Towards Good Practices for Multi-modal Fusion in Large-scale Video Classification

Jinlai Liu, Zehuan Yuan, and Changhu Wang

Bytedance AI Lab
{liujinlai.licio,yuanzehuan,wangchanghu}@bytedance.com

Abstract. Leveraging both visual frames and audio has been experimentally proven effective to improve large-scale video classification. Previous research on video classification mainly focuses on the analysis of visual content among extracted video frames and their temporal feature aggregation. In contrast, multimodal data fusion is achieved by simple operators like average and concatenation. Inspired by the success of bilinear pooling in the visual and language fusion, we introduce multi-modal factorized bilinear pooling (MFB) to fuse visual and audio representations. We combine MFB with different video-level features and explore its effectiveness in video classification. Experimental results on the challenging Youtube-8M v2 dataset demonstrate that MFB significantly outperforms simple fusion methods in large-scale video classification.

Keywords: Video classification; Multi-modal Learning; Bilinear Model

1 Introduction

Along with the dramatic increase in video applications and production, better video understanding techniques are urgently needed. As one of the fundamental video understanding tasks, multi-label video classification has attracted increasing attentions in both computer vision and machine learning communities. Video classification requires a system to recognize all involved objects, actions and even events in any video based on its available multimodal data such as visual frames and audio.

As deep learning has obtained a remarkable success in image classification [1–3], action recognition [4–6] and speech recognition [7, 8], video classification also benefit a lot from these powerful image, snippet, and audio representations. Since videos are composed of continuous frames, aggregating frame or snippet features into video-level representation also plays an important role during recognition process. Besides the direct aggregations such as temporal average or maxpooling, a few sophisticated temporal aggregation techniques are also proposed. For example, Abu-El-Haija *et al.* [9] proposes deep Bag-of-Frames pooling (DBoF) to sparsely aggregate frame features by ReLU activation. On the other hand, recurrent neural networks such as long short-term memeory (LSTM) [10] and gated recurrent unit (GRU) [11]) are applied to model temporal dynamics along frames.

Although much progress has been made in generating video-level visual representations, few work lies on integrating multimodal data which can supplement with each other and further reduce the ambiguity of visual information. Therefore, developing deep and fine-grained multimodal fusion techniques could be a key ingredient towards practical video classification systems. In this paper, we take the first step by introducing multi-modal bilinear factorized pooling into video classification, which has been extensively adopted in visual question answering [12–14]. We select three popular video-level representations, i.e, Average pooling, NetVLAD [15] and DBoF [9], to validate its effectiveness. Experimental results indicate that video classification can achieve a significant performance boost by leveraging the new pooling mechanism over video and audio features. In summary, our contributions are twofold:

- We first introduce multi-modal factorized bilinear pooling to integrate visual information and audio in large-scale video classification.
- We experimentally demonstrate that multi-modal factorized pooling significantly outperforms simple fusion methods over several video-level features.

2 Related Work

2.1 Video Classification

Large-scale datasets [16, 17] play a crucial role for deep neural network learning. In terms of video classification, Google recently releases the updated Youtube-8M dataset [9] with 8 millions videos totally. For each video, only visual and audio representations of multiple frames are made public. The approaches for video classification roughly follow two main branches. On the one hand, several architectures are introduced to extract powerful frame or snippet representations similar to image classification. Simonyan and Zisserman *et al.* first introduces deep convolutional neural networks to video action classification by performing frame-level classification [4]. In order to include more temporal information, 3D convolutional neural network and several variants [18–20] are proposed to generate representations of short snippets. The final video predictions can be estimated by late fusion or early fusion. On the other hand, researchers also direct their eyes to how to model long-term temporal dynamics when frame-level or snippet-level representation available.

Commonly used methods to model long-term temporal dynamics are various variants of Bag of Visual Words (BoVW) including Vector of Locally Aggregated Descriptors (VLAD) [21], Fisher Vector (FV) [22] and so on. But these handcrafted descriptors cannot be finetuned for the target task. Therefore, an end-to-end trainable NetVLAD [15] was proposed where a novel VLAD layer was plugged into a backbone convolutional neural network. Girdhar *et al.* proposed ActionVLAD that performs spatio-temporal learnable aggregation for video action classification. On the other hand, temporal models such as LSTM and GRU, are also widely used to aggregate frame-level features into a single representation due to its capability of capturing the temporal structures of videos.

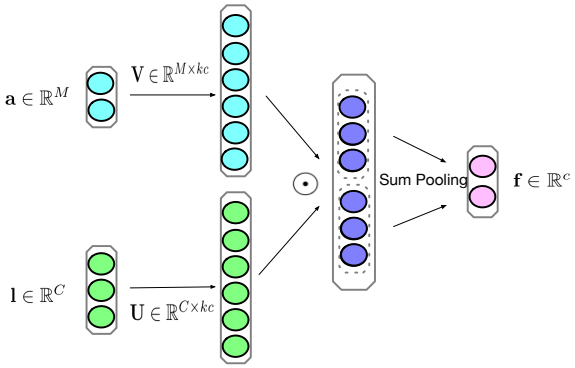

Fig. 1. The architecture of multi-modal factorized bilinear pooling.

2.2 Multimodal Learning

A simple attempt to integrate multimodal data is performing average or concatenation before input to final predictions. However, more fine-grained multimodal fusion models like bilinear pooling operations have been extensively explored and validated in visual and language multimodal learning. Lots of work has focused on addressing the huge number of model parameters and high computation cost in bilinear pooling. Multi-modal compact bilinear (MCB) [12] was proposed to employ tensor sketch algorithm to reduce the computation cost and the amount of parameters. Later, to overcome the high memory usage in MCB, multi-modal low-rank bilinear pooling (MLB) [13] adopted Hadamard product to combine cross-modal feature vectors. Furthermore, multimodal tucker fusion (Mutan) [23] and multi-modal bilinear factorized bilinear pooling (MFB) [14] were proposed to address rather huge dimensionality and boost training.

In the paper, inspired by the success of MFB in visual and language fusion, we apply MFB [14] into the video classification task by combining available visual and audio representations. The most related work to us is probably [24] which tried multi-modal compact bilinear pooling approach [12] in large-video video classification but failed to fit training data.

3 Multi-modal Bilinear Fusion

We apply multi-modal factorized bilinear pooling over video-level visual and audio features. However in practice, only frame-level or snippet-level representations are available. Therefore as mentioned above, three methods are exploited to aggregate frame-level features into a single video feature. In this section, we firstly review the MFB module and temporal aggregation models and then present our classification framework.

3.1 Multi-modal Factorized Bilinear Pooling

For any video, let $\mathbf{l} \in \mathbb{R}^C$ and $\mathbf{a} \in \mathbb{R}^M$ denote its visual feature and audio feature, respectively. M and C are their corresponding feature dimensions. Then the output of MFB over $\mathbf{l}$ and $\mathbf{a}$ is a new vector $\mathbf{f}$ with the i-th element formulated as

$$f_i = \mathbf{l}^T \mathbf{W}_i \mathbf{a}, \tag{1}$$

where $\mathbf{W}_i \in \mathbb{R}^{C \times M}$. In order to reduce the number of parameters and the rank of weight matrix $\mathbf{W}_i$, a novel low-rank bilinear model is proposed in [25]. Specifically, $\mathbf{W}_i$ is decomposed as the multiplication of two low-rank matrices $\mathbf{U}_i$ and $\mathbf{V}_i$, where $\mathbf{U}_i \in \mathbb{R}^{C \times k}$ and $\mathbf{V}_i \in \mathbb{R}^{M \times k}$. k is a predefined constant to control rank. Therefore,

$$\mathbf{f}_i = \mathbf{l}^T \mathbf{W}_i \mathbf{a} = \mathbf{l}^T \mathbf{U}_i \mathbf{V}_i^T \mathbf{a} = \mathbb{1}^T (\mathbf{U}_i^T \mathbf{l}) \odot (\mathbf{V}_i^T \mathbf{a}), \tag{2}$$

where $\mathbb{1} \in \mathbb{R}^k$ is an all-one vector and $\odot$ denotes Hadamard product. It is worthy noting that $\mathbb{1}$ is essentially a sum pooling operator as shown in Fig. 1. We follow the same normalization mechanism as in [25] except that the power normalization layer is replaced with a ReLU layer in our MFB module implementation.

3.2 Temporal Aggregation Model

In order to validate the general effectiveness of MFB in video classification, we experiment with video-level visual and audio features of three kinds obtained by average pooling, DBoF and NetVLAD over respective frame-level features. Let $\mathbf{L} \in \mathbb{R}^{N \times C}$ and $\mathbf{A} \in \mathbb{R}^{N \times M}$ denote frame-level visual and audio features for a given video with N frames, respectively. In our experiment, $C = 1024$ and $M = 128$. $\mathbf{L}$ and $\mathbf{A}$ are processed separately for each pooling mechanism.

Average Pooling (Avgpooling): The average pooling layer is simply averaging features across N frames, that is,

$$\mathbf{l} = \frac{1}{N} \sum_{i=1}^{n} \mathbf{L}_i, \mathbf{a} = \frac{1}{N} \sum_{i=1}^{n} \mathbf{A}_i. \tag{3}$$

DBoF: Deep Bag-of-Frames pooling extends the popular bag-of-words representations for video classification [26, 27] and is firstly proposed in [9]. Specifically, the feature of each frame is first fed into a fully connected layer(fc) to increase dimension, Max pooling is then used to aggregate these high-dimensional frame-level features into a fixed-length representation. Following [9], a rectified linear unit (RELU) and batch normalization layer (BN) is used to increase non-linearity and keep training stable.

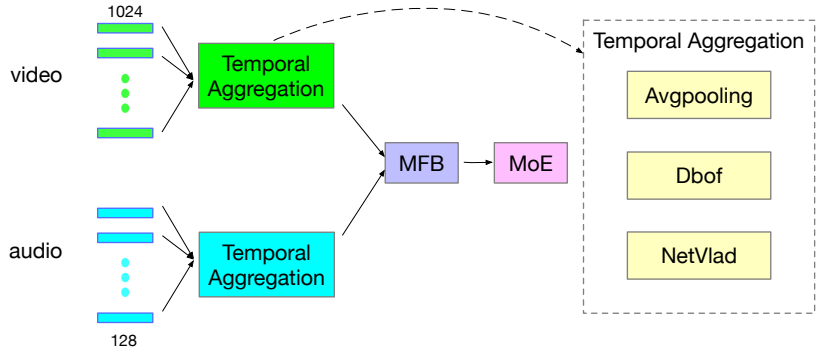

Fig. 2. The overall architecture of our MFB augmented video classification system.

NetVLAD: The NetVLAD [15, 6] employed the VLAD encoding [21] in deep convolutional neural networks. The whole architecture can be trained in an end-to-end way. Compared to VLAD encoding, the parameters of clusters are learned by backpropagation instead of k-means clustering. Assuming K clusters are used during training, NetVLAD assigns any descriptor $\mathbf{h}_i$ in $\mathbf{L}$ or $\mathbf{A}$ to the cluster k by a soft assignment weight

$$\alpha_k(\mathbf{h}_i) = \frac{\mathbf{w}_k^T \mathbf{h}_i + b_k}{\sum_{k'=1}^{K} e^{\mathbf{w}_{k'}^T \mathbf{h}_i + b_{k'}}}, \tag{4}$$

where $\mathbf{w}_{k'}$ and $\mathbf{b}_{k'}$ are trainable cluster weights. Compared to the hard assignment, $\alpha_k(\cdot)$ measures the distance between descriptors with the cluster k and thus maintains more information. With all assignments for descriptors, the final NetVLAD representation is a weighted sum of residuals relative to each cluster. For the cluster k:

$$VLAD[k] = \sum_{i=1}^{N} \alpha_k(\mathbf{h}_i)(\mathbf{h}_i - \mathbf{c}_k), \tag{5}$$

where $\mathbf{c}_k$ corresponds to the learnable center of the k-th cluster.

3.3 Video-level Multi-modal Fusion

In this section we will illustate that MFB module can be a plug-and-play layer to fuse aggregated visual and audio features. Fig. 2 shows the overall video-level fusion architecture. It mainly contains three parts. Firstly, the pre-extracted visual features $\mathbf{L}$ and audio features $\mathbf{A}$ are fed into two temporal aggregation modules separately. Each module outputs a single compact video-level representation and can be any one of the mentioned three aggregating mechanisms shown in the right side of figure. Next, MFB module fuse aggregated visual and audio features into a fixed-length representation. Finally, the classification module takes the resulting compact representation as input and outputs confidence

scores of each semantic label. Following [9], we adopt Mixture-of-Experts [28] as our classifier. The Mixture of Experts [28] classifier layer consists of m "expert networks" which take the global multimodal representation f as input and estimate a distribution over c classes. The final prediction $\mathbf{d}$ is defined as

$$\mathbf{d} = \sum_{i=1}^{m} \text{softmax}(\mathbf{g}_i) \odot sigmoid(\mathbf{e}_i), \tag{6}$$

$$\mathbf{g}_i = \mathbf{f}\mathbf{W}_{g,i} + \lambda||\mathbf{W}_{g,i}||_2, \tag{7}$$

$$\mathbf{e}_i = \mathbf{f}\mathbf{W}_{e,i} + \lambda||\mathbf{W}_{e,i}||_2, \tag{8}$$

where $\mathbf{W}_{g,i}, \mathbf{W}_{e,i}, i \in \{1, ..., m\}$ are trainable paramters and $O \in \mathbb{R}^c$. λ is the L2 penalty with the default value 1e-6. All our models are trained with 2-mixtures MoE.

4 Experiments

4.1 Implementation details

We implement our model based on Google starter code[1]. Each training is performed on a single V100 (16Gb) GPU. All our models are trained using Adam optimizer [29] with an initial learning rate set to 0.0002. The mini-batch size is set to 128. We found that cross entropy classification loss works well for maximizing the Global Average Precision (GAP). All model are trained with 250 000 steps. In order to observe timely model prediction, we evaluate the model on a subset of validate set every 10 000 training steps. For the cluster-based pooling, the cluster size K is set to 8 for NetVLAD and 2000 for DBoF. To have a fair comparison, 300 frames are sampled before aggregation. In addition, the dropout rate of MFB module is set to 0.1 in all our expriements.

4.2 Datasets and evaluation metrics

We conduct experiments on the recently updated Youtube-8M v2 dataset with improved machine-generated labels and higher-quality videos. It contains a total of 6.1 million videos, 3862 classes, 3 labels per video averagely. Visual and audio features are pre-extracted per frame. Visual features are obtained by Google Inception convolutional neural network pretrained on ImageNet [16], followed by PCA-compression to generate a vector with 1024 dimensions. The audio features are extracted from a VGG-inspired acoustic model described in [8]. In the official split, training, validataion and test have equal 3844 tfrecord shards. In practice, we use 3844 training shards and 3000 validation shards for training. We randomly select 200 shards from the rest of 844 validation shards (around 243 337 videos) to evaluate our model every 10 000 training steps. Results are evaluated using the Global Average Precision (GAP) metric at top 20 as used in the Youtube-8M Kaggle competition.

[1] https://github.com/google/youtube-8m

Table 1. Comparision study on Avgpooling feature

Model	GAP
Avgpooling + Audio Only	38.1
Avgpooling + Video Only	69.6
Avgpooling + Concatenation	74.2
Avgpooling + FC + Concatenation	81.8
Avgpooling + MFB	**83.3**

Table 2. Comparision study on NetVLAD feature

Model	GAP
NetVLAD + Audio Only	50.7
NetVLAD + Video Only	82.3
NetVLAD + Concatenation	85.0
NetVLAD + FC + Concatenation	84.6
NetVLAD + MFB	**85.5**

4.3 Results

In this section, we verify the effectiveness of MFB module by comparing its performance on the validation set with the simple concatenation fusion. We also conduct two comparative tests with single-modality input (only video or audio). To prove that the improvement of performance does not come from increasing parameters, we add another comparison with the same number of parameters as MFB. Specifically, the temporal aggregated video and audio representations are first projected using a fully connected layer respectively and then the projected video and audio vectors are concatenated to feed into the MoE classifier (For convenience, we call it FC Concatenation module later). The fully connected layers have the same parameter settings with those in MFB module. The superior GAP performance of MFB module on three temporal aggregation models is shown as follows.

Table 3. Comparision study on DBoF feature

Model	GAP
DBoF + Audio Only	48.9
DBoF + Video Only	81.8
DBoF + Concatenation	84.0
DBoF + FC + Concatenation	84.1
DBoF + MFB	**85.9**

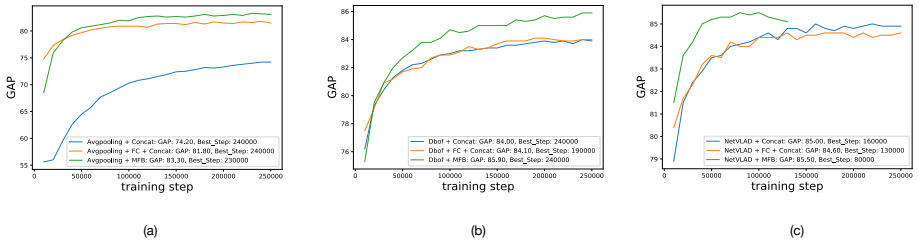

Fig. 3. (a) The GAP Performance of Avgpooling feature with different fusion modules (Concantenation, FC Concantenation, MFB). (b) The GAP Performance of DBoF feature. (c) The GAP Performance of NetVLAD feature.

The detailed results of MFB with Avgpooling features are shown in Tab 1. Firstly, the GAP performance of two modal fusion is far superior to single modality input (Video Only or Audio Only). In the NetVLAD and DBof video features, we can draw the same conclusion. Secondly, we can observe a significant increase in performance with the MFB module, which achieves a 9.1% higher GAP compared with the concatenation fusion baseline. Even if the concatenation is augmented with the same number of parameters as MFB, there is still a 1.5% gap. The main reason is probably that the simple fusion can not leverage high-order information across modalities.

In terms of NetVLAD video features, the MFB module improves the GAP from 85.0% to 85.5% compared to the concatenation module as shown in Tab 2. However surprisingly, adding fully connected layers performs worse, indicating that NetVLAD has been a quite good temporal model for single modal data aggregation. In some sense, increasing parameters will lead to overfitting . Therefore, it also proves that MFB contributes to the performance boost. For DBoF, the results are consistent with Avgpooling and NetVLAD, MFB module achieves the best GAP of 85.9%, around 1.8% higher than another two methods. We conclude that MFB encourages abundant cross-modal interactions and thus reduce the ambiguity of each modal data.

In order to give an intuitive observation on the advantage of MFB over simple fusion baselines, we illustrate the training processes of three fusion modules in Fig. 3, which shows the GAP performance on validation dataset as the training iteration increases. It is worthy noting that the experiment with the MFB module and NetVLAD features is early stopped at around 13 000 steps due to overfitting. Obviously, MFB module can not only increase the capability of video and audio fusion but also speed-up training.

5 Conclusions

In this paper, we first apply the multimodal factorized blinear pooling into large-scale video classification task. To validate its effectiveness and robustness,

we experiment on three kinds of video-level features obtained by Avgpooling, NetVLAD and DBoF. We conduct experiments on large-scale video classification benchmark Youtube-8M. Experimental results demonstrate that the carefully designed multimodal factorized bilinear pooling can achieve significantly better results than the popular fusion concatenation operator. Our future work mainly lies on directly combining multimodal factorized bilinear pooling with multimodal frame-level data.

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: International Conference on Neural Information Processing Systems. (2012) 1097–1105
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. (2016) 770–778
3. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: IEEE Conference on Computer Vision and Pattern Recognition. (2016) 2818–2826
4. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems. (2014) 568–576
5. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European Conference on Computer Vision, Springer (2016) 20–36
6. Girdhar, R., Ramanan, D., Gupta, A., Sivic, J., Russell, B.: Actionvlad: Learning spatio-temporal aggregation for action classification. (2017) 3165–3174
7. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine **29**(6) (2012) 82–97
8. Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R., Wilson, K.: Cnn architectures for large-scale audio classification. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). (2017)
9. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. (2016)
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8) (1997) 1735–1780
11. Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259 (2014)
12. Gao, Y., Beijbom, O., Zhang, N., Darrell, T.: Compact bilinear pooling. (2015)
13. Kim, J.H., On, K.W., Lim, W., Kim, J., Ha, J.W., Zhang, B.T.: Hadamard product for low-rank bilinear pooling. arXiv preprint arXiv:1610.04325 (2016)
14. Yu, Z., Yu, J., Fan, J., Tao, D.: Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. IEEE International Conference on Computer Vision (ICCV) (2017) 1839–1848

15. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 5297–5307
16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, Ieee (2009) 248–255
17. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 961–970
18. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 6450–6459
19. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3d residual networks. In: ICCV. (2017)
20. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, IEEE (2017) 4724–4733
21. Jgou, H., Douze, M., Schmid, C., Prez, P.: Aggregating local descriptors into a compact image representation. In: Computer Vision and Pattern Recognition. (2010) 3304–3311
22. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: IEEE Conference on Computer Vision and Pattern Recognition. (2007) 1–8
23. Ben-Younes, H., Cadene, R., Cord, M., Thome, N.: Mutan: Multimodal tucker fusion for visual question answering. In: Proc. IEEE Int. Conf. Comp. Vis. Volume 3. (2017)
24. Miech, A., Laptev, I., Sivic, J.: Learnable pooling with context gating for video classification. (2017)
25. Pirsiavash, H., Ramanan, D., Fowlkes, C.: Bilinear classifiers for visual recognition. In: International Conference on Neural Information Processing Systems. (2009) 1482–1490
26. Ng, Y.H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: Deep networks for video classification. **16**(4) (2015) 4694–4702
27. Wang, H., Ullah, M.M., Klser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: British Machine Vision Conference, BMVC 2009, London, UK, September 7-10, 2009. Proceedings. (2009)
28. Jordan, M.I., Jacobs, R.A.: Hierarchical mixtures of experts and the EM algorithm. Springer London (1994)
29. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)