Alibaba-AIC: Submission to Multi-Moments in Time Challenge 2019 *

Chao Li†, Shaobo Lin‡, Biao Wang†, Lei Zhang†, Xiansheng Hua†
† Artificial Intelligence Center, DAMO Academy, Alibaba Group, Hangzhou, China
‡ Sun Yat-sen University, Guangzhou, China
{lllcho.lc, wb.wangbiao, lei.zhang.lz, xiansheng.hxs}@alibaba-inc.com
linshb9@mail2.sysu.edu.cn

Abstract

This technical report describes the detail solution for Multi-Moments in Time Challenge 2019. We exploit and evaluate spatiotemporal feature learning model, such as C3D, (2+1)D convolution. And to learn the action class association in data, we use focal loss combined with label correlation loss as the loss function. Furthermore, we also explore the auditory modality to predict action categories. Our final submission to the challenge is an ensemble of the visual RGB model, Flow model and Audio model, achieving a mAP 67.0% on validation set.

1 Methodology

1.1 Objective Function

Different from the typical multi-class problem, multi-label classification usually requires additional efforts in learning the associated data/label information. To address this issue, we introduce a label correlation aware loss function proposed in [1]:

$$\Gamma(p, y) = \sum_{i=1}^{N} E_i \tag{1}$$

$$E_i = \frac{1}{|y_i^1||y_i^0|} \sum_{(s,t)\in y_i^1 \times y_i^0} exp(p_s - p_t) \tag{2}$$

where y_i^1 denotes the set of the positive labels in y_i for the instance x_i , and y_i^0 is that of the negative labels, p_s denotes the sth predicted score.

Considering the action categories in the dataset belong to the long tail distribution, we also introduce balanced Binary Cross Entropy loss, i.e. Focal loss[2].

$$F(p) = -(1-p)^{\gamma} log(p) \tag{3}$$

Thus, the objective function of our model can be formulate as follows:

$$L = \Gamma + \alpha F \tag{4}$$

*This work was done when Lin was visiting Alibaba as an intern

1.2 Visual Modality

In recent years, many convolutional structures are proposed to learn video feature, such as C3D [3], P3D or (2+1)D [4; 5], non-local networks [6]. For the trade-off of performance and computational cost, we use P3D, or (2+1)D, to learn visual feature from video clip. ResNet-101, Inception-v4 and Inception-ResNet-v2 and polyNet are used as the backbone models, which are pretrained on ImageNet. And to extract more motion feature, we also computed optical flow with a TV-L1 algorithm [7].

For each model, to obtain better generalization on test dataset, the Stochastic Weight Averaging (SWA) scheme [8] is adopted to ensemble the models, which are trained with cycle learning rate.

1.3 Auditory Modality

In this work, audio streams extracted from videos are also exploited for the task of action recognition. M34-res [9] and EnvNetv2 [10] learn semantic feature from the 1D raw audio waveforms in an end-to-end way. Moreover, log-mel spectrum is a powerful hand-tuned feature. In our work, we adopt the ResNet34 and VGG16 as the backbone models for action classification based on 2D log-mel feature. We train and ensemble the four state-of-the-art models on the Multi-Moments in Time dataset to get the results.

2 Experiments

The Multi-Moments in Time dataset [11] contains 1025862 training videos, 10000 validation videos, Excluding the videos without audio track, the auditory modality contains 550k training segments and 5487 validation segments. Totally 313 action categories are annotated.

For the visual RGB model, we randomly generate training samples from videos in training data. We random select 64 continuous frames from the video and then sample 8 frames by dropping 7 frames in every 8 frames. The spatial size is 224x224 pixels, randomly cropped from a scaled video whose shorter side is 288 pixels. In inference stage, we sample 6 clips evenly from a full-length video and compute the scores on them individually. The final prediction is the averaged scores of all clips. We deal with Flow model exactly the same as RGB model except that the input channel is set to 2.

For the training of audio models, all the sound data are downsampled to a frequency of 16kHz. We set two variables,

Table 1: performance on the validation set of the Multi-Moments in Time dataset.

Model	Modality	mAP(%)
ResNet-101	RGB	65.0
Inception-V4	RGB	63.7
Inception-ResNet-V2	RGB	63.4
PolyNet	RGB	63.2
Ensemble	RGB	66.2
ResNet-101	Flow	59.1
Inception-V4	Flow	54.5
Inception-ResNet-V2	Flow	55.7
Ensemble	Flow	59.2
M34-res	Audio	31.8
EnvNetv2	Audio	30.2
ResNet-34	Audio	30.2
VGG16	Audio	30.1
Ensemble	Audio	35.5
Ensemble	RGB+Flow+Audio	67.0

the audio length and the start time, which are both randomly selected in a certain range to obtain the training data. The range of the audio length is 1.5 to 2.5 seconds for training. By using these two variables, we can enhance the data. In the validation or testing stage, we fix the audio length to 2 seconds.

3 Conclusions

In the Multi-Moments in Time challenge 2019, we explore multiple modalities for the task of multi label video-based action recognition. By ensembling of the models based on visual RGB, Flow and Audio, we can fully extract the feature of the video in more patterns, which achieves superior performance over the single model . Extracting better general presentation based on the fundamental difference between Multi-Moments in Time dataset with other datasets is left for future work.

References

[1] Chih-Kuan Yeh, Wei-Chieh Wu, Wei-Jen Ko, and Yu-Chiang Frank Wang, "Learning deep latent space for multi-label classification," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[2] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar, "Focal loss for dense object detection," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[3] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," pp. 4489–4497, 2014.

[4] Zhaofan Qiu, Ting Yao, and Tao Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *The IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5534–5542.

[5] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann Lecun, and Manohar Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *CVPR*, 2018.

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

[7] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[8] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson, "Averaging weights leads to wider optima and better generalization," 2018.

[9] Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit Das, "Very deep convolutional neural networks for raw waveforms," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 421–425.

[10] Yuji Tokozume and Tatsuya Harada, "Learning environmental sounds with end-to-end convolutional neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 2721–2725.

[11] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al., "Moments in time dataset: one million videos for event understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–8, 2019.