Video Segments Retrieval System based on Attentive CNN

Jiazhao Li
University of Michigan
Ann Arbor, MI 48105
jiazhaol@umich.edu

Tianyu Zhang
University of Michigan
Ann Arbor, MI 48105
tyrozty@umich.com

Can Cui
University of Michigan
Ann Arbor, MI 48105
cancui@umich.edu

ABSTRACT

In the past few years, video retrieval has attracted a lot of attention. However, traditional methods simply relay on manual hash tags, title and description of video compared with free text query, which suffers from high bias and low accuracy. Hence, in video retrieval, how to understand video as well as the relationship between query text still needs more exploration. In this project [1], we proposes a novel video segments retrieval system based on architecture called Attentive Convolutional Nerual Network (ACNN). In our model, video clips are re-embedded by contextual information using attentive N-gram model and combined with query text features in cross-model. The probability of reference and offsets of start and end points are loss function to train model. Our results are promising compared with baseline model ACRN in [5].

KEYWORDS

Video Retrieval, Attention Mechanism, Convolutional Neural Network, Cross Feature Model

1 INSTRUCTION

With rapid advances in technologies and computation abilities in multimedia area, retrieval in vast amounts of digital information is becoming possible[2]. In the cross-modal retrieval problems, video retrieval based on text description has attracted a lot of attentions these years. Traditional methods search based on keywords (tags) or categories browsing can only retrieve one whole video and the result is highly influenced by manual tags or titles. However, in some real-world scenarios (e.g. robotic navigation, autonomous driving and surveillance), the untrimmed videos usually contain complex scenes and involve a large number of objects, attributes, actions and interactions, whereby only some parts of the complex scene convey the desired cues or match the description[5]. For instance, in a prepared soccer World Cup competition video lasting for several hours, one may only have interests in some specific moments, like "goal moments". Therefore, video moment retrieval, namely, localizing temporal moments of interest within a video is more useful yet challenging, as compared to simply retrieving an entire video.

Video moment retrieval aims to identify the specific start points and the end points within a video to precisely respond to the given query. The key problem is cross-modal similarity comparison between query texts and video moments. To better understand the video moment, the temporal contextual information is essential. Hence, N-grams model is used to considering the context information in different scales and the weights showing the importance of these vectors are decided by the attentive mechanism. Then, the moment candidates are embedded by the weighted sum of these vectors. Additionally, offsets of the start time and the end time of the moment as well as the similarity score based on combination of video moment embedding and query embedding are used for relevance estimation.

In this work, we investigated the video moments retrieval method proposed in [5], namely, Attentive Cross-Modal Retrieval Network (ACRN), to explore the attentive contextual combination of text features and visual features of moment candidates and adaptively set weights to the contextual information. Moreover, we developed the method of attentive re-embedding video clips with the N-gram model to improve the performance of video moment retrieval.

[1]This is just one course practice project of SI650 Information Retrieval, University of Michigan

2 RELATED WORK

2.1 Video moment retrieval

In previous work, some methods of retrieving temporal segments within a video based on query text have been proposed.[9] considered retrieving video segments from a home surveillance camera via text queries with a fixed set of spatial prepositions. Later, [2] proposed a joint video-language model to retrieve moments within a video based on texture queries. However, these methods can only verify the fixed size segments containing the corresponding moment or not, which still introduces irrelevant moments as noise. Although the densely segmentation of moments at different scales helps to retrieve the specific moments, it will be computationally expensive. To solve this problem, recently,[5] proposed a method to localize the specific moments by predicting the start and end time points of the desired video moments, but there is a large room to improve the retrieval performance of this method.

2.2 Attention mechanism

Nowadays, attention mechanism [12] has become a popular tool in sequence model, contributing to the impressive results in neural machine translation [6], video captioning [8] [11] and video question answering [13]. In specific video moment retrieval, contextual video moments might provide useful visual features for the better understanding of specific video moments, and the attention mechanism helps to focus on the useful visual contextual features by assigning larger weight[5]. [1] introduced a temporal memory attention model named ACRN to dynamically select context moments consistent with the input query and simultaneously memorize the context moment information. In our work, the attentive mechanism is applied to assign weights for N-grams embedded vectors of contextual information.

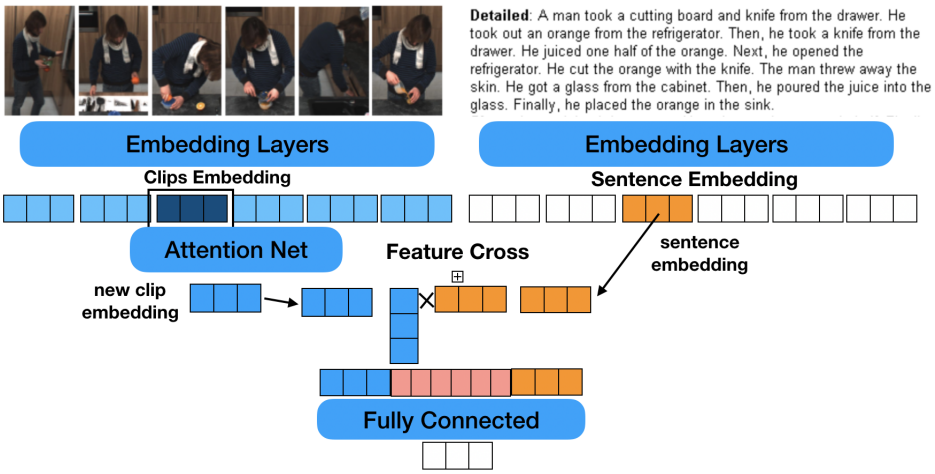

Figure 1: Workflow of model

3 METHODS

In this project we propose an attentive N-gram Convolutional Neural Network (ACNN), shown in Figure1 to capture hierarchical contextual representation for a video clip. Then, the latent relationship between matched text features and video features can be learned by a Neural Network. In this step the cross feature was proposed, aiming to learn not only the basic features but the higher order interaction features as well. Finally the multiple loss is stacked to take all consideration into account in the training process.

3.1 Data Pre-processing

In this project, a retrieval system is designed to retrieve the related video clips based on the information comes from the text queries. In other word, the query is the sentence, and the document are the video clips that match the description of the sentence. Thus, it is necessary to define a function to compute the similarity between the video clips and sentence queries as most retrieval systems do. Unlike traditional text retrieval systems, which can calculate the similarity based on the word level similarity or document level similarity, this video retrieval system has a challenge to measure the similarity between different data format. To build the relevance between the video clips and sentence queries, the time overlap becomes the key criteria in this video retrieval system. In detail, in the dataset of TACos, each sentence and video clip are labeled by their start time and end time.

An intuitive rule for the reference between the sentence queries and video chips is introduced into this retrieval system. Taking two conditions into consideration, the reference between sentence queries and video clips are measure different E.q. 1.

$$label = \begin{cases} 1, & \frac{Interaction of Union}{length of clip} > 0.5 \\ 1, & \frac{Not Interaction of Union}{length of clip} < 0.15 \\ 0, & else \end{cases} \quad (1)$$

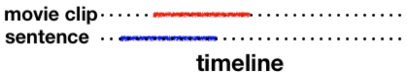

Figure 2: Timeline

The first condition is that the length of the sentence is longer than the length of the video clip. And another condition is that the length of the sentence is much more smaller than that of the video clip.

Based on this criteria, the related sentence and video clip pair can be generated as the positive sample in for the future learning. However, to make our machine learning model robust enough to measure the relevance and irrelevance, the negative samples is also indispensable in the training process. In this step, the negative samples are generated by using the interaction for the queries and clips. In Figure 3, the diagonal elements are the matched pairs, which are the positive samples for the training, and the others are the mismatching pairs, which can be fed into learning model as negative samples.

sentence

video clip	S1	S2	S3	S4	S5	S6	S7
C1	1	1					
C2		1	1				
C3			1	1	1		
C4						1	
C5							1

Figure 3: Sampling Generation

3.2 Video attentive representation

From the former description, human usually understand the videos through synthesizing the contextual information. Hence, In this project an attentive N-gram Convolutional Neural Network is built

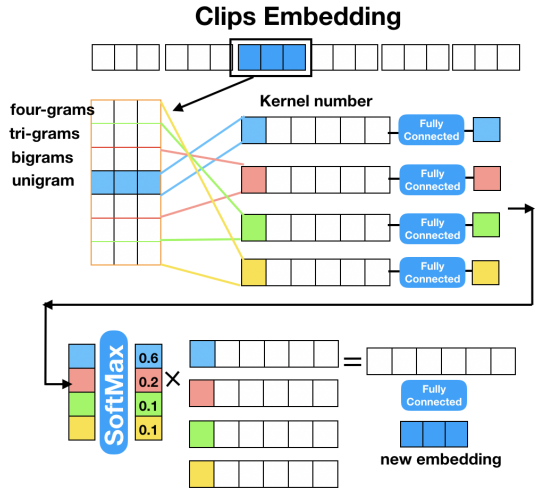

Figure 4: Attention Net

to capture the context information and reconstruct a video clip embedding representation for training step. Obviously, most of the clips have the tight relationship between the clips near it on the scale of time. Inspired by the idea of N-gram in the Natural Language Processing, a hierarchical Convolutional Neural Network is introduced to learn this contextual information. From the Figure 4, the kernels with different size in CNN are able to combine different ranges information centered by this video clip.

$$\hat{m}_c = \sum_i a_i * m_i \tag{2}$$

So, in this step this model collects the contextual features in different levels. However, not all of those contextual features equally contribute to the final video clips representation. For the sake of augmenting the importance of the key contextual features, the attention mechanism is employed in this model. So, after the N-gram CNN, the contextual are compress into vectors with the dimension of number of kernels. Then the final clip representation should be E.q.2 , where a_i is the attention score for feature m_i. To get the attention score for each feature vector, we parameterize the attention score with a multi-layer perceptron, which is called attention network in our model. In detail this network is defined as E.q.3 and E.q.4:

$$a'_i = ReLU(W^T * m_i + b) \tag{3}$$

$$a_i = \frac{exp(a'_i)}{\sum_j exp(a'_j)} \tag{4}$$

Where the $W \in R^{k1}$ and $b \in R^1$ are model parameters, and k denotes the number of kernels which is similar with the dimension of the feature vector. From the equation, the attention score are normalized through the softmax to avoid the gradient explosion in reweight the features. In this attention network, the parameters can be updated in back propagation, and the more informative feature tends to be assigned higher attention score as the model is trained step by step.

3.3 Cross modal and loss definition

In this part the weighted sum video clips features and sentence embedding will be feed into our similarity computing neural network pair by pair. In this network are combined by three part: feature dimension alignment, cross modal and output. Since the video clips and query sentences are preprocessed with different method, their input dimensions are different. In the first part, a fully connected layer is applied to capture the information from the video and sentence, respectively E.q.5 and E.q.6.

$$m_c = W_c * \hat{m}_c + b_c \tag{5}$$

$$m_q = W_q * \hat{m}_q + b_q \tag{6}$$

where $W_c \in R^{k \times t}$, $W_q \in R^{q \times t}$, $b_c \in R^t$ and $b_q \in R^q$ Also, based on the E.q.7, in the cross modal layer, this model will learn both one order features and higher order interaction[5] features E.q.7

$$f_c = [m_c, m_c \otimes m_q, 1] \tag{7}$$

And the last layer is the output layer. Traditionally, the retrieval system merely take the relevance score as the final output. In this video retrieval system, the output are combined by three parts: the relevance score, start time offset and the end time offset. Since the positive and negative are labeled by the rules declared in data processing, the training samples are label by 1 and 0. In this way, this model cannot measure the intensity of the similarity. To make this model more sensitive, the time offset is introduced to the loss function. Hence the model will minimize the both time offset and the dissimilarity score.

3.4 Loss Function

Loss function consists of two parts: probability of reference and start and end offsets, shown in E.q. 11.

In reference part, shown in E.q. 8, positive instances are belong to set P and negative instances are belong to set N.

$$L_{simi} = \alpha_1 \sum_{(c,q) \in P} log(1 + exp(-s_{cq})) + \alpha_2 \sum_{(c,q) \in N} log(1 + exp(s_{cq})) \tag{8}$$

As the multi-scale temporal sliding window is adopted to segment videos, different moment candidates have different durations. Hence for each moment-query pair, we need to not only judge whether the moment is relevant to the query, but also decide the localization offsets compared to the golden moment. Here we used the moment boundary adjustment strategy presented in [5]. Formally, we denote the offset values for the start and end points as follows:

$$\begin{cases} \delta^*_s = t_s - \tau_s \\ \delta^*_e = t_e - \tau_e \end{cases} \tag{9}$$

where (t_s, t_e) is the start and end points of the given query, and (τ_s, τ_e) is the start and end points of a candidate moment in P. Meanwhile, we use $\delta^* = [\delta^*_s, \delta^*_e]$ to denote the ground truth localization offsets. Based on the ground truth offsets, we can adaptively adjust the alignment points of the current moments to match the exact temporal duration. Towards this end, we design a location offset regression modal as:

$$L_{loc} = \sum_{X(c,q) \in P} R(\delta^*_s - \delta_s) + R(\delta^*_e - \delta_e) \tag{10}$$

where P is the set of positive moment-query pairs and R is the L1 norm function. We devise the optimization framework consisting of the alignment loss and the localization regression loss processes:x

$$L = L_{simi} + \lambda L_{loc} \tag{11}$$

where is a hyper-parameter to balance the two losses.

4 EXPERIMENT AND RESULTS

4.1 Data description

4.1.1 TACoS. The orginal dataset Max Planck Institute for Informatics (MPII Cooking 2) is built by[7] and contains 127 videos. Each video is associated with two type of annotations. One is the fine-grained activity label with temporal annotation (i.e., the start and end points). The other is natural language descriptions for the temporal annotations. The dataset is used in [3] for temporal activity localization, dubbed as TACoS. We briefly describe the dataset construction process. In paper [3], each training video is sampled by multi-scale temporal sliding windows with size of [64, 128, 256, 512] frames and 80% overlap. As for the testing samples, they are coarsely sampled using sliding windows with size of [128, 256] frames. For a sliding window moment c from C with temporal annotation (τ_s , τ_e) and a query description q with temporal annotation (t_s, t_e), they are aligned as a pair of training sample if they satisfy the following conditions: 1) the Intersection over Union (IoU) is larger than 0.5; 2) the non Intersection over Length (nIoL) is smaller than 0.15; and 3) one sliding window moment can be aligned with only one query description. In the dataset we used, there are 75 training videos, 25 testing videos, and 26,963 training moment-query pairs satisfying the above conditions. Besides, they utilized 3D ConvNets (C3D) [10] as the moment-level visual encoder and Skip-thoughts[4] as the query description embedding extractor. Therefore, the dimension of the visual embedding and the query description embedding are 4,096 and 4,800, respectively.

4.2 Evaluation

4.2.1 Evaluation metrics. To thoroughly measure our model and the baselines, we used "R@n, IoU=m" proposed by [5] as the evaluation metric. To be more specific, given a query, it is the percentage of top-n results having IoU larger than m. In the following, we used R(n,m) to denote "R@n, IoU=m". This metric itself is on the query level, so the overall performance is the average among all the queries for a query q_i.

4.2.2 Baseline. : ACRN[5]: In this model, perceptron layer is used to form moment-query embedding and attention net computes the scores of context moment-query to re-embed attentive embedding.

4.3 Results

Firstly, we compared our result, shown in Table 2 with original paper, shown in Table 1. Our model has significantly improvement at R@1 in bold while the results are similar in R@5 and R@10.

Additionally, we ploted the training process of 5K iterations for R@1 and IoU = 0.5 evaluation shown in Figure 5. We noticed there is some over-fitting at the end of training for ACNN model since the size of dataset is not big enough.

Table 1: Recall of baseline

	IoU=0.1	IoU=0.2	IoU=0.3	IoU=0.4	IoU=0.5
R@10	0.712	0.618	0.533	0.489	0.328
R@5	0.588	0.504	0.435	0.340	0.274
R@1	0.211	0.182	0.146	0.109	0.077

Table 2: Recall of ACNN

	IoU=0.1	IoU=0.2	IoU=0.3	IoU=0.4	IoU=0.5
R@10	0.719	0.606	0.519	0.445	0.347
R@5	0.588	0.504	0.426	0.373	0.271
R@1	**0.287**	**0.266**	**0.202**	**0.165**	**0.131**

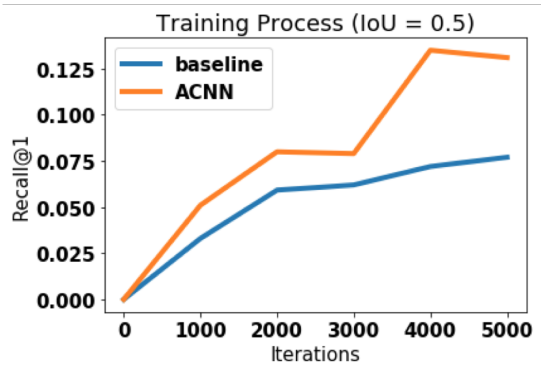

Figure 5: Training Process R@1

5 DISCUSSION AND FUTURE WORK

5.1 Discussion

In this project, it is obvious that the contextual representation of video clips plays an pivotal role in synthesizing more extension of information. Also, the hierarchical Convolutional Neural Network is useful in capturing multi-level contextual information. Moreover the attention mechanism can 'pay more attention' to the key information by assigning more weights to the related parts. Finally, the cross modal not only introduces general features in lower dimension, but takes the higher order interaction features into consideration. Usually, in measuring the similarity between the queries and documents, the information hidden in the higher dimension is indispensable.

5.2 Future Work

Though our model has shown promising results, some work still need to be done in the future to make our work more convincing. 1) Without too much time, our result is not compared with result of non-attention model as additional baseline. 2) Additionally, hyper-parameters (number of kernels, size of kernels) are not fine-tuning to get best performance, which implies our model still has some potential improvements. 3) Finally, testing result is a bit over-fitting since our model is too complicated and dataset is not big enough. Another much larger dataset called DiDeMo could be used to train

our model in next step.

ACKNOWLEDGMENTS

The authors would like to thank Prof. Qiaozhu Mei and GSI Shiyan Yan for their supervision and support.

REFERENCES

[1] Da Cao, Xiangnan He, Liqiang Nie, Xiaochi Wei, Xia Hu, Shunxiang Wu, and Tat-Seng Chua. 2017. Cross-platform app recommendation by jointly modeling ratings and texts. *ACM Transactions on Information Systems (TOIS)* (2017), 37.

[2] Zhang Hongjiang, Wu Jianhua, Zhong Di, and W.Smoliar Stephen. 1997. An integrated system for content-based video retrieval and browsing. *Pattern Recognition* (1997).

[3] Gao Jiyang, Sun Chen, Yang Zhenheng, and Nevatia Ram. 2017. TALL: Temporal Activity Localization via Language Query. *CoRR* (2017).

[4] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought Vectors. *NIPS'15* (2015).

[5] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive Moment Retrieval in Videos. *The 41st International ACM SIGIR Conference on Research* (2018).

[6] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

[7] Rohrbach Marcus, Rohrbach Anna, Regneri Michaela, Amin Sikandar, Andriluka Mykhaylo, Pinkal Manfred, and Schiele Bernt. 2015. Recognizing Fine-Grained and Composite Activities using Hand-Centric Features and Script Data. *International Journal of Computer Vision* (Feb. 2015).

[8] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. 2017. Seeing Bot. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2017).

[9] Stefanie Tellex and Deb Roy. 2009. Towards surveillance video search by natural language query. *Proceedings of the ACM International Conference on Image and Video Retrieval* (2009).

[10] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. *ICCV '15* (2015).

[11] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and languag. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016).

[12] Kelvin xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *International conference on machine learning* (2015), 2048–2057.

[13] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016).