Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Computer Vision Lab

Deep Visual Semantic Embedding for Video Thumbnail Selection

Master Thesis Defence

30-08-2016

Arun Balajee Vasudevan

Supervisors: Michael Gygli, Anna Volokitin, Dr. Achanta Radhakrishna, Prof. Luc Van Gool, Prof. Sabine Susstrunk

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Problem

- Given a query and video
- Extract query relevant video thumbnails

Semantic Information:
Query: Future Cars

Video Thumbnail Extractor

Thumbnails

1. Improving Video Search

- Do you get what the video is about?

Not a good thumbnail !

Annoying Animals

Future Cars

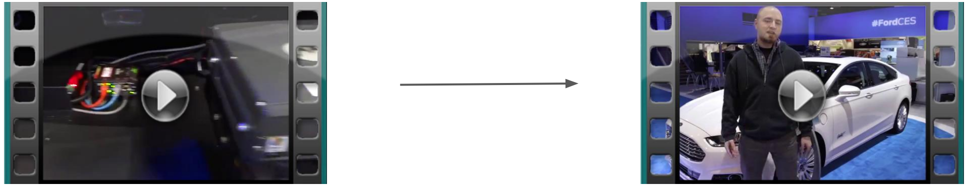
Baby Antelope

Improving Video Search

Phone videos aren’t titled well. Thumbnails play a major part for search

Name	Size	Type	Modified
2015-02-06 11.52.55.mp4	13.5 MB	Video	10:36

In general, a Bad thumbnail make even a good video to be unattractive and make it hard to judge relevance



A good thumbnails

- Increase views for videos
- Thumbnails recommended for users

Improving Video Search

- Jerry in fancy dress in “Tom and Jerry Show”
- Do you remember the dress and don’t remember the video?

  Yes!

  Query: Jerry in Fancy dress

  - Titles may be episode numbers
  - Thumbnails may be Tom chasing Jerry
  - How to get the right video?




2. Video Level Search

- Have you tried to revisit any movies?

Jeans_movie
Length: 02:22:36

Date modified: 27-05-2015 23:29
Size: 839 MB

- Where is “Eiffel Tower” in the movie?
  - Search the whole video !

Snapshot at:
00:56:36

Source: https://www.youtube.com/watch?v=hS7fy3Q3ss4

3. GIFs

- Thumbnails are just keyframes of video. But how thumbnails can be shown as a sequential event?

Beach

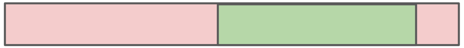
Height Jump

- GIFs getting popularity these days
  - Save as a shorter versions
  - Highlights of video

Use Cases

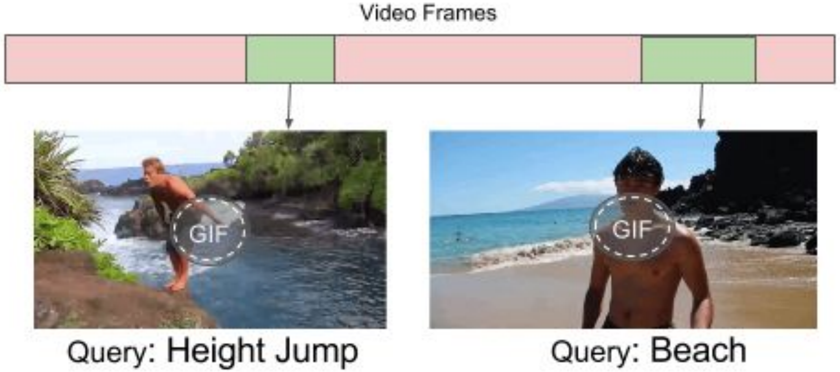
- Improving Video Search
- Video level search
- GIF generation from videos
- Query adaptive video summarization

Problem- Query Adaptive Video Thumbnails



Aftershocks earthquake

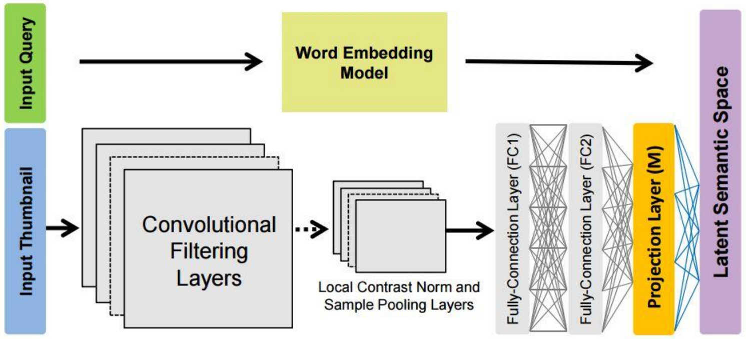
During earthquake



Relevant

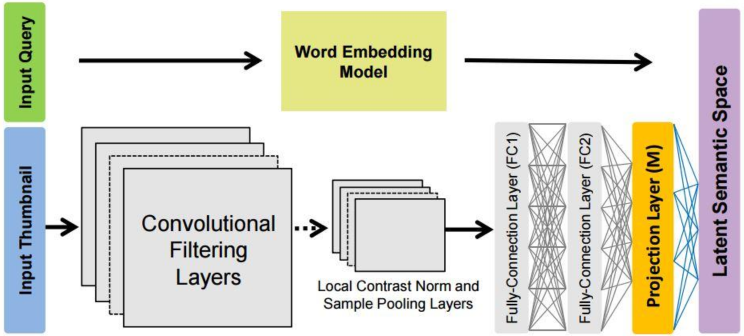
Non-Relevant

Baseline[Liu et al CVPR 2015]



- Input Query - Text Queries
- Query Embedding Model - GloVe [Pennington et al.]
  - Average of all words in the query
- Convolutional Neural Network model
  - AlexNet
  - A fully connected layer added
- Use Bing image search data (query, image, # of clicks) to learn a joint embedding space for images and text
- Compute frame relevance as cosine similarity between the query or title embedding and the frame embedding

Baseline[Liu et al]



Limitations

- Average model for query modelling
- (query+, image+, query-)

("Cat" ,  , "New York")

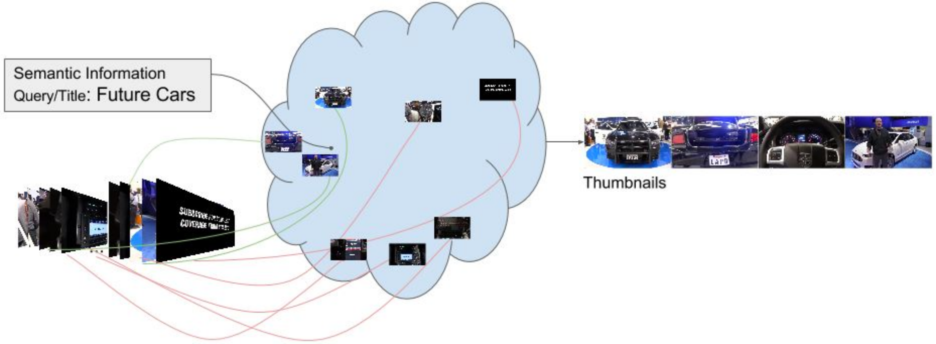
- Inference:
  - Thumbnail with maximum proximity to one of the query word

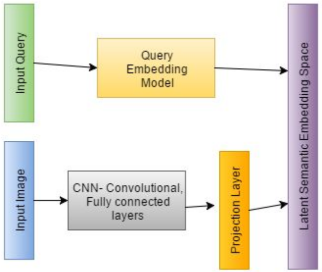
Baseline with our improvements

- Input Query - Text Queries
- Query Embedding Model - word2vec
    - LSTM with fixed length embedding
- Convolutional Neural Network model
    - Finetuned on VGG-19
    - A fully connected layer added
- Training Data: (query+, image+, image-)

("cat",  , )

- Latent Semantic Embedding Space- Both models project to a common vector space

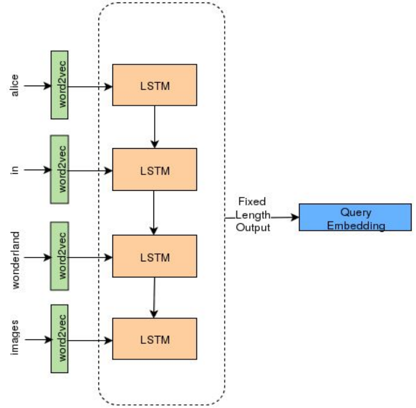


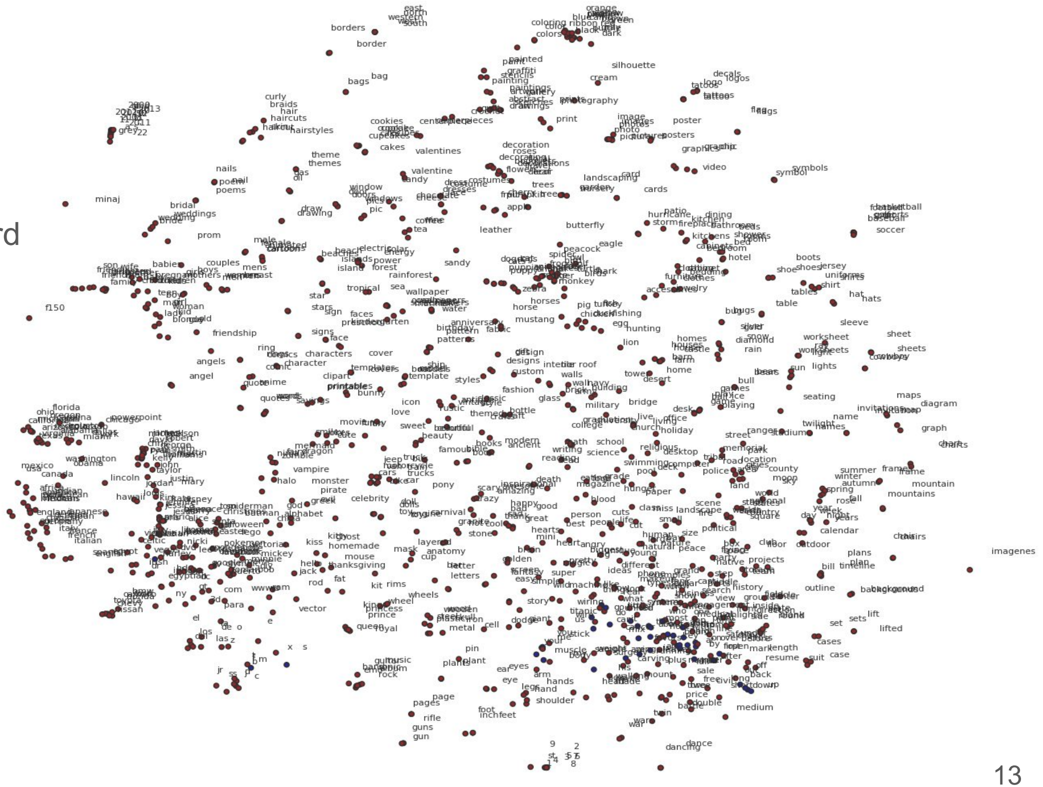


Query Modeling

Word representation- word2vec model pre-trained on 100B words Google news dataset

1. Average model as in [Liu et al CVPR 2015]
2. LSTM model
   a. Memory network used for sequence modeling
   b. Learns the importance of each query word
   c. Takes input as a sequence of words
   d. Yield a fixed length output at the end of sequence input





Parameters

Convolutional Layers	
Learning Rate	0.1
#Convolutional Layers	5
#Fully Connected Layers	3
Output Dimension	300
Weight Regularization: λ	0.001
Dropouts in Fully Conn. Layers	0.5
Batchsize	128

Long Short Term Memory	
Learning Rate	0.01
#Hidden Layers	1
Hidden layer dimension	300
Output Dimension	300
Weight Regularization: λ	0
Dropouts in Hidden Layers	0
Clip Gradient	5

Dimensions of Latent semantic embedding space: 300

Training Data: MSR Clickture dataset

MSR Clickture Dataset

- One year bing image search data of query, image and clicks
- <query, image+, image->
  - Triplets extracted for training
  - Image+ - maximally clicked images for the query
  - Image- - any random image with cosine similarity of queries <0.8

Dataset	
Unique #queries Unique #Images #queries (>1 inst)	73.6M 40M 3.85M

- All words used for word2vec vocabulary learning

Visual Semantic Embedding

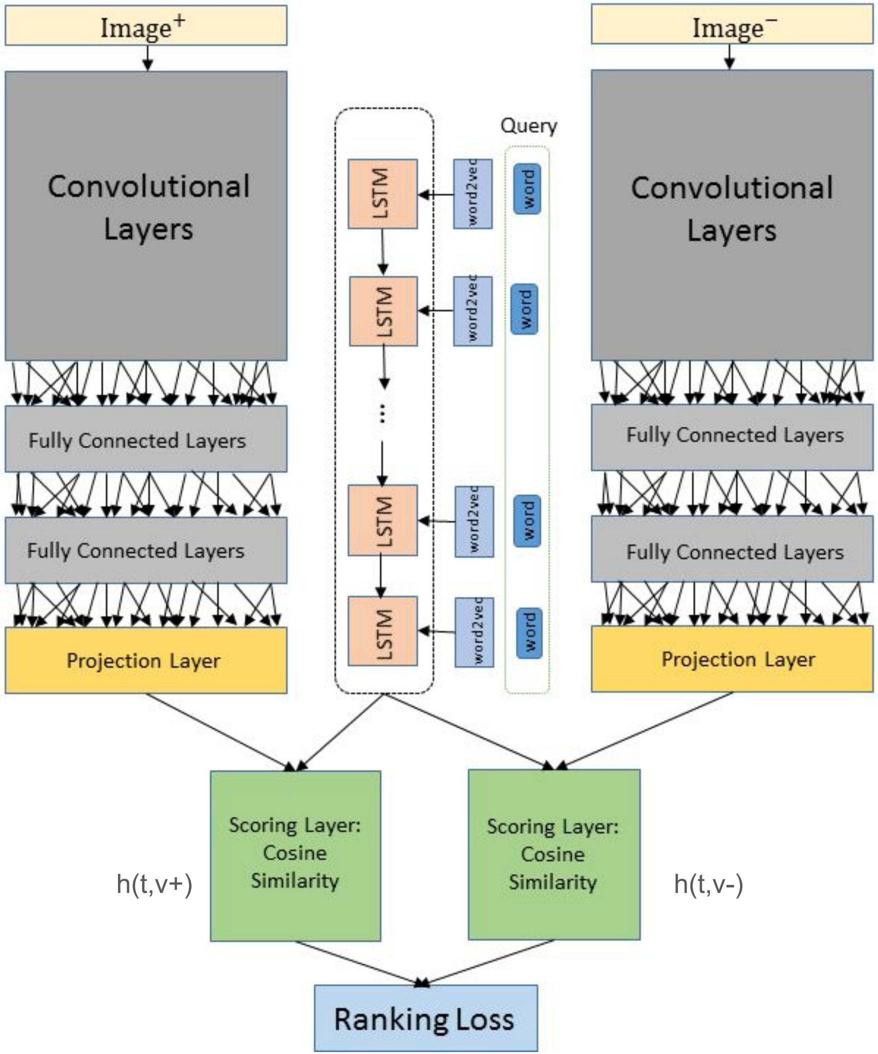
Q = {f(q(i))|i=1...#Query words}

f: word2vec

$$t = LSTM_{\theta_l}(Q)$$

$$v = W_m(CNN_{\theta_c(I)}) + b_m$$

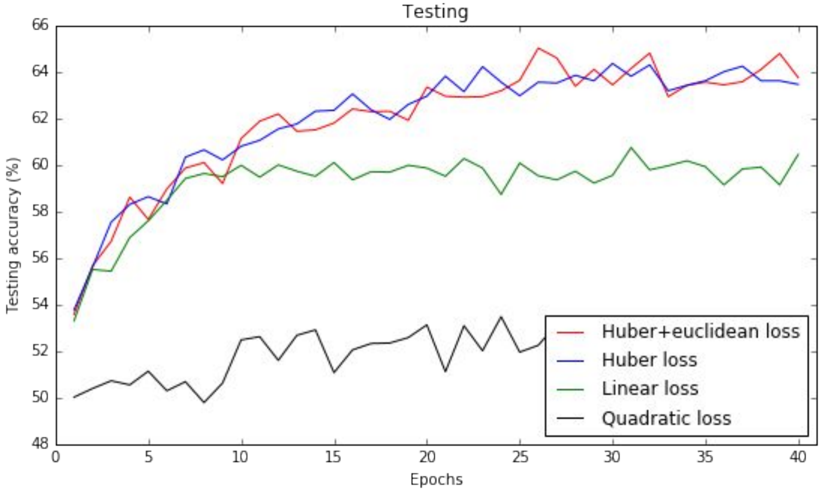
$$h(t, v^+) > h(t, v^-)$$

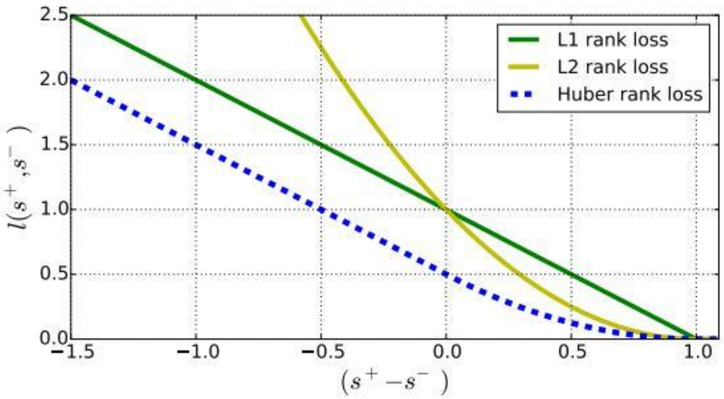


Loss Function Comparison

1. L1 rank loss
2. Huber loss

$$l_p(t, v^+, v^-) = max(0, \gamma - \hat{\mathbf{v}^+}\hat{\mathbf{t}} + \hat{\mathbf{v}^-}\hat{\mathbf{t}})^p$$





Taken from [Gygli et al. CVPR 2016]

$$l_{Huber}(t, v^+, v^-) = \begin{cases} \frac{1}{2} l_2(t, v^+, v^-), & \text{if } u \leq \delta \\ \delta l_1(t, v^+, v^-) - \frac{1}{2}\delta^2, & \text{otherwise} \end{cases}$$

Evaluation

- 749 query-video pair from MSR Evaluation dataset
- For each video, 20 candidate thumbnails extracted using video attributes
- Each thumbnail is labelled: VG, G, F, B, VB (V:very,G:good,B:bad,F:fine)
- Hit@1: hit ratio for the highest ranked or first selected thumbnail

Mean Average Precision:

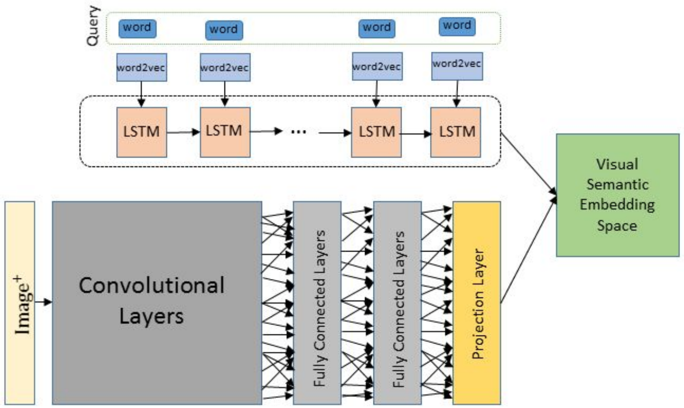
$$MAP = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} Precision(R_{jk})$$

|Q|: Query set

m: Candidate Thumbnails

Precision(R): Average precision at the position of returned kth positive thumbnails

Experiments

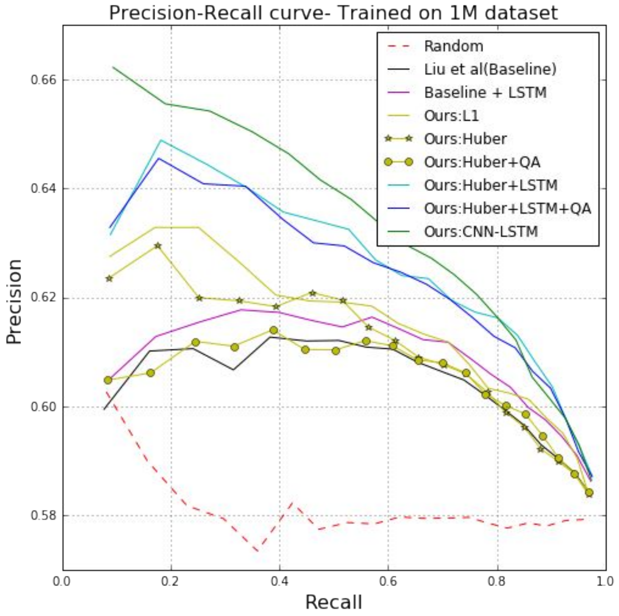


- Baseline
  - Average query word [Liu et al.]
- Linear projection + LSTM
  - All layers of VGG are kept unchanged
  - LSTM trained from scratch
- CNN-LSTM
  - Projection layer is learnt finetuning the previous fully connected layers of VGG
- Query Agnostic Model
  - Rank frames that are aesthetically close to a photograph higher than an ordinary less composed video frame

Performance Evaluation

1. MSR Evaluation Dataset
   749 query-video pairs

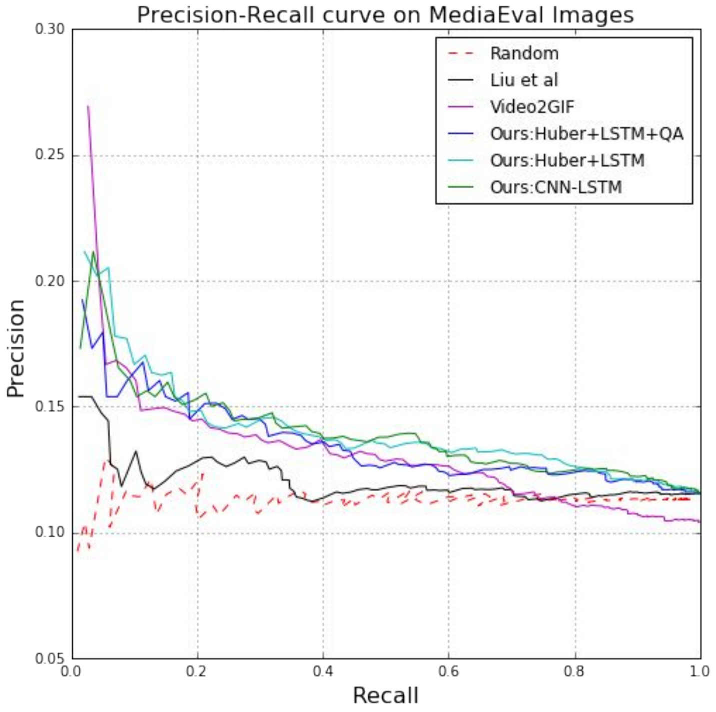
Method	Hit@1:VG	Hit@1:VG/G	Correlation	mAP
Random	28.2 ± 1.5	57.17 ± 1.5	-	-
Liu et al. (Baseline)	33.00	59.81	0.112	0.603
Baseline+LSTM	32.03	60.48	0.138	0.607
Ours L1	32.42	62.61	0.139	0.611
Ours Huber	32.61	62.21	0.132	0.608
Ours Huber + QA	31.83	61.81	0.149	0.603
Ours Huber + LSTM	32.42	63.15	0.178	0.621
Ours Huber+LSTM+QA	35.93	63.28	**0.183**	0.619
Ours CNN-LSTM	**37.11**	**66.22**	0.179	**0.626**



Performance Evaluation

2. MediaEval Dataset
   52 query-video pairs

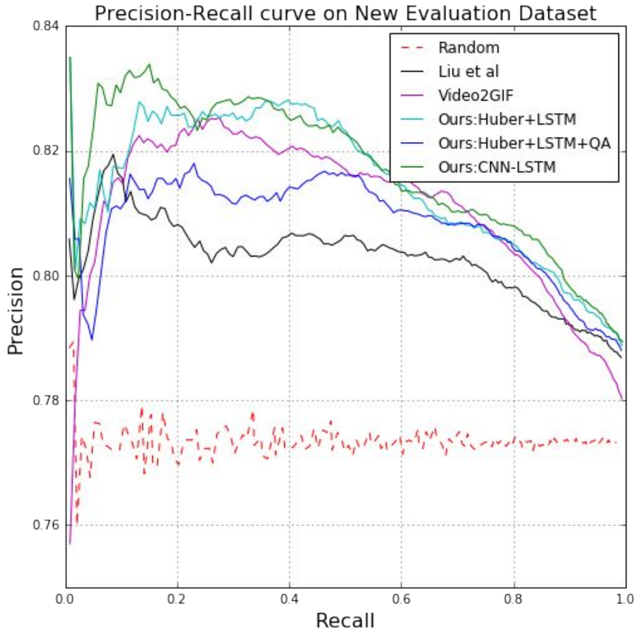
Method	Hit@1:VG	Correlation	mAP
Random	10.48 ± 4.4	-	-
Liu et al. (Baseline)[1]	15.38	0.0217	0.1603
Video2GIF [5]	**25.0**	0.0672	0.1893
Ours Huber + LSTM	21.15	0.0671	**0.1896**
Ours Huber+LSTM+QA	19.23	0.0602	0.1811
Ours CNN-LSTM	17.03	**0.0715**	0.1863



Performance Evaluation

3. RAD Dataset
   100 query-video pairs

Method	Hit@1:VG	Hit@1:VG/G	Correlation	mAP
Random	28.1 ± 4.5	77.05 ± 3.5	-	0.773
Liu et al. (Baseline)[1]	28.12	79.61	0.112	0.80
Video2GIF [5]	28.98	74.76	**0.197**	0.806
Ours Huber + LSTM	29.68	82.52	0.190	0.810
Ours Huber+LSTM+QA	31.25	80.58	0.189	0.804
Ours CNN-LSTM	**35.93**	**82.52**	0.196	**0.812**



Query Relevance Results

Query: Justin Bieber behind the scenes

Query: Chris brown-turn up the music

Query: Rainy September Ride in a Rhino 700

Query: The Best Surprise Military Homecomings

Diversity: Sub-modular maximization

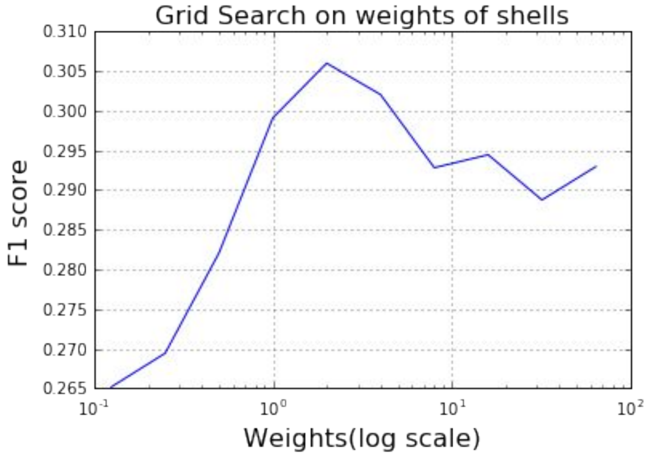
- As in [Gygli et al CVPR 2015], submodular functions need to be defined for relevance and diversity separately.
- Learn the weights for each submodular function and maximize the summarization objective:

$$\mathbf{y}^* = \arg\max_{\mathbf{y} \in \mathbf{Y_V}} \mathbf{w}^{\mathrm{T}} \mathbf{f_x}(\mathbf{y})$$

f1: Relevance Shell

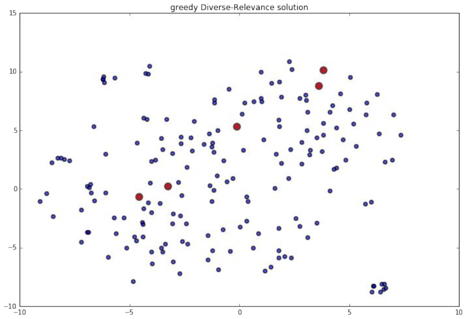
f2: Diversity Shell

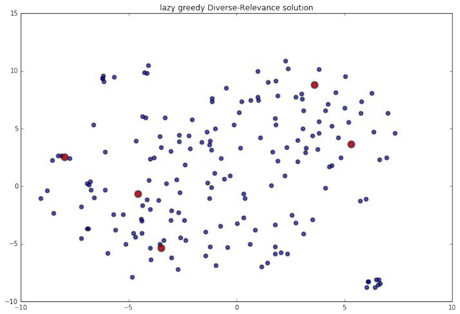
W = [1, 2]



Diversity Results

Query: Anaconda snake





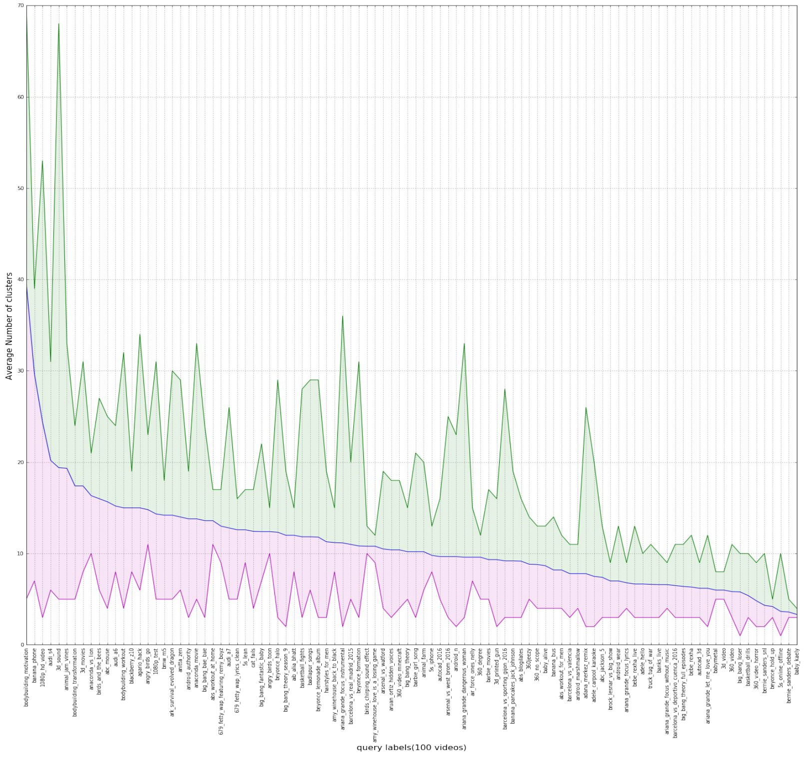
Relevance And Diversity (RAD) Dataset

- Query relevant selected frames are not diverse
- MSR evaluation dataset has less provision for diversity evaluation
- Creating new dataset cater to diversity and relevance in AMT
- Tasks:
  - Data - Uniformly sampled video frames
  - Relevance Task - Annotating each video frame as VG, G, NG, Trash based on its query relevance
  - Diversity Task - Clustering the video frames based on visual similarity
    - # of clusters is arbitrary

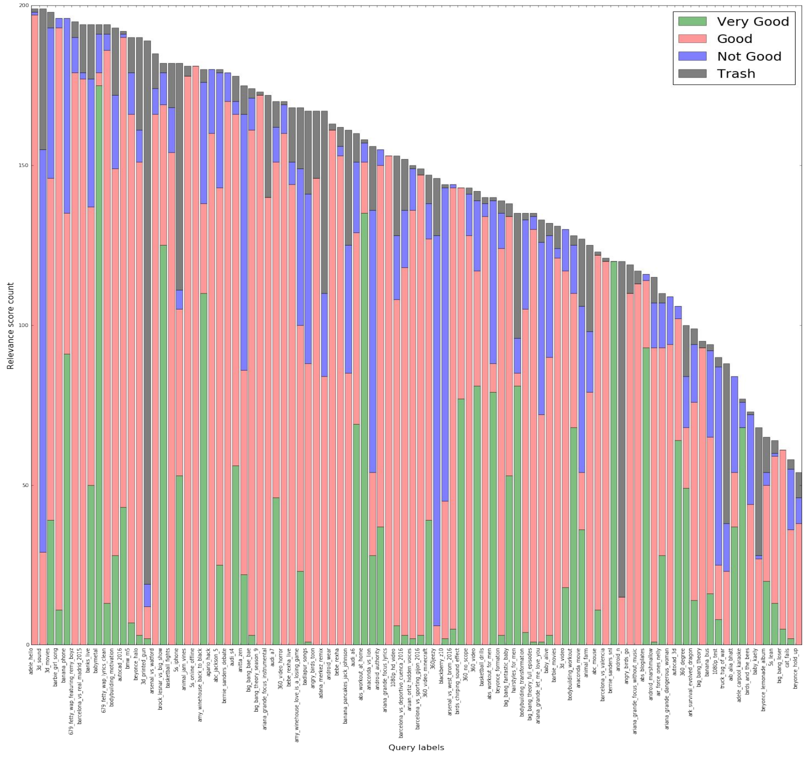
https://people.ee.ethz.ch/~arunv/div_rel_annotator?video_id=cat_fails

Visualize the HIT assignment:
https://people.ee.ethz.ch/~arunv/div_rel_annotator/static/visualize/?visualizeId=3S4AW7T80CP7VW6AK65JFG2VBLTL4L

Relevance And Diversity (RAD) Dataset



Distribution of number of clusters



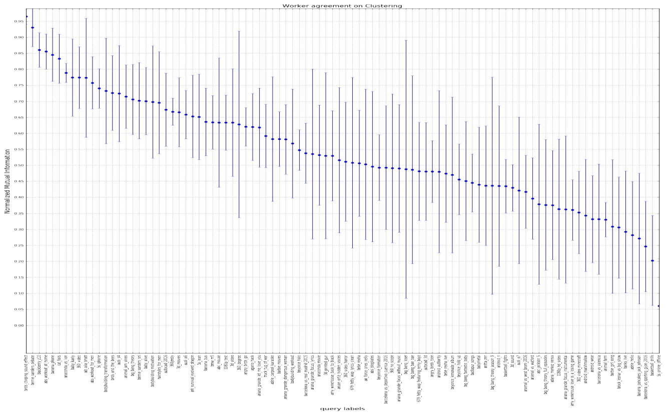
Distribution of Relevance scores over the dataset

RAD Stats

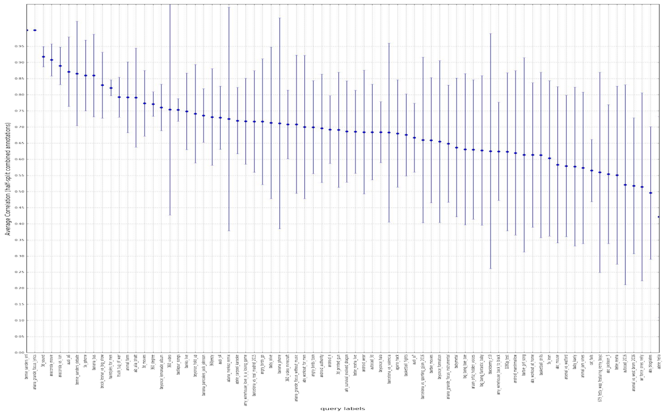
Videos Annotated: 100
Annotators: 48 (Trusted)
Annotations per video: 5
Relevance Labels:
- VG: 16.73%
- G: 61.61%
- NG: 13.58%
- Trash: 8.08%

Avg Spearman’s Rank Correlation scores: 0.69
Avg Normalized Mutual Information: 0.54

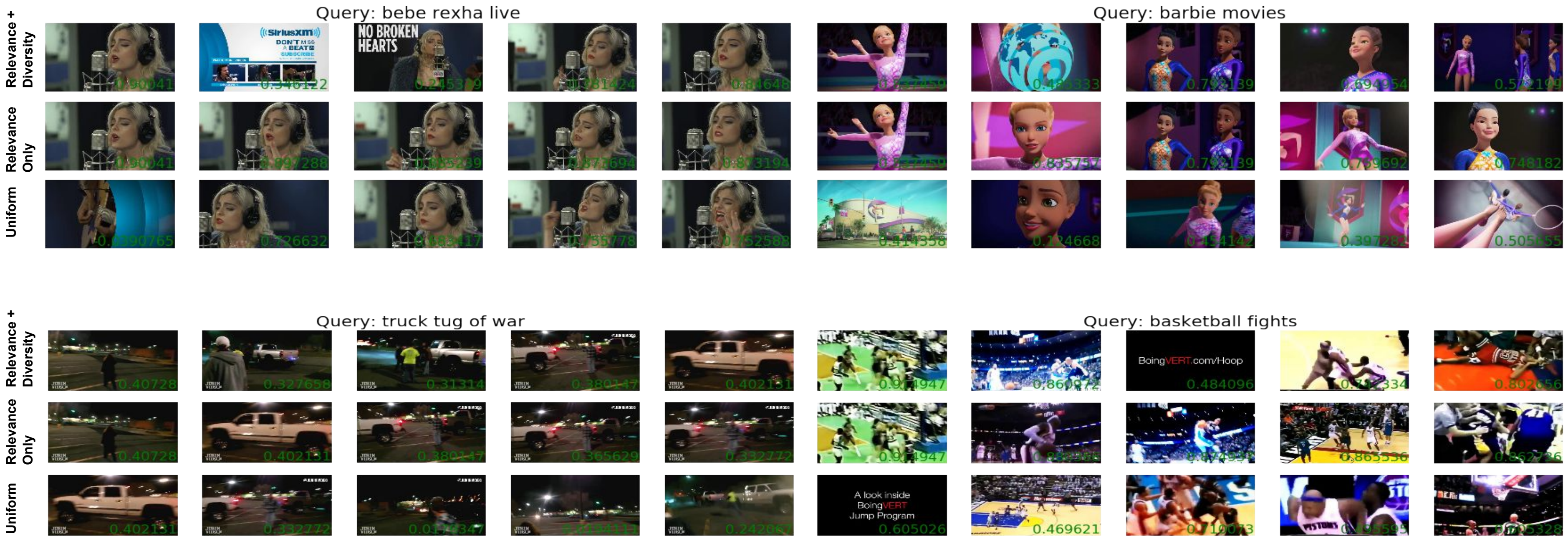
Worker agreement in clustering



Worker agreement in relevance annotations



Qualitative Results



Conclusion

- The **improvements** on Deep visual semantic embedding model using CNN-LSTM architecture and a better objective with training triplets significantly improved our results on the extraction of query relevant thumbnails
- **RAD dataset**- new dataset comprising of 100 query-video pairs with query relevance annotations for all the frames and cluster groupings of the frames based on visual similarity. This dataset caters to the evaluation of selection of diversified set of query relevant thumbnails for videos.
- Query Relevant Video Summarization in form of keyframes- we propose a model based on deep networks and submodular mixtures to make a subset selection of diversified query relevant thumbnails from the video.

References

1. Wu Liu, Tao Mei, Yongdong Zhang, Cherry Che, and Jiebo Luo. Multi-task deep visual-semantic embedding for video thumbnail selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3707–3715, 2015.
2. Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pages 2121–2129, 2013.
3. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105,2012.
4. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
5. Gygli, Michael, Yale Song, and Liangliang Cao. "Video2GIF: Automatic Generation of Animated GIFs from Video." arXiv preprint arXiv:1605.04850 (2016).

Thank you all for your time