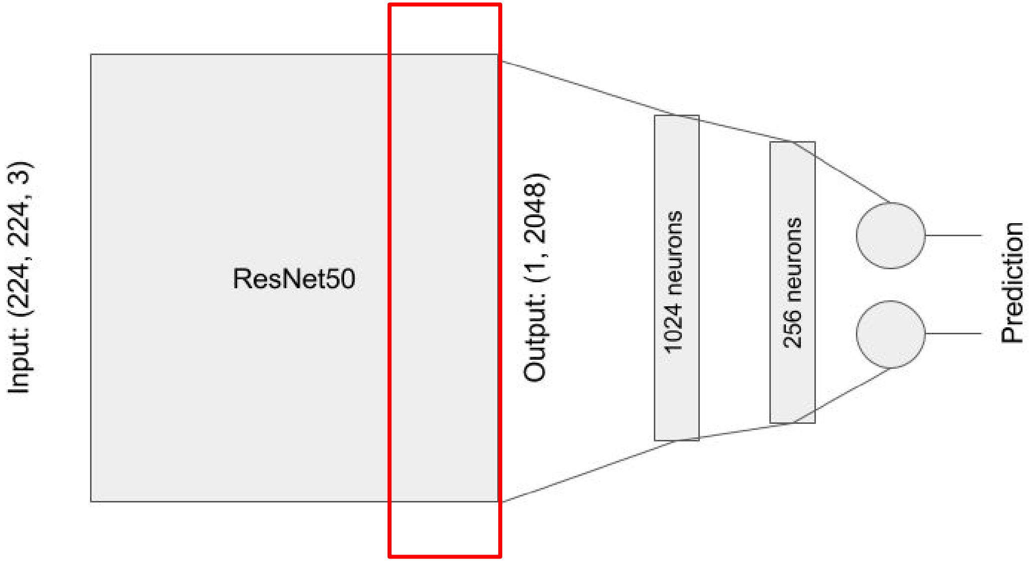

Input: (224, 224, 3)
ResNet50
Output: (1, 2048)
Dropout 0.5
1024 neurons
Dropout 0.5
256 neurons
Dropout 0.5
Prediction

Unbalanced classes

- Class weights



Train last layers



Outline

- Motivation
- Predicting image interestingness
- **Results**
- Predicting video interestingness
- Results

Evaluation metric

- Mean Average Precision (MAP)

$$mAP = \frac{1}{M}\sum_{m=1}^{M} AP(m)$$

For both subtasks

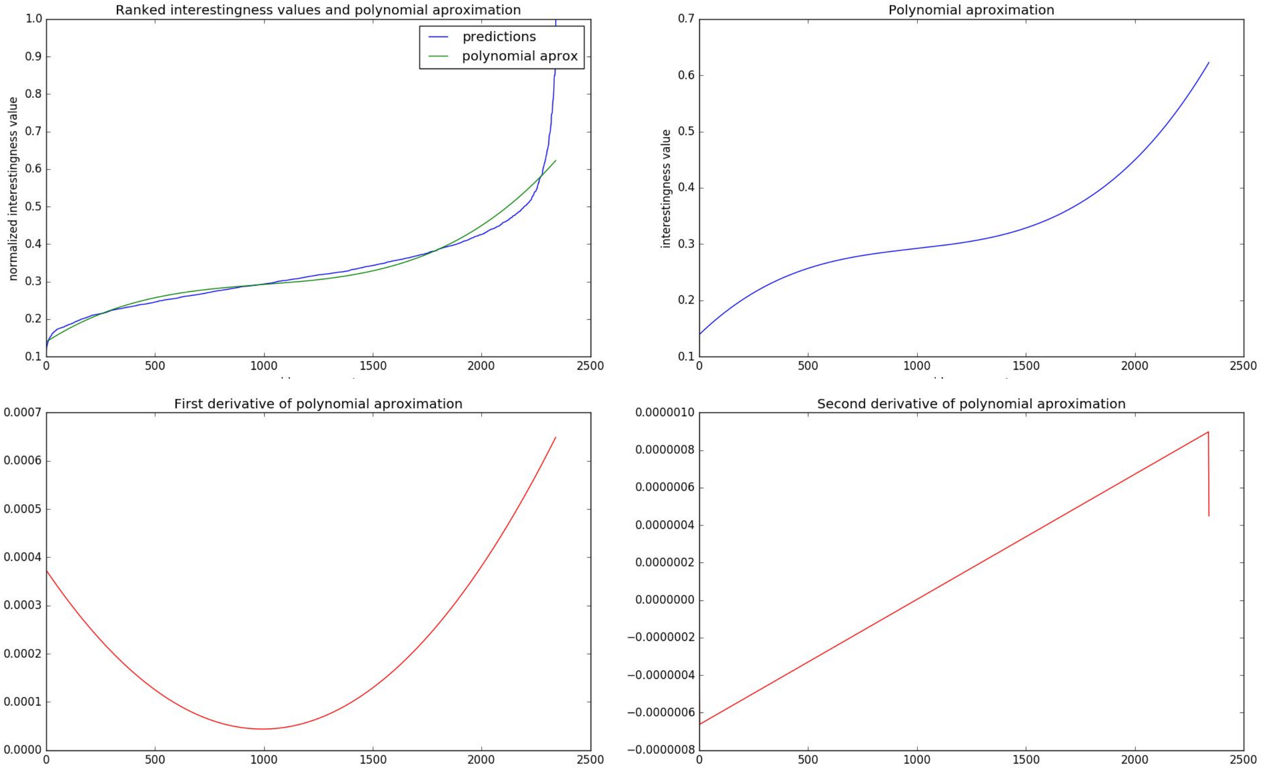
Results: Image interestingness

2016	MAP
Baseline	0.1655
Top result	0.2336

Threshold: 0.5

Id	MAP	Architecture
25	0.1392	train new layers and 2 last layers from ResNet
27	**0.1728**	augment just class 1 and balanced
30	0.1478	dropout of 0.5
31	0.1177	Class weights + dropout + horizontal flip
37	0.1564	Class weights + dropout + flip, shift, zoom
39	0.1402	Class weights + dropout + flip, shift, zoom + 2 ResNet layers

Threshold


Ranked interestingness values and polynomial aproximation
normalized interestingness value
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
500
1000
1500
2000
2500
predictions
polynomial aprox
Polynomial aproximation
interestingness value
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
500
1000
1500
2000
2500
First derivative of polynomial aproximation
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0
500
1000
1500
2000
2500
Second derivative of polynomial aproximation
0.0000010
0.0000008
0.0000006
0.0000004
0.0000002
0.0000000
−0.0000002
−0.0000004
−0.0000006
−0.0000008
0
500
1000
1500
2000
2500

Results: Image interestingness

2016	MAP
Baseline	0.1655
Top result	0.2336

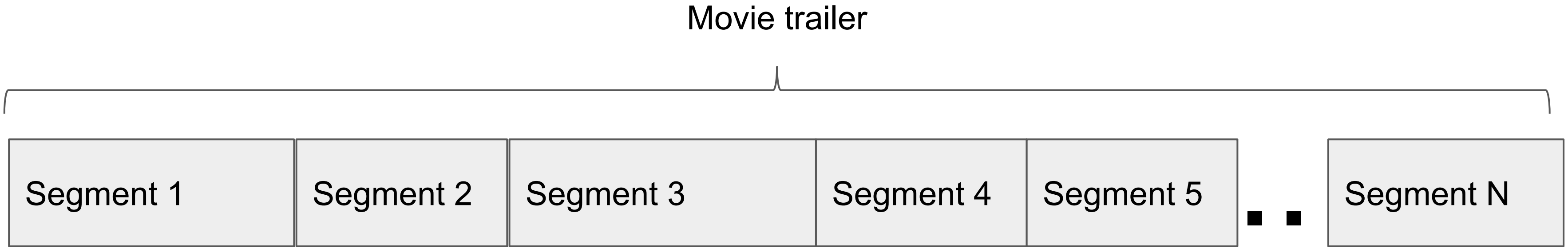
	Static Threshold	Dynamic threshold	
Id	MAP	threshold	MAP
25	0.1392	0.1577	0.1932
27	**0.1728**	0.4875	0.1909
30	0.1478	0.1572	0.2243
31	0.1177	0.5066	**0.2396**
37	0.1564	0.5295	**0.2362**
39	0.1402	0.1336	0.1795

Outline

- Predicting image interestingness
- Results
- **Predicting video interestingness**
- Results

Dataset 2016: Segments

- 52 movie trailers - development
- 26 movie trailers - testing

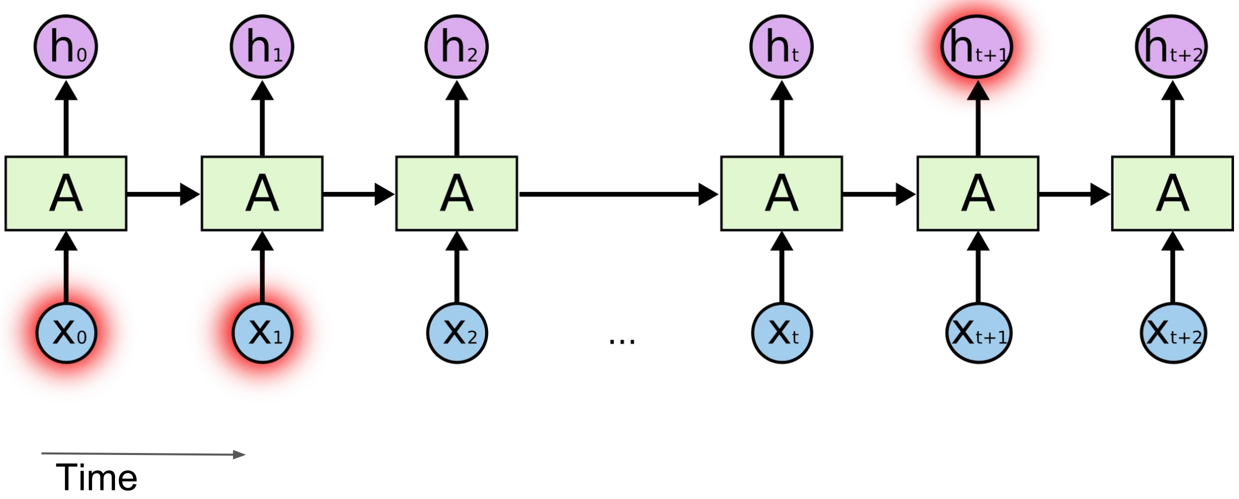


Movie trailer

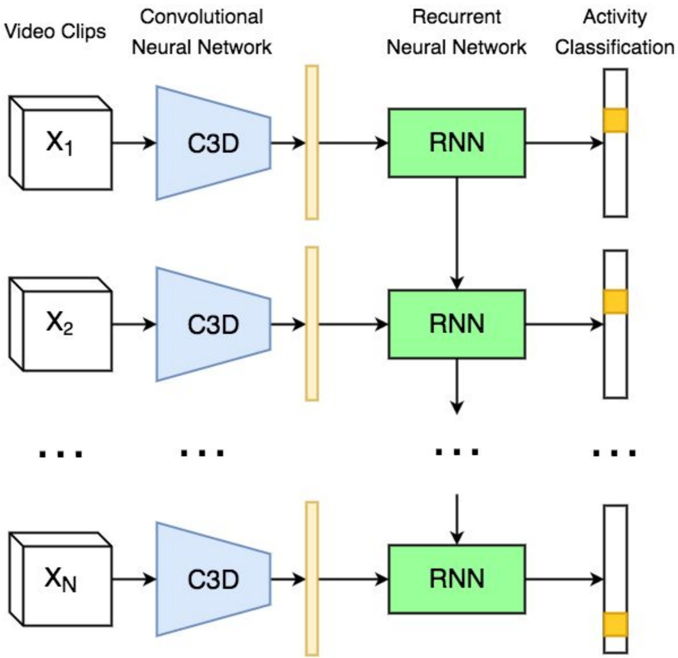
Segment 1 Segment 2 Segment 3 Segment 4 Segment 5 .. Segment N

Predicting video interestingness

- Extract features: C3D
- Training LSTM network



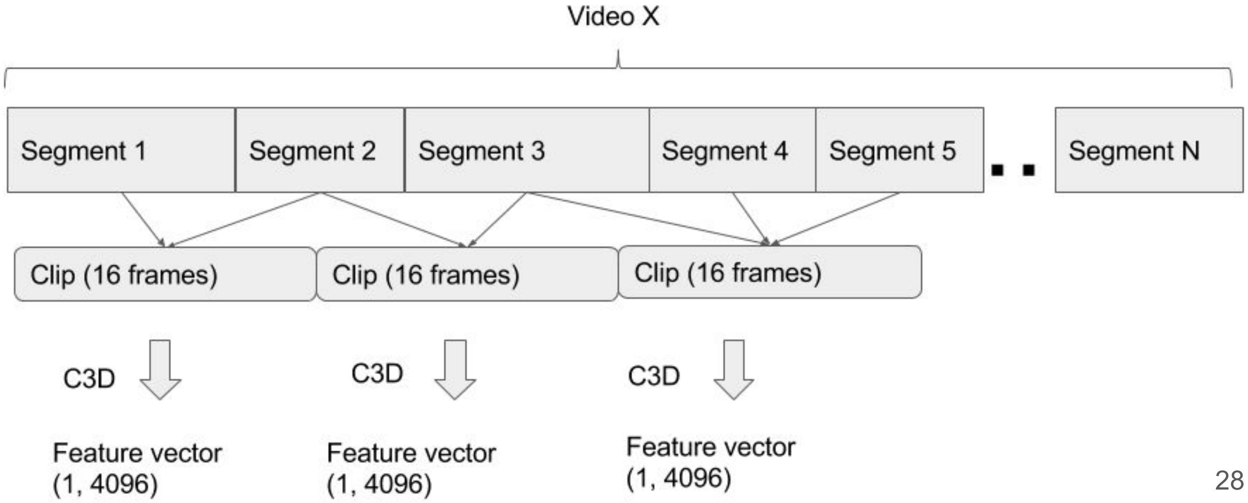
3D Convolutional network



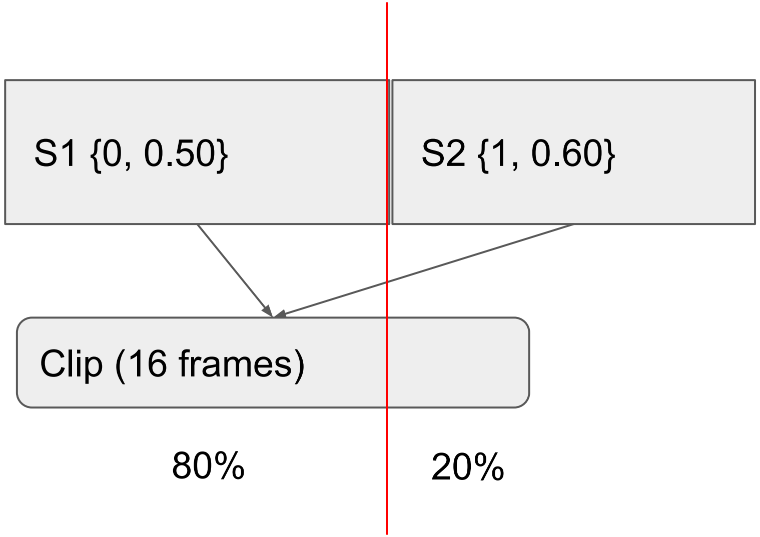
Montes, Alberto, Amaia Salvador, and Xavier Giro-i-Nieto. "Temporal activity detection in untrimmed videos with recurrent neural networks." *NIPS Workshop Large Scale Computer Vision Systems 2016*

Extract features

- Preprocess
  - Clips
- Feature extraction
  - 3D convolutional network
- Label mapping
  - Feature vector

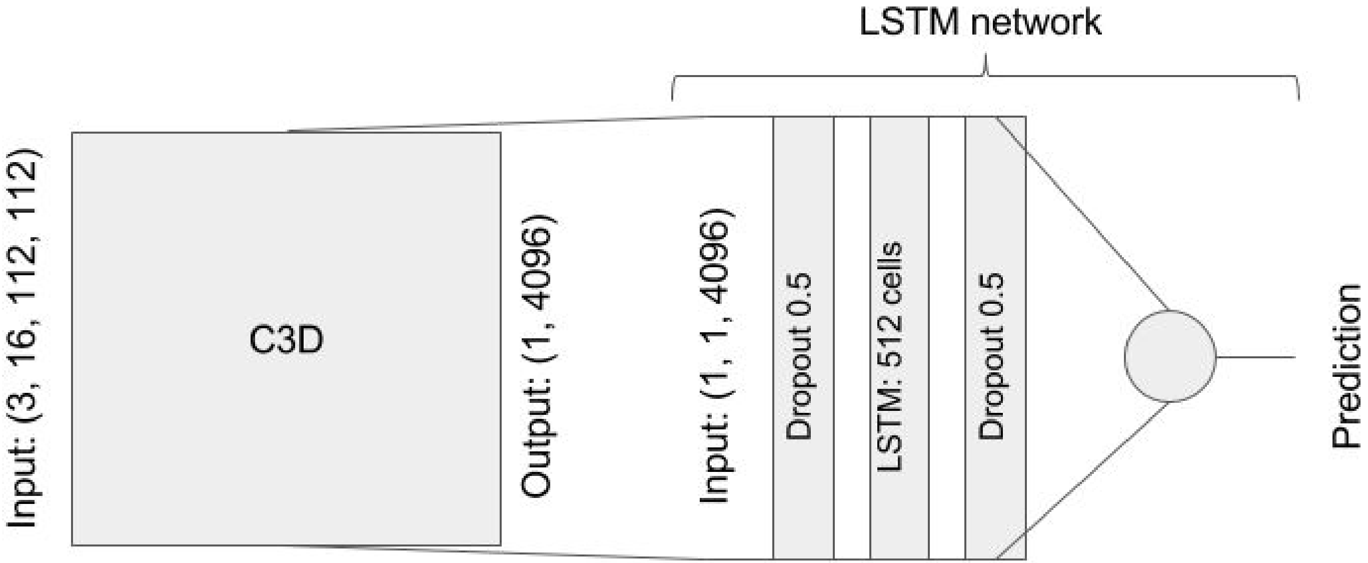


Label mapping



0.8 x 0.5 + 0.2 x 0.6 = 0.52

Fine-tuning LSTM



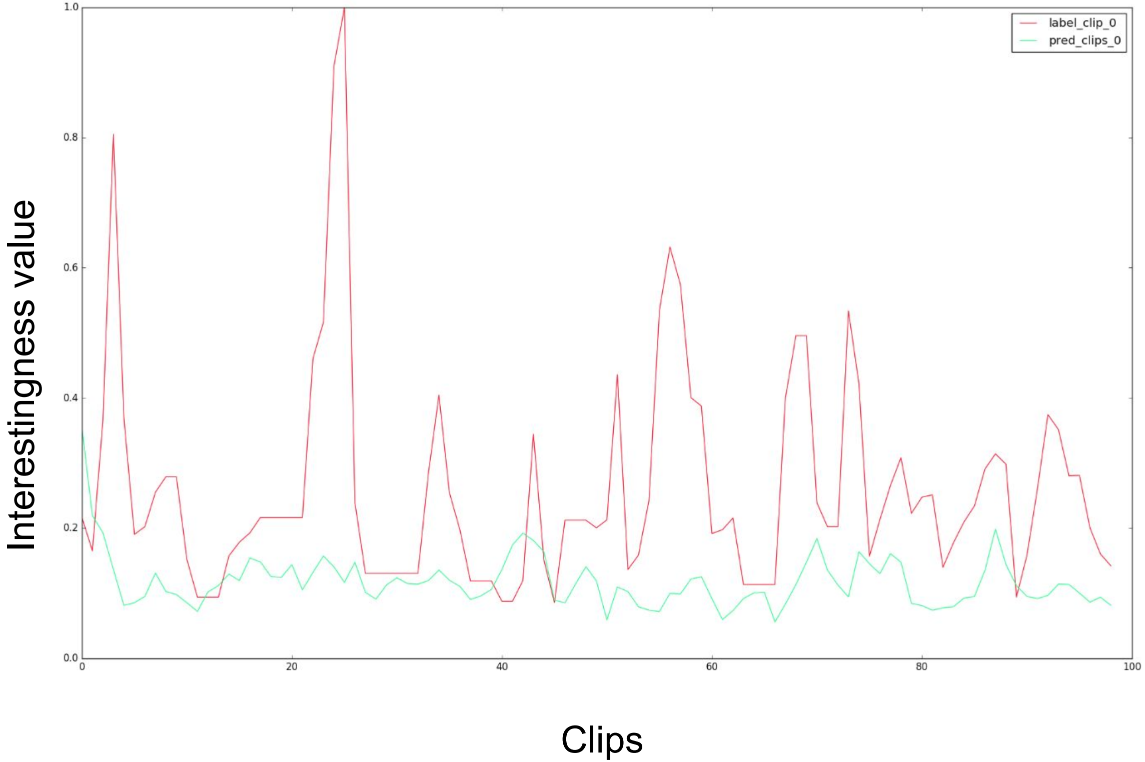
Outline

- Predicting image interestingness
- Results
- Predicting video interestingness
- **Results**

Results: Video interestingness

2016	MAP
Baseline	0.1496
Top result	0.1815
Technicolor	0.1365

Id	MAP
65	0.1541



Shen, Yuesong, Claire-Hélène Demarty, and Ngoc QK Duong. "Technicolor@ MediaEval 2016 Predicting Media Interestingness Task." *MediaEval* (2016).

Conclusions

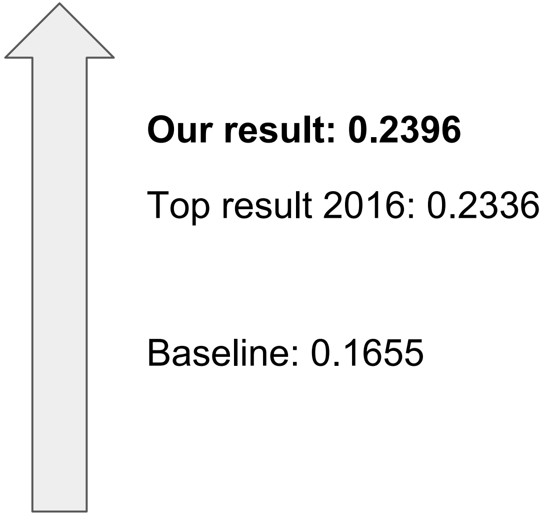
Predicting image interestingness	**MAP**
Class weights + dropout + horizontal flip	0.2396
Class weights + dropout + flip, shift, zoom	0.2362

Conclusions

Static Threshold	Dynamic threshold
MAP	MAP
0.1392	0.1932
0.1728	0.1909
0.1478	0.2243
0.1177	**0.2396**
0.1564	**0.2362**
0.1402	0.1795

Conclusions

Image

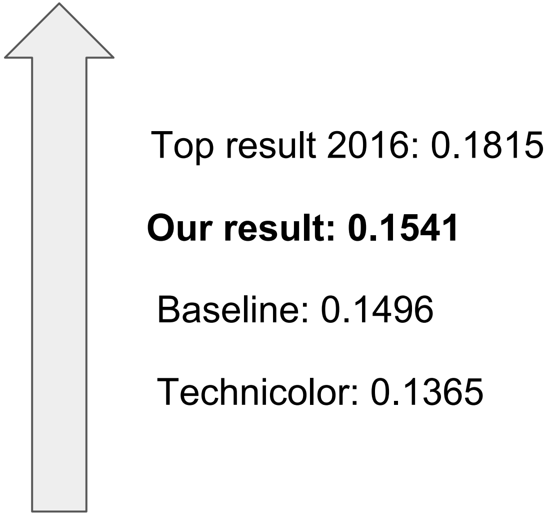


Our result: 0.2396

Top result 2016: 0.2336

Baseline: 0.1655

Video



Top result 2016: 0.1815

Our result: 0.1541

Baseline: 0.1496

Technicolor: 0.1365

GitHub

https://github.com/lluccardoner/MediaInterestingness