MULTI-TEACHER KNOWLEDGE DISTILLATION FOR COMPRESSED VIDEO ACTION RECOGNITION ON DEEP NEURAL NETWORKS

Meng-Chieh Wu, Ching-Te Chiu, Kun-Hsuan Wu

National Tsing-Hua University, Hsinchu, Taiwan
k64311@gmail.com, and chiusms@cs.nthu.edu.tw

ABSTRACT

Recently, convolutional neural networks (CNNs) have seen great progress in classifying images. Action recognition is different from still image classification; video data contains temporal information that plays an important role in video understanding. Currently, most CNN-based approaches for action recognition have excessive computational costs, with an explosion of parameters and computation time. The currently most efficient method trains a deep network directly on compressed video containing the motion information. However, this method has a large number of parameters. We propose a multi-teacher knowledge distillation framework for compressed video action recognition to compress this model. With this framework, the model is compressed by transferring the knowledge from multiple teachers to a single small student model. With multi-teacher knowledge distillation, students learn better than with single-teacher knowledge distillation. Experiments show that we can reach a 2.4× compression rate in a number of parameters and a 1.2× computation reduction with 1.79% loss of accuracy on the UCF-101 dataset and 0.35% loss of accuracy on the HMDB51 dataset.

Index Terms— Deep Convolutional Model Compression, Action Recognition, Knowledge Distillation, Transfer Learning

1. INTRODUCTION

Human action recognition has been an active research topic in computer vision because of its wide range of applications, such as smart-home and driver monitoring. Implementation of these applications using VLSI or embedded computing systems has low-power and real-time requirements. Recently, convolutional neural networks (CNNs) have seen great progress in classifying images; ConvNets have also been considered to solve action recognition problems. Most current CNN-based approaches for action recognition are based on the two-stream [1] and 3D convolutional (C3D) [2] approaches.

For two-stream-based approaches [1, 3, 4, 5, 6], the input to the spatial and temporal streams is RGB frames and stacks of multiple-frame dense optical flow fields, respectively. Using dense optical flow information for action recognition usually has good accuracy, but it has excessive computational costs.

C3D-based approaches [2, 7, 8] learn spatio-temporal features with clips of multiple continuous frames; their architecture contains 3D convolution and fully connected layers, which cause an explosion of parameters and computation time.

These methods are unable to perform action recognition efficiently. Some approaches explored other robust deep video representations [9, 10], such as CoViAR [10], to train a deep network directly on the compressed video. Video compression techniques (such as MPEG and H.264) retain only a few frames completely and reconstruct other frames on the basis of offsets from the complete images, called motion vectors and residual error. They avoid calculating the dense optical flow due to the motion vector and still achieve good performance. They also achieve the best efficiency, while requiring a far smaller amount of data. However, CoViAR has excessive storage size because of the number of parameters. For embedded mobile applications, their size consumes excessive storage/memory and computational resources. Therefore, model size reduction becomes crucial.

In our work, we propose a multi-teacher knowledge distillation framework to compress the CoViAR model. We teach the student with the comprehensive knowledge by integrating multiple teachers' knowledge to improve the accuracy after compression.

2. MULTI-TEACHER KNOWLEDGE DISTILLATION FOR COMPRESSED VIDEO ACTION RECOGNITION ON DEEP NEURAL NETWORKS

2.1. Knowledge Distillation

Distillation [11] is a technique that transfers knowledge from a cumbersome model to a small model; we call them the teacher and student models, respectively. The stu-

dent model has richer knowledge than a "vanilla" student model, but has fewer parameters and complexity than the original teacher model.

For video-level tasks, CoViAR [10] uses a sparse sampling strategy [3] on an input video, where the samples distribute uniformly along the temporal dimension, aggregating information from the samples during training. In our proposed multi-teacher knowledge distillation framework, the logits vector produced by the student network for an input video $v_i, i = 1, ..., N$ is represented by $(z_s)_i$, where the dimension of vector $(z_s)_i = [(z_s)_i^1, ..., (z_s)_i^C]$ is the number of categories C. The softmax layer converts the logits vector $(z_s)_i$ to a probability distribution $(q_s)_i = [(q_s)_i^1, ..., (q_s)_i^C]$,

$$(q_s)_i = Softmax((z_s)_i) \tag{1}$$

, where

$$(q_s)_i^j = \frac{\exp((z_s)_i^j)}{\sum_k \exp((z_s)_i^k)} \ , \text{ for } j = 1, ..., C \tag{2}$$

On the other hand, the logits vector produced by the teacher network for an input video $v_i, i = 1, ..., N$ is represented by $(z_t)_i$, where the dimension of vector $(z_t)_i = ((z_t)_i^1, ..., (z_t)_i^C)$ is the number of categories C. By introducing a parameter called temperature T, the generalized softmax layer $GSoftmax$ converts the logits vector $(z_t)_i$ to soft probability distribution $(q_t^T)_i = [(q_t^T)_i^1, ..., (q_t^T)_i^C]$,

$$(q_t^T)_i = GSoftmax((z_t)_i, T) \tag{3}$$

, where

$$(q_t^T)_i^j = \frac{\exp((z_t)_i^j/T)}{\sum_k \exp((z_t)_i^k/T)} \ , \text{ for } j = 1, ..., C \tag{4}$$

Distillation uses the class probabilities produced by the teacher model as "soft labels" for training the student model.

There are two objective functions when training the student model. The first objective function $\mathcal{L}_1$ minimizes the cross entropy with the soft labels $(q_t^T)_i$ and the soft probability $(q_s^T)_i$ produced by the student model. $(q_s^T)_i$ is computed by $GSoftmax$ with the same temperature T as the teacher model,

$$(q_s^T)_i = GSoftmax((z_s)_i, T) \tag{5}$$

, where

$$(q_s^T)_i^j = \frac{\exp((z_s)_i^j/T)}{\sum_k \exp((z_s)_i^k/T)} \ , \text{ for } j = 1, ..., C \tag{6}$$

The first objective function $\mathcal{L}_1$ is

$$\arg\min_W \mathcal{L}_1(W) = \arg\min_W -\frac{1}{NC} \sum_{i=1}^{N} \sum_{c=1}^{C} (q_t^T)_i^c \ln (q_s^T)_i^c \tag{7}$$

where $(q_s^T)_i^c$ produced by the student is the probability that the ith video belongs to the cth class, $(q_t^T)_i^c$ is the soft label produced by the teacher, W is the weights of the student, N is the number of training videos, and C is the number of total classes.

The second objective function $\mathcal{L}_2$ minimizes the cross entropy with the hard labels y_{true} and the probability $(q_s)_i$ produced by the student.

$$\arg\min_W \mathcal{L}_2(W) = \arg\min_W -\frac{1}{NC} \sum_{i=1}^{N} \sum_{c=1}^{C} (y_{true})_i^c \ln (q_s)_i^c \tag{8}$$

where $(q_s)_i^c$ produced by the student is the probability that the ith video belongs to the cth class, $(y_{true})_i^c$ is the hard label information, and $(y_{true})_i^c = 1$ if the ith video belongs to the cth class, otherwise $(y_{true})_i^c = 0$. W is the weights of the student, N is the number of training videos, and C is the number of total classes.

The overall objective function $\mathcal{L}$ is a weighted average of two different objective functions.

$$\arg\min_W \mathcal{L}(W) = \arg\min_W \lambda T^2 \mathcal{L}_1(W) + (1 - \lambda)\mathcal{L}_2(W) \tag{9}$$

where W is the weights of the student and λ is a relative weight.

2.2. Distilling on a Given Input I-frame

In CoViAR architecture, the spatial network ResNet-152 spends more time than other temporal networks. For this reason, we decided to compress the spatial network to a smaller model. According to the ResNet architecture for ImageNet [12], the number of parameters of ResNet-152 is approximately 58.2 million, and for ResNet-18 is approximately 11.2 million; the computational cost of ResNet-152 is 11.3 GFLOPs, and that of ResNet-18 is 1.8 GFLOPs. The spatial network has a 5.2-fold compression rate and 6.28-fold computation reduction because of model compression from 152 layers to 18 layers.

In our proposed multi-teacher distillation, we teach the student more comprehensive knowledge which from multiple teachers with different input types in an attempt to increase accuracy. The teacher candidates are from CoViAR separated models. For the case where all three input types (I-frame image, motion vector, and residual) are selected as teachers, we integrate the knowledge from multiple teachers and teach the student this comprehensive knowledge in the form of the soft label. The soft label is a weighted average of different soft probability distributions from multiple teachers. For the three-teachers case, the teachers $t1$, $t2$ and $t3$ produce soft probability distributions q_{t1}^T, q_{t2}^T and q_{t3}^T with the $GSoftmax$ layer and the same temperature T. The soft

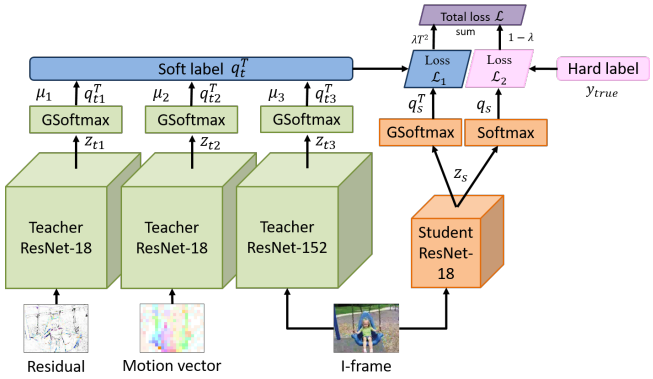

Fig. 1. Architecture of multi-teacher distillation on an I-frame image.

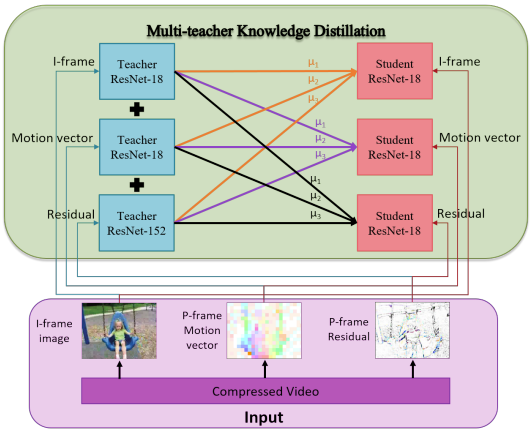

Fig. 2. Architecture of the multi-teacher to multi-student mode.

label q_t^T is a weighted average of q_{t1}^T, q_{t2}^T, and q_{t3}^T:

$$q_t^T = \frac{\mu_1 \times q_{t1}^T + \mu_2 \times q_{t2}^T + \mu_3 \times q_{t3}^T}{\mu_1 + \mu_2 + \mu_3} \quad (10)$$

where q_{t1}^T, q_{t2}^T, and q_{t3}^T are weighted by μ_1, μ_2 and μ_3, respectively. The architecture of three-teacher distillation on an I-frame image is shown in Fig. 1.

2.3. Distilling on a Given Input P-frame

Although the temporal networks that take the motion vector and residual as input have the smallest model size in the ResNet series, they can also transfer more comprehensive knowledge by distillation. The architecture of three-teacher distillation on a given motion vector or residual is similar to Fig. 1, while the difference is that we take motion vector or residual as input of student network.

2.4. Multi-teacher to Multi-student mode

We utilize the knowledge distillation technique not only to compress the spatial network model but also to promote the performance of temporal networks by multi-teacher knowledge distillation. Fig. 2 shows that comprehensive knowledge from multiple teachers was taught to different students with different input types separately, and then the results of the separate students was fused for final prediction.

3. EXPERIMENTAL RESULTS

We implemented our proposed architecture by using the open-source PyTorch framework [13]. Our models were pre-trained on the ILSVRC2012-CLS dataset [12], and we optimized our architecture by using the mini-batch and Adam optimizer [14] with a weight decay of 0.0001, eps of 0.001, initial learning rate of 0.003 for the input I-frame image, learning rate of 0.01 for the input motion vector, and learning rate of 0.005 for the input residual, which is divided by 10 when the accuracy plateaus. We trained and evaluated on a server with a 3.50-GHz Intel i7-7800K CPU, 16 GB memory, and NVIDIA GeForce GTX 1080 GPU.

3.1. Data Preprocessing

Following CoViAR [10], We used the MPEG-4 video coding format, which has on average 11 P-frames for every I-frame. The input data (images, motion vectors, and residuals) were extracted from the resized encoded videos.

There are two restrictions on the input source while distilling. First, the extracted data for teachers and student must be from the same frame or from the same group of pictures (GOP). Second, the extracted data for teachers and student must have the same data augmentation process which is following CoViAR [10], because different preprocessing processes may affect teachers' observations.

3.2. Dataset and Evaluation protocol

We evaluated our method on two action recognition datasets, UCF-101 [15] and HMDB51 [16]. UCF-101 contains 13,320 videos from 101 action categories. HMDB51 contains 6,766 videos from 51 action categories. Each video in both datasets is annotated with one action label. Each dataset has three training/testing splits for evaluation. We report the average performance of the three testing splits.

During testing, we uniformly sampled 25 frames, each with flips plus five corner crops, and then averaged the

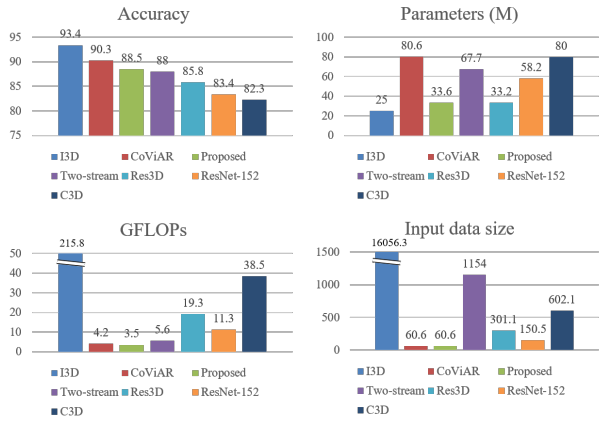

Fig. 3. Network computation complexity and accuracy on UCF-101.

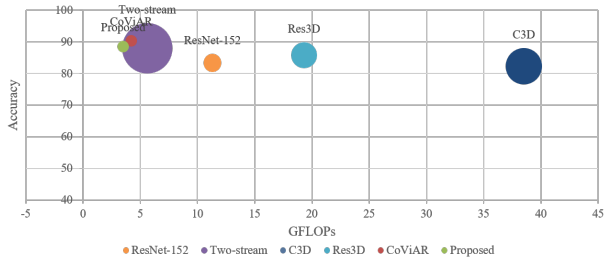

Fig. 4. Network computation complexity and accuracy on UCF-101. Node size denotes the input data size.

scores for final prediction following CoViAR [10].

3.3. Training and Testing Results

We compare the results of multi-teacher distillation with the uncompressed CoViAR [10] model in Table 1. By multi-teacher knowledge distillation, we compressed the spatial network but observed a 3.08% loss in accuracy; the temporal network on the input motion vector has a 3.82% increase in accuracy, and another temporal network on the input residual has a 3.04% increase in accuracy. The final result after compression is shown in Table 2; we achieved a 2.4x compression rate on the number of parameters with a 1.79% loss in accuracy on the UCF-101 dataset and a 0.35% loss in accuracy on the HMDB51 dataset.

Figs. 3 and 4 summarize the results. Our model achieves the best efficiency and has the fewest parameters, while having a far smaller input data size. Note that some of the state-of-the-art methods in Figs. 4 can achieve higher accuracy with large-scale video datasets. For fair comparison, we used the accuracies observed for training only on the UCF-101 dataset.

Table 1. Accuracy of the multi-teacher to multi-student mode. The top half of the table is the baseline CoViAR; the bottom half of the table is the accuracy of fusing the distillations on different inputs.

Baseline CoViAR	Input	Architecture	UCF-101	HMDB51
	Iframe	ResNet-152	87.25	51.13
	mv	ResNet-18	67.02	34.47
	residual	ResNet-18	80.88	40.85
	Fusion		90.29	56.51
Multi-teacher Distillation	Student	Architecture	UCF-101	HMDB51
	Iframe	ResNet-18	84.17 (-3.08)	48.91 (-2.22)
	mv	ResNet-18	70.84 (+3.82)	44.86 (+10.39)
	residual	ResNet-18	83.92 (+3.04)	48.39 (+7.54)
	Fusion		88.50 (-1.79)	56.16 (-0.35)

Table 2. Final results after compression.

	UCF-101	HMDB51	Parameters (M)	GFLOPs	inference time (ms)
CoViAR	90.29	56.51	80.64	4.2	12.88
Proposed	88.50 (-1.79)	56.16 (-0.35)	33.6 (0.42x)	3.5 (0.83x)	6.88 (0.53x)

4. CONCLUSION

In this study, we compressed the model, which is currently the most efficient method for action recognition, and improved the overall speed by using knowledge distillation technology to transfer its knowledge to a small model. The small model has richer knowledge than the "vanilla" small model, yet has fewer parameters and less complexity than the original cumbersome models. We also propose a multi-teacher knowledge distillation framework for compressed video action recognition to improve accuracy after compression. We integrated the knowledge from different teachers; the comprehensive knowledge can promote the performance of the student. We explored multi-teacher knowledge distillation with various combinations of different teachers to further observe its impact. Experiments show that we can reach a 2.4× compression rate and 1.2× computation reduction with approximately 1.79% loss of accuracy on the UCF-101 dataset and 0.35% loss of accuracy on the HMDB51 dataset. Our approach achieves the best efficiency and has the fewest parameters, while having a far smaller input data size.

5. REFERENCES

[1] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in Advances in neural information processing systems, 2014, pp. 568–576.

[2] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional net-

works," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.

[3] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in European Conference on Computer Vision. Springer, 2016, pp. 20–36.

[4] Zhenzhong Lan, Yi Zhu, Alexander G Hauptmann, and Shawn Newsam, "Deep local video feature for action recognition," in Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on. IEEE, 2017, pp. 1219–1225.

[5] Jiagang Zhu, Wei Zou, and Zheng Zhu, "End-to-end video-level representation learning for action recognition," arXiv preprint arXiv:1711.04161, 2017.

[6] Ali Diba, Vivek Sharma, and Luc Van Gool, "Deep temporal linear encoding networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, vol. 1.

[7] Du Tran, Jamie Ray, Zheng Shou, Shih-Fu Chang, and Manohar Paluri, "Convnet architecture search for spatiotemporal feature learning," arXiv preprint arXiv:1708.05038, 2017.

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

[9] Bowen Zhang, Limin Wang, Zhe Wang, Yu Qiao, and Hanli Wang, "Real-time action recognition with enhanced motion vector cnns," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2718–2726.

[10] Chao-Yuan Wu, Manzil Zaheer, Hexiang Hu, R Manmatha, Alexander J Smola, and Philipp Krähenbühl, "Compressed video action recognition," in CVPR, 2018.

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. Ieee, 2009, pp. 248–255.

[13] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in pytorch," 2017.

[14] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[15] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," arXiv preprint arXiv:1212.0402, 2012.

[16] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre, "Hmdb: a large video database for human motion recognition," in Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011, pp. 2556–2563.