Alibaba-Venus at ActivityNet Challenge 2018 - Task C Trimmed Event Recognition (Moments in Time)

Chen Chen, Xueyong Wei, Xiaowei Zhao and Yang Liu
Alibaba Group, Hangzhou, China
{chenen.cc, xueyong.wxy, zhiquan.zxw, panjun.ly}@alibaba-inc.com

ABSTRACT

In this paper, we present a solution to Moments in Time (MIT) [1] Challenge. Current methods for trimmed video recognition often utilize inflated 3D (I3D) [2] to capture spatial-temporal features. First, we explore off-the-shelf structures like non-local [3], I3D, TRN [4] and their variants. After a plenty of experiments, we find that for MIT, a strong 2D convolution backbone following temporal relation network performs better than I3D network. We then add attention module based on TRN to learn a weight for each relation so that the model can capture the important moment better. We also design uniform sampling over videos and relation restriction policy to further enhance testing performance.

1 INTRODUCTION

Video understanding is a challenging task in computer vision and has significant attention during these years with more and more large-scale video datasets. Compared with image classification, video classification needs to model temporal information and more modalities can be extracted in videos like acoustic, motion, ASR etc. Multi-modalities are mutual complement to each other in many cases.

The recent challenge "Moments in Time Challenge" provides a platform to explore new approaches for short video understanding. The dataset has 339 categories which cover dynamic events unfolding within three seconds. The training/validation/test set has 802264/33900/67800 trimmed videos respectively. The evaluation metric is the average of top1 and top5 accuracy. The organizers provide raw videos and a preprocessed version which normalize videos to resolution 256x256 at 30fps. Participants are allowed to utilize any modality.

2 APPROACH

2.1 Modality Preparation

2.1.1 Visual image preprocessing.

We use preprocessed videos officially provided with resolution 256x256 and 30fps. We extract frames to jpeg format with best quality by using FFmpeg. After checking hundreds of videos, we found a lot of videos have vertical/horizontal black borders like movie style. We remove the black borders by some OpenCV tool and rescale it back to the resolution 256x256. We train/test models by using videos with and without black borders respectively.

2.1.2 Motion Features.

We use traditional TVL1 features which is implemented in OpenCV. It costs 2 weeks to extract motion features for all the MIT video data in a 2 gpu (M40) machine. Horizontal and vertical components are saved as gray image files and we concatenate them to an image with 2 channels during training.

2.1.3 Acoustic Features.

Audio contains a lot of information that helps to classify videos. We extract audio feature by a VGG like acoustic model trained on AudioSet [5] which consists of 632 audio event classes and over 2 million labeled 10-second sound clips. The process is the same as that in Youtube-8M, Google has released the extraction code in tensorflow model github.

2.2 Network Architecture

In this section, we describe all the networks involved.

2.2.1 NetVLAD aggregation with acoustic feature.

Acoustic feature pre-trained on AudioSet for each video has a dimension of 3x128. We use NetVLAD as that in [6] to aggregate acoustic features through time. It learns VLAD encoding followed by fully connect, mixture of experts and context gating.

2.2.2 Non local network.

We use off-the-shelf non local network, and train it with settings of both 32 and 64 sampled frames. The implementation of non-local network decodes video file during training, so we only do experiments on RGB modality.

2.2.3 Inflated 3d model.

I3D and its variant has achieved state of the art performance on datasets like kinetics. It's natural to apply it here in MIT dataset. We use two backbones. One is the origin Inception-V1 pre-trained on kinetics. The other backbone is Inception-V3 inflated ourselves. We inflate the convolution kernel of size 3x3, 5x5, 3x1, 1x3, 7x1, 1x7 into 3x3x3, 3x5x5, 3x3x1, 3x1x3, 3x7x1 and 3x1x7. We drop every other frames, the input video data dimension is 45x224x224. The spatial size is randomly cropped from a scaled video whose shorter side is randomly sampled in [240, 256]. We also randomly flip the whole video horizontally as an augmentation. We use 8 P100 cards to train this model, the batch size is 32. In testing, we use multi-crops (4 corners and center crop together with horizontal flipping) and average to get the final score.

2.2.4 Temporal Relation Network.

TRN achieves advanced performance on three video datasets, Something-Something, Jester, and Charades. These datasets all depends on temporal relational reasoning and MIT has similar character. We employ InceptionV3, InceptionResnetV2 and SENet-154 [7] as backbones for MultiScale TRN and build attention module based on squeeze & excitation module to learn the weighted relations leveraging the global relation distribution instead of simply accumulating them. In testing, we uniformly sample frames over whole video and utilize multi-crops. Also, we analyze the impact of different relations and select them explicitly. We find the following restriction will improve the performance

slightly. In 2-frames relation, the minimum relation sampling distance should be 2. In 3-frames relation, the distance should be in range [2, 3]. In 4-frames relation, the distance should be in range [2, 4].

We also try to combine I3D and TRN together. First, we split the video frames into 5 segments, each segment has 18 frames. Then, we apply 3D convolution model to each segment and will get a representation vector. Finally, TRN builds the relationship between the 5 segments. We apply this model to both RGB and Flow with Inception-V1 backbone, and we rescale the input resolution to 184x184 to reduce the complexity. The batch size is 64.

2.3 Ensemble

We use class-wise weighted ensemble. We calculate average precisions for each model and then normalize the weight for each class through models. After this operation, the ensemble model will take the different ability for each single model on each class into consideration. For example, when dealing with "clapping", acoustic model will have a predominant weight. In the final submission, we ensemble 13 models and the result is showed in next section**.**

3 EXPERIMENT

3.1 Experiment Results

We test on 3 modalities with different models. The input resolution is 224x224 except the case in I3D (184x184). We use multi-crop testing in all cases. Details are listed in Table 1.

We notice that in MIT dataset, 2D convolution following temporal relations works better than 3D convolution networks including I3D, non-local network and their variants. In temporal relation testing, uniform sampling policy over the whole videos works well. We use 8 segments here (90 frames) and average the score of 11 uniformly sampled clips. With the test enhancement, the baseline performance greatly improves from 28.61/54.65 to 29.67/55.74. The backbone is also of great importance, we compare InceptionV3, Inception Resnet V2, and SENet-154. SENet-154 is the best backbone in cost of high complexity and long training time. We spend 6 days to train SENet-154 TRN model. Actually, we also try Nasnet but fail to get a good result due to small batch size (only 8). Attentional TRN and restricting distance between consecutive sampled relations also help which means that more effective relations are selected. . The best single RGB model (32.21/59.05) is attentional temporal relation network with backbone senet154, and test using uniform sampling, multi-crop and manually restricted relation policy. Our acoustic model using AudioSet pre-trained features following netVLAD aggregation layer is better than baseline SoundNet metric. The final class-wise weighted ensemble consists of 13 models listed in the table which achieves top1/top5 (%) 36.23/64.56 on validation set. Since the ensemble weights depend on validation set, it makes more sense to check it on test set. We verify it on test server and find the weighted ensemble is better than average ensemble by a considerable margin about 0.2.

Table 1: *Experimental results on Validation Set (model with * is used in ensemble. Test enhancement means uniform sampling and multi-crop. ATRN is attentional temporal relation network)*

Models	Modality	Backbone	Top1/Top5
Non-local 32 frames *	RGB	Resnet50	27.04/54.02
Non-local 64 frames	RGB	Resnet50	26.44/53.11
I3D *	RGB	InceptionV3	27.62/53.89
I3D + TRN *	RGB	InceptionV1	28.25/54.83
I3D + TRN *	Flow	InceptionV1	18.00/39.17
TRN without test enhancement	RGB	InceptionV3	28.61/54.65
TRN without test enhancement, with relation restricted	RGB	InceptionV3	28.82/54.72
TRN *	RGB	InceptionV3	29.67/55.74
TRN black borders removed *	RGB	InceptionV3	29.59/55.86
TRN	Flow	InceptionV3	16.55/37.04
TRN *	RGB	InResnetV2	29.33/56.57
TRN (without test enhancement)	RGB	SENet-154	31.10/58.08
TRN	RGB	SENet-154	31.89/58.82
ATRN *	RGB	SENet-154	32.09/58.91
ATRN black borders removed *	RGB	SENet-154	31.97/59.26
ATRN relation restricted *	RGB	SENet-154	32.21/59.05
ATRN 512 dim bottleneck	RGB	SENet-154	31.63/58.92
ATRN vertical flipped input *	RGB	SENet-154	31.32/58.84
ATRN	Flow	SENet-154	17.92/39.60
netVLAD 64clusters *	Audio	VGG	9.00/19.51
netVLAD 128clusters *	Audio	VGG	8.90/20.23
13 models ensemble	None	None	36.23/64.56

4 CONCLUSIONS

In summary, we have tried off-the-shelf models for video recognition. To our surprise, temporal relation on top of 2d convolution works better than inflated 3d models in Moments in Time. It may be the case that MIT dataset is more complicated than traditional trimmed activity datasets like kinetics in terms of 1) Events are not limited to human related, more objects and scenes are involved, deep 2d convolution networks have stronger representation ability. 2) Big inner-class difference, for example, "fencing" has two totally different meanings which makes it harder to train 3d models. Furthermore, based on TRN, we add

attention module on relations, try stronger backbone and design effective uniform sampling test which greatly improves the performance.

REFERENCES

[1] Mathew Monfort, Bolei Zhou, Sarah Adel Bargal, Alex Andonian, Tom Yan, Kandan Ramakrishnan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, Aude Oliva. Moments in Time Dataset: one million videos for event understanding. 2018.

[2] Xiaolong Wang and Ross Girshick and Abhinav Gupta and Kaiming He. Non-local Neural Networks. In CVPR 2018.

[3] Joao Carreira and Andrew Zisserman. Quo Vaids, Action Recognition? A new model and the kinetics dataset. In CVPR 2017.

[4] B. Zhou, A. Andonian, and A. Torralba. Temporal Relational Reasoning in Videos. arXiv:1711.08496, 2017.

[5] Jort F. Gemmeke and Daniel P.W. Ellis and Dylan Freedman and Aren Jansen and Wade Lawrence and R. Channing Moore and Manoj Plakal and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In ICASSP 2017.

[6] Antoine Miech, Ivan Laptev and Josef Sivic. Learnable pooling with context gating for video classification. arXiv preprint arXiv:1706:06905, 2017.

[7] Jie Hu, Li Shen, Gang Sun. Squeeze and Excitation Networks. arXiv preprint arXiv:1709.01507, 2017.