Team SPEEDY Multi Moments in Time Challenge 2019 Technical Report

Yang Liu* Samuel Albanie* Qingchao Chen* Andrew Zisserman
University of Oxford, UK
{yangl, albanie, az}@robots.ox.ac.uk, qingchao.chen@eng.ox.ac.uk

Abstract

In this report, we present a solution to the Multi Moments in Time Challenge 2019. At present, the dominant research paradigm in video classification pursues improvement through end-to-end learning of effective video representations. This is undoubtedly a useful research direction, but it is not the only route towards better video understanding. In this report, we focus on an alternative 'speedy-expert' approach: features are first pre-extracted from a wide range of pretrained models (the experts) and cached as an intermediate representation that can then be used to train the final action recognition system. Our best single model achieved a performance of 0.628 mean average precision on the Multi Moment validation set.

1. Introduction

The focus of this report is video classification. In contrast to images, video contains temporal information and valuable cues from multiple modalities [10], such as an accompanying audio track. This has several consequences: video signals are richer than their static counterparts, but also more challenging to represent. Recently, the release of large-scale video classification datasets such as Kinetics [9] and Multi Moments in Time [13] has enabled researchers to train and explore different model designs for automatic video understanding and analysis.

The dominant research paradigm in this domain uses end-to-end training models to learn robust video representations for classification. This is undoubtedly a useful research direction, but we note that it is by no means the sole path towards improving video understanding. In this challenge, we focused on an attractive lightweight alternative: using collections of existing pretrained models as "speedy experts", offering representations which have been specialised for semantically relevant machine perception tasks. The primary motivation for our approach is that we can reasonably approximate the discriminative content of a video with a set of semantic projections of the data provided by different experts (in scene, audio, etc). Effectively, this approximation enables us to exploit knowledge (and datasets) from existing individual sources and significantly reduce the computational burden of training a new video classification system.

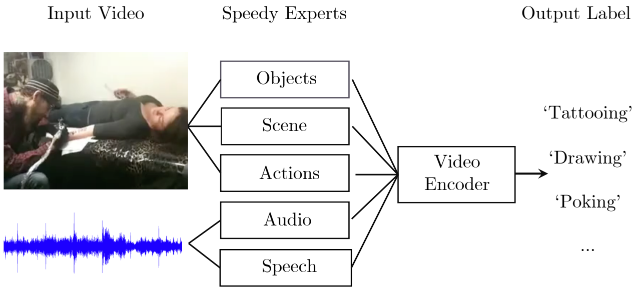

Figure 1. The overall network architecture.

2. Speedy Experts

In this section, we briefly introduce our method for integrating multi-modal features extracted from various expert models. We extracted multi-modal features $\{X_i\}_{i \in \mathcal{M}}$ from the video clips with models pretrained on various datasets including static images, videos and audios (see Sec. 3 for details). X_i denotes the features from the i^{th} modality and $\mathcal{M}$ indicates the set of all modalities. To integrate these features, we first transform them into the same dimension, calculate the context vector $\frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} X_i$ using the average pooling operation and finally select the useful information $\bar{X}_i$ within X_i based on the context vector. After this feature selection operation for each modality, we concatenate the new feature set $\{\bar{X}_i\}_{i \in \mathcal{M}}$ and train a classifier by minimizing a multi-label binary cross-entropy loss. The proposed Speedy experts framework is shown in Figure 1.

3. Experiments

In our experiments, we use the following speedy experts in the form of pretrained semantic embeddings:

Appearance frame-level embeddings of the visual data are generated with a SENet-154 model [7] (pretrained on ImageNet for the task of image classification) from

[*] contributed equally and courageously.

frames extracted at 5fps, where each frame is resized to 224 × 224 pixels. Features are collected from the final global average pooling layer, and have a dimensionality of 2048. We also extracted additional features with a DenseNet-161 [8], Inception-ResNetv2 [15] and an Instagram-pretrained ResNext-101 [11].

Scene embeddings of 2208 dimensions are extracted from 224×224 pixel centre crops of frames extracted at 1fps using DenseNet-161 [8] and ResNet50 [6] models pretrained on Places365 [17].

Motion embeddings are generated using the I3D inception model following the procedure described by [1]. Frames extracted at 25fps and processed with a window length of 64 frames and a stride of 25 frames. Each frame is first resized to a height of 256 pixels (preserving aspect ratio), before a 224 × 224 centre crop is passed to the model. Each temporal window produces a (1024x7)-matrix of features. We further included R(2+1)D embeddings pretrained on Instagram-videos [4], as well as features extracted from the pretrained ResNet-3d50 and TRN [16] models provided by the Moments-in-Time dataset organisers [14].

Audio embeddings are obtained with a VGGish model, trained for audio classification on the YouTube-8m dataset [3]. To produce the input for this model, the audio stream of each video is re-sampled to a 16kHz mono signal, converted to an STFT with a window size of 25ms and a hop of 10ms with a Hann window, then mapped to a 64 bin log mel-spectrogram. Finally, the features are parsed into non-overlapping 0.96s collections of frames (each collection comprises 96 frames, each of 10ms duration), which is mapped to a 128-dimensional feature vector.

Speech to Text The audio stream of each video is re-sampled to a 16kHz mono signal. We then obtained transcripts of the spoken speech for using a Deep speech model [5] pretrained on The Wall Street Journal, Switchboard and Fisher [2] corpora and Baidu Speech dataset[5]. We encode each word using the Google News[1] trained word2vec word embeddings [12]. Finally, all the word embeddings in each sentence are aggregated using NetVLAD.

We evaluated our model on the validation set of the Multi Moment in time dataset. Multiple models were trained using features from various combinations of modalities using the expert aggregation mechanism described in Sec. 2. Our single strongest model—which combined object, scene, motion and audio features—achieved a mean average precision of 0.628 on the validation set. Lastly, we combined several models by ensembling their logits on the validation set—the 5-fold cross validation mean average precision of this ensemble on the validation set was 0.658.

4. Conclusion

In our submission to the Multi Moments in Time Challenge 2019, we explore multiple modalities for the task of video classification. At the core of our method is a technique we term speedy expert, a learnable mechanism for exploiting contextual information from a collection of modality signals to render them maximally discriminative for the video classification task. Our approach achieves competitive results with significantly fewer parameters and computing resources than required in end-to-end training. A more thorough evaluation of the architecture is left for future work.

5. Acknowledgements

We would like to sincerely thank Ashish Thandavan for his considerable help and support with the computing resources used for our submission. We would also like to thank Arsha Nagrani for inspiring our early work on video understanding. Lastly, we would also like to thank Chuhan Zhang and Tengda Han for their comprehensive nutritional support during the challenge.

References

[1] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 2

[2] C. Cieri, D. Miller, and K. Walker. The fisher corpus: a resource for the next generations of speech-to-text. In *LREC*, volume 4, pages 69–71, 2004. 2

[3] S. H. et al. Cnn architectures for large-scale audio classification. In *ICASSP*. 2017. 2

[4] D. Ghadiyaram, D. Tran, and D. Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12046–12055, 2019. 2

[5] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014. 2

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[7] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. Squeeze-and-excitation networks. *TPAMI*, 2019. 1

[8] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[9] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[10] Y. Liu, S. Albanie, A. Nagrani, and A. Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *British Machine Vision Conference*. 1

[1] GoogleNews-vectors-negative300.bin.gz can be found at: https://code.google.com/archive/p/word2vec/

[11] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 181–196, 2018. 2

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 2

[13] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfruend, C. Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–8, 2019. 1

[14] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, Y. Yan, L. Brown, Q. Fan, D. Gutfreund, C. Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2

[15] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 2

[16] B. Zhou, A. Andonian, A. Oliva, and A. Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018. 2

[17] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *TPAMI*, 2017. 2