Generating Video Thumbnails Using Deep Neural Networks

Jeroen Wijering

JWPLAYER

About JW Player

The leading video platform for media.

10%

Of all video views on the open web

25k

Events captured every second

NOS
W
CNBC
VICE
NASA
IMDb
group nine media
Discovery NETWORKS
FANDOM
REFINERY29
BBC Worldwide
Red Bull TV
The Weather Channel
EUROSPORT
talpa
BLIZZARD
Tastemade
The Washington Post
Jamie Oliver
VIACOM
apartment therapy
sky
BONNIER
BUSINESS INSIDER
FAST COMPANY
FOX
NATIONAL GEOGRAPHIC
OLYMPIC
nu nl
PBS

DEMO

Metadata Browser

JW Enrich

A video recognition engine to grow audience engagement:

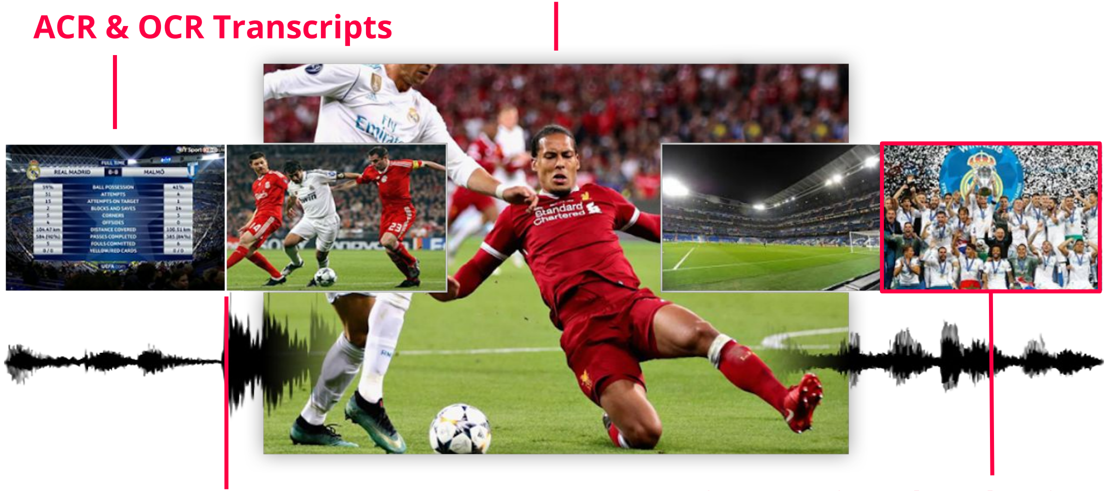
01 | **In-Video Search**

02 | **Visual Previews**

03 | **Recommendations**

04 | **Trends Analytics**

05 | **Full API Coverage**



Audio & Visual Tags

ACR & OCR Transcripts

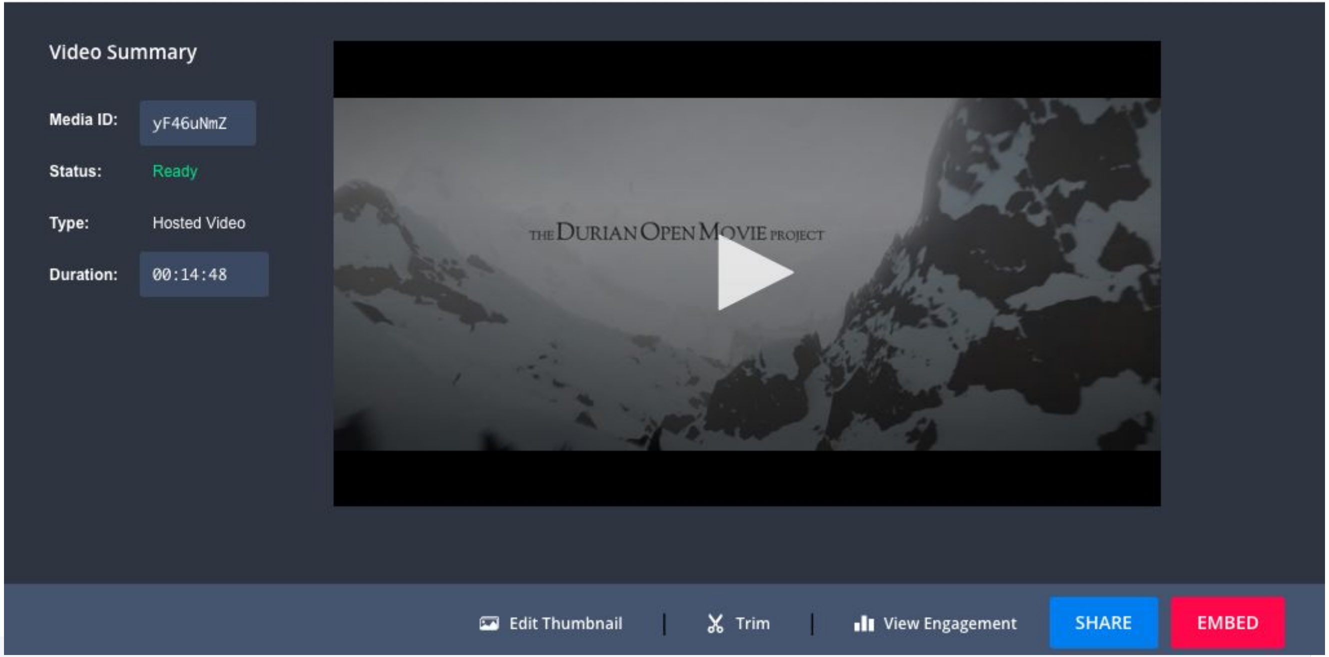
Shot & Scene Transitions

Static & Motion Thumbnails

Thumbnails: the first impression and promotion of your videos

~60% Of Editors Don’t Design Thumbnails

Defaulting to an unreliable, 10s frame capture



Video Summary

Media ID: yF46uNmZ

Status: Ready

Type: Hosted Video

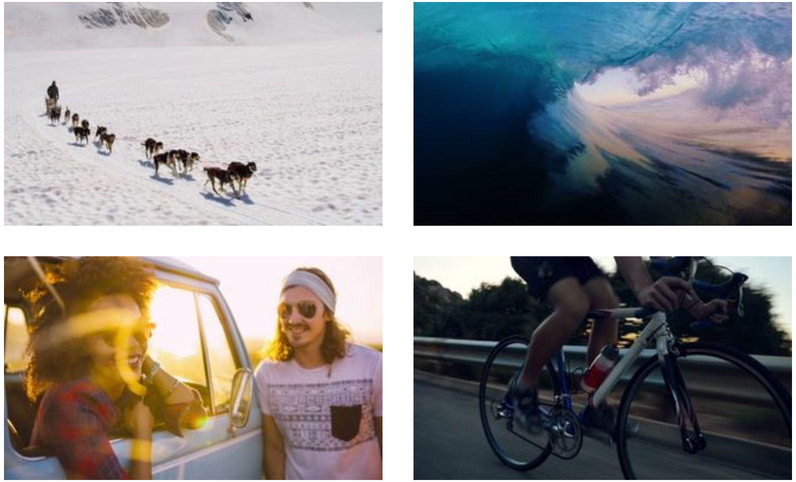
Duration: 00:14:48

What is a good thumbnail?

Good thumbnails are subjective to the viewer!

Common properties:

- Subject not blurry
- Balanced brightness and contrast
- Well framed objects
- Relevant to the subject



Mac and Cheese Hot Dog

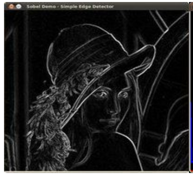
How do we build a model that automatically picks good thumbnails?

Manually creating a model is hard

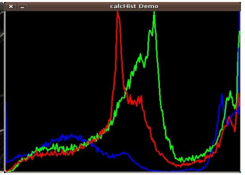
- Which features to extract?
- How to describe those features?
- How to weight individual features?
- How to penalize overfitting of models?
- Many techniques: SIFT, SURF, HOG?

Need to be an expert in Computer Vision :-(

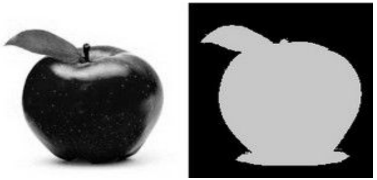
So Many Image Features...



Edge Detection



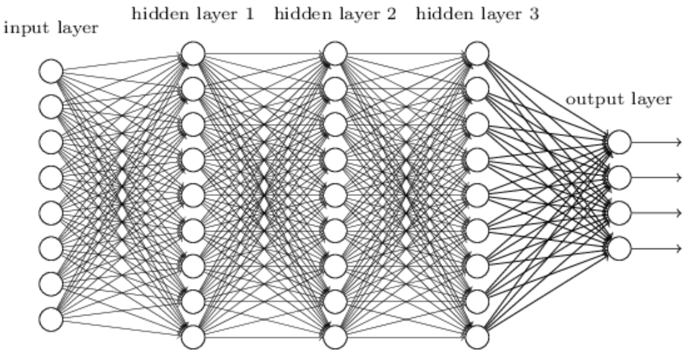
Color Histogram



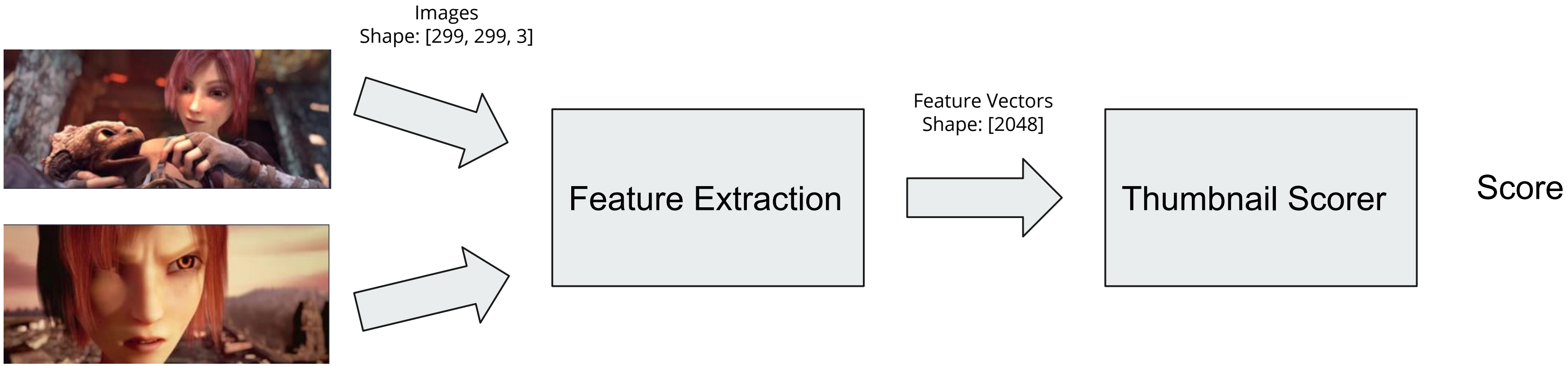
Pixel Segmentation

Deep Learning

- Learn features implicitly
- Learn from examples
- Techniques to avoid overfitting
- Successful in a wide variety of applications:
  - Image classification
  - Sentiment analysis
  - Text Translation
  - Audio transcription



Thumbnail Selection using Deep Learning



Images
Shape: [299, 299, 3]

Feature Extraction

Feature Vectors
Shape: [2048]

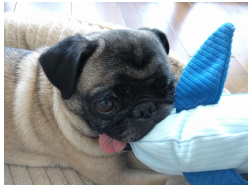
Thumbnail Scorer

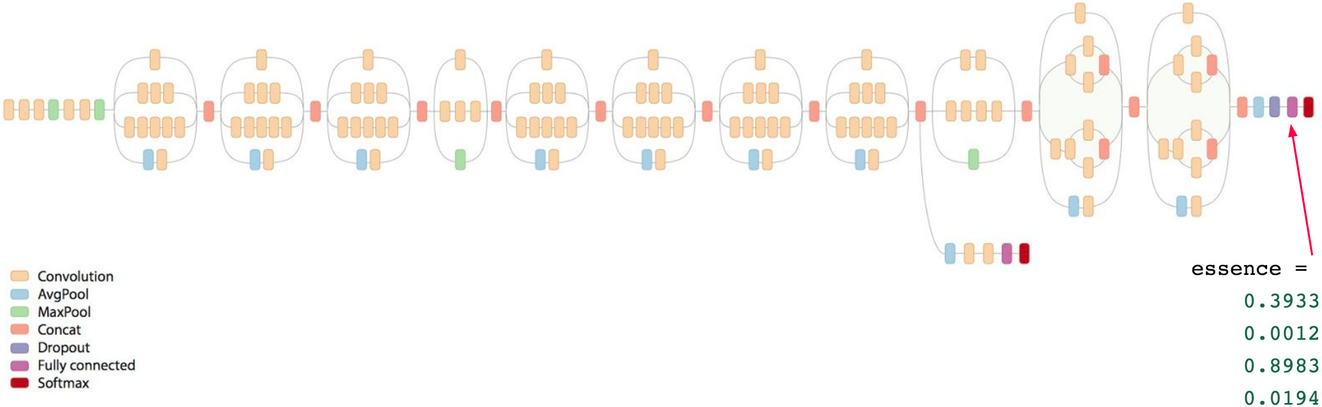
Score

Feature Extraction

- The process of reducing the amount of resources required to describe a large set of data.
- Images usually contain a lot of redundant information.
- Before we can efficiently process the information captured by an image we need to get rid of redundant information.
- Feature extraction can be done in many ways, but often done using ConvNets.

Inception V3 Architecture





Dog (0.80)
Cat (0.05)
Rat (0.01)
...

```
essence = [
    0.3933345622,
    0.0012100923,
    0.8983323001,
    0.0194860482,
    ...
]
```

https://research.googleblog.com/2016/03/train-your-own-image-classifier-with.html

Framing the problem

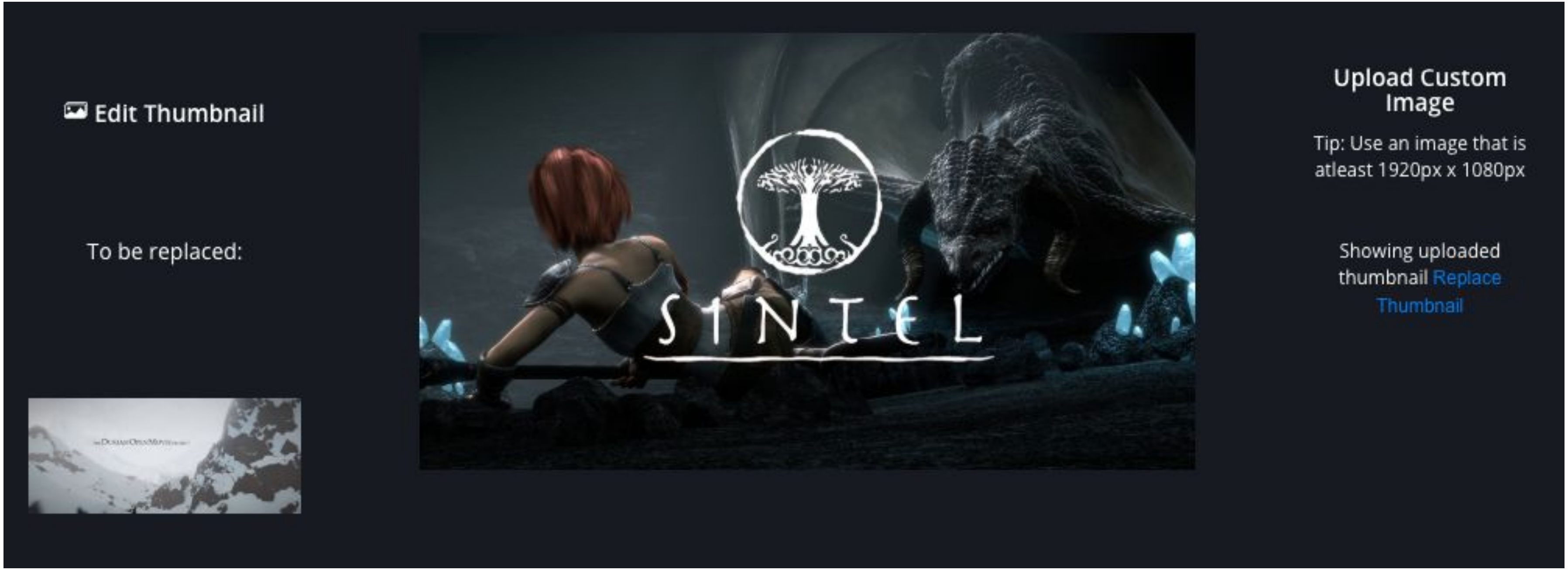
1. Classification

2. Machine-learned Ranking

JW Player Thumbnail Datasets

- Custom Uploads
  - Poster Images uploaded by an editor.

Custom Upload



Edit Thumbnail

To be replaced:

Upload Custom Image

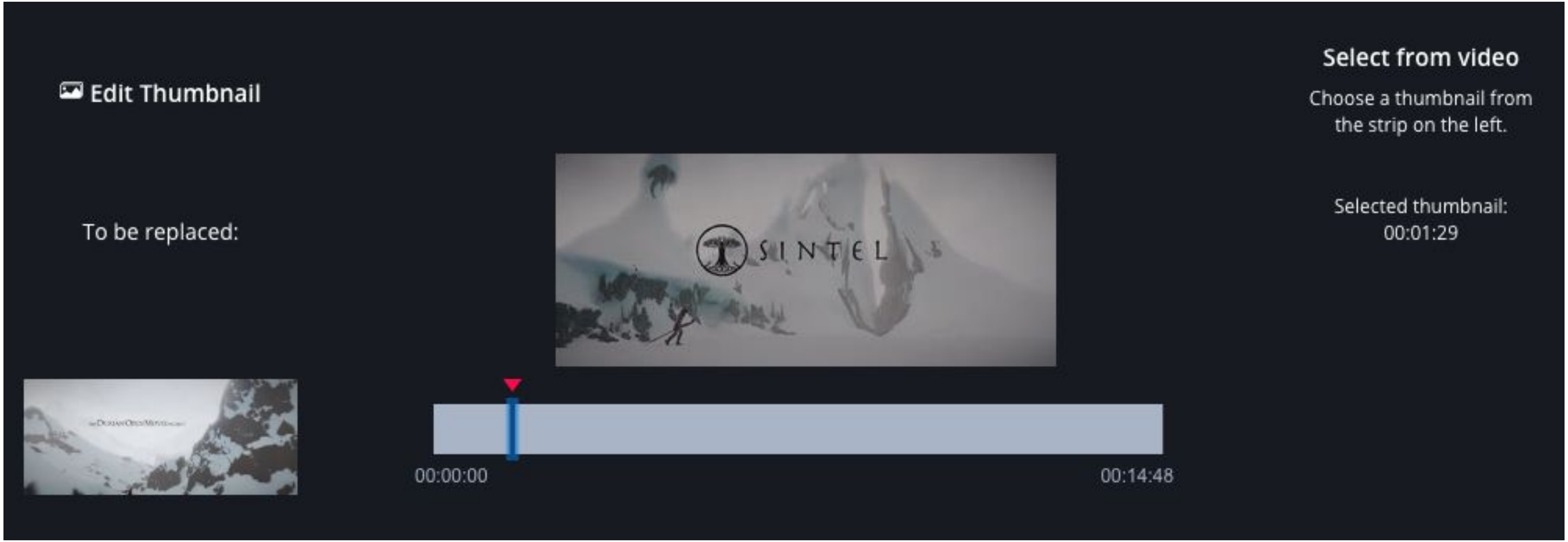
Tip: Use an image that is atleast 1920px x 1080px

Showing uploaded thumbnail Replace Thumbnail

JW Player Thumbnail Datasets

- Custom Uploads
  - Poster Images uploaded by an editor.
- Thumbnail Index Poster Images
  - Poster Images selected from a list of frames sampled from the video by the editor.

Thumbnail Index



Edit Thumbnail

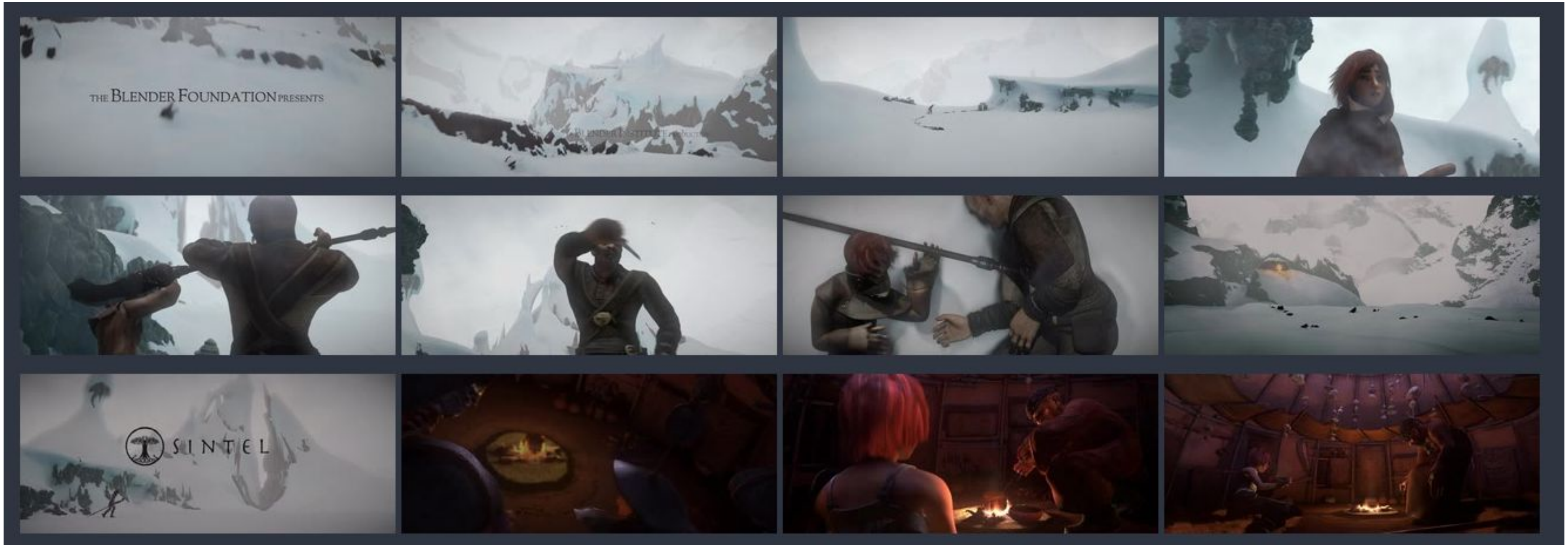
To be replaced:

Select from video

Choose a thumbnail from the strip on the left.

Selected thumbnail: 00:01:29

Thumbnail Index



Framing the problem a Classification task

Teach a model to make decisions like an editor

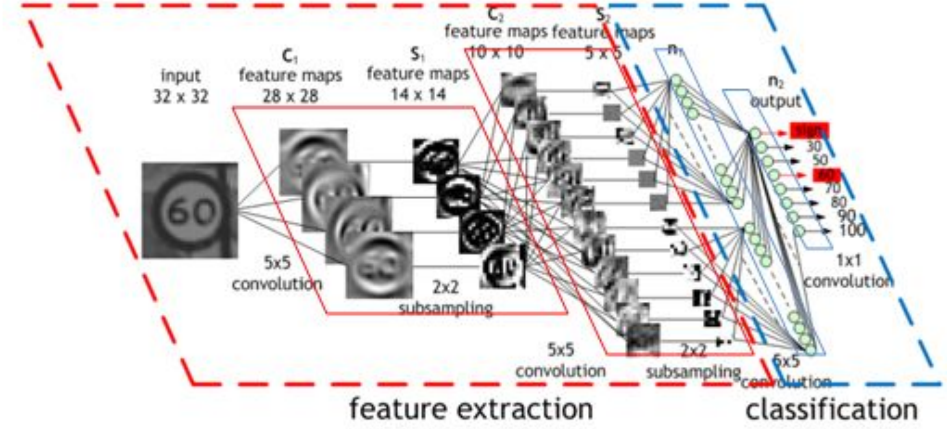
- Leverage editorial decisions made across JW Network
- We consider thumbnails that have been hand selected by editorial staff as “Good” thumbnails
- “Bad” thumbnails are default frames when no editorial choice has been made

Training set: 20.000 images
Model predicts: *how likely an image is to be of editor quality*

We start with a pre-trained version of Google's *Inception* Neural Network...

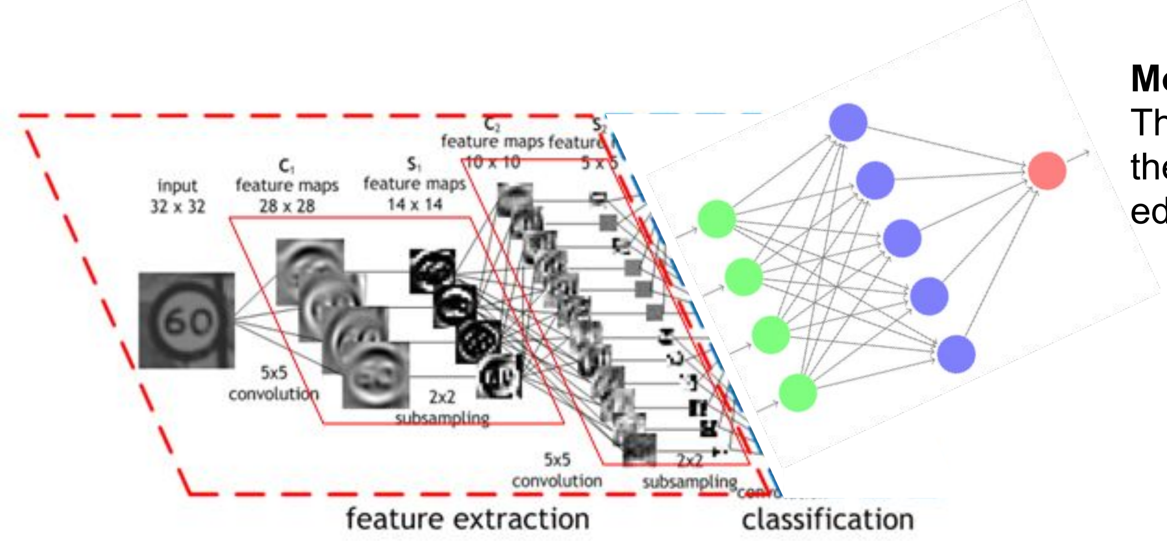
IMAGENET

1,000,000 images, 1,000 categories

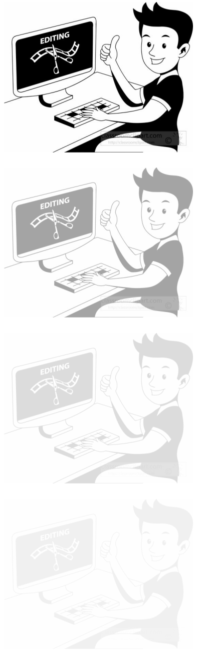


Business office

... and retrain the Final Layer for our Thumbnail Task



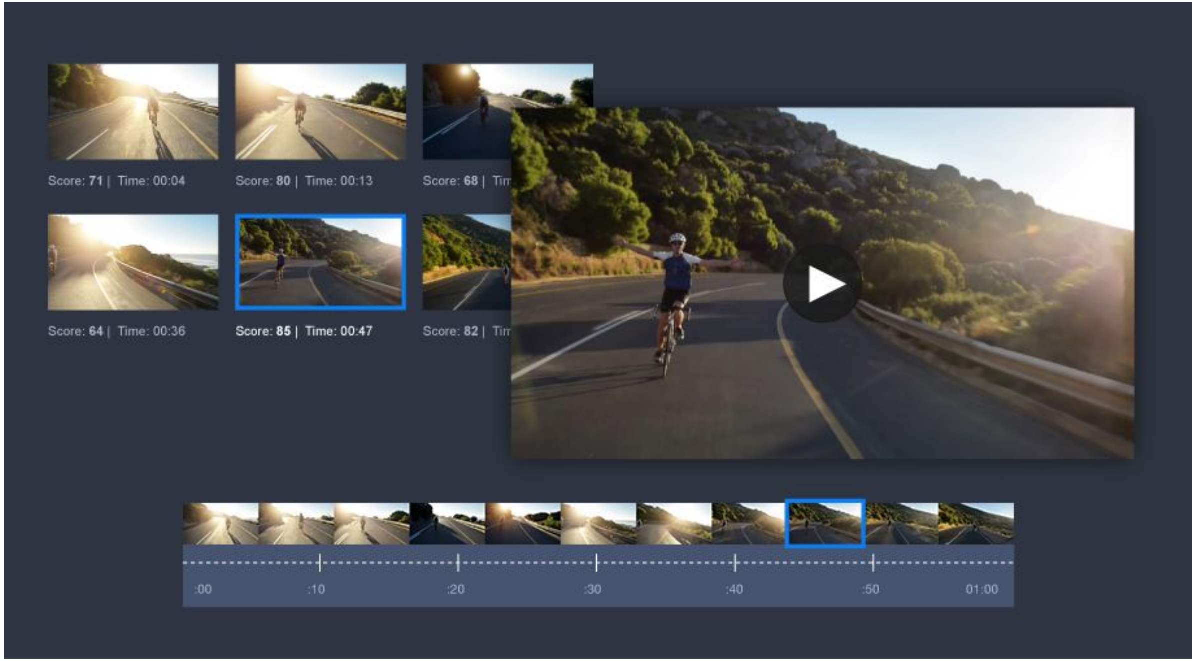
Model output: Thumbnail score**,** the likelihood it’s editor worthy



Framing the problem as a Ranking Problem

- Similar to framing as a classification problem, except:
  - We make pairs of images where one image is always the image selected by the publisher and the other a randomly sampled negative .
  - The final layer outputs only a single score per image.
  - A pairwise cross-entropy loss function is used with the goal to minimize the number of inversions in the ranking.

How does this work for videos?



Score: 71 | Time: 00:04

Score: 80 | Time: 00:13

Score: 68 | Tim

Score: 64 | Time: 00:36

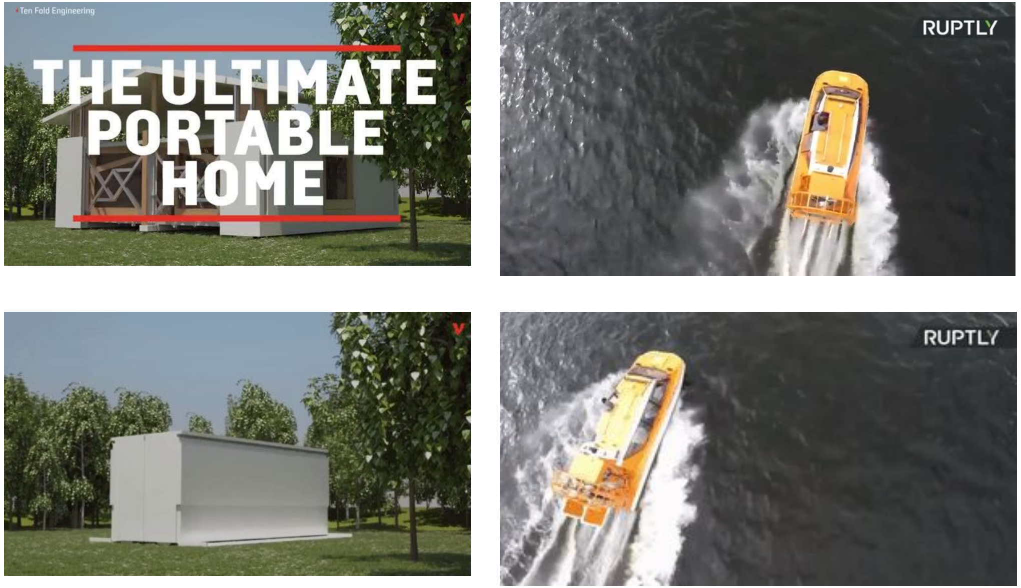
Score: 85 | Time: 00:47

Score: 82 | Tim

What if we could optimize click-through rates by displaying a preview of the video rather than a static image?

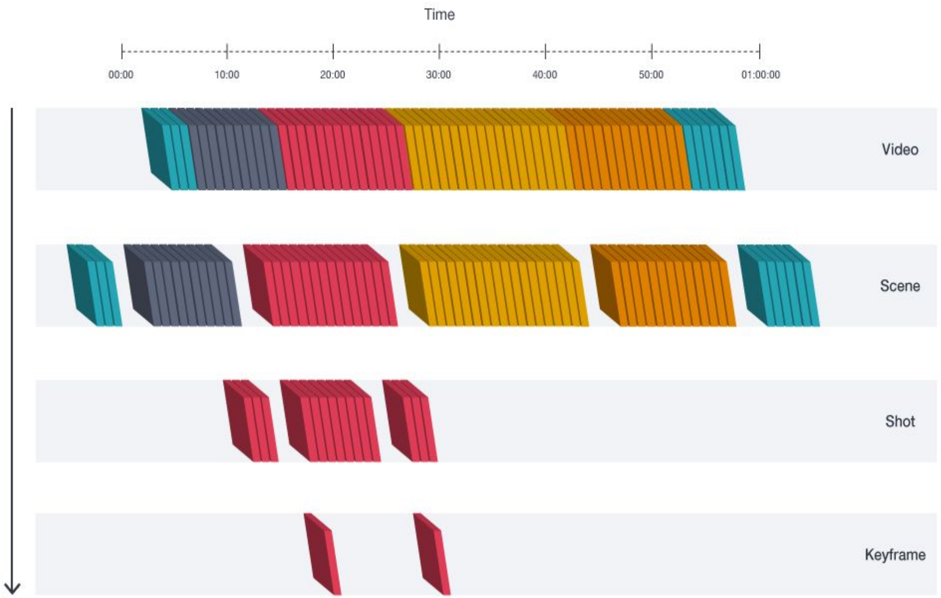
Animated Thumbnails

- Short "GIF-like" video previews
- Consist of 1-3 shots
- Small in file size
- Optimized for CTR, not accurate summarization.



Generating an Animated Thumbnail

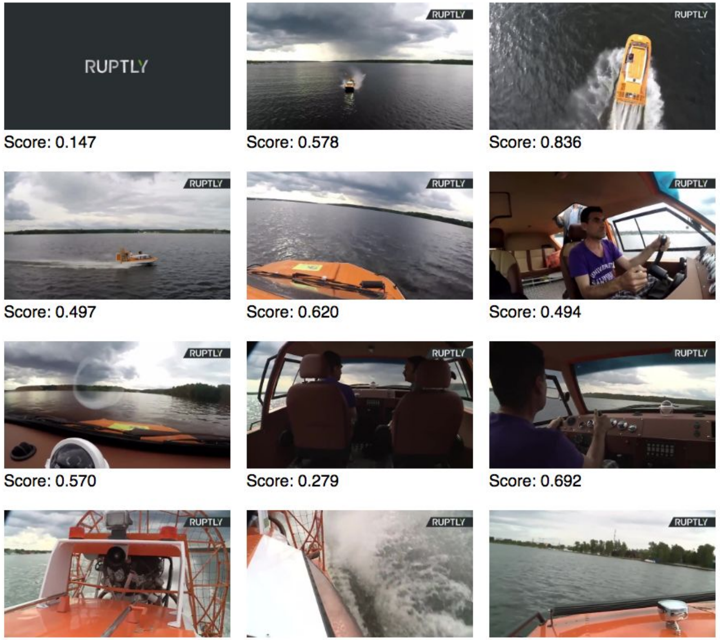
1. Partition a video into a list of shots
2. Sample a representative frame for every shot



Video Structure

Generating an Animated Thumbnail

3. Score frames using our intelligent thumbnail model
4. Calculate a moving average of thumbnail scores over the timeline of the video
5. Sample the window with the highest average of the original video



A/B Testing shows 5-30% increases in click-through versus “manual” thumbnails: Success!

Are We Done Yet? Never...

- Animated thumbs are now continuous
    - Cluster shots & find top 3.
- Bias towards “still” thumbnails
    - Include motion information
- Preview live streams?
    - Need to re-think the model

Thank You. Questions?

Jeroen Wijering
jeroen@jwplayer.com