Qiniu Submission to ActivityNet Challenge 2018

Zhang Xiaoteng, Bao Yixin, Zhang Feiyun, Hu Kai, Wang Yicheng,
Zhu Liang, He Qinzhu, Lin Yining, Shao Jie and Peng Yao
Qiniu AtLab
Shanghai, China
shaojie@qiniu.com

Abstract

In this paper, we introduce our submissions for the tasks of trimmed activity recognition (Kinetics)[8] and trimmed event recognition (Moments in Time)[9] for Activitynet Challenge 2018. In the two tasks, non-local neural networks and temporal segment networks are implemented as our base models. Multi-modal cues such as RGB image, optical flow and acoustic signal have also been used in our method. We also propose new non-local-based models for further improvement on the recognition accuracy. The final submissions after ensembling the models achieve 83.5% top-1 accuracy and 96.8% top-5 accuracy on the Kinetics validation sets, 35.81% top-1 accuracy and 62.59% top-5 accuracy on the MIT validation sets.

1. Introduction

Activity Recognition in videos has drawn increasing attention from the research community in recent years. The state-of-the-art benchmark datasets such as ActivityNet, Kinetics, Moments in Times have contributed to the progress in video understanding.

In Activitynet Challenge 2018, we mainly focused on two trimmed video recognition tasks based on Kinetics and Moments in Times datasets respectively. The Kinetic dataset consists of approximately 500,000 video clips, and covers 600 human action classes. Each clip lasts around 10 seconds and is labeled with a single class. Similarly, the Moments in Times dataset is also a trimmed dataset, including a collection of 339 classes of one million labeled 3 second videos. The videos not only involve people, but also describe animals, objects or natural phenomena, which are more complex and ambiguous than the videos in Kinetics.

To recognize actions and events in videos, recent approaches based on deep convolution neural networks have achieved state-of-the-art results. To address the challenge, our solution follows the strategy of non-local neural network and temporal segment network. Particularly, we learn models with multi-modality information of the videos, including RGB, optical flow and audio. We find that these models are complementary with each other. Our final result is an ensemble of these models, and achieves 83.5% top-1 accuracy and 96.8% top-5 accuracy on the Kinetics validation set, 35.81% top-1 accuracy and 62.59% top-5 accuracy on the MIT validation sets.

2. Our Methods

2.1. Temporal Segment Networks

One of our base model is temporal segment network (TSN)[11]. TSN models long-term temporal information by evenly sampling fixed number of clips from the entire videos. Each sampled clips contain one or several frames / flow stacks, and produce the prediction separately. The video-level prediction is given by the averaged softmax scores of all clips.

We experiment with several state-of-the-art network architectures, such as ResNet, ResNeXt, Inception[10], Inception-ResNet, SENet[7], DPN[2]. These models are pretrained on ImageNet, and have good initial weights for further training. Table 1 and 2 show our TSN results on Kinetics and Moments in Times dataset.

Models	Top-1 acc(RGB)	Top-1 acc(Flow)
DPN107	75.95	69.60
ResNext101	75.43	None
SE-ResNet152	73.88	None
InceptionV4	73.51	68.76
ResNet152	72.04	67.13
InceptionV3	68.52	64.08

Table 1. Performance of TSN on Kinetics

2.2. Acoustic Model

While most motions can be recognized from visual information, sound contains information in another dimen-

Models	Top-1 acc(RGB)	Top-1 acc(Flow)
DPN107	31.06	None
ResNet152	30.21	None
ResNet269	None	18.53*
ResNet101	None	22.82

Table 2. Performance of TSN on Moments in Times dataset. *: Due to time limit, the training of these models was not finished.

sion. We use audio channels as complementary information to visual information to recognize certain classes, especially for the actions with better distinguishability on sound, whistling and barking for example.

We use the raw audio as input into the pre-trained VGGish model[6][4], and extract $n \times 128$ dimension features to do classification (n is the length of audio). Besides, we extract MFCC features from raw audio and train with SE-ResNet-50 (Squeeze-and-Excitation Network) and ResNet-50 (Deep residual network[5]). After ensembling with visual models, we achieve 0.7% improvement in top 1 error rate.

Table 3 shows our acoustic results on on Kinetics and Moments in Times dataset. Figure 1 shows 15 classes with best top 1 accuracy in MIT validation dataset and2 a shows 25 classes with best top 1 accuracy in Kinetics validation dataset.

Models	Kinetics	MIT
MFCC+ SENet-50	7.73	16.8
VGGish	7.83	17.12
Audio Ensemble	8.83	19.02

Table 3. Performance of acoustic models on Kinetics and MIT dataset.

2.3. Non-local Neural Networks

Non-local Neural Networks[12] extract long-term temporal information which have demonstrated the significance of non-local modeling for the tasks of video classification, object detection and so on.

Notation Image data and feature map data are generally three-dimensional: channel, height and width (in practice there is one more dimension: batch). They are represented as C-dimensional vectors with 2-dimensional index

$$\boldsymbol{X} = \{\boldsymbol{x}_i | i = (h, w) \in \mathbb{D}^2, \boldsymbol{x}_i \in \mathbb{R}^C\} \tag{1}$$

where $\mathbb{D}^2 = \{1, 2, \cdots, H\} \times \{1, 2, \cdots, W\}$. Video data get one more dimension time and are represented as C-dimensional vectors with 3-dimensional index

$$\boldsymbol{X} = \{\boldsymbol{x}_i | i = (t, h, w) \in \mathbb{D}^3, \boldsymbol{x}_i \in \mathbb{R}^C\} \tag{2}$$

where $\mathbb{D}^3 = \{1, 2, \cdots, T\} \times \mathbb{D}^2$.

Non-local Operation Non-local operation define a generic non-local operation in deep neural networks as:

$$\boldsymbol{y}_i = \sum_{j \in \mathbb{D}^3} f(\boldsymbol{x}_i, \boldsymbol{x}_j) g(\boldsymbol{x}_j). \tag{3}$$

Here function f representing the relation between position i and j. Many visions of function f such as $f(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left[\theta(\boldsymbol{x}_i)^T \varphi(\boldsymbol{x}_j)\right]$ are discussed, but performed almost the same.

As done in[1][3], a 2D $k \times k$ kernel can be inflated as a 3D $t \times k \times k$ kernel that spans t frames, in our experiments we used 32 frames. So this kernel can be initialized from 2D models(pretrained on Imagenet), each of the t planes in the $t \times k \times k$ kernel is initizlized by pretrained $k \times k$ weights, rescaled by 1/t. Each video we sample 64 consecutive frames from the original full-length video and then dropping every other frame. The non-local operation computes the response at a position as a weighted sum of the features at all positions with Embedding Gaussian. We used 5 non-local blocks added to i3d baseline. Table 4 shows our non-local results on Kinetics and Moments in Times dataset.

Models	Kinetics	MIT
Res50 baseline	78.63	30.83
Res50 non-local	80.80	32.96
Res101 baseline	79.58	31.33
Res101 non-local	81.96	33.69

Table 4. Performance of nonlocal NN on Kinetics and MIT dataset.

3. Relation-driven Models

We are interested in two questions. Firstly, *non-local* operations would be important for relation learning, but *global* operations may be unnecessary. If position i is far away from j, then $f(\boldsymbol{x}_i, \boldsymbol{x}_j) \approx 0$. Second quesion is that an unsupervised function may not be able to learn relations.

Mask Non-local To answer the first question, we compared the performance of non-local opeations and mask non-local opeations:

$$\boldsymbol{y}_i = \sum_{j \in \mathbb{D}^3} \mathbb{I}_{\mathbb{D}_i}(j) f(\boldsymbol{x}_i, \boldsymbol{x}_j) g(\boldsymbol{x}_j). \tag{4}$$

Here $\mathbb{D}_i$ is the $\delta-$ neighbourhood of $i = (t_i, h_i, w_i)$. Say:

$$\mathbb{D}_i = [t_i - \delta_t, t_i + \delta_t] \times [h_i - \delta_h, h_i + \delta_h] \times [w_i - \delta_w, w_i + \delta_w]. \tag{5}$$

$\mathbb{I}_{\mathbb{D}_i}(j)$ is the mask function. Say:

$$\mathbb{I}_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}. \tag{6}$$

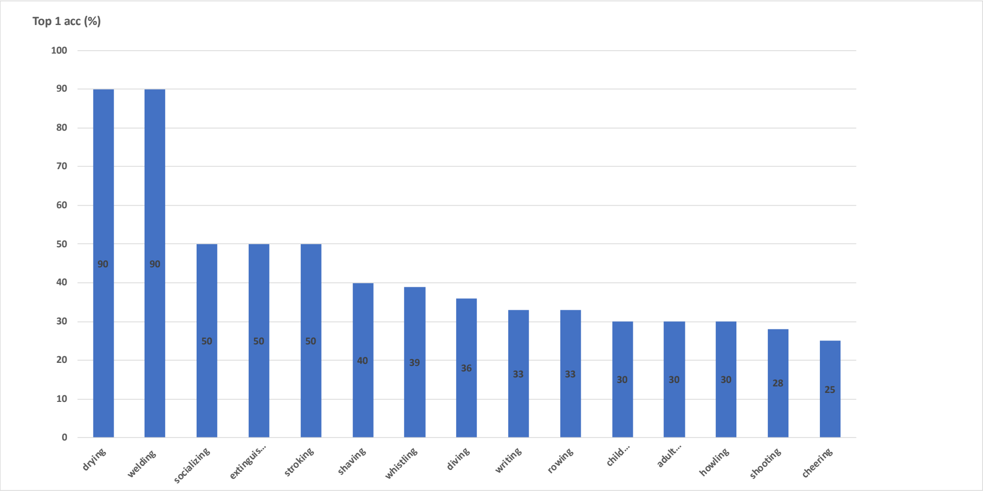

Figure 1. Acoustic Model: 15 classes with best top 1 accuracy in MIT validation dataset

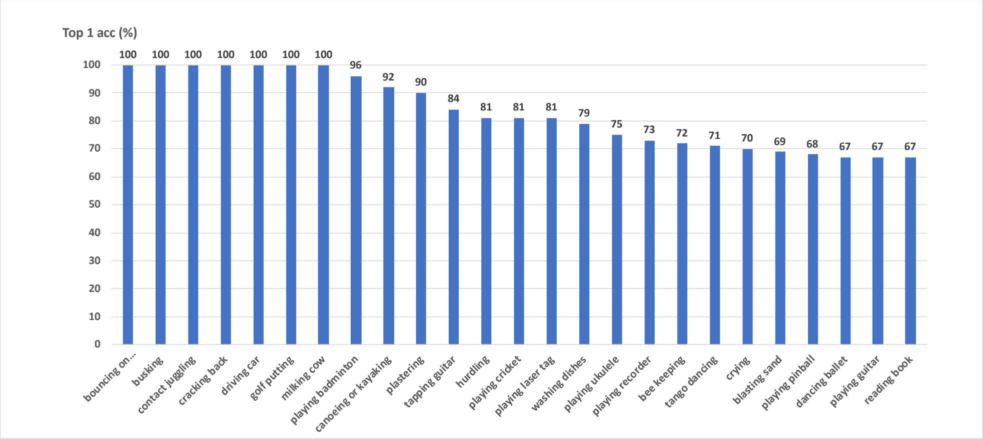

Figure 2. Acoustic Model: 25 classes with best top 1 accuracy in Kinetics validation dataset.

δ_t	δ_h	δ_w	top-1 acc
$+\infty$	$+\infty$	$+\infty$	80.80
$+\infty$	$\frac{3}{7}H$	$\frac{3}{7}W$	81.26
$+\infty$	$\frac{3}{28}H$	$\frac{3}{28}W$	80.63
$\frac{1}{2}T$	$\frac{3}{7}H$	$\frac{3}{7}W$	81.65
$\frac{1}{2}T$	$\frac{3}{28}H$	$\frac{3}{28}W$	80.93

Table 5. Perfomance for different settings of δ neighbourhood.

Table 5 shows mask nonlocal's performance on Kinetics. $+\infty$ means non-local operation in the dimention. Note that the first setting is the non-local baseline.

Learning Relations in Video Common convolution layers use invariant kernels for feature extraction at all positions in the feature map. It's limited for learning relations between different positions on the feature map. Nonlocal operations compute a feature-map-wise relation matrix to represent the kernel so that different positions get different but related feature extractions. The problem is that an unsupervised function may not be able for relations learning. We proposed a new model to learn the relation patten.

The network contains a network-in-network with a $(2t_0 + 1) \times (2h_0 + 1) \times (2w_0 + 1)$ size receptive field. The network-in-network computes a $(2t_1 + 1) \times (2h_1 + 1) \times (2w_1 + 1)$-dimensional relation vector $r^{(i)}$ for any position $i = (t, h, w)$ at the feature map ($t_1 < t_0 < \frac{1}{2}T$, $h_1 < h_0 < \frac{1}{2}H$ and $w_1 < w_0 < \frac{1}{2}W$).

The learnable relation vector $r^{(i)}$ represent the relation between position i and its neighbourhood

$$\mathbb{D}_i = [t_i \pm t_1] \times [h_i \pm h_1] \times [w_i \pm w_1],$$
$$\boldsymbol{y}_i = \sum_{j \in \mathbb{D}_i} r_j^{(i)} g(\boldsymbol{x}_j). \tag{7}$$

Here $r_j^{(i)}$ is the j^{th} element of $r^{(i)}$. Figure 3 shows our network structure. Note that, by using mask non-local's initialization, our network can get better results than what table 5 shows. But due to time limit and training from scratch, we haven't finished the experiments.

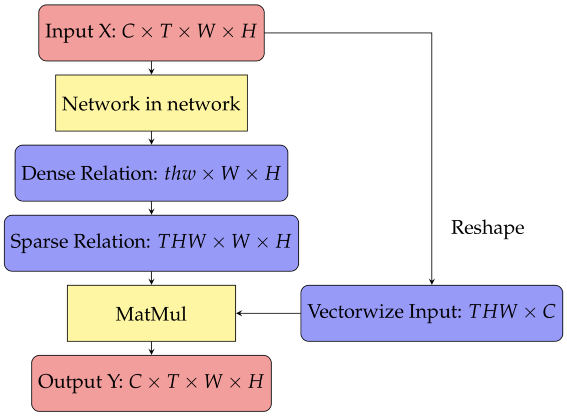
Input X: $C \times T \times W \times H$
Network in network
Dense Relation: $thw \times W \times H$
Sparse Relation: $THW \times W \times H$
Reshape
MatMul
Vectorwize Input: $THW \times C$
Output Y: $C \times T \times W \times H$

Figure 3. Our network structure for learning relations

Reference

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733. IEEE, 2017.

[2] Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks. In *Advances in Neural Information Processing Systems*, pages 4470–4478, 2017.

[3] Christoph Feichtenhofer, Axel Pinz, and Richard Wildes. Spatiotemporal residual networks for video action recognition. In *Advances in neural information processing systems*, pages 3468–3476, 2016.

[4] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 776–780. IEEE, 2017.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 131–135. IEEE, 2017.

[7] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*, 2017.

[8] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[9] Mathew Monfort, Bolei Zhou, Sarah Adel Bargal, Tom Yan, Alex Andonian, Kandan Ramakrishnan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding.

[10] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12, 2017.

[11] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *arXiv preprint arXiv:1705.02953*, 2017.

[12] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *CVPR*, 2018.