Data Analytics
Stiftung Universität Hildesheim
Marienburger Platz 22
31141 Hildesheim
Dr. Josif Grabocka

Seminar Data Analytics : Advances in Deep Learning
Highlight Detection with Pairwise Deep Ranking for First-Person Video Summarization
Semester 2

Abdul Rehman Liaqat

271336, Liaqat@uni-hidesheim.de

Abstract

To extract the useful knowledge data from a video, it is necessary to detect the important segments and summarize the videos. In this report, first detailed working of the one proposed method of detecting important video segments, called highlight detection and summarizing videos with these highlights, is shared. Afterwards, the method is tested against different prevalent methods using one common dataset. Lastly, results, possible merits, demerits, usage and points of improvements of the highlight detection and video summarization method are shared.

Contents

1 Introduction

1.1 Motivation

As they say:

> A picture is worth thousand words

Videos are the biggest source of unstructured Big Data. On youtube only, 300 hours of video per minute is uploaded [4] and these videos are from more than 88 countries and 76 languages. Each video is recorded in different frame rates ranging from 24 to 120 FPS. Only these many facts are enough to prove the variety, velocity and volume of data.
To tap this source of data, it is first necessary to filter out the important data. This filtered out video data, which can be called "Summary of Video" can be later used by both humans and machines.

1.2 Problem statement and goal

So, the problem of video summarization can be thought as :

> Converting a video V1 with total playback time T, to a video V2 with playback time t such that V2 relays **Important information** of V1 completely while keeping:
>
> $T >>> t$

There is not one specific definition of **Important Information** . It is dependent upon the type of the video, quality of the video and audience of the video.There are following three main procedures which can be used to define important information.

1. Key Frame / Shot Based
2. Structure Based
3. Highlight

2 Related Work

2.1 Key Frame or Shot Based

According to this type of procedure, the basic unit of important information is a key frame. Thus each frame (or group of frames) is searched for the specific characteristics. If any frame contains required characteristics, it is extracted and considered as part of video summary. These characteristics can be any of the following:

1. Visual Importance: which means to maximize the diversity of things inside a frame. Each frame is searched for maximum number of objects or change in scenes, the ones containing the most diversity are selected.

2. Important Objects: This one is pretty simple. Each frame is searched for a specific object (or group of object) such as human face.

3. Movement detection: In this case, movement of objects is detected among some consecutive frames. Later on the frames containing the movement are selected as characteristic key frames for video summarization.

2.2 Structure Based

In structure based important information extraction, prior domain knowledge of object under observance is used. For example in a sport video such as football, it is known that a goal is accompanied by crowd cheering and player gathering at one place. So such frames are extracted out and included in the final video summary.

2.3 Highlight Based

In the third kind, which is also the one proposed by the author, a small video segment is assigned a numeric value and the video segments containing the highest values are selected. This numeric value represent the score of highlights in each segment.**Highlight** is nothing but a video segment considered important enough to be part of final video summary. A video segment is nothing but a small clip of actual video.

3 Proposed Approach

Important Information is extracted using Highlight based video segment selection which is later used to create a summary of video. The main part to be solved through deep learning architecture is detecting the highlight segments of a video. Supervised data consists of video segments duly scored by human observers. An architecture is then trained on these segments to score highlighted video segments higher than the non-highlighted ones. Later on this architecture is used predict the highlight score of unknown segments.
As described earlier, learning to **highlight score** of each video segment is the main problem to be solved by deep learning architecture. If considered independently, we can say that highlight score of each video segment will be dependent upon:

1. Spatial Information of Segment: Spatial information means the knowledge embedded in independent frames. In other words, a segment is further distributed into frames and each frame is considered as a picture whose content is understood by one part of the architecture.

2. Temporal Information of Segment: Temporal information represents the knowledge hidden in the interaction of different frames. Thus one part of architecture extracts this information of interactions.

This way the architecture consists of two different streams of sub-architecture which will encode both spatial and temporal information independently and later on output scores for both streams (spatial and temporal). Both scores will be combined and ranked against other segments of the video. Top N ranks will be selected and considered as important information for video summarization.

3.1 Architecture : Prediction

A video is passed through following steps to extract highlights:

1. Division of a video into smaller segments. This division can be uniform such that each video segment is of the same length or non-uniform such that each segment consists of some kind of movement of object under consideration. For example a video is 10 minutes long. It is divided into 10 smaller segments each 1 minute long named $S1, S2, ..., S10$.

2. Each video segments is fed to both streams in parallel.

3. Spatial stream consists of "AlexNet" architecture [1], Average pooling and fully connected layers orderly. "AlexNet" is pre-trained and fully connected layers are trained. There are total of 6 fully connected layers each named as $F1000, F512, F256, F128,$ $F64, F1$ and contains $1000, 512, 256, 128, 64, 1$ number of neurons respectively. The final $F1$ layer will output one value. Intuitively, it can be said that AlexNet and average pooling is used to encode any segment containing M frames into M vectors where each vector has the dimension of $\mathbb{R}^{1000}$. Each vector, after entering fully connected layers, will be transformed into one value at the output based upon the values of the input vector.

4. Temporal stream, similar to spatial stream, has three parts: $C3D$ architecture [2], Average Pooling and fully connected layers.Temporal stream exactly same width and length of fully connected layers as spatial stream. Temporal stream will also output one value for each segment.

5. Output of both streams are simply averaged with some constant ratio which acts as a hyper-parameter. Intuitively, it means that we assign specific weight value to each stream such that the one having more say in the selection of highlight has higher value. For example for one person videos, highlights are more dependent upon the interaction among consecutive frames hence temporal stream is given more weightage. More specifically the final highlight score for a segment s_i will be:

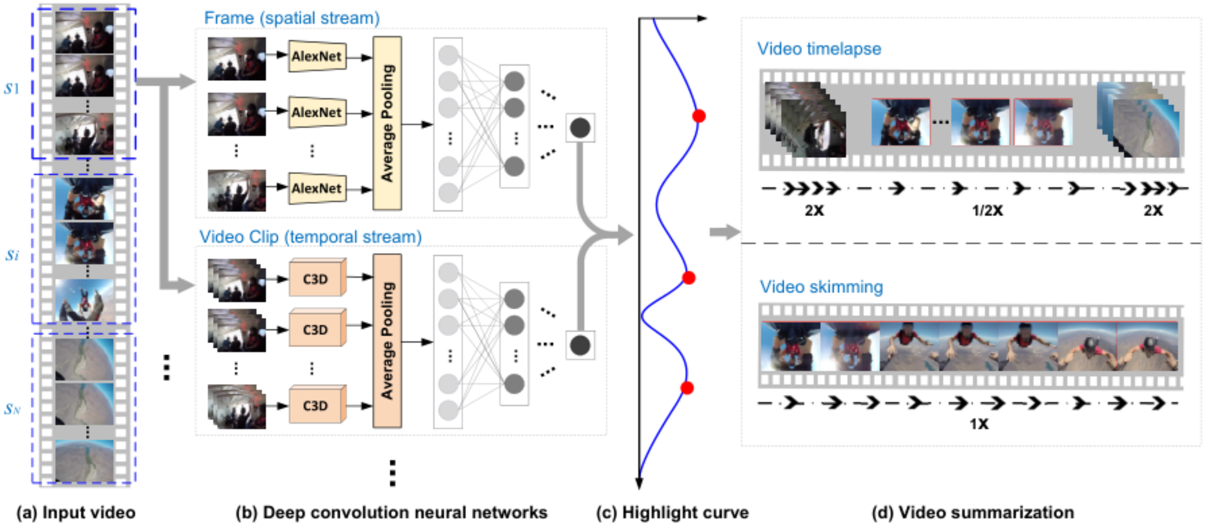
$$Score(s_i) = (1 - \omega) \times SDCNN(s_i) + (\omega) \times TDCNN(s_i)$$

Where $Score(s_i)$ is scoring function, ω is weighting variable, $SDCNN(s_i)$ is spatial stream output function (or spatial deep convolutional neural network) and $TDCNN(s_i)$ is temporal stream.

6. Now that each segment has one highlight score assigned to it, segments with high score are considered as highlights of the video and used to create summary of video.

7. Lastly, two different types of video summaries are created from the detected highlight segments. These two types are: Video Timelapse and Video skimming. As names suggest, video skimming summary only contains highlighted segments while in video timelapse highlighted segments has way slower playback time than non-highlighted ones.

Same steps can be shown in figure form as following:



(a) Input video (b) Deep convolution neural networks (c) Highlight curve (d) Video summarization

Figure 1: Prediction and video summary creation steps [3]

3.2 Architecture : Training

As described in the section above, only fully connected layers are trained to output numerical value of each video segment such that value of highlight segment is higher than non-highlight segment. This means our objective function will be a pairwise ranking loss. Similarly, each stream acts independently to score the highlight value of each segment till the very second last step, thus both streams can be trained independently with same training data and objective function. More specifically objective function is margin ranking loss for each pair which can be formally written as:

$$min \sum_{(h_i,n_i) \in TrainingData} max(0, 1 - f(h_i) + f(n_i)) \quad (1)$$

Training of the architecture goes through following steps:

1. Since each stream is trained independently, both have their own differences while training. For both streams one video segment is taken into consideration at a time. In spatial stream, each video segment is further divided into frames and each frame is fed as input one by one while in temporal stream, the whole segment is fed as input.

2. Each frame and video segment will pass through AlexNet architecture and C3D architecture for spatial and temporal stream respectively.

3. After both AlexNet and C3D a layer of average pooling is used. This average pooling layer will fuse all kind of frames into one vector of length 1000. Thus for each segment fully connected layers has only one vector. For both AlexNet and C3D pre-trained values are used.

4. Now we are at the stage where actual training is required. Let's take spatial stream first. Our training data is in the form of pairs of highlight and non-highlight segments. Each highlight and non-hilghlight member of pair is fed to two parallel spatial streams. In the forward pass, both parallel spatial streams will create one value each. This value is fed to margin ranking loss function described above, which is minimized by back-propagating error of wrong ranking.

5. Dropout with a probability of 0.5 is used to avoid over-fitting.

6. Since only last 6 fully connected layers are trained, it is a rather fast process.

7. Temporal stream also goes through exactly same kind of training.

8. After training, each stream can predict highlight score for any highlighted segment h_i and non-highlighted segment n_i in the training data such as:

$$Score(h_i) > Score(n_i), \forall (h_i, n_i) \in TrainingData$$

4 Video Summarization

Now that highlight score for each segment is achieved. The step of creating video summary is rather easy. As shared before there are two following methods of video summarization used:

- Video Skimming: In this case, those video segments are selected which maximize the following function.

$$max_b \sum_{i=1}^{c} b_i f(s_i) \;\; s.t. \sum_{i=1}^{c} b_i |s_i| \leq L \tag{2}$$

- Video Timelapse: In second option, all video segments are included but highlighted segments have slower playback rate. To create a summary of playback time L, playback rate r for non-highlight segments or $\frac{1}{r}$ for highlight segments can be calculated as following:
  $r = [\frac{L}{2L_h}] + \sqrt{Y}$
  where $Y = \frac{L^2-4L_vL_h+L_h^2}{L_{4h}^2}$
  and $L = Playbacktimeofvideosummary$
  $L_v = Playbacktimeoforiginalvideo$
  $L_h = Originaltotalplaybacktimeofhighlightedvideosegments$

5 Experiments

5.1 Data

To train, test and compare the proposed methods with other methods, data was collected from youtube. More specifically, Youtube videos which belong to any of the categories, filmed by the player himself, filmed using GoPro and there were no editing traces were selected. Each category had 40 videos and each video was 2 minutes to 15 minutes long.

Further following steps were taken to create highlight and non-highlight pairs:

1. Split each video into five seconds segments which will be marked by human evaluators as
   - Highlight (Score 3)
   - Normal (Score 2)
   - Boring (Score 1)
2. Total of 12 evaluators were invited. Each evaluator had different educational background and was outdoor sports enthusiast.
3. Any video was marked by three different evaluators.
4. Only segments with total score equal or greater than 8 was selected as highlights.
5. Lastly, test and training datasets were created from the evaluated segments. The dataset for highlight detection is in the form of pairs of highlight and non-highlight segment. Total 105k pairs were formed.

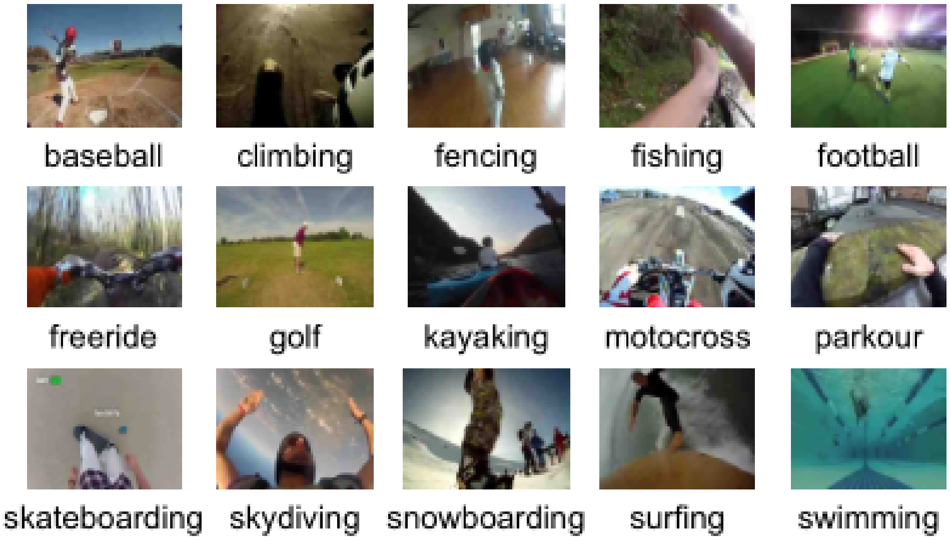


Figure 2: All 15 video categories with representative frames [3]

6 Results

Each comparative model was tested on the same test data which was collected using the procedure given above.

6.1 Results for Highlight detection

To measure the quality of highlight detection following different approaches were tested against the proposed method:

- Rule-Based:First color change based division into segments, then further segmentation on motion. Sub-segments with the most length are highlights. This method is called **Rule** in short form.

- Importance-Based: Two different sub-methods based upon the pre-processing of the video segments. First which is called **Imp+IDT**, it is a linear SVM with the input as the vector obtained through improved dense trajectories motion features and PCA on top of it. Second is named **Imp+DCNN**: In this method input is the average of output vector of AlexNet of each video frame in every video segment. Linear SVM classifier is used to detect highlighted segments afterwards.

- Latent Ranking Based: Same as before but instead of linear SVM, latent ranking based SVM is used. Method with input from improved dense trajectories motion features is named as **LR+IDT** while the input from DCNN is named as **LR+DCNN**

- Deep Convolution Neural Network: Proposed method but further tested for each component architecture. In other words, testing results on a single spatial stream is named as **S-DCNN**, on a single temporal stream is named as **T-DCNN** and the whole proposed method as it is named as (Temporal Spatial DCNN) **TS-DCNN**.

To further reduce the size of data for processing, only three frames per second are chosen, thus a five second video segment has only 15 frames. In total there are $105K$ pairs in the training set. Three following different evaluation metrices are used:

- mAP: Mean average precision of detecting highlights on the test data.
- NDCG@1: Due to the ranking problem, Normalized Discounted Cumulative Gain (NDCG@d) for the ranking depth d, is also taken as evaluation metric to consider the multi-level highlight score. NDCG@1 is the average NDCG of all video segments in test data with $d = 1$
- NDCG@5: Similarly, NDCG@5 is the average NDCG of all video segments in the test data with $d = 5$

After using test data with all described approaches, following comparison graph is obtained:

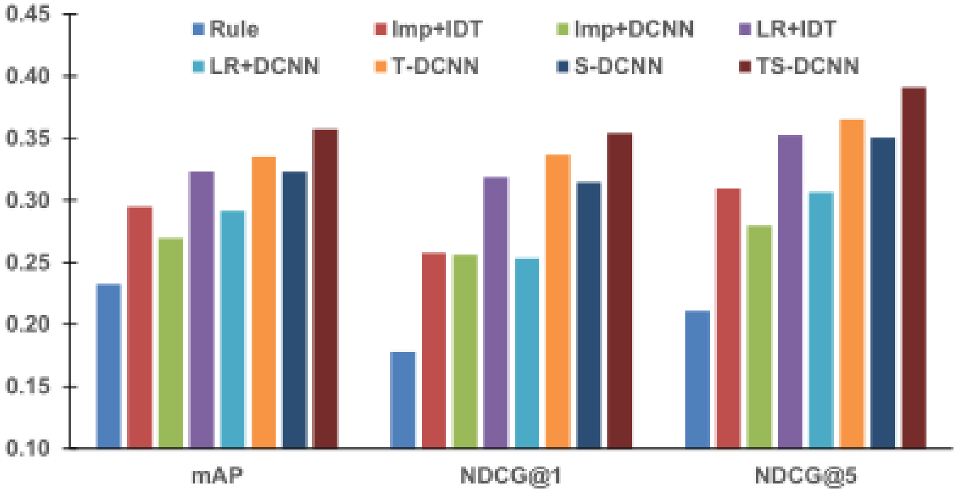


Figure 3: Performance comparison of different approaches on different evaluation metrices [3]

If we look at the performance of each method on different evaluation metrices, most of the time following trend is seen:

$$TS{-}DCNN > T{-}DCNN > S{-}DCNN \geq LR{+}IDT > Imp{+}IDT > LR{+}DCNN > Imp{+}DCNN \tag{3}$$

which implies that:

1. Proposed TS-DCNN, T-DCNN and S-DCNN are outperforming other methods across all evalution methods.
2. T-DCNN is performing better than S-DCNN showing that in the given dataset capturing interaction among different consecutive frames is more important than static spatial features.
3. LR(latent ranking SVM) methods are always performing better than Imp(Importance based Linear SVM) meaning that the problem is inherently a ranking problem.

Another way to prove that T-DCNN is more useful than S-DCNN is the following graph which shows TS-DCNN with different ω values across different evaluation metrices and showing best results at $\omega = 0.3$ proving that to get the best results 70% of T-DCNN and 30% of S-DCNN is needed.

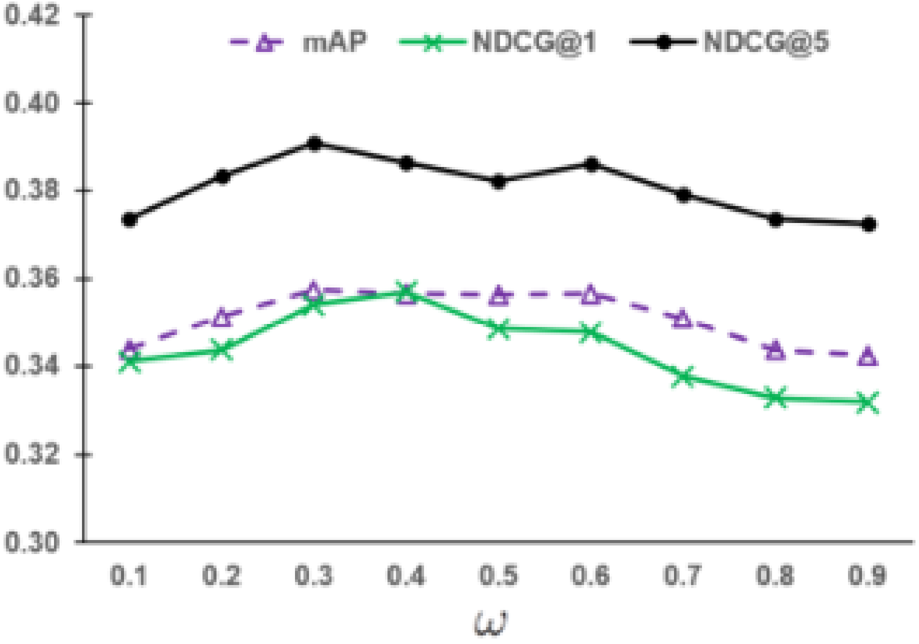


Figure 4: Change in performance of TS-DCNN with different values of ω [3]

One more thing is the prediction timing. Even after such a performance increase prediction is done in reasonable timing. The prediction timing for 5 minutes original video with Rule,LR+IDT,LR+DCNN,S-DCNN and TS-DCNN is 25s,5h,65s,72s and 360s respectively.

6.2 Results for Video summarization

Just like Highlight detection, video summarization was also compared against different following methods:

- Uniform Sampling: shortened as UNI, As name suggests in this process every Kth frame is taken to create a video.
- Importance-driven Summary: shortened as IMP, in this method highlights are detected using Imp+DCNN and later used in creating video summary.
- Interestingness-driven Summary:shortened as INT, this summary is created by measuring the highlight score of each segment by summing quality, motion and person detection features of each frame.
- Highlight-driven Summary: This is the proposed method of detecting highlights. After scoring highlights, two different summaries are created by video skimming and video timelapse; also both are shortened as HD-VS and HD-VT respectively.

Since video summarization is highly subjective as described in the introduction section, video summarization is marked by human evaluators. During the process, each evaluator will watch the original video then watches all the summaries. Later on each evaluator will mark summaries against following two criteria:

1. Coverage: meaning the progress of video along time.
2. Presentation: meaning the gist idea of video.

The result as the percentage of users who prefer HD-VT and HD-VS over all other methods can be seen in 1 and 2 respectively. Obviously, proposed highlight detection based methods are always preferred over others. In comparison of HD-VS (video skimming) and HD-VT (video timelapse), HD-VS is preferred when it comes to coverage but HD-VT is selected for presentation. Reason for such results is also pretty obvious as video skimming contains only the most important segments thus easier to convey the gist of whole video. On the other hand video timelapse summary has the whole video hence easier to understand the progress of events and so preferred for presentation.

—	UNI	IMP	INT	HD-VS
Coverage	91.4%	80.1%	74.3%	68.6%
Presentation	85.7%	60.2%	64.8%	34.3%

Table 1: % of users preferring HD-VT over other methods [3]

—	UNI	IMP	INT	HD-VT
Coverage	87.2%	77.1%	71.4%	31.4%
Presentation	88.6%	74.3%	82.9%	65.7%

Table 2: % of users preferring HD-VS over other methods [3]

7 Discussion

With the evaluation experiments and their results, it has been proved that the proposed method of highlight detection is outperforming other prevalent methods. Also human observers prefer video summaries created from the highlights detected using proposed methods and created using video timelaps and video skimming methods.

7.1 Apparent good points of algorithm

Upon examination, following good points were found in the proposed method:

- Modifiability: The proposed method of highlight detection is very easy to modify. First possible modification is the pre-trained models (AlexNet and C3D). If any new better architecture like both is available, both can be easily replaced. Secondly, if any new type of data stream is to be considered,it can be easily added to the existing ones. For example, if audio of the given videos are also to be considered while detecting highlights, a new stream with the similar architecture to other streams can be added.

- Adaptability to subject of video: Even for videos in the same categories, such as the first person of sports categories, can have different performance optimization value of ω. This means that with the change in subject of videos it is highly likely that different value of ω (which actually describes the percentage important of spatial and temporal stream on highlight detection) is required. With the proposed method, a suitable weighted score of both streams can be selected by changing the value of ω thus making this method highly adaptable according to subject of videos. This way generality of video summarization across all types of videos can be achieved.

7.2 Possible negative points of algorithm

Just like good points there are possible negative points and points of confusion in the research paper and proposed method which are:

- In the temporal stream, video segments are provided in the input of C3D as it is but according to the provided architecture image multiple video segments are provided. Since only one video segment is considered at a time, thus the reason of confusion which is not yet clear.

- Secondly the reason of selecting marginal ranking loss as objective function is not clearly defined. Clearly, given function is not continuous while a better continuous objective function such as pairwise ranking loss can be used which is also better approximation to NDCG@d.

- Lastly, the evaluation is highly subjective and provided with a few numbers of human evaluators. Both the number and type of evaluators should be increased.

7.3 Possible usage

From a usage point of view, there are a few fields where video summarization can be of huge value. Few are as following:

- Video Captions: Video skimming summary as video captions can help convey the message of the video without actually watching it completely. Thus saving time of the user and highly useful features of websites containing videos.

- Security: In case of theft, it is very easy to detect the possible important video segment instead of watching the footage of whole recording.

- Running highlight detection: It is possible to use the highlight detection method to detect only the important part of any live video recording and making decisions on the basis of selected segments while system is in run-time. For example in autonomous driving, it is time consuming to extract each piece of information from vast amount of incoming data. Instead if each incoming video frame is detected for highlights, only useful video frames will be passed through for further processing.

7.4 Possible future work

Video highlight detection and summarization is a highly active and vast area of research. The proposed method also has a few areas of further research. One area is to test the method across different types of videos based upon categories. Other is to include more input resources such as audio to check any improvement in the highlight detection. Also the proposed method should be tested with different objective functions especially with pairwise ranking loss.

8 References

‘

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[3] T. Yao, T. Mei, and Y. Rui, “Highlight detection with pairwise deep ranking for first-person video summarization,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 982–990.

[4] youtube. (2017) Youtube by the numbers. [Online]. Available: https://www.youtube.com/yt/about/press/