Context Modulated Dynamic Networks for Actor and Action Video Segmentation with Language Queries

Hao Wang,[1,2 *] **Cheng Deng,**[1 †] **Fan Ma,**[2] **Yi Yang**[2]
[1]School of Electronic Engineering, Xidian University, Xi'an 710071, China
[2]ReLER, University of Technology Sydney, Australia
{haowang.xidian, chdeng.xd}@gmail.com, fan.ma@student.uts.edu.au, Yi.Yang@uts.edu.au

Abstract

Actor and action video segmentation with language queries aims to segment out the expression referred objects in the video. This process requires comprehensive language reasoning and fine-grained video understanding. Previous methods mainly leverage dynamic convolutional networks to match visual and semantic representations. However, the dynamic convolution neglects spatial context when processing each region in the frame and is thus challenging to segment similar objects in the complex scenarios. To address such limitation, we construct a context modulated dynamic convolutional network. Specifically, we propose a context modulated dynamic convolutional operation in the proposed framework. The kernels for the specific region are generated from both language sentences and surrounding context features. Moreover, we devise a temporal encoder to incorporate motions into the visual features to further match the query descriptions. Extensive experiments on two benchmark datasets, Actor-Action Dataset Sentences (A2D Sentences) and J-HMDB Sentences, demonstrate that our proposed approach notably outperforms state-of-the-art methods.

Introduction

Video understanding has attracted ever-increasing attention in computer vision community, and it serves as a basic foundation for human action recognition (Yang et al. 2016b; Carreira and Zisserman 2017; Yang et al. 2016a; Xie et al. 2019; Yang et al. 2016c), action detection and localization (Huang et al. 2018), and video object segmentation (Xu et al. 2015; Kalogeiton et al. 2017). However, these works emphasize on the coarse grained video understanding, failing to adequately interact with the high-level semantics. Attempting to study the combination of video content and human language, which is a more practical and natural way of human-computer interaction, Gavrilyuk et al. (2018) introduced a challenging task of actor and action video segmentation with language queries, as illustrated in Figure 1.

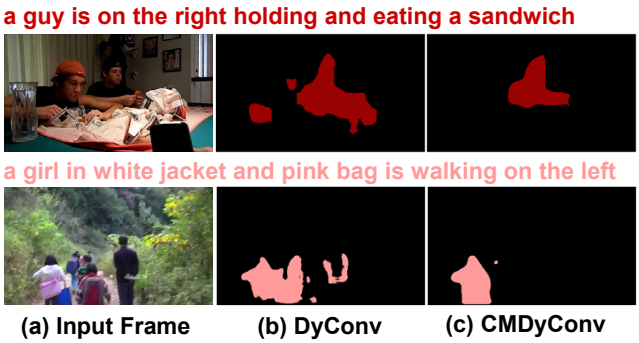

Figure 1: The first column shows the input frames. The second and the last columns demonstrate the results of state-of-the-art method and our approach, which reflects the importance of context modeling. Here, the input sentence and colored mask share the same color if they are corresponding.

Actor and action video segmentation via language is to segment out the related objects in the query sentences. Recently, many approaches (Hu, Rohrbach, and Darrell 2016; Li et al. 2017b; Gavrilyuk et al. 2018; Shi et al. 2018; Margffoy-Tuay et al. 2018; Opazo et al. 2019) have been exploited to learn the complex interactions between heterogeneous modalities. These approaches can be roughly categorized into two paradigms: heterogeneous feature fusion and dynamic convolution methods. The former framework managed to explore the relationship between different inputs by merging their features (Shi et al. 2018). Dynamic convolution explicitly models the correlation between visual and semantic information via convolution computation, which has been investigated in an increasing number of works (Margffoy-Tuay et al. 2018; Opazo et al. 2019). The kernels in dynamic convolution methods are generated from the embedding of the query sentence to segment objects in the video. However, the generated spatial-irrelevant kernel makes it hard to model the contextual visual information. It will cause ambiguities if there are similar objects in the video for the given query. As shown in Figure 1(b), the model with dynamic convolution fails to segment target objects in query sentences. To address this, we propose a

*This work was done when Hao Wang was visiting the ReLER Lab, UTS
†Corresponding author

context modulated dynamic network to integrate spatial visual context and language for computing the segmentation mask. The convolutional kernels for the specific patch in our network are adjusted with both query language and the surrounding features. To incorporate the context more efficiently and effectively, we employ a split-transform-merge strategy (Zhang et al. 2018) to reduce computation and the deformable convolution to expand context range.

Temporal evolution has also played a crucial role in video representation. It encourages the model to narrow the segment area with the query sentence. As noted in (Brox and Malik 2010), it is also a salient cue for determining the object boundary in video segmentation. For example, some works (Gavrilyuk et al. 2018; Ji et al. 2018) utilize optical flow for better segmentation performance. However, the calculation of optical flow is time-consuming even on graphics processing units. We employ the convolutional long short-term memory network (Xingjian et al. 2015) to encode relative motions in feature maps. In this way, we can integrate the temporal features into the context modulated convolution to segment the target object in videos. Overall, the main contributions of our approach can be summarized as follows:

- We propose a novel context modulated dynamic convolutional network with groupwise kernel predication, to incorporate context information into the correlation learning of heterogeneous modalities, for more effective actor and action video segmentation with language queries;
- We frame a simple yet effective pixel-level temporal evolution encoder to explicitly capture motion information, obtaining further performance improvement on actor and action video segmentation;
- Extensive experiments on two popular video segmentation datasets, A2D Sentences and J-HMDB Sentences, demonstrate that our proposed approach significantly outperforms state-of-the-art methods.

Related Work

Actor and Action Video Segmentation. Towards understanding different actions performed by different actors in the video, Xu et al. (2015) released the Actor-Action Dataset (A2D) with various actor-action tuples and introduced an interesting task of actor and action video segmentation. Early works, based on the supervoxels features, mostly utilized graphical models to group the spatial and temporal information for segmentation. Xu et al. (2015) proposed a trilayer approach, which consisted of separate classifiers for actor, action, joint actor-action nodes and conditional classifiers between single actor or action node and joint actor-action nodes. Xu et al. (2016) utilized a graphical model to encourage adaptive and long-range interaction of video parts. Yan et al. (2017) proposed a robust multi-task ranking model for weakly supervised actor and action video segmentation. Recently, deep learning has been successfully applied in various fields (Margffoy-Tuay et al. 2018; Wei et al. 2019) as its powerful capability of feature extraction. Kalogeiton et al. (2017) jointly learned the actor-action detector on single frames and then performed segmentation via existing methods. Gavrilyuk et al. (2018) further annotated A2D with corresponding sentences and extended the task to actor and action video segmentation with language queries. They adopted dynamic convolutions to learn the correlation between video contents and input sentences. Wang et al. (2019) utilized "concatenation-convolution" to align the visual and semantic inputs by performing convolution on the concatenation of heterogeneous features. These approaches rely on two-stage architecture (Kalogeiton et al. 2017), traditional dynamic convolution (Gavrilyuk et al. 2018) or simple "concatenation-convolution" (Wang et al. 2019), while our proposed approach is end-to-end and can model contextual information during correlation learning, leading to better segmentation performance.

Actor and Action Localization with Language Queries. To understand how vision and language interact with each other, numerous works have been explored on actor and action localization with language queries. Yamaguchi et al. (2017) obtained candidate tubes and conducted multi-modal retrieval between tubes and textual descriptions for spatio-temporal person retrieval. Li et al. (2017a) proposed a recurrent neural network with gated attention mechanism to search person with natural language description. Chen et al. (2019) utilized cross-gated attended recurrent network to explore fine-grained interaction and self interactor to perform cross-frame matching, for localizing natural language in videos. Hendricks et al. (2017) integrated global and local video features to localize moments in video with input sentences. Mithun et al. (2019) learned latent alignment between video and sentence representation to address weakly supervised video moment retrieval from text queries. Different from above works, we prefer pixel-wise segmentation for actor and its action in video, which is more challenging than the task of localization.

Dynamic Networks. Contrasted to the traditional network, the weights of layers in dynamic network (Xu et al. 2016) are not fixed after training, namely its weights can be generated dynamically conditioned on various inputs. Noh et al. (2016) predicted the weights of fully connected layer with hashing technique to avoid introduce huge amount of parameters. Li et al. (2017b) generated all the weights from language or vision input for tracking person, which was difficult to train. As a generalization of conditional normalization, Perez et al. (2018) predicated the parameters of the feature-wise affine transformation from input sentences to conduct visual reasoning. These methods usually lack the capability of context modeling while our proposed approach can incorporate contextual information into correlation learning, resulting in better interaction between different modalities.

Methodology

Given a video and language query, our approach aims to segment out the actor and its action relevant to the input query. The architecture of our approach is illustrated in Figure 2, which consists of multi-modal feature encoder, context modulated dynamic network and temporal evolution encoder.

Multi-modal Feature Encoder

Video Feature Encoder. To effectively encode appearance and motion information for actor-action pair, we adopt 3D

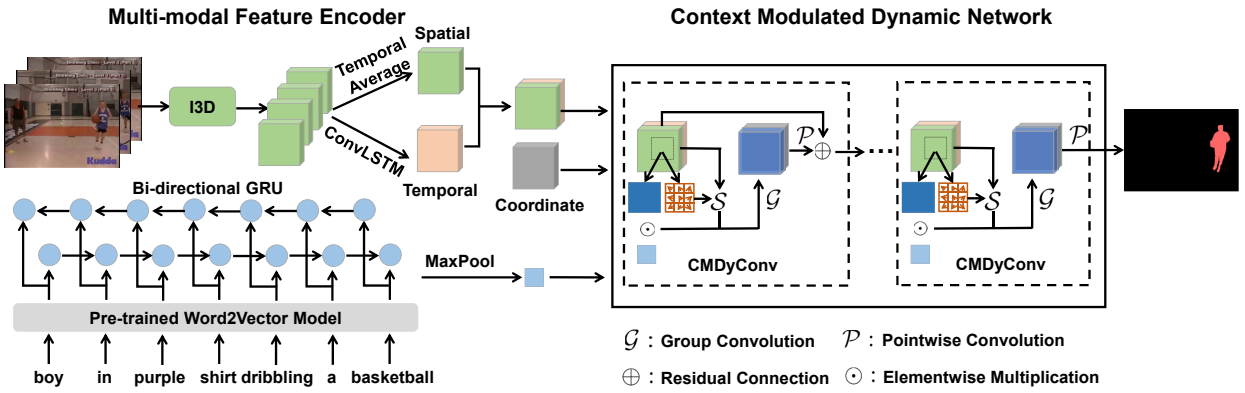

Figure 2: Our proposed Context Modulated Dynamic Convolutional (CMDyConv) networks, which consists of multi-modal feature encoder and context modulated dynamic network. There are also a temporal evolution encoder (i.e., ConvLSTM) to explicit model motion information. After extracting video and text features, our proposed approach performs context modulated dynamic convolution three times and generates the final segmentation output.

convolution network to learn video representation. However, 3D convolution networks are difficult to train from scratch, especially without a large scale video dataset for training. In order to benefit from the 2D convolution networks pre-trained on ImageNet (Deng et al. 2009), Carreira et al. (2017) proposed Inflated 3D ConvNet (I3D) and obtained state-of-the-art performance on video classification. Here, we adopt I3D as basic video feature encoder. Specifically, the clip features are generated by temporal average pooling on extracted intermediate output and followed L_2 normalization on each pixel. Following (Gavrilyuk et al. 2018), given a video clip $V \in \mathbb{R}^{3\times N_V\times H_V\times W_V}$, the video encoder can be defined as

$$F_V = L_2(\mathrm{Avg}(\mathrm{pI3D}(V))), \tag{1}$$

where L_2 and Avg stand for L_2 normalization and temporal average pooling, pI3D stands for the partial layers of I3D feature extractor. N_V, H_V, and W_V denote the number (i.e., 16), the height (i.e., 512), and the width (i.e., 512) of sampled input clip, respectively. For video representation $F_V \in \mathbb{R}^{H_F\times W_F\times D_F}$, H_F, W_F and D_F denote the height (i.e., 32), the width (i.e., 32) and the dimension (i.e., 832) of extracted video feature map, respectively.

We upsample the feature map in a progressive manner like in (Gavrilyuk et al. 2018). The small and medium resolution (i.e., $32\times32\times832$ and $128\times128\times256$) video features are denoted as $F_V^S \in \mathbb{R}^{H_F^S\times W_F^S\times D_F^S}$ and $F_V^M \in \mathbb{R}^{H_F^M\times W_F^M\times D_F^M}$, which can be formulated as

$$F_V^M = \mathrm{Decon}(F_V^S), \tag{2}$$

where Decon stands for deconvolution network and F_V^S is the same as F_V above. In order to keep the trade-off between segmentation performance and computation cost, we conduct the interaction at the medium resolution. For simplicity, we define the whole video encoder as

$$F_V^M = \mathrm{Enc}_V(V;\theta_V), \tag{3}$$

where $\mathrm{Enc_V}$ is the video encoder parameterized with θ_V.

Text Feature Encoder. To extract sentence features, we first translate each word to vector via the word2vec model pre-trained on the Google News Dataset (Mikolov et al. 2013). Then temporal information of whole sentence is captured by recurrent neural networks as in (Hu, Rohrbach, and Darrell 2016; Li et al. 2017b). After that, sentence features are obtained by conducting temporal max-pooling along word orders. Specifically, each word is encoded as 300-dimensional word vector and the sentence can be regarded as a feature matrix $S \in \mathbb{R}^{N_W\times D_W}$. N_W is the maximum length (i.e., 20) of words in the sentences and D_W is the feature dimension (i.e., 300) of word vector. It should be noted that longer sentences are truncated and shorter sentences are padded with zeros to obtain the fixed dimensional sentence representation. Therefore, the text encoder can be defined as

$$F_T = \mathrm{Max}(\mathrm{Enc}_T(S;\theta_T)), \tag{4}$$

where Max is the temporal max-pooling along word orders, Enc_T is text encoder parameterized with θ_T. $F_T \in \mathbb{R}^{D_T}$ is the extracted sentence representation with dimension D_T.

Context Modulated Dynamic Network

The comparison between traditional dynamic convolution and our proposed approach is illustrated in Figure 3. Following (Gavrilyuk et al. 2018), given video representation $F_V^M \in \mathbb{R}^{H_F^M\times W_F^M\times D_F^M}$ and sentence representation $F_T \in \mathbb{R}^{D_T}$, the prediction of dynamic weights from the input sentence can be formulated as

$$k = \delta(W_k F_T + b_k), \tag{5}$$

where δ is the $\tanh$ function and $k \in \mathbb{R}^{D_F^M}$ is the predicted dynamic weights with the same number of channels as F_V^M. This process can be implemented efficiently with fully connected layers (i.e., W_k and b_k). Then the dynamic weights are convolved with F_V^M to generate pixel-wise segmentation response map $R^M \in \mathbb{R}^{H_F^M\times W_F^M\times 1}$ as

$$R^M = k * F_V^M = \sum_{i=0}^{D_F^M-1} k_i \cdot (F_V^M)_i, \tag{6}$$

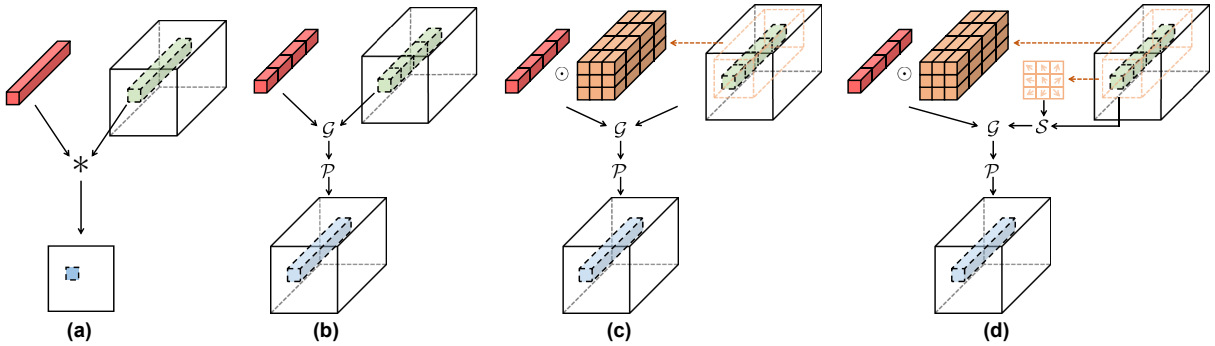

Figure 3: The comparison between the traditional dynamic convolution and our proposed method. (a) is the traditional dynamic convolution. (b) is our elaborate dynamic convolution using group convolution $\mathcal{G}$ and pointwise convolution $\mathcal{P}$. (c) is the proposed context modulated dynamic convolution and takes the neighbor pixels into consideration. (d) is the deformable version of (c), which attempts to address geometric deformation by predicting 2D offsets during sampling procedure $\mathcal{S}$.

where the subscript i denotes the i-th channel. However, traditional dynamic convolution will introduce numerous parameters, which makes it impossible to scale to the entire network for obtaining better performance.

Context Modulated Dynamic Convolution. To address this problem, we frame a novel dynamic convolution with groupwise kernel prediction. Concretely, we utilize group convolution (e.g., the number of groups equals to the number of output channels) to greatly reduce the number of weights to be predicted. Then we adopt pointwise convolution to aggregate information across groups and increase the channels to original input for residual connection simultaneously. Formally, the elaborate dynamic convolution can be defined as

$$R^M = \mathcal{P}(\mathcal{G}(k, F_V^M)) = \mathcal{P}(\sum_{i=0}^{N-1}\sum_{j=0}^{G-1} k_{i,j} \cdot (F_V^M)_{i,j}), \quad (7)$$

where $\mathcal{P}(\cdot)$ and $\mathcal{G}(,)$ stand for pointwise and group convolution. $N = D_F^M/G$ and G are the number (e.g., 4) of channels in each subgroup and the number (e.g., 64) of total groups. The subscripts i, j denotes the i-th channel in j-th group. Thus, $R^M \in \mathbb{R}^{H_F^M \times W_F^M \times D_F^M}$ has the same number of channels with input, without predicting more dynamic weights.

Besides, human tends to describe objects with the spatial qualifiers, such as "bottom left" and "in the middle". To identify them, we concatenate the spatial coordinate features $C \in \mathbb{R}^{H_F^M \times W_F^M \times D_C^M}$ along channel dimension with the outputs after group convolution, which can be formulated as

$$R^M = \mathcal{P}([\mathcal{G}(k, F_V^M), C]), \quad (8)$$

where $[,]$ denotes the concatenation operation. Compared with the direct interaction (i.e., $[F_V^M, C]$ and k) in traditional dynamic convolution, our proposed progressive interaction method achieves better performance. One possible reason is that it makes the training simpler than direct interaction.

Based on the proposed dynamic convolution, we stack it multiple times and employ residual connection in each one to learn complex correlations between different modalities. Nevertheless, the input sentence does not possess any explicit spatial information, consequently generating spatial-irrelevant dynamic weights and then leading to unsatisfactory segmentation performance. To equip with the capability of context modeling in dynamic convolution, we endow the operation with learnable spatial-revelant filters to incorporate neighbor information into correlation learning. In other words, the final weights should not only query-dependent but also content-dependent, which means

$$k = \mathcal{K}(F_T, F_V^M). \quad (9)$$

For simplicity, we process them separately and update the above equation to

$$k = k^T \cdot k^V = \mathcal{K}^{\mathcal{T}}(F_T) \cdot \mathcal{K}^{\mathcal{V}}(F_V^M), \quad (10)$$

where k^T stands for a fully connected layer to generate weights from input query and k^V denotes a convolutional layer to predict weights from input video. Then the context modulated dynamic convolution can be formulated as

$$R_p^M = \mathcal{P}(\sum_{i=0}^{N-1}\sum_{j=0}^{G-1}\sum_{l\in\Omega(p)} k_{i,j}^T \cdot k_{i,j,l}^V \cdot (F_V^M)_{i,j,l}), \quad (11)$$

where $\Omega(p)$ denotes the sliding window at location p and the subscripts i, j, l means the l-th pixel in $\Omega(p)$ of i-th channel in j-th group. As the geometric variations such as the pose and part deformation can greatly affect the learning of visual information, Dai et al. (2017) introduced deformable convolution to enhance the transformation modeling capability of existing CNNs. Hence, we adopt an extra branch to learn the 2D offsets during spatial sampling and finally formulate our proposed context modulated dynamic convolution as

$$R_p^M = \mathcal{P}(\sum_{i=0}^{N-1}\sum_{j=0}^{G-1}\sum_{l\in\Omega(p)} k_{i,j}^T \cdot k_{i,j,l}^V \cdot (F_V^M)_{i,j,l+\Delta l}), \quad (12)$$

where Δl is the learnable 2D offsets. For simplicity, the procedure of correlation learning can be defined as

$$R^M = \mathrm{CMDyConv}(F_V, F_T; \theta_D), \quad (13)$$

where CMDyConv is the context modulated dynamic convolution network parameterized with θ_D. As a result, the

spaital-irrelevant dynamic weights predicted from input sentence are modulated with the spatial-revelant weights generated from input video to take contextual information into consideration when learning the correlation, achieving better segmentation performance.

Temporal Evolution Encoder

It is well-known that temporal evolution plays a crucial role in video representation. Although 3D convolution can implicitly capture spatio-temporal information to some extent, explicit modeling of temporal dynamics (Carreira and Zisserman 2017; Ji et al. 2018) still brings further performance improvement in many fields. Optical flow is most commonly used for modeling motion information. However, the calculation of optical flow is extremely time-consuming even on GPUs. Hence, we adopt convolutional long short-term memory network (ConvLSTM or CLSTM) to learn pixel-level temporal evolution. The key equations of ConvLSTM are summarized as

$$\begin{aligned} i_t &= \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1}) \\ f_t &= \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1}) \\ g_t &= \tanh(W_{xc} * X_t + W_{hc} * H_{t-1}) \\ C_t &= f_t \circ C_{t-1} + i_t \circ g_t \\ o_t &= \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_t) \\ H_t &= o_t \circ \tanh(C_t), \end{aligned} \tag{14}$$

where X_t, H_t and C_t are inputs, hidden states and cell outputs. i_t, f_t, g_t, o_t, $*$, $\circ$ and $W_{\cdot,\cdot}$ denote input gate, forget gate, cell gate, output gate, convolution, elementwise multiplication and learnable weights, respectively. We utilize the hidden state at last timestep as temporal representation. After feeding video clip into video encoder, we enhance video feature in equation (2) with explicit temporal evolution encoding in equation (14) and update it to

$$F_V^M = [\mathrm{Decon}(F_V^S), \mathrm{Decon}(\mathrm{CLSTM}(\mathrm{pI3D}(V)))], \tag{15}$$

where $[,]$ is the concatenation operation.

Training and Inference

Given video clip V and input sentence S, our proposed approach generates pixel-wise segmentation mask R^L, which is supervised with the binary ground-truth Y. Specifically, the segmentation loss $\mathcal{L}$ can be formulated as

$$\mathcal{L} = -\lambda Y \log(\sigma(R^L)) - (1 - Y) \log(1 - \sigma(R^L)), \tag{16}$$

where R^L is the bilinear interpolation of R^M and σ is the sigmoid function. Here, we weight the foreground with λ to pay more attention on them than the background. During inference, we take a pixel as foreground when its value is higher than the half of the maximum value in response map.

Experiment

In this section, we first introduce the dataset statistics and implementation details used in all experiments. Then we compare our proposed approach with state-of-the-art methods to verify its effectiveness. Finally, we give a detailed ablation studies of each component in our model and provide qualitative visualizations of experimental results.

Datasets and Evaluation Criteria

A2D Sentences is released in (Gavrilyuk et al. 2018) by additionally providing corresponding human natural descriptions on A2D (Xu et al. 2015). There are total 3,782 videos collected from YouTube in original dataset, which contain 8 action classes (i.e., climbing, crawling, eating, flying, jumping, rolling, running and walking) performed by 7 actor classes (i.e., adult, baby, ball, bird, car, cat and dog). For pixel-wise masks, 3 to 5 frames in each video are labeled to evaluate segmentation performance. Then A2D is extended with 6,655 sentences to suite for the task of actor and action video segmentation from a sentence. We split the whole dataset into 3,017 training videos, 737 testing videos and 28 unlabeled videos, as followed in (Gavrilyuk et al. 2018).

J-HMDB Sentences contains total 928 videos, which are annotated with the pixel-wise 2D articulated human puppet masks. It is also extended from J-HMDB dataset (Jhuang et al. 2013) by annotating 928 natural language sentences to describe the actor and its action in the video.

To evaluate the segmentation performance, we adopt the common criteria of Intersection-over-Union (IoU) and precision. Specifically, the overall IoU computes the ratio of total intersection area divided by the total union area on the whole dataset while mean IoU calculates the ration of each sample first and obtains the average results on entire dataset. The precision@t reports the percentage of testing samples whose IoU scores are higher than threshold t. We also compute the mean average precision over different thresholds from 0.50 to 0.95 with the step 0.05.

Implementation Details

For multi-modal feature encoder, we adopt pre-trained I3D model as video encoder and Gated Recurrent Unit (GRU) (Chung et al. 2014) as text encoder. Specifically, we extract the intermediate features of I3D model at $\mathrm{maxpool3d\_4a}$ and then fine-tune all layers of $\mathrm{mixed\_4}*$. For temporal evolution encoder, we utilize convolutional LSTM with the same hidden dimension as input. The deconvolution module consists of a deconvolution layer (i.e., kernel size 8 and stride 4) and a convolutional layer (i.e., kernel size 3). The maximum length of sentences is set to 20 and the dimension of word vector is 300. The dimension of sentence features are 600 as we set the hidden dimension of GRU as 300 and adopt its bi-directional variant. For context modulated dynamic network, we utilize a fully connected layer to generate weights from input queries and a convolutional layer to predict weights from input video. The kernel size in each context modulated dynamic convolution module is 3 and we stack this module 3 times with residual connection. The last module generate segmentation response map directly.

The experiments are trained with the PyTorch package on 4 GPUs. The optimizer is Adam with the learning rate 1×10^{-3}. The batch size is 36 and the number of maximum training epochs is 10. We divide the learning rate by 10 after every 6 epochs. The weight of foreground pixels λ is set to 1.5 by observing segmentation performance. We fix the annotated frame in the middle of the sampled clip.

Table 1: Segmentation results on A2D Sentences. The approaches marked by † fine-tune the layer mixed_4f while the others marked by ‡ fine-tune the layers from mixed_4b to mixed_4f on the A2D Sentences. It should be noted that the results of (Gavrilyuk et al. 2018) are obtained on two streams - RGB and Optical Flow while ours only take RGB frames as input.

Method	Overlap					mAP	IoU	
	P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	0.5:0.95	Overall	Mean
(Hu, Rohrbach, and Darrell 2016)	7.7	3.9	0.8	0.0	0.0	2.0	21.3	12.8
(Li et al. 2017b)	10.8	6.2	2.0	0.3	0.0	3.3	24.8	14.4
(Hu, Rohrbach, and Darrell 2016) †	34.8	23.6	13.3	3.3	0.1	13.2	47.4	35.0
(Li et al. 2017b) †	38.7	29.0	17.5	6.6	0.1	16.3	51.5	35.4
(Gavrilyuk et al. 2018) †	50.0	37.6	23.1	9.4	0.4	21.5	55.1	42.6
(Gavrilyuk et al. 2018) ‡	53.8	43.7	31.8	17.1	2.1	26.9	57.4	48.1
Ours	**60.7**	**52.5**	**40.5**	**23.5**	**4.5**	**33.3**	**62.3**	**53.1**

Table 2: The generalization ability of each method on J-HMDB Sentences with the best model trained on A2D Sentences.

Method	Overlap					mAP	IoU	
	P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	0.5:0.95	Overall	Mean
(Hu, Rohrbach, and Darrell 2016) †	63.3	35.0	8.5	0.2	0.0	17.8	54.6	52.8
(Li et al. 2017b) †	57.8	33.5	10.3	0.6	0.0	17.3	52.9	49.1
(Gavrilyuk et al. 2018) †	69.9	46.0	17.3	1.4	0.0	23.3	54.1	54.2
(Gavrilyuk et al. 2018) ‡	71.2	51.8	26.4	3.0	0.0	26.7	**55.5**	57.0
Ours	**74.2**	**58.7**	**31.6**	**4.7**	0.0	**30.1**	55.4	**57.6**

Table 3: Segmentation results on A2D Sentences for ablation studies. Context modulated dynamic convolution and its deformable version are abbreviated as "CMDyConv" and "DCMDyConv", respectively.

Method	Overlap					mAP	IoU	
	P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	0.5:0.95	Overall	Mean
(Gavrilyuk et al. 2018) ‡	53.8	43.7	31.8	17.1	2.1	26.9	57.4	48.1
Baseline (DyConv)	55.7	46.4	33.8	18.9	2.3	28.5	60.2	49.3
Baseline + Our DyConv	55.4	47.0	35.2	20.0	2.8	29.3	60.1	50.1
Baseline + CMDyConv	57.5	48.4	36.5	20.6	3.0	30.2	61.3	51.0
Baseline + DCMDyConv	60.4	**52.7**	**40.9**	23.4	3.8	**33.3**	**62.6**	53.0
Baseline + DCMDyConv + Temporal	**60.7**	52.5	40.5	**23.5**	**4.5**	**33.3**	62.3	**53.1**

Comparison with State-of-the-art Methods

The comparison with state-of-the-art methods is demonstrated in Table 1. Following (Gavrilyuk et al. 2018), we first evaluate the models (Hu, Rohrbach, and Darrell 2016; Li et al. 2017b) pre-trained on ReferIt dataset (Kazemzadeh et al. 2014) and then evaluate the fine-tuned version on A2D sentences. Experimental results in the first four rows show that the model remarkably improves the performance by fine-tuning on video segmentation task. Besides, we get more layers fine-tuned and take the order of words into consideration, obtaining extra performance improvement in the sixth row. Our approach achieves state-of-the-art performance on all metrics, especially at higher IoU thresholds. The qualitative results on A2D sentences are illustrated in Figure 4. Our approach generates fine-grained segmentation of objects, e.g., the arms and legs of actor in the first sample. The results of the second video show that with the help of context, our approach obtains better segmentation outputs. For more complex situation, our approach obviously outperforms the existing state-of-the-art, e.g., segmenting out the inconspicuous sitting person in the third video. To evaluate the generalization ability of our model, we conduct experiments on J-HMDB sentences with the best model trained on A2D sentences. We uniformly sample 3 frames in each video and report the results in Table 2. Our approach significantly outperforms existing state-of-the-art methods on most metrics. For the result of precision@0.9, one possible reason is that the feature encoder can not produce fine representations without any fine-tuning on the target dataset.

Ablation Studies. To verify the effectiveness of each component, detailed ablation studies are conducted in Table 3. The significant differences between *Baseline* and (Gavrilyuk et al. 2018) are that we adopt bi-GRU as text encoder, batch normalization to prevent overfitting and interaction on medium resolution. First, *Baseline* utilize the traditional dynamic convolution while *Baseline + Our DyConv* takes our elaborate dynamic convolution. The results obviously show the effectiveness of groupwise kernel predic-

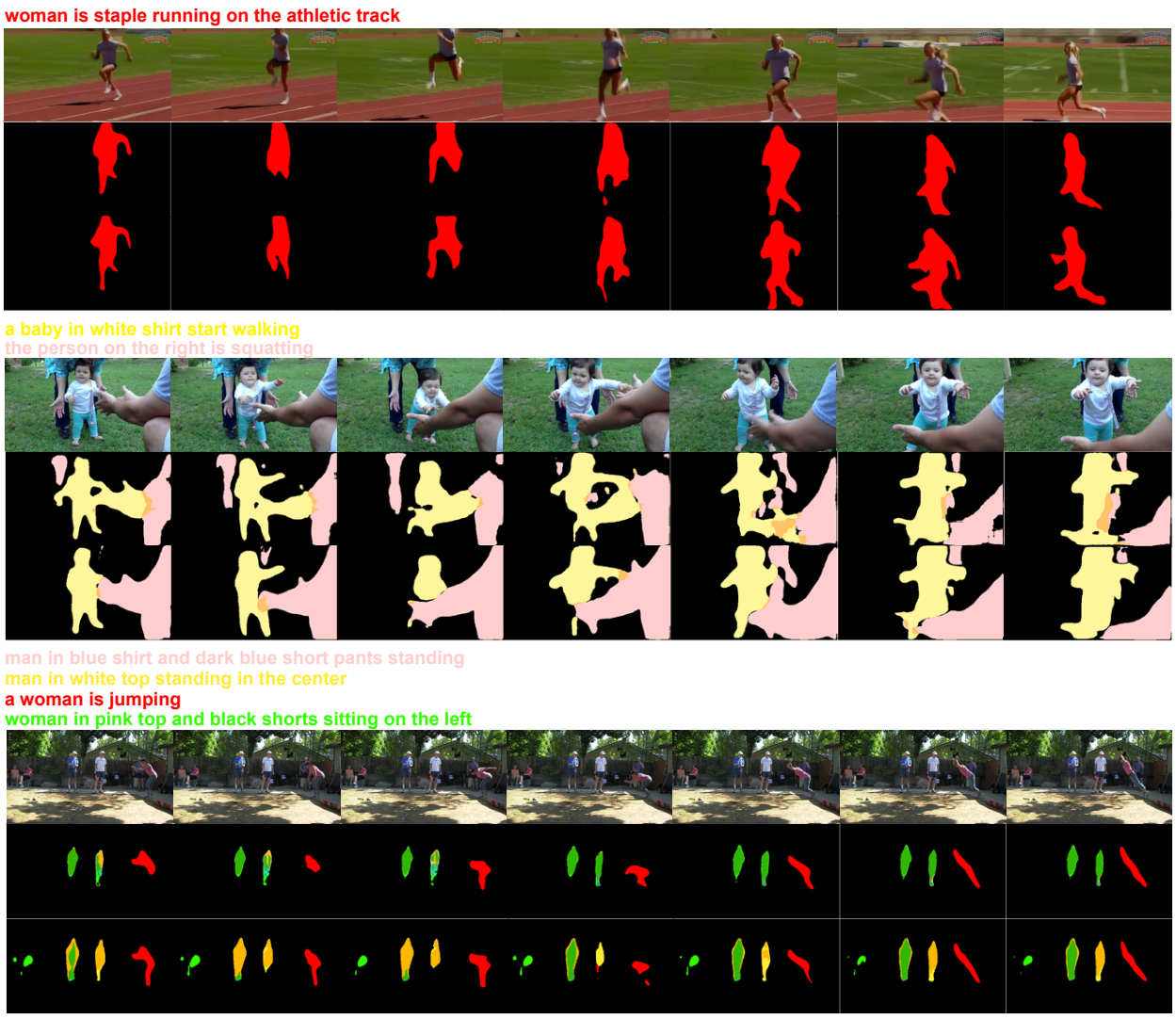

Figure 4: Qualitative results on A2D sentences. The first row shows the input frames of sample video. The second row is the segmentation results of (Gavrilyuk et al. 2018) and the third row illustrates the outputs from our approach. The input sentence and colored mask share the same color if they are corresponding. The mixtures of colors denotes the overlap of prediction.

tion and progressive interaction. Then we evaluate the *Baseline+CMDyConv* by stacking this module 3 times, which reflects the contextual information and the depth of interaction module are greatly beneficial for this task. Next, we adopt the deformable version *Baseline+DCMDyConv* to address the geometric deformations, achieving further increase on semantic segmentation. Finally, we integrate temporal modeling into *Baseline+DCMDyConv+Temporal* and the results on most metrics especially precision@0.9 show the explicit temporal modeling can bring extra improvement. Besides, the results of direct interaction on mAP@0.5:0.95, Overall IoU and Mean IoU are 32.5, 61.6 and 52.5, respectively.

Conclusion

In this paper, we have proposed a context modulated dynamic convolution network for actor and action video segmentation with language queries, which takes contextual information into consideration during the correlation learning. Our approach achieves significant improvement on segmentation performance. In the future, we should denote more efforts to learn the fine-grained corresponding between vision and language.

Acknowledgements

Our work was supported by National Natural Science Foundation of China under Grant 61572388 and 61703327, the Key R&D Program-The Key Industry Innovation Chain of Shaanxi under Grant 2018ZDXM-GY-176, the National Key R&D Program of China under Grant 2016YFE0200400 and 2017YFE0104100.

References

Brox, T., and Malik, J. 2010. Object segmentation by long term analysis of point trajectories. In *ECCV*, 282–295.

Carreira, J., and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 6299–6308.

Chen, J.; Ma, L.; Chen, X.; Jie, Z.; and Luo, J. 2019. Localizing natural language in videos. In *AAAI*, 8175–8182.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv:1412.3555*.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *ICCV*, 764–773.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Gavrilyuk, K.; Ghodrati, A.; Li, Z.; and Snoek, C. G. 2018. Actor and action video segmentation from a sentence. In *CVPR*, 5958–5966.

Hendricks, L. A.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In *ICCV*, 5803–5812.

Hu, R.; Rohrbach, M.; and Darrell, T. 2016. Segmentation from natural language expressions. In *ECCV*, 108–124.

Huang, J.; Li, N.; Zhang, T.; Li, G.; Huang, T.; and Gao, W. 2018. Sap: Self-adaptive proposal model for temporal action detection based on reinforcement learning. In *AAAI*.

Jhuang, H.; Gall, J.; Zuffi, S.; Schmid, C.; and Black, M. J. 2013. Towards understanding action recognition. In *ICCV*, 3192–3199.

Ji, J.; Buch, S.; Soto, A.; and Carlos Niebles, J. 2018. End-to-end joint semantic segmentation of actors and actions in video. In *ECCV*, 702–717.

Kalogeiton, V.; Weinzaepfel, P.; Ferrari, V.; and Schmid, C. 2017. Joint learning of object and action detectors. In *ICCV*, 4163–4172.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 787–798.

Li, S.; Xiao, T.; Li, H.; Zhou, B.; Yue, D.; and Wang, X. 2017a. Person search with natural language description. In *CVPR*, 1970–1979.

Li, Z.; Tao, R.; Gavves, E.; Snoek, C. G.; and Smeulders, A. W. 2017b. Tracking by natural language specification. In *CVPR*, 6495–6503.

Margffoy-Tuay, E.; Pérez, J. C.; Botero, E.; and Arbeláez, P. 2018. Dynamic multimodal instance segmentation guided by natural language queries. In *ECCV*, 630–645.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, 3111–3119.

Mithun, N. C.; Paul, S.; and Roy-Chowdhury, A. K. 2019. Weakly supervised video moment retrieval from text queries. In *CVPR*, 11592–11601.

Noh, H.; Hongsuck Seo, P.; and Han, B. 2016. Image question answering using convolutional neural network with dynamic parameter prediction. In *CVPR*, 30–38.

Opazo, C. R.; Marrese-Taylor, E.; Saleh, F. S.; Li, H.; and Gould, S. 2019. Proposal-free temporal moment localization of a natural-language query in video using guided attention. *arXiv:1908.07236*.

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *AAAI*.

Shi, H.; Li, H.; Meng, F.; and Wu, Q. 2018. Key-word-aware network for referring expression image segmentation. In *ECCV*, 38–54.

Wang, H.; Deng, C.; Yan, J.; and Tao, D. 2019. Asymmetric cross-guided attention network for actor and action video segmentation from natural language query. In *ICCV*, 3939–3948.

Wei, K.; Yang, M.; Wang, H.; Deng, C.; and Liu, X. 2019. Adversarial fine-grained composition learning for unseen attribute-object recognition. In *ICCV*, 3741–3749.

Xie, D.; Deng, C.; Wang, H.; Li, C.; and Tao, D. 2019. Semantic adversarial network with multi-scale pyramid attention for video classification. In *AAAI*, 9030–9037.

Xingjian, S.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; and Woo, W.-c. 2015. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NeurIPS*, 802–810.

Xu, C., and Corso, J. J. 2016. Actor-action semantic segmentation with grouping process models. In *CVPR*, 3083–3092.

Xu, C.; Hsieh, S.-H.; Xiong, C.; and Corso, J. J. 2015. Can humans fly? action understanding with multiple classes of actors. In *CVPR*, 2264–2273.

Xu, J.; Bert, D. B.; Tinne, T.; and Luc V, G. 2016. Dynamic filter networks. In *NeurIPS*, 667–675.

Yamaguchi, M.; Saito, K.; Ushiku, Y.; and Harada, T. 2017. Spatio-temporal person retrieval via natural language queries. In *ICCV*, 1453–1462.

Yan, Y.; Xu, C.; Cai, D.; and Corso, J. J. 2017. Weakly supervised actor-action segmentation via robust multi-task ranking. In *CVPR*, 1298–1307.

Yang, Y.; Deng, C.; Gao, S.; Liu, W.; Tao, D.; and Gao, X. 2016a. Discriminative multi-instance multitask learning for 3d action recognition. *IEEE Trans. Multimedia* 19(3):519–529.

Yang, Y.; Deng, C.; Tao, D.; Zhang, S.; Liu, W.; and Gao, X. 2016b. Latent max-margin multitask learning with skelets for 3-d action recognition. *IEEE Trans. Cybern.* 47(2):439–448.

Yang, Y.; Liu, R.; Deng, C.; and Gao, X. 2016c. Multi-task human action recognition via exploring super-category. *IEEE Trans. Signal Process.* 124:36–44.

Zhang, X.; Zhou, X.; Lin, M.; and Sun, J. 2018. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*, 6848–6856.