Moment Localization with two different approaches using Residual Connection

Eunseok Yoo, Sora Lee, Hayoung Lee
University of Gachon
sunnycloud56@gachon.ac.kr, ehrqor39@gachon.ac.kr, hhzet11@gachon.ac.kr

Abstract

As various types of unstructured data, such as life logging, increase with the development of networks and smart devices, research on multimodal learning through vision and language is drawing more attention. In particular, we noted the moment localization that determines the temporal moment corresponding to the natural language query, where many studies are being conducted. Accordingly, two improvement measures are proposed by analyzing the (2D-TAN) model that proposed the 2D thermal map. Based on the application of the residual block and the concept of DenseNet, we experimented with a model that combines the hidden layers of all previous blocks as input, and for the Charades-STA dataset, performance improvement of up to 5 points compared to the performance of the previous model was confirmed.

1. Introduction

1.1 The importance of Vision and Language

Understanding vision and language is an important issue in computer vision. Its value is increasingly attracting attention because it has excellent potential for applications such as robotics. Due to this influence, it is performing well in tasks such as video captioning and VQA.

1.2 Moment localization

Moment localization is also expected to be a promising research field, and many studies have continued.

Previous studies aimed to localize the notable point of view of video in a predefined series of motions. However, in order to search for specific events in vast amounts of video data, manual processing of video segments and alignment with pairs of natural language sentences are inevitable. In addition, limiting the precision of the query set involved in the search was the biggest problem.

Accordingly, the researchers proposed and studied moment localization using natural language.

Figure 1: An example of language-based moment localization.

1.3 Moment localization with natural language

The task of moment localization using natural language is was introduced by Gao et al. [1] and Hendricks et al. [2] (Hendricks et al. 2017; Gao et al. 2017)

Based on a given natural language sentence, a momentary segment is extracted from the untrimmed video to determine the position of the beginning and end.

As shown in Figure 1, when a query in untrimmed video and natural language is given, it is a task to find a temporal moment that best corresponds to the sentence 'a personal openness cabinet'.

Previous studies first selected moment candidates in Input video. Then, it was determined whether each moment candidate matched the query sentence and was the target. The method ignores the temporal dependence because it considers each different moment candidate. This method is difficult to predict the exact time boundary of moment.

To solve this problem, Zhang et al. (Zhang et al. 2019) [3] proposed a new network. It is to localize the target moment on the two-dimensional temporal map. The network may recognize more moment context information when predicting whether one moment is related to another target segment.

The previous model by [3] is Temporal Advent Network consists of eight convolution layers. (Refer to Figure 1) However, we expect that the corresponding convolution layer will not be able to fully extract the association between the natural

language query and the 2D temporal map. This is because the more input data goes through the layer, the more hidden layer information at the beginning may be lost. Therefore, two improved models are proposed to prevent Vanishing Gradient.

- Replace the convolution layer of the Temporal Advent Network with a residual block He et al. (He et al.2016). [4]
- Replace the convolution layer of the Temporal Advent Network with a residual block. Unlike the previous method of receiving only the output of the previous block as input, the result of adding the outputs of all passed blocks is received as input.

The proposed model maintains sufficient information on the past hidden layer in a simple way, allowing more contextual reference of the original video feature and the natural language query. It proves improved performance compared to previous models through results.

2. Related Work

We want to successfully implement moment localization based on the 2D-TAN model as our baseline model. Therefore, we would like to introduce 2D-TAN model, moment localization, and residual block and DenseNet, which are methods to increase performance.

2.1. 2D Temporal Adjacent Networks (2D-TAN) Model

Moment localization using natural language is a very difficult task due to the flexibility and complexity of momentary description. Therefore, moment localization using natural language aims to find a temporal segment in the untrimmed video as queried by a given natural language sentence. [1]

Most of the previous language-query moment localization models follow two steps. [2] First, a moment candidate is selected from an input image having a sliding window. After that, it is checked whether each moment's candidate matches the query statement to determine whether it is a target moment or not. This pipeline has the disadvantage of ignoring time dependence because it considers different moment candidates individually.

Therefore, the core idea of the 2D Temporal adjacent networks (2D-TAN) model is to localize the target moment in a two-dimensional thermal map where one dimension represents the start time of one moment and the other represents the end time, like figure 2. Therefore, it represents the adjacent relationship and can treat the moments of various videos in different lengths. In this way, 2D-TAN will be able to recognize more moment context

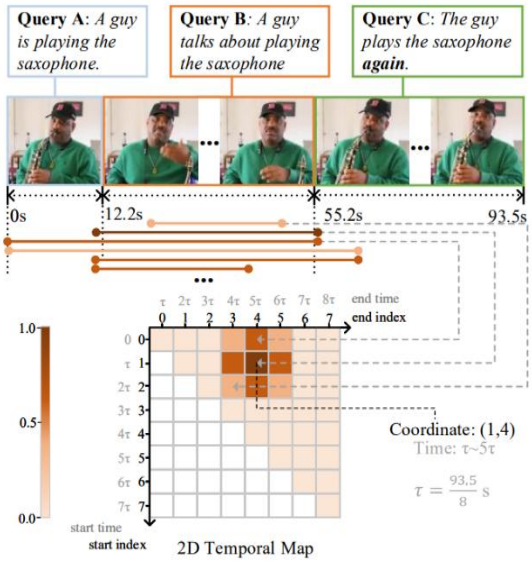

Figure 2: Examples of localizing moments with natural language in an untrimmed video in 2d-TAN. In the two-dimensional temporal map, the black vertical and horizontal axes represent the start and end frame indices while the corresponding gray axes represent the corresponding start and end time in the video. The values in the 2D map, highlighted by red color, indicate the matching scores between the moment candidates and the target moment.

information when predicting whether the moment is related to other temporal segments. As a result, 2D-TAN can learn differential features to distinguish moments. [3]

2.2. Moment Localization

Localizing the moment of the video by referring to the expression is a very difficult task because it aims to predict the start time, end time, and label of the activity instance in the untrimmed video. In addition, it is necessary not only to properly understand video content, but also to be able to match the meaning of video and language. (Zheng et al. 2016) [5]

Visual Context Understanding. Context information is effective for visual content modeling. The previous method integrates the thermal context in two ways: using the entire video as a global context [1] and the surrounding clip as a local context [2]. However, since these methods model the context with one-dimensional sliding window, moment longer than window size is omitted. In addition, there is a disadvantage that long-range temporal differences across multiple windows do not appear.

On the other hand, the 2D Tan model selects candidates from the entire input video rather than the window. Therefore, it has the advantage that the model can recognize more context information and learn differential features by being able to select

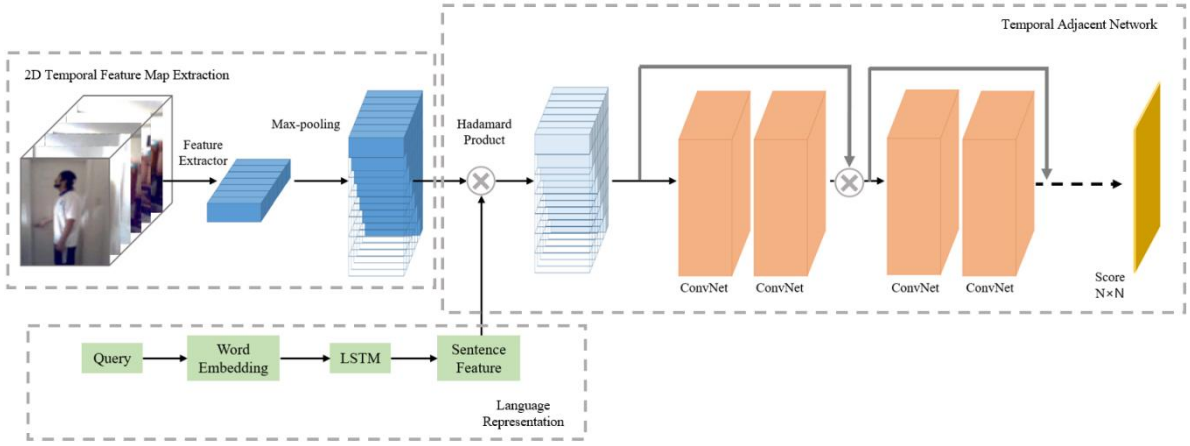

Figure 3: The framework of our proposed model 1. It consists of a text encoder for language representation, a 2D temporal feature map extractor for video representation and a temporal adjacent network consists of Residual block layer for moment localization.

segments of free length as candidates. Furthermore, unlike the previous model that modeled the context for visual features, 2dtan models the context for the fused features of images and language. [3]

Video and Language Cross-Modality Alignment. There are two main ways to model video and language alignment. First, the attention mechanism is a method of fitting visual features related to the query text description by the attention module. (Vaswani et al. 2018) [6] And another method is a sequential modeling that matches a language to an image by an RNN. (Chen et al. 2018) [7]

The 2D-TAN model adopts a simple multiplication operation for visual and language function fusion instead of using complex attrition modules. In addition, unlike the RNN-based method in which context information is gradually aggregated in the clip representation, the 2D-TAN explicitly models the context in the moment representation over the 2D convolution network. [3]

2.3. Residual Block

In the Convolutional Neural Network (CNN), there have been many cases of excellent performance because the deeper the layer, the richer the feature. However, simply stacking layers deeply caused problems such as vanishing gradient and overfitting to fade hidden layers close to the input layer, which is not very helpful for learning. Therefore, there is a limit about performance only by increasing the depth of the network.

To solve the above problem, [4] made the model simple by applying a method of bringing input as it is using a residual block and adding only the remaining residual information to the learned function before delivering parameters in block units. Therefore, learning became easier than learning the whole. In addition, when learning the whole, each weight layer is separated, so the difficulty of convergence has increased by learning for each layer, and the convergence can be better by using the residual block.

Accordingly, we constructed the model by replacing the 8 convolution layers used in the 2D-TAN model with 4 residual blocks. Each residual block is composed of two convolution layers to reduce the difficulty of learning and improve complexity and performance.

2.4. DenseNet

DenseNet is a model with higher performance with fewer parameters than ResNet, where Residual blocks are gathered. It connects the feature maps of all layers and performs concatenation on the feature maps of all layers that appear after the feature maps of the previous layer. Therefore, since the size of the feature map must be the same, and the number of channels may increase as you continue to connect, the number of feature maps in each layer uses a very small value. Through this, DenseNet can prevent loss of information, such as alleviating the vanishing gradient problem, and learn by connecting feature maps of various layers to see the effect of regularization. And above all, because of the use of a small number of channels, the number of parameters to be learned and the amount of computation are greatly reduced. (Huang et al. 2017) [8]

Based on the 2D-Tan model, we propose a model that utilizes the residual block and the dense block. The existing dense block concatenates all previously passed blocks of inputs, but we want to construct a model that adds previous inputs using a residual block. This not only reduces the complexity of the model, but

also seeks to obtain good performance by reducing the number of parameters.

3. Approach

In this section, we introduce the process of moment localization with natural language. The process consists of three steps: language presentation, video presentation and moment localization. After introducing the added model, we will explain the improvements over the previous model.

3.1. Language Representation

We first extract the features of the input query statement. For each word in the input sentence, an embedding vector is generated through the GloVe word2vec model. (Pennington et al. 2014) [9] Word embedding is then sequentially entered into the 3-layer LSTM network (Hochreiter et al. 1997) [10], and the last hidden state is used as feature presentation of the input sentence.

3.2. Video Representation

In this section, features of the input video stream are extracted and features are divided into two-dimensional temporal maps. Then, fixed-interval sampling is performed on the video clip and N video clips are generated. For each sampled video clip, feature is extracted using a pre-trained CNN model.

We create a feature map of moment candidates with video clip features. For each moment candidate, we max-pool the corresponding clip features across a specific time span and obtain its feature.

We structure the entire sample moments into a 2D temporal map. The feature map consists of three dimensions, the first two dimensions represent the start and end points of each clip, and the third dimension represents the feature dimension. For more information, please refer to [3].

3.3. Moment Localization

Based on Language feature and video feature, we predict the most appropriate moment for query among all candidates.

First, the 2D temporal feature map and the encoded sentence feature are fused. Specifically, we project these two cross-domain features into a unified subspace by fully-connected layers, and then fuse them through Hadamard product and normalization.

The Temporal Adjacent Network for 2D feature map consists of L convolution layers with kernel size of K. However, we improved this in consideration of the loss of meaningful information as the previous network passes through the convolution layer.

Moment localization with Residual block. We constructed a Temporal adjective network by reflecting the Residual block of [4]. Refer to Figure 4 for the structure of the Residual block. Each block consists of two convolution networks, which are $F(x)$. The input of the next Residual block is the sum of the inputs of $F(x)$ and x, which are the outputs of the previous block. It represented as

$$H(x) = F(x) + x \quad (1)$$

A temporal adjective network is configured by replacing the convolution layer of the previous model with 4-layer residual blocks. Refer to Figure 3 for the overall structure of the model.

It has a simple form by applying a method of

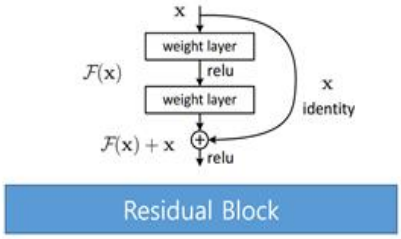

Figure 4: Structure of the Residual Block

additionally adding only residual information to the learned function before transferring the block-by-block parameters. Therefore, learning became easier than learning the whole. In addition, in the case of the previous method, since all weight layers are separated when learning the whole, the difficulty of convergence increased by learning for each layer, and convergence became easier by using residual blocks. By directly inserting information about the previous input x each block, the information on the original video and the natural language query can be consistently maintained.

Moment localization with Dense layer. We applied the Dense layer by advancing one step further from the network of Moment localization with residual block. In the Dense layer, like Residual blocks, one block consists of two convolution networks. Unlike the Residual block, which adds the outputs of the previous block, $F(x)$ and x, it adds the input of all passed blocks. It represented as

$$H(x) = F(x_{n-1}) + \sum_{i=1}^{n-1} x_i \quad (2)$$

A temporal adjective network is configured by replacing the convolution layer of the previous model with 4 Dense layers. Refer to Figure 5 for the overall structure of the model.

Dense Layer concatenates the feature map of the previous layer to the feature map of all layers that appear thereafter. Through this configuration, the effect of regularization can also be seen because it prevents loss of information, such as alleviating the

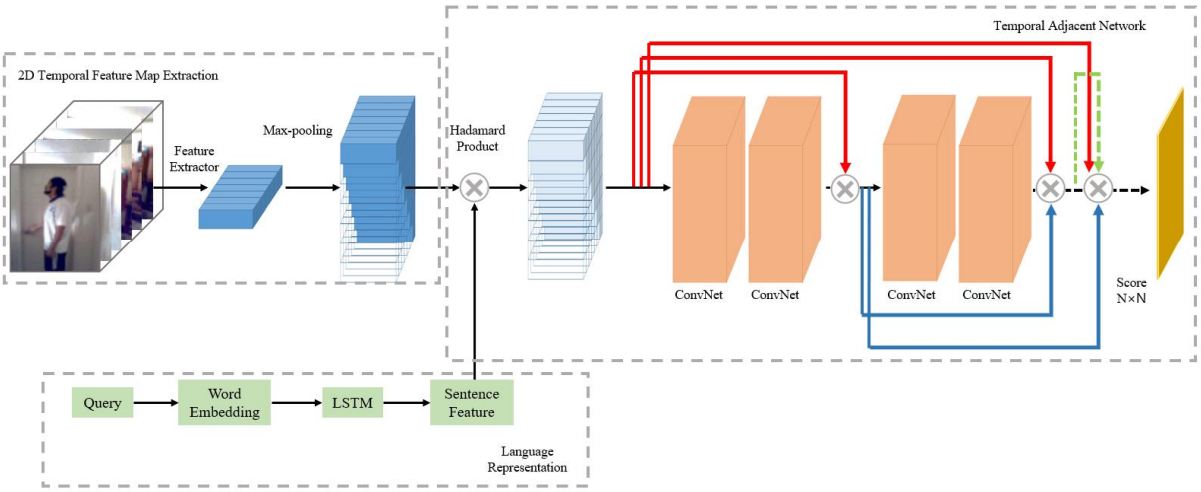

Figure 5: The framework of our proposed model 1. It consists of a text encoder for language representation, a 2D temporal feature map extractor for video representation and a temporal adjacent network consists of Dense layer for moment localization.

vanishing gradient problem, and learns by connecting feature maps of various layers.

3.4. Loss Function

We use the loss function of [3] equally. Details are as follows. During the training of model, [3] adopted a scaled IoU value as the supervision signal, rather than a hard binary score.

The IOU score o_i is scaled by two thresholds t_{min} and t_{max} as

$$y_i = \begin{cases} 0 & o_i \leq t_{min} \\ \frac{o_i - t_{min}}{t_{max} - t_{min}} & t_{min} < o_i < t_{max} \\ 1 & o_i \geq t_{max} \end{cases} \quad (3)$$

The network is trained by binary cross entropy loss as

$$Loss = \frac{1}{C} \sum_{i=1}^{C} y_i \log p_i + (1 - y_i)(\log 1 - p_i) \quad (4)$$

where p_i is the output score of a moment and C is the total number of valid candidates.

4. Dataset

We use three public large-scale datasets for experiments: Charades-STA, ActivityNet Captions, TACoS.

4.1. Charades-STA

Charades-STA dataset consisting of 9,848 videos of various daily indoor activities such as drinking coffee or sitting in a chair and wearing shoes. Originally, the Charades dataset was designed for action registration and localization, but through language description in [1], the name was modified to Charades-STA by adding temporal annotation (labeling of moment's start and end time) to this dataset. Charades-STA contains 12,408 pairs of moment-sentences in training data and 3,720 pairs of moment-sentences in test data.

4.2. ActivityNet Captions

ActivityNet dataset is currently the largest data set in the field of moment localization, consisting of 19,209 videos, and the content is diverse and open. ActivityNet Captions is built on ActivityNet v1.3, which consists of 20k YouTube untrimmed videos with 100,000 capture annotations, with an average length of 120 seconds. Originally designed for video captioning work, these two tasks were recently introduced in moment localization work using natural language because they are reversible. After setting up the experiment (Zhang et al. 2019) [11], we use val_1 as a validation set and val_2 as a test set, and val_2 has 37417, 17505 and 17031 moment-sentence pairs for training, validation, and testing, respectively.

4.3. TACoS

This video consists of 127 videos selected from MPII cooking composite activities video corpus (Rohrbach et al. 2012) [10] and contains various activities taking place in the kitchen. All images are associated with granular activity labels with temporal location and natural language description with temporal location. Natural language description was constructed by cross-sourcing annotators explaining the contents of video clips by sentence. The standard split [12] consists of 9790, 4436, and 4001 pairs of moment-sentences for training, validation, and test, respectively.

5. Experiment

This section introduces the evaluation metric, environmental setting, and the performance of our proposed model and the results of comparative experiments with existing models.

5.1. Evaluation Metric

As the previous work, we evaluate the model by calculating Rank n @ m [1], which means that the percentage of language queries with one or more accurate moment retrieval in the top-n moments obtained. A retrieved moment is determined to be correct when the Intersection over Union (IoU) with the ground truth moment is greater than m. Three datasets have slightly different settings for n and m. To be specific, we decided the results as $n \in \{1, 5\}$ with $m \in \{0.5, 0.7\}$ for Charades-STA, $n \in \{1, 5\}$ with $m \in \{0.3, 0.5, 0.7\}$ for ActivityNet Captions, and $n \in \{1, 5\}$ with $m \in \{0.1, 0.3, 0.5\}$ for TACoS.

5.2. Implementation Details

We use Adam with learning rate of 1×10^{-5} batch size of 32, and maximum epoch of 20 for optimization. Also, 3-layer LSTM is used for language encoding. We adopt an 8-layer convolution network with kernel size 5. The existing 2D-TAN model [3] consists of an 8-layer convolution network with kernel size of 5, but we replaced it with 4-residual blocks. Each residual block consists of 2-convolution layers and Relu function. In addition, in the existing model, all 8-layer convolution network had kernel size of 5 and only the first layer had padding 16. In the model we propose, padding is 2, which is contained in all 8-convolution layers in the residual block. However, like the existing model, the number of channels in the kernel is always the same, and the input feature map is 512 just like the dimension. Therefore, both input size and output size are the same. The scaling thresholds t_{min} and t_{max} are set to 0.5 and 1.0 for Charades-STA and ActivityNet Captions, and 0.3 and 0.7 for TACoS dataset.

5.3. Experiment Result

We evaluate our proposed approach, moment localization with residual block (MoL+R) and moment localization with dense layer (MoL+D) on three benchmark datasets. Then compare those with the baseline model 2D-TAN approach. The results are summarized in Table 1-3.

When comparing the two models that we propose with 2D-TAN, it can be seen that they are all performing better except for ActivityNet Captions. To be specific, in the case of Charades-STA, our proposed model surpasses by more than 5 points compared to 2D-TAN in term of Rank5@0.5. Also, in the case of TACoS, our proposed model surpasses by more than 3points compared to 2D-TAN in term of Rank5@0.7. In this way, it validates that the MoL+R and D that we propose are more able to localize the moment boundary more precisely.

However, in ActivityNet Captions, except for in term of Rank1@0.3, it can be seen that the results are very slightly lower than 2D-TAN. Thinking about the cause, it consists of a much larger dataset than the other two datasets, which takes much more time to train. But we ran out of time, so we just got the following disappointing performance. However, the results obtained show slightly narrow difference, less than 1 point in all areas. Thus, the models that we propose are considered meaningful.

Our proposed approaches are much simpler because they add only residual information to the learned function, compared to the 2D-TAN using only the convolution layer. So, our proposed approaches are much easier to learn than 2D-TAN. In addition, convergence occurs better if the residual block is used than the 2D-TAN, where each weight layer is separated and learning must be performed for each layer. Thus, by using the residual block, we were able to simplify the model to lower the difficulty of learning and obtain better performance.

Method		2D-TAN	**MoL+R**	**MoL+D**
Rank1@	0.5	39.70	41.02	**42.04**
	0.7	23.31	23.33	**24.46**
Rank5@	0.5	80.32	84.33	**85.94**
	0.7	51.26	50.03	**52.18**

Table 1: Performance comparison on Charades-STA. The values highlighted by bold fonts indicate the top methods, respectively. The remaining tables follow the same notations.

Method		2D-TAN	**MoL+R**	**MoL+D**
Rank1@	0.3	59.45	**60.71**	60.63
	0.5	**44.51**	44.08	44.33
	0.7	**26.54**	26.05	26.43
Rank5@	0.3	**85.53**	85.20	85.23
	0.5	**77.13**	76.50	76.40
	0.7	**61.96**	60.61	60.83

Table 2: Performance comparison on ActivityNet Captions.

Method		2D-TAN	**MoL+R**	**MoL+D**
Rank1@	0.1	47.59	47.86	**48.09**
	0.3	37.29	**38.09**	36.49
	0.5	25.32	**26.27**	25.12
Rank5@	0.1	70.31	72.48	**73.11**
	0.3	57.81	**60.93**	57.79
	0.5	45.04	**47.54**	45.51

Table 3: Performance comparison on TACoS.

Furthermore, comparing our two approaches, MoL+D, which adds input from a previous block using only a residual block, performs much better than MoL+R, which adds input from a previous block to all inputs from a previous block like a residual block. Dense Layer is characterized by connecting feature maps of all layers with fewer parameters than networks configured by collecting residual blocks. Because such a small number of channels is used, the number of parameters to be learned and the amount of computation are small, indicating better performance than ResNet. So, MoL+D is similar to DenseNet, but because of the use of residual block, it has the advantages of both methods, showing higher performance. Overall, MoL+D shows superior results, compared to MoL+R in all the three datasets. In particular, Charades-STA shows higher results in all fields, especially by more than 2 points in term of Rank5%0.7. Thus, it validates that MoL+D, which adds not only the previous block but also the residual information of all blocks, makes better use of information than MoL+R, which simply adds the residual information of the previous block.

6. Conclusion

In this paper, we experimented with the improvement of performance by improving the existing Convolution Neural Net in Moment localization task. We proposed two networks. First, the Temporal adjacent network using the Residual block. This replaces the existing convolution layer with four residual blocks. Second, the Temporal adjacent network using the Dense network. This replaces the existing Convolution layer with the Dense network. Unlike previous networks that had problems with convergence, convergence has become easier. It also relives the problem of vanishing gradient and prevents loss of initial information. Our model is simple and performs better than the previous model in three datasets. Afterwards, we would like to devise a method that can more fluently extract feature information from max-pooling of the video presentation section.

Reference

[1] Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. TALL: Temporal activity localization via language query. IEEE International Conference on Computer Vision (ICCV). 2017.

[2] Hendricks, L. A.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. Localizing moments in video with natural language. IEEE International Conference on Computer Vision (ICCV). 2017.

[3] Zhang, S.; Peng, H.; Fu, J.; and Luo, J. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. AAAI Conference on Artificial Intelligence. 34. 2020.

[4] He, K.; Zhang, X.; Ren, S.; and Sun, J. Deep residual learning for image recognition. IEEE conference on computer vision and pattern recognition (pp. 770-778). 2016.

[5] Zheng, Shou; Dongang, Wang; and Shih-Fu, Chang. Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1049-1058, 2016.

[6] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. Attention is all you need. Neural Information Processing Systems 30(NeuIPS). 2017.

[7] Chen, J.; Chen, X.; Ma, L.; Jie, Z.; and Chua, T.-S. Temporally grounding natural sentence in video. Conference on Empirical Methods in Natural Language Processing (EMNLP). pp.162-171, 2018.

[8] Huang, G.; Z. Liu, Z.; Van Der Maaten, L.; and Weinberger, K., Q. Densely Connected Convolutional Networks, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261-2269, 2017.

[9] Pennington, J.; Socher, R.; and Manning, C. D. Glove: Gloval vectors for word representation. Conference on empirical methods in natural language processing (EMNLP) pp. 1532-1543. 2014.

[10] Hochreiter, S.; and Jürgen S. Long short-term memory. Neural computation 9.8 pp. 1735-1780. 1997.

[11] Zhang, Z.; Lin, Z.; Zhao, Z.; and Xiao, Z. Cross-modal interaction networks for query-based moment retrieval in videos. Special Interest Group on Information Retrieval (SIGIR). pp.655-664. 2019.

[12] Rohrbach, M.; Regneri, M.; Andriluka, M.; Amin, S.; Pinkal, M.; and Schiele, B. Script data for attribute-based recognition of composite activities. European Conference on Computer Vision (ECCV). pp.144-157. 2012.