Computer Vision and Image Understanding
journal homepage: www.elsevier.com

# Cascade Multi-Head Attention Networks for Action Recognition

Jiaze Wang[a], Xiaojiang Peng[*,a], Yu Qiao[†,a,]

[a]*Shenzhen Key Lab of Computer Vision and Pattern Recognition, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China*

ABSTRACT

Long-term temporal information yields crucial cues for video action understanding. Previous researches always rely on sequential models such as recurrent networks, memory units, segmental models, self-attention mechanism to integrate the local temporal features for long-term temporal modeling. Recurrent or memory networks record temporal patterns (or relations) by memory units, which are proved to be difficult to capture long-term information in machine translation. Self-attention mechanisms directly aggregate all local information with attention weights which is more straightforward and efficient than the former. However, the attention weights from self-attention ignore the relations between local information and global information which may lead to unreliable attention. To this end, we propose a new attention network architecture, termed as Cascade multi-head ATtention Network (CATNet), which constructs video representations with two-level attentions, namely multi-head local self-attentions and relation based global attentions. Starting from the segment features generated by backbone networks, CATNet first learns multiple attention weights for each segment to capture the importance of local features in a self-attention manner. With the local attention weights, CATNet integrates local features into several global representations, and then learns the second level attention for the global information by a relation manner. Extensive experiments on Kinetics, HMDB51, and UCF101 show that our CATNet boosts the baseline network with a large margin. With only RGB information, we respectively achieve 75.8%, 75.2%, and 96.0% on these three datasets, which are comparable or superior to the state of the arts.



## 1. Introduction

Video-based action recognition is one of the prime issues in computer vision. With the availability of powerful parallel machines and massive increase in training data, deep convolutional neural networks (CNNs) achieve superhuman performance on image based tasks (Krizhevsky et al., 2012; Schroff et al., 2015). However, the progress of CNNs for video analysis still lags their image counterparts due to much more complex spatial-temporal structures.

Earlier works address the action recognition by extracting local hand-crafted features and aggregating them into video representations (Peng et al., 2016; Wang and Schmid, 2013). The two-stream convolutional networks (Simonyan and Zisserman, 2014a) first shed lights on deep CNNs based video action recognition which separately train a spatial network based on RGB frames and a temporal network based on optical flows. Recent researches mainly focus on two directions: (a) 3D convolution based spatiotemporal feature learning and (b) temporal modeling with frame-level/clip-level features.

For spatiotemporal feature learning, instead of using 2D CNNs on frames or optical flows as in (Simonyan and Zisserman, 2014a), most of the methods extend traditional 2D CNNs to 3D or the variations of 3D. Tran *et al.* (Tran et al., 2015) introduce a deep 3-dimensional convolutional network to learn C3D (Convolutional 3D) features for short clips. C3D is trained from scrach on very large video datasets. To inherit the advantages of well-trained deep 2D CNNs, Carreira *et al.* (Carreira and Zisserman, 2017) propose an Inflated 3D ConvNet (I3D) which initializes 3D ConvNets by inflating very deep image classification ConvNets. Tran *et al.* (Tran et al., 2018) factorize the 3D convolutional filters into separate spatial and temporal components, and present a 'R(2+1)D' block for spatiotempo-

*Equally-contributed first author: Xiaojiang Peng.
†Corresponding author: Yu Qiao. e-mail: yu.qiao@siat.ac.cn (Yu Qiao)

ral feature learning. Wang *et al.* (Wang et al., 2018b) propose a spacetime non-local operation which makes it possible for a certain position's response to pay attention to all other positions. Overall, these methods only model spatiotemporal information on short clips, and a further aggregation is needed to obtain video-level results.

Temporal modeling methods aim to model long-term temporal information based on frame-level or clip-level features to get better video representations. Yue *et al.* (Yue-Hei Ng et al., 2015) explore five feature-pooling strategies and introduce Long Short-Term Memory (LSTM) cells to model long-term information. Instead of using dense frame-level features, Wang *et al.* (Wang et al., 2016) propose a temporal segment network (TSN) to train video-level representation directly where a training sample is constructed by several uniformly-sampled frames from a video. Inspired by the multi-head attention of recent self-attention (Vaswani et al., 2017) method in machine translation, Long *et al.* (Long et al., 2018) introduce a similar 'attention clusters' method which aggregates all the frame-level features (extracted offline) into several weighted mean vectors by groups of attention weights. These attention clusters represent multiple aspects of a video which may be objects, motions, etc. However, the relation between these components is not fully considered which may provide rich temporal information (Santoro et al., 2017; Zhou et al., 2017). These self-attention weights may be not reliable due to that they are achieved by local features without accessing to the whole video representation.

To further exploit the attention mechanism in both local and global level, in this paper, we propose a Cascade multi-head ATtention Network (CATNet), which constructs video representations with multi-head local attention and global attention in an end-to-end framework. With local frame-level or clip-level features, the CATNet first aggregates them into multiple global features by the multi-head attention module. The multi-head attention module captures different aspects of a given video. The CATNet then models the relations between these global features and a global video representation by the global attention module. The global attention module captures more reliable attention with the help of the global video representation. To validate our CATNet, we conduct extensive experiments on three widely-used datasets: Kinetics, UCF101, and HMDB51. Our method achieves performance superior or comparable to that of recent state-of-the-art methods.

## 2. Related Works

Video action recognition has drawn great attention in the past years. Many works have devoted to designing CNNs for action recognition. In this section, we briefly review the existing related works.

**Action Recognition.** Researchers adapt CNNs which achieve great success on image recognition tasks to videos mainly by the following three ways:

*Recurrent neural network.* RNNs have been applied in many papers to model long-term relation in action recognition. Long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) plays a significant role in RNN in the recent years. Different architectures to combine image information across a video in a long time are proposed to handle full length videos (Yue-Hei Ng et al., 2015). An end to end trainable RNN which can be compositional in spatial and temporal layers is proposed to produce a variable-length prediction (Donahue et al., 2015). Unsupervised manner has also been studied by the model uses an encoder LSTM to map an input sequence (Srivastava et al., 2015). Lattice-LSTM (Sun et al., 2017) which extends LSTM by learning independent hidden state transitions of memory cells for individual spatial locations and multi-modal training procedure is proposed in the work. VideoLSTM (Li et al., 2018b) which introduces convolutions to exploit the spatial correlations and shallow convolutional for motion information is proposed. However, the RNN methods' performance is unsatisfactory in action recognition, which means RNN may not be the best solution for the task. Our proposed method explores another way of generating global features and feature aggregation.

*Spatio-temporal convolutions.* Spatio-temporal convolution is first presented to extract features from both the spatial and the temporal dimensions by using 3D convolutions (Ji et al., 2013). Then convolutional 3D (C3D) feature has made great progress in action recognition (Tran et al., 2015). It is so general that inspire many works on C3D feature (Molchanov et al., 2016; Diba et al., 2016; Camgoz et al., 2016; Wang et al., 2018a). Two-stream 3D-convNet Fusion is proposed to recognize actions in arbitrary size and length videos with multiple features (Wang et al., 2018a). Temporal linear encoding which captures the appearance and motion throughout entire videos further improved the performance of 3D CNNs (Diba et al., 2017). Compared to 2D CNNs, 3D CNNs convert spatial filters to spatio-temporal ones and therefore greatly increase parameters. Many works are proposed to solve the problem (Sun et al., 2015; Qiu et al., 2017; Zhou et al., 2018). Two-Stream Inflated 3D ConvNet (I3D) that is based on 2D ConvNet inflation achieved great success on action recognition and proved that Kinetics have sufficient data for training 3D CNNs (Carreira and Zisserman, 2017). The architectures of various 3D CNNs have been explored and obtain good results (Hara et al., 2018). On this foundation, R(2+1)D improve the accuracy significantly by factorizing the 3D convolutional filters into separate spatial and temporal components without reducing parameters (Tran et al., 2018). In the paper, we use ResNet I3D as our baseline model and generate features with spatio-temporal information

*Multiple streams.* Many works focus on improving the network performance by processing multiple streams. Two streams approaches have achieved good results on action recognition, which using RGB and optical flow as inputs (Feichtenhofer et al., 2016b; Simonyan and Zisserman, 2014b). Human pose and skeleton images are also very important cues for action recognition (Chéron et al., 2015; Choutas et al., 2018; Yang et al., 2018). What' more, researchers have tried many other ways to capture motion information, like Optical Flow guided Feature (OFF) (Sun et al., 2018) and features from compressed data (Wu et al., 2018). Although these inputs can bring improvement, but also increase the amount of calculation at the same time. To guarantee the practicability of our model, we

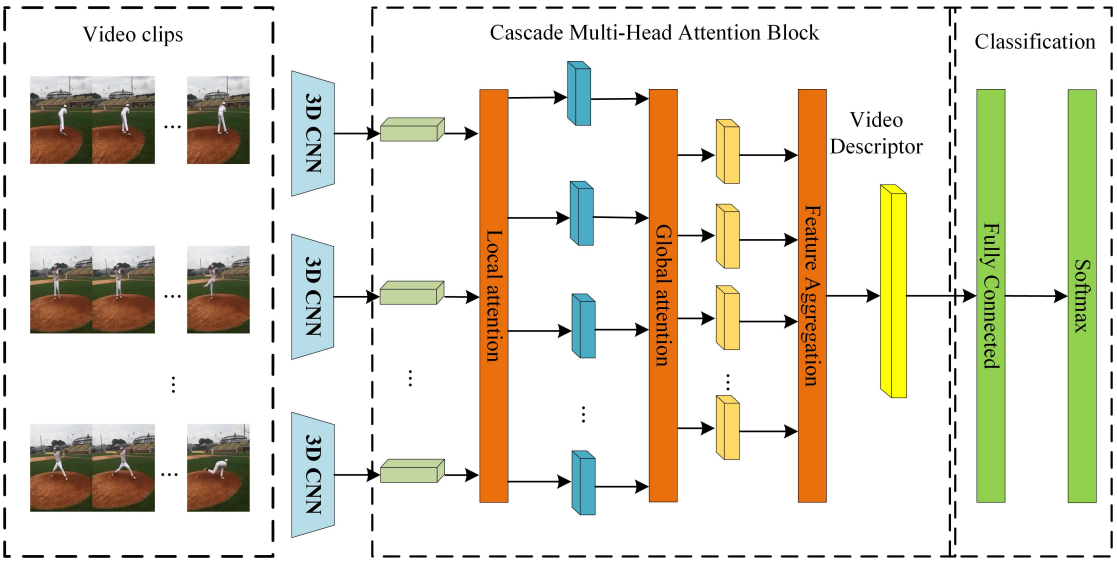


**Fig. 1. Cascade Multi-Head Attention Networks: one video is sampled into $K$ clips with the length of $L$. These clips are fed into 3D CNNs for feature extraction. The clip features are further fed into a local attention module and a global attention module to get video representations.**

focus on designing an architecture with only RGB inputs.

**Attention Mechanisms.** Attention mechanisms are first presented on the basis of REINFORCE algorithm. An RNN based on attention mechanisms which only processed the selected regions at high resolution is proposed by Mnih et al (Mnih et al., 2014). However, due to the binary choices when training, the method doesn't perform so well as expected. Then, researchers propose soft attention mechanisms which use weighted averages to replace hard selections (Bahdanau et al., 2014). Next, many self-attention models are proposed for different tasks, like LSTM for machine reading (Cheng et al., 2016), multi-head attention for machine translation (Vaswani et al., 2017) and attention clusters for video classification (Long et al., 2018). However, these works only focus on how to generate global features, but fail to effectively exploit the relation information between different features.

Recently, relational reasoning module is proposed for visual question answering and get super-human performances (Santoro et al., 2017). Inspired by this work, researchers propose a relational reasoning module for action recognition to learn various temporal relations in videos in a supervised learning setting (Zhou et al., 2017). A general attention neural cell which estimates both the attention probablity at each spatial location and each video segment is proposed for action recognition (Li et al., 2018a). On the basis of previous works, we propose an architecture that can generate global features by multi-head attention module and analyze the relation between features by relational reasoning module.

## 3. Cascade Multi-Head Attention Network

In this section, we first overview our cascade multi-head attention network and then detail the local attention module and the global attention module.

### *3.1. Overview of Our CATNet*

As we discussed in Introduction, existing temporal modeling methods such as TSN, Self-Attention Clusters, and LSTM lack the ability of modeling relations between local components and the video representation. This is mainly due to their limited access to the global video representation within the network. Inspired by the Multicolumn Networks in face recognition (Xie and Zisserman, 2018), we propose the cascade multi-head attention network (CATNet) for action recognition to capture the relational information, which is illustrated in Figure 1.

Similar to most of temporal modeling methods, our CATNet operates on a sequence of short clips sampled from the entire video. For one video, we sample $K$ clips with the length of $L$, and extract features based on recent spatiotemporal networks. Specially, we use an I3D network as the backbone feature extraction module. These clip features are then fed into the local attention module which is a multi-head attention structure. The outputs of the local attention module are weighted mean features with different attention weights. We further average these features to generate a global video representation for the next global attention module. We then combine the global video representation with different weighted mean features for relation reasoning. The global attention module learns attention for these fused vectors and generates the final video representation by feature aggregation. Softmax is finally used for classification.

### *3.2. Local Attention Module*

Formally, we denote the $K$ clips as $I_1, I_2, \cdots, I_K$ for a video, and the feature extraction network $r(\cdot;\theta_1)$ with $\theta$ as its parameters, then we obtain clip-level features set $X$ as,

$$X = [\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_K] = [r(I_1;\theta_1), \cdots, r(I_K;\theta_1)], \quad (1)$$

where each video clip $I_i \in R^{H\times W\times 3\times L}$, $H$ and $W$ refer to the height and width of input video clips respectively, $L$ is the length of the video clip, and the local feature $\mathbf{x}_i \in R^M$

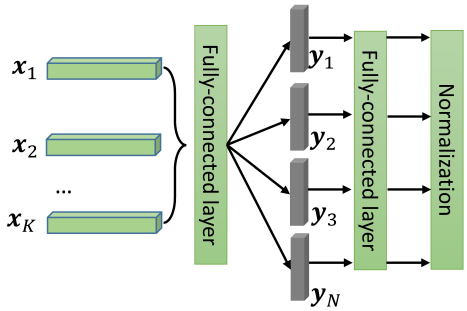


**Fig. 2. Local Attention Module. It consists of two fully-connected (FC) layers with the first FC aims to learn self-attention weights and the second FC to learn diverse attention features.**

The local attention module is shown in Figure 2. It consists of two fully-connected (FC) layers. The first FC aims to learn self-attention weights, the second FC and the normalization aims to learn diverse attention features. One of the attention weights for an input $\mathbf{x}_j$ is defined as follows,

$$\alpha_{ij} = \frac{e^{\mathbf{x}_j^T \mathbf{w}_i}}{\sum_k e^{\mathbf{x}_k^T \mathbf{w}_i}}. \tag{2}$$

Each output $\mathbf{y}_i$ of the first FC (see the gray boxes of Figure 2) is the weighted sum of original features with $i-th$ head attention, which is defined as follows,

$$\mathbf{y}_i = \sum_{j=1}^{K} \alpha_{ij} \mathbf{x}_j. \tag{3}$$

As discussed in (Long et al., 2018; Vaswani et al., 2017), a single self-attention weighted global feature may only reflect a certain aspect of the video. To this end, multi-head attention mechanism is used to capture more aspects. As shown in Figure 2, we use $N$ heads, i.e. $\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N$.

As found in the experiments, these self-attention global features always tend to focus on similar signals. To increase the diversity of self-attention features, the linear transformation and L2-normalization is used to shift the weighted sum in the feature space. The scale invariance can be ensured by the using of linear transformation and L2-normalization, and the weighted sum is shifted in the feature space. Thus these operations make the features different from each other and have different distributions, and scale invariance helps to optimize the whole network.

$$Y' = [\frac{y_1{}'}{\sqrt{N}||y_1{}'||_2}, \frac{y_2{}'}{\sqrt{N}||y_2{}'||_2}, \cdots, \frac{y_N{}'}{\sqrt{N}||y_N{}'||_2}] \tag{4}$$

where $y'$ is the linear transformation of $y$ with a FC layer.

### 3.3. Global Attention Module

To estimate the relation between each $Y_i'$ and the video representation, we first approximate a video representation as follows,

$$G = \frac{\sum_{i=1}^{N} \mathbf{y}_i{}'}{N}, \tag{5}$$

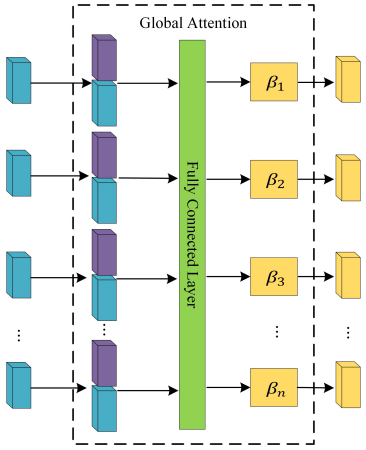


**Fig. 3. Global Attention Module. It learns attention weights based on a global video representation.**

and then combine them with operator $C(\mathbf{y}_i', G)$. Luong *et al.* (Luong et al., 2015) investigate three different alternatives for this operator, we choose the concatenation operation in our work empirically.

The global attention module is illustrated in Figure 3. It takes as input the self-attention features and the approximate video representation, and then learns attention weights with a FC layer connected to the concatenation of video representation and self-attention features. We call this module as global attention module since it operates on global features. Each attention weight $\beta_i$ can be formulated as follows,

$$\beta_i = \text{sigmoid}(\mathbf{w}^T[\mathbf{y}_i'; G]), \tag{6}$$

where $\mathbf{w}$ is the parameter of a FC layer, $[\mathbf{y}_i'; G]$ denotes the concatenation of $\mathbf{y}_i'$ and $G$.

### 3.4. Feature Aggregation

With the global attention weights, we investigate two aggregation methods to obtain the final video representation.

**Summation**. The summation method achieves the weighted mean of all the inputs as follows,

$$V = \frac{\sum_{i=1}^{n} \beta_i \mathbf{y}_i{}'}{\sum_{i=1}^{N} \beta_i} \tag{7}$$

**Concatenation**. Since the number of local attention features is fixed as $K$, we can concatenate them with global attention weights as follows,

$$V = [\beta_1 \mathbf{y}_1{}', \beta_2 \mathbf{y}_2{}', \cdots, \beta_n \mathbf{y}_n{}'] \tag{8}$$

## 4. Experiments

### 4.1. Experiment Settings

**Datasets.** We perform comprehensive studies on the challenging Kinetics dataset. We also report results on the UCF101 dataset and HMDB51 dataset to prove the generality of our models.

*Kinetics* contains around 246 K training videos and 20K validation videos. Each video clip lasts around 10s. It is a challenging action recognition task with 400 categories. Our models are trained on training set and test on validation set. Due to the database has a large amount of data and many categories. So the experiment results on Kinetics are more convincing and stable. We perform comprehensive studies and reasonable analysis on Kinetics.

*UCF101* dataset (Soomro et al., 2012) contains 101 action classes which can be divided into five types: human-object interaction, body-motion only, human-human interaction, playing musical instruments, and sports. Totally, it has 13,320 video clips with a fixed frame rate and resolution as 25 FPS and 320 240, respectively. It has three train/test splits, we perform our model on all three splits and report the average test results.

*HMDB51* (Kuehne et al., 2011) contains 6849 clips divided into 51 action categories. The video clips are collected from YouTube and the average duration of each video is about 3 seconds. Like UCF101, it has three train/test splits, and we report the average test results.

**Local Feature Extraction.** To reduce the amount of computation during training, we exact local features first instead of training end to end. The input video is divided into 10 clips. For each clip of Kinetics, we sample 32 frames with a stride of 4 frames. For each clip of UCF101 and HMDB51, we sample 32 successive frames. We initialize ResNet-50 I3D with a model pre-trained on ImageNet and fine-tune it by video clips of the training set. Then we use the fine-tuned model to extract the local features for each video clip.

**Data augmentation.** Data augmentation is an important part in training CNNs, which can increase the diversity of data and avoid overfitting. When training ResNet-50 I3D, just like (Simonyan and Zisserman, 2014b), the spatial size is [224, 224] pixels, randomly cropped from a scaled video whose shorter side is randomly sampled in [256, 320] pixels. For each clip, it will be cropped three times, thus we can get 30 local features for a single video. Due to not all videos clips contain enough useful information, we only need several important clips to understand the video. When we train CATNet, we can randomly choose a few features to train and use all features when testing. In this way we can reduce the training time and ensure the accuracy of the model at the same time.

**Training configuration.** In the experiments, we use 32 attention units and concatenation for feature aggregation as default. Mini-batch stochastic gradient descent algorithm is used to learn network parameters. We set the batch size to 64 and momentum to 0.9. We train our network on an 8-GPU machine, so each GPU has 8 video clips. The initial learning rate is set to 0.01 and decreases to its 0.1 every 30 epochs. The maximum training epoch is set to 90.

### *4.2. Exploration of Our CATNet on Kinetics*

In this section, we first compare our CATNet to the baseline with default settings, and then evaluate the feature aggregation strategies as well as the number of self-attention in local attention module, and we finally give some analysis and compare our CATNet to several state-of-the-arts methods on Kinetics.

**Table 1. Comparison to the baseline model on Kinetics validation set. The second FC layer and normalization in local attention module are denoted as FC-Norm.**

| Attention Module | Top-1(%) | Top-5(%) |
|---|---|---|
| ResNet-50 I3D(Baseline) | 73.0 | 91.0 |
| Local Attention Module w/o FC-Norm | 74.6 | 91.5 |
| Local Attention Module w/ FC-Norm | 75.1 | 91.7 |
| Our CATNet | **75.8** | **92.0** |

**Table 2. Accuracy of varied number of self-attention features and different aggregation strategies on Kinetics validation set.**

| n | Concatenation | | Summation | |
|---|---|---|---|---|
| | Top-1(%) | Top-5(%) | Top-1(%) | Top-5(%) |
| 1 | 74.9 | 91.7 | 75 | 91.5 |
| 4 | 75 | 91.6 | 75 | 91.7 |
| 8 | 75.2 | 91.7 | 75.1 | 91.6 |
| 16 | 75.1 | 91.8 | 75.1 | 91.6 |
| 32 | **75.8** | **92.0** | 75.1 | 91.7 |
| 64 | 75.6 | 91.9 | 75.2 | 91.7 |

**Comparison to baseline**. We use the ResNet50-I3D network as our baseline model and train it from scratch on Kinetics. As shown in Table 1, our baseline obtains 73% in top-1 accuracy. To evaluate the effect of each module, we test the performance of local attention module w/o FC-Norm, local attention module w/ FC-Norm, and our CATNet respectively. As we can see from Table 1, the local attention module w/o FC-Norm boosts the baseline significantly, which indicates multi-head attention mechanism is useful for action recognition and can recognize actions effectively. Our local attention module with FC-Norm outperforms the baseline model by 2.1% on top-1 accuracy and 0.7% on top-5 accuracy, which proves the availability of linear transformation and normalization of local attention module. Adding the global attention module further improves the top-1 accuracy by 0.7%. Overall, our CATNet outperforms the baseline by 2.8% on top-1 accuracy and 1.0% on top-5 accuracy.

**Evaluation of multi-head and aggregation**. Table 2 presents the performance of our CATNet with varied number of self-attention features and different aggregation strategies on Kinetics. For the aggregation strategy, summation is not sensitive to the variation of the multi-head structure (*i.e.* different $n$) and consistently improves the baseline method. On the contrary, concatenation is sensitive to $n$ and slightly superior to summation. For the multi-head structure with concatenation, increasing the number of self-attention features slightly improves the performance and saturates after 32.

**Comparison to state of the arts**. We compare our CATNet to several state-of-the-art methods in Table 3. Considering the RGB stream, our CATNet outperforms I3D by 3.7% and outperforms R(2+1)D by 1.5%. However, CATNet is slightly worse than NL I3D (0.7%), which may be caused by that NL I3D uses a better baseline model with 75.2% top-1 accuracy. 'Attention Clusters' is very similar with our local attention module but differs in the operation used for generating diverse self-attention features. Our CATNet is slightly better than the 'Attention Clusters'.

**Table 3. Comparison with state-of-the-art methods on Kinetics validation set.**

| Method | Top-1(%) | Top-5(%) |
|---|---|---|
| Two-Stream (Carreira and Zisserman, 2017) | 62.2 | - |
| ConvNet+LSTM (Carreira and Zisserman, 2017) | 63.3 | - |
| I3D-RGB (Carreira and Zisserman, 2017) | 72.1 | 90.3 |
| R(2+1)D-RGB (Tran et al., 2018) | 74.3 | 91.4 |
| Attention Clusters-RGB (Long et al., 2018) | 75.0 | 91.9 |
| NL I3D (Wang et al., 2018b) | **76.5** | **92.6** |
| ResNet-50 I3D (our baseline) | 73.0 | 91.5 |
| CATNet (our method) | 75.8 | 92.0 |

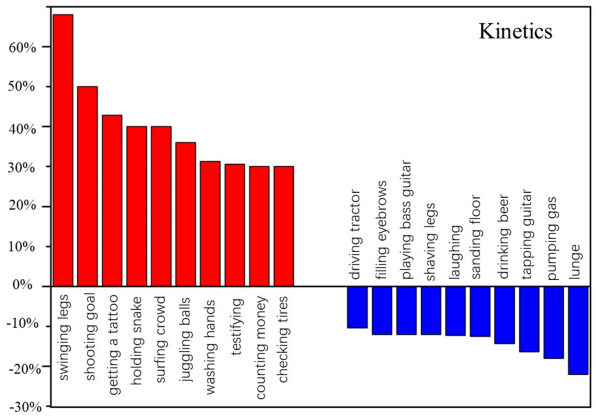


**Fig. 4. Per-category top-1 accuracy improvement compared to the baseline model on Kinetics. Red bars: the top-10 classes with significant gains. Blue bars: the top-10 classes with performance degradation.**

**Analysis and visualization**. We analyze the effect of our CATNet in Figure 4, Figure 5 and Figure 6. Figure 4 shows the per-category difference (in top-1 accuracy) of the top-10 performance classes which are better or worse than the baseline model. We observe that our CATNet mainly improves the accuracy of those categories with strong motions, such as swimming legs and shooting goal. Specially, our CATNet improves the top-1 accuracy of swimming legs by 68%, and it also brings obvious improvement for 20.8% of all the categories and degradation for around 10% of the categories in our observation.

As shown in Figure 5, we calculate the accuracy of each category and sort them in a descending order. We notice that 50.8% categories have an accuracy higher than 80% and 90.5% categories have an accuracy higher than 50%. What's more, we find out that category with high accuracy has clear motion pattern or the action duration lasts longer. Therefore, in the future studies, how to recognize actions with short duration and invisible motion is a challenging issue.

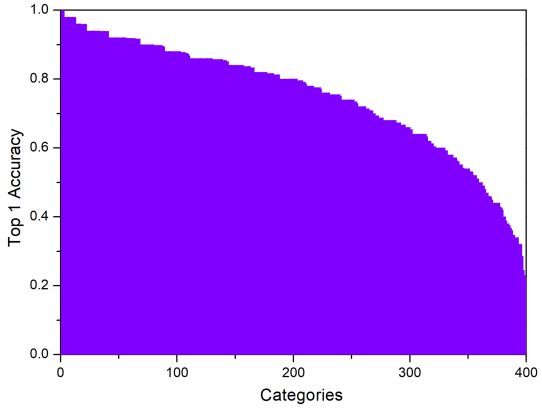


**Fig. 5. Accuracy distribution of CATNet on Kinetics validation set. The category is sorted by test accuracy in a descending order.**

In order to better illustrate what local attention module has learned, we visualize some examples of local attention weights on the validation data of Kinetics in Figure 6. In this illustration, we summarize the multi-head attention weights for each clip (i.e. $\sum_i \alpha_{ij}$ for $I_j$ in Eq. (1)) and visualize a video clip by 3 frames with a stride of 10. Each row represents one video. The first 3 images show frames with highest attention weights, and the middle 4 images are frames with medium weights while the last 4 images are frames with lowest weights. We see that the local attention module of our CATNet is able to highlight diverse important clips and to avoid irrelevant frames corresponding to static background or nonaction poses.

### *4.3. Experiments on UCF101 and HMDB51*

To validate the generality and effectiveness of our CATNet, we conduct experiments on UCF101 and HMDB51 with the above pre-trained CATNet. We transfer the model trained on Kinetics to these two datasets by finetuning.

We compare our CATNet to the state-of-the-art methods in Table 4. With the RGB stream only, our method achieves performance comparable or superior to these methods on both datasets. On UCF101, CATNet outperforms ARTNet by 1.7% and I3D-RGB by 0.4%. Our method can outperform STAN which proposes a unified spatio-temporal attention architecture by 2.4%. And we can outputform VideoLSTM which introduce soft-attention to LSTM by 3.8% on UCF101 and by 1.5% on HMDB51 .It even outperforms TSN with both RGB and flow streams by more than 1.4%. 'Attention Clusters' uses a similar attention scheme with our local attention module, but it is even inferior to our CATNet (with RGB only) with both RGB and flow streams. Overall, the good performance demonstrates the effectiveness of our CATNet.

## 5. Conclusions

In this paper, we present the Cascade multi-head ATtention Network (CATNet), a novel video-level architecture to construct video representations for action recognition. CATNet is a high-performance network to integrate clip-level features with multi-head attention module and to capture relations between different video aspects with global attention module. We conduct extensive experiments on three popular action recognition datasets: Kinetics, UCF101, and HMDB51. Our proposed CATNet achieves performance comparable or superior to state-of-the-art methods on these datasets. We hope our work can inspire new studies on feature integration and CATNet can be an important network architecture in the future.

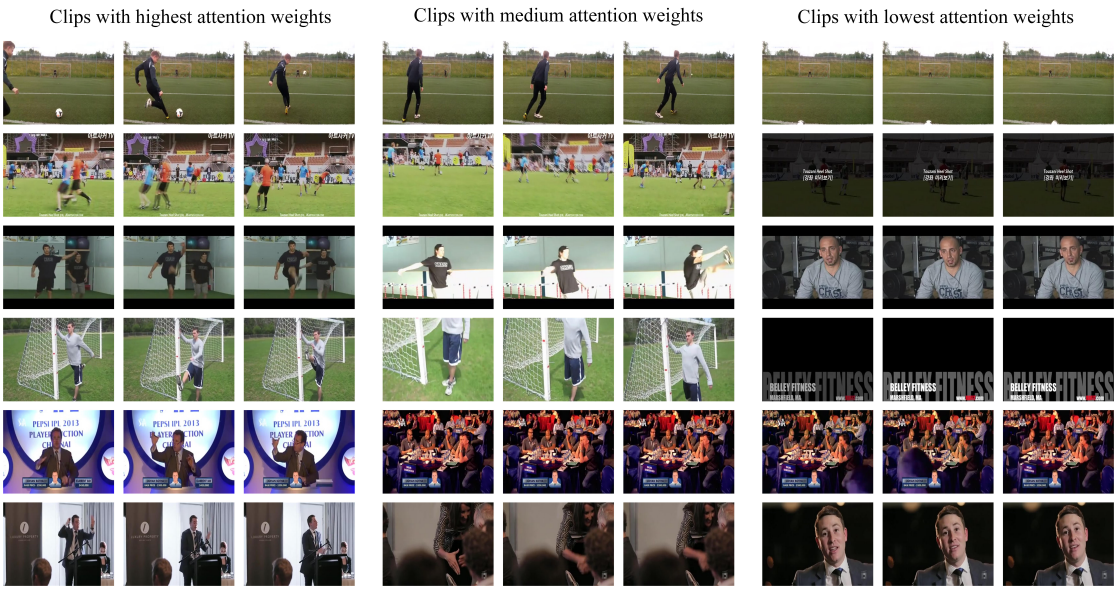


**Fig. 6. Visualization of local attention weights on the validation data of Kinetics.The above two videos are from shooting goal (soccer) category, the middle two videos are from swinging legs category and the last two videos are from auctioning category.**

**Table 4. Comparison with state-of-the-art on UCF101 and HMDB51.**

| Method | Modality | UCF101(%) | HMDB51(%) |
|---|---|---|---|
| Two-Stream (Simonyan and Zisserman, 2014a) | RGB+flow | 88.0 | 59.4 |
| VideoLSTM + iDT + Objects(Li et al., 2018b) | RGB+flow | 92.2 | 73.7 |
| Two-Stream Fusion +IDT (Feichtenhofer et al., 2016b) | RGB+flow | 93.5 | 69.2 |
| Spatiotemp. ResNet (Feichtenhofer et al., 2016a) | RGB+flow | 93.4 | 66.4 |
| STAN(Li et al., 2018a) | RGB+flow+clip | 93.6 | - |
| TSN (Wang et al., 2016) | RGB+flow | 94.2 | 69.4 |
| Attention Clusters (Long et al., 2018) | RGB+flow | 94.6 | 69.2 |
| C3D (Tran et al., 2015) | RGB | 82.3 | 51.6 |
| Res3D (Tran et al., 2017) | RGB | 85.8 | 54.9 |
| ARTNet (Wang et al., 2017) | RGB | 94.3 | 70.9 |
| I3D-RGB (Carreira and Zisserman, 2017) | RGB | 95.6 | 74.8 |
| R(2+1)D-RGB (Tran et al., 2018) | RGB | **96.8** | 74.5 |
| ResNet-50 I3D (our baseline) | RGB | 93.9 | 73.6 |
| CATNet (our method) | RGB | 96.0 | **75.2** |

## References


Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 .

Camgoz, N.C., Hadfield, S., Koller, O., Bowden, R., 2016. Using convolutional 3d neural networks for user-independent continuous gesture recognition, in: Pattern Recognition (ICPR), 2016 23rd International Conference on, IEEE. pp. 49–54.

Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? a new model and the kinetics dataset, in: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, IEEE. pp. 4724–4733.

Cheng, J., Dong, L., Lapata, M., 2016. Long short-term memory-networks for machine reading. arXiv preprint arXiv:1601.06733 .

Chéron, G., Laptev, I., Schmid, C., 2015. P-cnn: Pose-based cnn features for action recognition, in: Proceedings of the IEEE international conference on computer vision, pp. 3218–3226.

Choutas, V., Weinzaepfel, P., Revaud, J., Schmid, C., 2018. Potion: Pose motion representation for action recognition, in: CVPR 2018.

Diba, A., Pazandeh, A.M., Van Gool, L., 2016. Efficient two-stream motion and appearance 3d cnns for video classification. arXiv preprint arXiv:1608.08851 .

Diba, A., Sharma, V., Van Gool, L., 2017. Deep temporal linear encoding networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T., 2015. Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2625–2634.

Feichtenhofer, C., Pinz, A., Wildes, R., 2016a. Spatiotemporal residual networks for video action recognition, in: Advances in neural information processing systems, pp. 3468–3476.

Feichtenhofer, C., Pinz, A., Zisserman, A., 2016b. Convolutional two-stream network fusion for video action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1933–1941.

Hara, K., Kataoka, H., Satoh, Y., 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 18–22.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.

Ji, S., Xu, W., Yang, M., Yu, K., 2013. 3d convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence 35, 221–231.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with

deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T., 2011. Hmdb: a large video database for human motion recognition, in: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE. pp. 2556–2563.

Li, D., Yao, T., Duan, L.Y., Mei, T., Rui, Y., 2018a. Unified spatio-temporal attention networks for action recognition in videos. IEEE Transactions on Multimedia 21, 416–428.

Li, Z., Gavrilyuk, K., Gavves, E., Jain, M., Snoek, C.G., 2018b. Videolstm convolves, attends and flows for action recognition. Computer Vision and Image Understanding 166, 41–50.

Long, X., Gan, C., de Melo, G., Wu, J., Liu, X., Wen, S., 2018. Attention clusters: Purely attention based local feature integration for video classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7834–7843.

Luong, M.T., Pham, H., Manning, C.D., 2015. Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 .

Mnih, V., Heess, N., Graves, A., et al., 2014. Recurrent models of visual attention, in: Advances in neural information processing systems, pp. 2204–2212.

Molchanov, P., Yang, X., Gupta, S., Kim, K., Tyree, S., Kautz, J., 2016. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4207–4215.

Peng, X., Wang, L., Wang, X., Qiao, Y., 2016. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. Computer Vision and Image Understanding 150, 109–125.

Qiu, Z., Yao, T., Mei, T., 2017. Learning spatio-temporal representation with pseudo-3d residual networks, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE. pp. 5534–5542.

Santoro, A., Raposo, D., Barrett, D.G., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrap, T., 2017. A simple neural network module for relational reasoning, in: Advances in neural information processing systems, pp. 4967–4976.

Schroff, F., Kalenichenko, D., Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815–823.

Simonyan, K., Zisserman, A., 2014a. Two-stream convolutional networks for action recognition in videos, in: Advances in neural information processing systems, pp. 568–576.

Simonyan, K., Zisserman, A., 2014b. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Soomro, K., Zamir, A.R., Shah, M., 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 .

Srivastava, N., Mansimov, E., Salakhudinov, R., 2015. Unsupervised learning of video representations using lstms, in: International conference on machine learning, pp. 843–852.

Sun, L., Jia, K., Chen, K., Yeung, D.Y., Shi, B.E., Savarese, S., 2017. Lattice long short-term memory for human action recognition., in: ICCV, pp. 2166–2175.

Sun, L., Jia, K., Yeung, D.Y., Shi, B.E., 2015. Human action recognition using factorized spatio-temporal convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 4597–4605.

Sun, S., Kuang, Z., Sheng, L., Ouyang, W., Zhang, W., 2018. Optical flow guided feature: A fast and robust motion representation for video action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1390–1399.

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

Tran, D., Ray, J., Shou, Z., Chang, S.F., Paluri, M., 2017. Convnet architecture search for spatiotemporal feature learning. arXiv preprint arXiv:1708.05038 .

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M., 2018. A closer look at spatiotemporal convolutions for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6450–6459.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Advances in Neural Information Processing Systems, pp. 5998–6008.

Wang, H., Schmid, C., 2013. Action recognition with improved trajectories, in: Proceedings of the IEEE international conference on computer vision, pp. 3551–3558.

Wang, L., Li, W., Li, W., Van Gool, L., 2017. Appearance-and-relation networks for video classification. arXiv preprint arXiv:1711.09125 .

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L., 2016. Temporal segment networks: Towards good practices for deep action recognition, in: European Conference on Computer Vision, Springer. pp. 20–36.

Wang, X., Gao, L., Wang, P., Sun, X., Liu, X., 2018a. Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length. IEEE Transactions on Multimedia 20, 634–644.

Wang, X., Girshick, R., Gupta, A., He, K., 2018b. Non-local neural networks, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Wu, C.Y., Zaheer, M., Hu, H., Manmatha, R., Smola, A.J., Krähenbühl, P., 2018. Compressed video action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6026–6035.

Xie, W., Zisserman, A., 2018. Multicolumn networks for face recognition. arXiv preprint arXiv:1807.09192 .

Yang, Z., Li, Y., Yang, J., Luo, J., 2018. Action recognition with visual attention on skeleton images, in: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE. pp. 3309–3314.

Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G., 2015. Beyond short snippets: Deep networks for video classification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4694–4702.

Zhou, B., Andonian, A., Torralba, A., 2017. Temporal relational reasoning in videos. arXiv preprint arXiv:1711.08496 .

Zhou, Y., Sun, X., Zha, Z.J., Zeng, W., 2018. Mict: Mixed 3d/2d convolutional tube for human action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 449–458.