RANKING VIDEO SEGMENTS WITH LSTM AND DETERMINANTAL POINT PROCESSES

Juan Liu[†] *Zhengyang Wu*[⋆] *Fuxin Li*[†]

Oregon State University[†] Magic Leap[⋆] Oregon State University[†]

ABSTRACT

This paper presents a novel approach to ranking unsupervised video segment proposals. This is important in order to obtain a small set of object proposals that cover diverse objects in the video. This paper utilizes a convolutional-recurrent network framework for learning a score function of video segments. A modified Determinantal Point Processes (DPP) approach is used on the results to sample a diverse and meaningful set of objects from the video. Given a pool of video object proposals, our approach can rapidly find a small and diverse set of objects which could then be used for subsequent processing. Results on the challenging VSB-100 dataset demonstrate the performance of our approach.

Index Terms— Video Segmentation, LSTM, DPP

1. INTRODUCTION

Video segmentation aims to locate and track object silhouettes in a video. It is an important problem with applications including mobile robotic guidance, robot navigation and manipulation, healthcare monitoring and surveillance, video search, non-rigid structure from motion and many others. Interest in video segmentation has increased in recent years and as a result, much progress has been realized along with advances in both methods and datasets [1, 2, 3].

We focus on a particular video segmentation task, the problem of unsupervised object discovery, where the video is given with no annotations and an algorithm is supposed to find consistent objects in the video. This is related to the object proposals problem [4, 5, 6, 7] in image segmentation. However, because of the intrinsic ambiguities in images, $100-500$ proposals are required to cover most objects, which could be excessive for subsequent processing. Because of the strong temporal consistency and motion continuity in videos, there is hope to greatly reduce the number of proposals while still cover as many objects as possible.

This paper explores such prospects by training an "objectness" regressor on video object proposals to rank them. Such a score function would allow selecting a small subset of video object proposals based on the long-term behavior of each proposal, such as temporal consistency of appearance and motion, as well as boundary coherence. Different video object proposals have different lengths (number of frames), and hence it is difficult to represent them with a fixed-length feature vector. Therefore, we adopt a recurrent network, namely the Long-Short Term Memory[8] model, in order to directly map input proposals of different length into a single output.

A long-standing problem of objectness ranking is the difficulty in outputting a small and diverse set of proposals. This problem manifests because similar proposals are likely to generate similar ranking scores, hence when one selects a subset of proposals based on top scores, the algorithm is very likely to select dozens of very similar proposals, which is not aligned with our goal to discover as many objects as possible from an unannotated video. Approaches have been proposed such as the maximal marginal relevance[4] or the diverse M-best inference [9]. However they mostly rely on greedy approaches and have problems especially when all the proposals that are sufficiently different from each other have been selected at least once. In order to solve this problem, in this paper we experiment with Determinantal Point Processes (DPP), the random point process for modelling repulsion. DPP is a natural model for subset selection problems where diversity is desired [10]. However as far as we know, this is the first time DPP is used in ranking object proposals in images or video. Our experiments show that DPP outperforms other approaches for diversification, including direct ranking, random selection and MMR-diversification.

DPPs are stochastic and do not necessarily select the top-scoring segments at the first few samples. This leads to sub-optimal performance when the number of selected proposals is very small. We propose a deterministic version of DPP which is based on greedily select the DPP solution with maximal likelihood. Such a simple modification to the original DPP led to significant performance improvements, especially when only $1-50$ proposals are required to be selected.

2. RELATED WORK

Video segmentation has received a lot of interest in recent years[1, 11, 12, 13, 14, 15]. The definition of specific video segmentation problem varies. This paper focuses on the problem definition of creating overlapping video object proposals to accommodate as many object definitions as possible, which have been pursued in e.g. [16, 3, 14].

Objectness was proposed in [17, 18] as an approach to measure whether a bounding box or segment closely resem-

bles a natural object, without using category-specific information. Objectness is usually trained using a classification ($\geq 50\%$ overlap with ground truth or not) or regression on the intersection-over-union (IoU) overlap metric between the object and the ground truth. There has been relatively few works on training objectness scores on video segments, due to their non-uniform length. [19] trains a ranking function on video segments for ranking them, but it works only on moving objects, there is no diversification and only CNN is used.

Deep image and sequence recognition techniques, in the forms of convolutional neural networks and recurrent neural networks, have greatly improved in the last few years[20, 21]. In video processing, a current trend is to combine convolutional layers with recurrent layers to deal with the variable length in videos [22]. Approaches have been proposed for action recognition [23], video captioning [22], etc. Less success has been achieved in video segmentation, where the high spatial precision required are at odds with many deep models that are not be spatially accurate enough.

Diversification has been a major issue in information retrieval for a long time[24], dating back to the creation of search algorithms (such as PageRank). Maximal marginal relevance [24] has been the go-to approach for many years. This approach decreases the score of an object if it has a large similarity with any of the objects that have already been retrieved. It was used in [17] and other subsequent work. However, because it greedily decreases the score, after each diverse object has been selected at least once, scores across all proposals will receive similar penalties, limiting the diversification capabilities of the algorithm. DPP has recently been proposed as an alternative approach to this and has been used in the information retrieval literature. In computer vision literature, DPP has also been used recently in other contexts such as [25, 26], but never in the video segmentation context. [27] jointly trains an LSTM and DPP in the video summarization setting and saw improvements in performance.

3. TECHNICAL APPROACH

In this section, we first review the LSTM and DPP approaches, then introduce the LSTM-DPP model learned from the training dataset overlap scores. Furthermore, we apply overlap prediction scores on each video in the test dataset, and finally combined with our modified DPP algorithm to obtain segmentation in diversity.

3.1. Long Short Term Memory Networks

Long Short Term Memory Networks were introduced by [8], in which the main framework is the same as RNNs, but the internal circular module is designed in a different structure. The key to LSTM networks is the cell state and four controlling gates which are called as the forget gate, the input gate, the update gate and the output gate.

At time t, the update formula is

$$\begin{aligned} f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\ i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t &= tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \\ C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t \\ o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t &= o_t * tanh(C_t) \end{aligned} \quad (1)$$

where $\sigma(\cdot)$ is the sigmoid function. This takes into account both the memory at the previous timestep C_{t-1}, the forget gate f_t, the current "gradient" to the memory $\tilde{C}_t$, and the input gate i_t. If f_t is 0 and i_t is 1, the previous memory is "forgotten" and the memory is decided from the current input only. On the other hand, if $f_t = 1$ and $i_t = 0$, the current input is deemed useless and the memory is not updated. The flexibility of such an approach and the stability in the gradient from having a fixed C_{t-1} makes LSTMs much easier to train than conventional RNNs and usually obtain better results, making them a better choice in many recurrent models.

3.2. Determinantal Point Processes

Determinantal Point Processes was first proposed for machine learning algorithms by Kulesza and Taskar in [10] as a useful tool to find diverse sets of high-quality search results based on an elegant probalilistic model. DPP is formulated as such: Assume the goal is to select a maximally diverse subset from a finite set $\mathcal{Y} = \{y_1, \ldots, y_N\}$, a point process $\mathcal{P}$ is defined as a probability measure on $2^{\mathcal{Y}}$, the set of all subsets of $\mathcal{Y}$. For every $A \subseteq \mathcal{Y}$, $\mathcal{P}(A \subseteq Y) = det(K_A)$, where the symmetric $N \times N$ matrix $K = [K_{ij}]$, with $K_{ij} = k(y_i, y_j)$ describing the similarity between elements y_i and y_j, and $k_{ii} = \mathcal{P}(i \in Y)$ is the marginal probability of y_i. This prompts us to select a subset which leads to a high principle minor. As an extreme case, DPP will never select the same element twice as the determinant of the minor goes to 0, which means the probability $\mathcal{P}(i \in Y) = 0$.

3.3. Our LSTM-DPP Approach

The following framework (Figure 1) describes our algorithm that combines LSTM and the Determinantal Point Process. We extract video segment proposals using the approach described in [14], which gives us several hundreds of proposals per video. These video proposals contains one binary figure-ground mask for each frame the object is present in the video. For each mask, we obtain its bounding box, resize it to 224×224 and run it through a pretrained VGG-16 model to convert it into a 4096-dimensional vector by extracting the second-to-last layer features. The vectors for all the frames where the object is present are then sent to the LSTM in a temporal sequence to predict an overlap score S_i at the end of the video segment (Fig. 2).

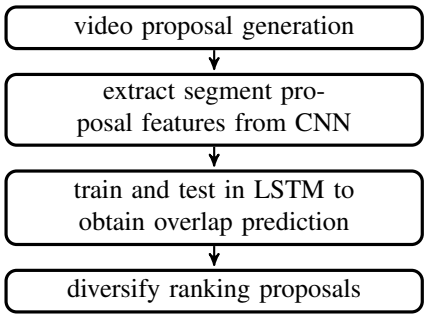

Fig. 1: The main algorithm for ranking video segments

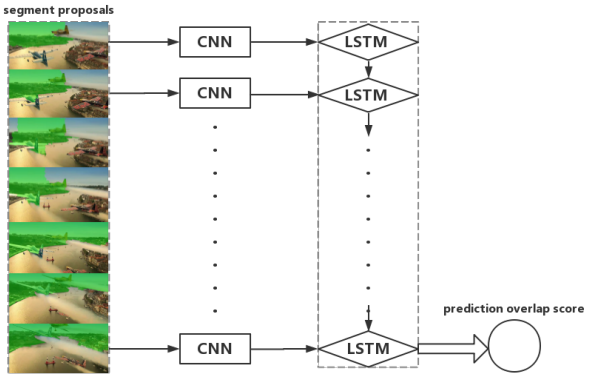

Fig. 2: Our algorithm initially takes video segment proposals (generated from a video segmentation algorithm) from the training set as input and uses a pretrained CNN model [28] to extract convolutional features for the segment. The result is taken as input to the LSTM, which learns to predict the overlap between each segment and ground truth. During testing time, each proposal is first tested on CNN+LSTM, then a determinantal point process algorithm uses the prediction scores to generate a set of diverse and high-quality segmentations.

The scores S_i on each video segment are then used as the marginal probabilities of inclusion for individual segments in the DPP framework. We compute an overlap matrix O between all video segment pairs by averaging their IoU overlaps on each frame. For the frames where the object is not present, the overlap on that frame is 0. Given the overlap-matrix O and predicted score vector S, we define our specific K-matrix as following: $K_{ij} = O_{ij}\sqrt{S_i}\sqrt{S_j}$, with $i,\ j \in N$. We can rewrite it as: $K = S^{\frac{1}{2}}OS^{\frac{1}{2}}$.

Such a DPP kernel matrix will tend to select proposals that generate a high overlap score from the LSTM, while at the same time attempts to sample diverse proposals. We make one further modification to the DPP algorithm based on our initial experiments, which shows that while the original DPP algorithm can sample a diverse subset of video segments, it is not always able to capture the first few highly-ranked ones. This is due to the sampling procedure being too random and thus may not select several top-ranked segments until it has sampled several dozens of segments. Therefore, we modified the original DPP randomized elimination part to deterministic: For $i \in [1, k]$, where k is the number of items we want to select, each time we pick the item by the max probability rather than randomly choosing it weighted by its probability. The equation for our modified DPP-deterministic algorithm is: $Y_i = \max\{\frac{1}{V}\sum_{\boldsymbol{v}\in V}(\boldsymbol{v}^T\boldsymbol{e}_i)^2\}$. The algorithm is shown in the above algorithm box.

Algorithm DPP-Deterministic for diversifying the ranking

Input: O ▷ overlap matrix; n ▷ number of video segments k ▷ number of desired segments; S ▷ LSTM outputted scores

Output:
Y ▷ a vector containing k segments

function DPP_DETERM(O, k, S)
  for $i = 1 : n$ **do**
    for $j = 1 : n$ **do**
      $K_{ij} = O_{ij}\sqrt{S_i S_j}$
    end for
  end for
  $\{(\boldsymbol{v}_n, \lambda_n)\}_{n=1}^{N} \leftarrow$ eigendecomposition of K
  $V \leftarrow \{\boldsymbol{v}_k\}_{n=1}^{k}$ ▷ eigenvectors corresponding to the k-largest eigenvalues
  $Y \leftarrow \emptyset$
  while $|V| > 0$ **do**
    From $\mathcal{Y}$, select $\arg\max_i\{\frac{1}{V}\sum_{\boldsymbol{v}\in V}(\boldsymbol{v}^T\boldsymbol{e}_i)^2\}$
    $Y \leftarrow Y \cup i$
    $V \leftarrow V_\perp$ ▷ an orthonormal basis for the subspace of V orthogonal to e_i
  end while
end function

4. EXPERIMENTS

Our experiments are performed on the challenging VSB-100 dataset from [2] in which the video sequences have pixel-level ground truth annotations once every 20 frames and each annotated frame contains ground truth objects annotated independently by 4 annotators. This is a challenging dataset: some of its videos have many ground truth objects that share similar color or texture and overlaps significantly with each other such as group dance scenes in `salsa` and `ballet`. Besides, it is category-agnostic as it contains ground truths from many diverse categories of objects and non-objects, such as person, backpacks, the ground, the void, mountains (Fig. 5), the sea, etc. Therefore, there is hope that the ranking model learned from this dataset will not be biased towards certain pre-defined categories as in other semantic segmentation datasets such as the ImageNet VID. We trained on 20 videos and tested on 32 videos. For each video we extracted features on the foreground segment in each frame through Convolutional Neural Network with the pretrained model imagenet-vgg-f, using the MatConvNet toolbox in MATLAB.

For all frames of each segment proposal in each video, we used binary masks to capture only the foreground object and resized to 224*224. The result is then input to the VGG network and a 4096-dimensional feature is obtained from the second-to-last layer of the network (before classification layer). The LSTM is trained with a loss function of Mean Squared Error, with a learning rate of 0.01.

We use the evaluation metric from the paper [14]. Only ground truths with agreements from at least two human annotators are used. We report the overall average score per object per video on all the 32 videos in the testing set, and also compare with other ranking algorithms (Figure 3).

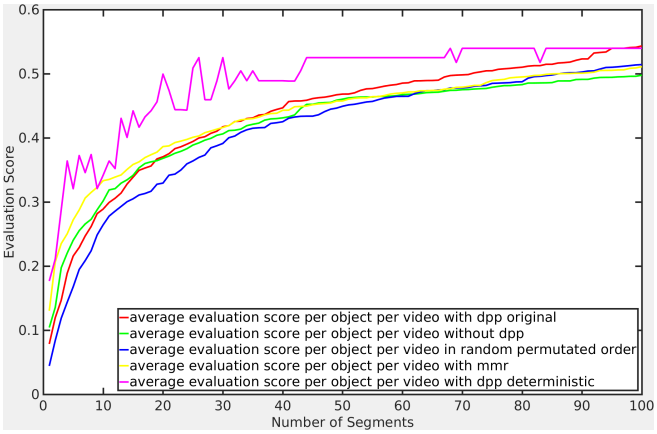

Fig. 3: Our dpp_deterministic algorithm was compared with the dpp_original algorithm, random permutation ranking, and LSTM-only ranking without dpp, and also the Maximal Marginal Relevance (MMR) algorithm. Our score jitters because DPP eigendecomposition is dependent on the number of items to be chosen and different eigenvectors dramatically change the proposals selected.

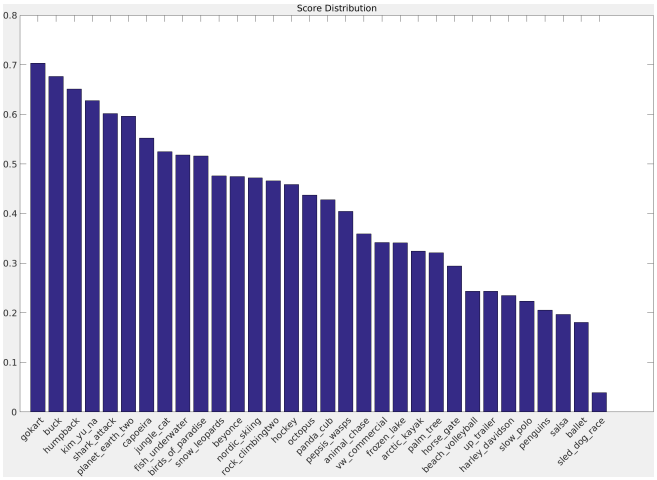

Fig. 4: The score distribution for top-10 retrieved segments for all the tested videos

From the evaluation score curve in the figure based on 100 objects discovered in video, our lstm_dpp algorithm is significantly better than other diversification methods. As an example, Fig. 5 shows the top 10 segment proposals we have retrieved for the video "rock_climbingtwo", which reflects a high-quality and diverse set of objects, including both persons, the head of one person, one backpack on the ground, the cliff and the shadowed void. One can see that our result covers most of the objects in this video. The diversity of the retrieved results also shows that it is not governed by specific semantic categories, but discovers objects in a category-agnostic manner.

5. CONCLUSION

This paper proposes a novel algorithm that combines determinantal point processes with a LSTM network that ranks video segments to obtain a diverse and high-quality subset from a pool of video segment proposals. The results show a significant reduction in the amount of objects required to achieve a good coverage of all ground truth objects in the video. This demonstrates the state-of-the-art performance of our LSTM-DPP approach. In future work, our novel approach can be implemented in large-scale video segmentation for higher-level object discovery tasks in various applications.

Acknowledgements

This work is supported by the NSF grant IIS-1464371. We thank Alrik Firl for helping on the grammar checking.

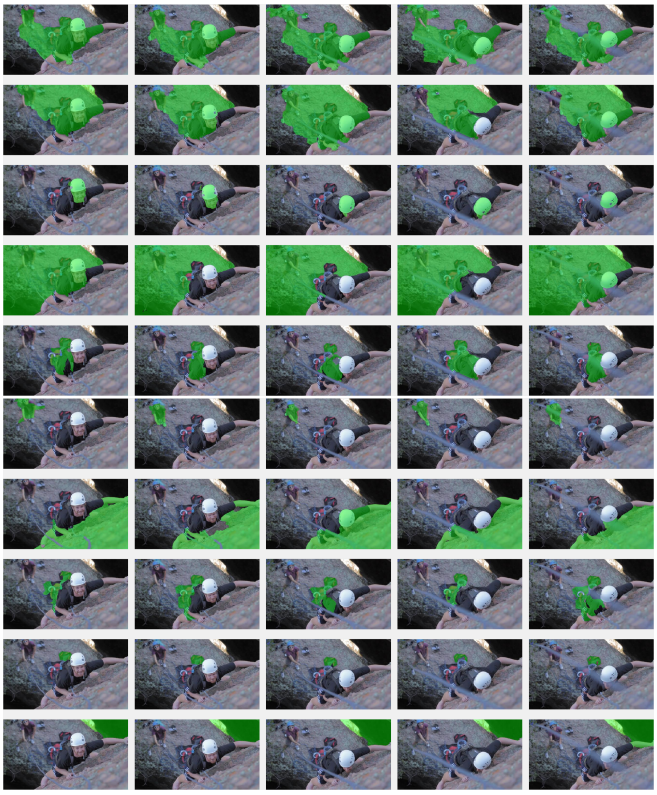

Fig. 5: 10 video segments objects discovered in 1st, 21st, 41st, 61st, 81st frames of video "rock_climbingtwo", columns represent the frames and rows represent the segments selected by algorithm.

6. REFERENCES

[1] Naveen Shankar Nagaraja, Frank R. Schmidt, and Thomas Brox, "Video segmentation with just a few strokes," in *ICCV*, 2015. 1

[2] Fabio Galasso, Naveen Shankar Nagaraja, Tatiana Jimenez Cardenas, Thomas Brox, and Bernt Schiele, "A unified video segmentation benchmark: Annotation, metrics and analysis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3527–3534. 1, 3

[3] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M. Rehg, "Video segmentation by tracking many figure-ground segments," in *International Conference on Computer Vision*, 2013. 1

[4] Joao Carreira and Cristian Sminchisescu, "Cpmc: Automatic object segmentation using constrained parametric min-cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1312–1328, 2012. 1

[5] JRR Uijlings, KEA van de Sande, T Gevers, and AWM Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013. 1

[6] Ahmad Humayun, Fuxin Li, and James M. Rehg, "RIGOR: Reusing Inference in Graph Cuts for generating Object Regions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 1

[7] Ahmad Humayun, Fuxin Li, and James M. Rehg, "The Middle Child Problem: Revisiting Parametric Min-cut and Seeds for Object Proposals," in *Computer Vision (ICCV), IEEE International Conference on*. Dec 2015, pp. 1600–1608, IEEE. 1

[8] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997. 1, 2

[9] Dhruv Batra, Payman Yadollahpour, Abner Guzman-Rivera, and Gregory Shakhnarovich, "Diverse m-best solutions in markov random fields," in *European Conference on Computer Vision*. Springer, 2012, pp. 1–16. 1

[10] Alex Kulesza and Ben Taskar, "Determinantal point processes for machine learning," *arXiv preprint arXiv:1207.6083*, 2012. 1, 2

[11] Anna Khoreva, Fabio Galasso, Matthias Hein, and Bernt Schiele, "Classifier based graph construction for video segmentation," in *CVPR*, 2015. 1

[12] Buyu Liu and Xuming He, "Multiclass semantic video segmentation with object-level active inference," in *CVPR*, 2015. 1

[13] Chenglong Li, Liang Lin, Wangmeng Zuo, Shuicheng Yan, and Jin Tang, "Sold: Sub-optimal low-rank decomposition for efficient video segmentation," in *CVPR*, 2015. 1

[14] Zhengyang Wu, Fuxin Li, Rahul Sukthankar, and James M Rehg, "Robust video segment proposals with painless occlusion handling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4194–4203. 1, 2, 4

[15] Brian Taylor, Vasiliy Karasev, and Stefano Soatto, "Causal video object segmentation from persistence of occlusions," in *CVPR*, 2015. 1

[16] Dong Zhang, Omar Javed, and Mubarak Shah, "Video object segmentation through spatially accurate and temporally dense extraction of primary object regions," in *CVPR*, 2013. 1

[17] João Carreira and Cristian Sminchisescu, "Constrained parametric min cuts for automatic object segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 1, 2

[18] "What is an object?," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 1

[19] Katerina Fragkiadaki, Pablo Arbelaez, Panna Felsen, and Jitendra Malik, "Learning to segment moving objects in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4083–4090. 2

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440. 2

[21] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587. 2

[22] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2625–2634. 2

[23] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici, "Beyond short snippets: Deep networks for video classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4694–4702. 2

[24] Jaime Carbonell and Jade Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998, pp. 335–336. 2

[25] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha, "Diverse sequential subset selection for supervised video summarization," in *Advances in Neural Information Processing Systems*, 2014, pp. 2069–2077. 2

[26] Aidean Sharghi, Boqing Gong, and Mubarak Shah, "Query-focused extractive video summarization," in *European Conference on Computer Vision*. Springer, 2016, pp. 3–19. 2

[27] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman, "Video summarization with long short-term memory," in *European Conference on Computer Vision*, 2016. 2

[28] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 3