CROSS-MODAL KNOWLEDGE DISTILLATION FOR ACTION RECOGNITION

Fida Mohammad Thoker   *Jürgen Gall*

Institute of Computer Science III, University of Bonn, 53117 Bonn, Germany

fmthoker@gmail.com   gall@iai.uni-bonn.de *

ABSTRACT

In this work, we address the problem how a network for action recognition that has been trained on a modality like RGB videos can be adapted to recognize actions for another modality like sequences of 3D human poses. To this end, we extract the knowledge of the trained teacher network for the source modality and transfer it to a small ensemble of student networks for the target modality. For the cross-modal knowledge distillation, we do not require any annotated data. Instead we use pairs of sequences of both modalities as supervision, which are straightforward to acquire. In contrast to previous works for knowledge distillation that use a KL-loss, we show that the cross-entropy loss together with mutual learning of a small ensemble of student networks performs better. In fact, the proposed approach for cross-modal knowledge distillation nearly achieves the accuracy of a student network trained with full supervision.

***Index Terms*—** Knowledge Distillation, Action Recognition, Transfer Learning, Cross-Modality Action Recognition.

1. INTRODUCTION

Action recognition is addressed in many works and in particular deep learning methods have been proposed for various modalities like RGB videos [1, 2, 3, 4] or skeleton data [5, 6, 7, 8]. Deep learning methods for action recognition, however, require large annotated datasets. This poses a problem if the modality required for an application differs from the modality of an already annotated dataset. While acquiring data is usually not a bottleneck, annotating a dataset is very time consuming. It is therefore desirable to transfer the knowledge from a network that has learned on the already annotated dataset to a network for the new modality.

For the cross-modal knowledge transfer, we assume that we have already trained a deep learning model for action recognition. This model is also called teacher network and we aim to distill [9, 10, 11] and transfer the knowledge of the teacher network to the student network for the target modality. For the transfer, we assume that we have paired videos of both modalities that are not annotated. This assumption is not a constraint for most applications since acquiring videos with two different sensors at the same time is straightforward.

In this work, we focus on the knowledge transfer from RGB videos to sequences of 3D skeleton poses [12] since skeleton and RGB data are very different modalities in terms of data structure. To transfer the knowledge from the teacher network to the student network, we propose a different loss than the KullbackLeibler (KL) divergence loss, which was used in [10]. Instead of the KL-loss, we propose the cross-entropy for the transfer from the teacher to the student and train not one student network, but multiple student networks. Using an additional mutual loss for the student networks regularizes the transfers and increases the action recognition accuracy on the target modality.

We evaluate the approach on the NTU RGB+D dataset [12] using ST-GCN [6] and HCN [7] as network architectures for the student network. The experimental evaluation shows that the proposed approach outperforms an approach that uses the KL-loss as distillation and that it nearly achieves the accuracy of a student network trained with full supervision.

2. RELATED WORKS

There is a large body of works on action recognition from 3D human pose data [13]. More recently, most approaches use either recurrent neural networks to learn spatio-temporal features from sequences of skeleton data [14, 15, 12] or convolutional neural networks for classifying the skeleton sequences [16, 8]. In [6], a spatio-temporal graph convolutional network has been proposed to learn both spatial and temporal features directly from the skeleton data. A convolutional neural network is also used in [7] to learn co-occurrence features. It combines different levels of contextual information for learning co-occurrence features in a hierarchical manner. Both raw skeleton coordinates and their temporal differences are used within a two-stream framework.

Knowledge distillation has been originally proposed to compress ensemble classifiers into a smaller network without any significant loss of performance [9, 10]. In [11], the approach has been extended to compress large networks and they showed that softening the softmax predictions of a network by a high temperature conveys important information, also called dark knowledge. Recently, knowledge distillation

*The work has been supported by the ERC Starting Grant ARCA (677650).

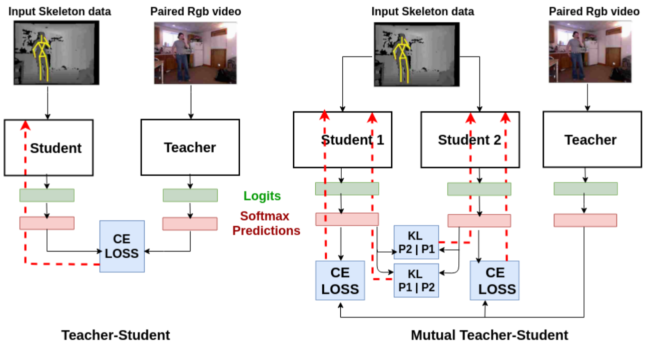

Fig. 1: *Left:* The teacher network, which has been previously trained for RGB videos, provides the supervision for the student network for skeleton data. For training the student network, unlabeled pairs for both modality and the cross-entropy loss are used. *Right:* Instead of one student network, two or more student networks can be trained together using mutual learning such that each student learns from the supervision of the teacher as well as the other student. The red dashed lines denote back-propagation for the corresponding loss functions. (Best viewed in color)

has been proposed for multi-modal action recognition. For instance, [17] use a graph-based distillation method for action recognition that is able to distill information from multiple modalities during training. Similarly, [18] proposed a multi-modal action recognition framework that uses multiple data modalities at training time. While these works analyze if the networks can be better trained using full supervision if additional modalities including the modality of the test data are available during training, we address the problem if the modality of the annotated training set differs from the modality of the test set. In [4], a 3D convolutional neural network is initialized by transferring the knowledge of a pre-trained 2D CNN. Cross-modal distillation has been also used for other tasks such as object detection [19], emotion recognition [20], or human pose estimation [21].

3. CROSS-MODAL ACTION RECOGNITION

For cross-modal action recognition, we assume that a teacher network has been already trained on RGB videos. We now aim to train the student network for another modality, namely sequences of 3D human poses. For training the student network, we use pairs of RGB videos and human pose sequences. The pairs are not annotated and were therefore not part of the training data for the teacher network.

3.1. Teacher-Student Network

The training of the student network is illustrated in Fig. 1(a). The trained teacher network predicts for a training pair from the source modality the target class probabilities, where the vector of all class probabilities is denoted by P_T. The parameters of the student network are then optimized such that the class probabilities P_S estimated by the student for the target modality matches P_T. In [10], the KullbackLeibler (KL) divergence has been proposed as loss for knowledge transfer between two networks of the same modality:

$$\mathcal{KL}(\mathrm{P}_\mathrm{S}^\tau, \mathrm{P}_\mathrm{T}^\tau) = \sum_c \mathrm{P}_\mathrm{S}^\tau(c) \log \frac{\mathrm{P}_\mathrm{S}^\tau(c)}{\mathrm{P}_\mathrm{T}^\tau(c)} \tag{1}$$

where $\mathrm{P}_\mathrm{S}^\tau$ and $\mathrm{P}_\mathrm{T}^\tau$ are softmax predictions of the student and teacher networks both softened with temperature τ:

$$\mathrm{P}^\tau(c) = \frac{exp(\frac{z_c}{\tau})}{\sum_d exp(\frac{z_d}{\tau})}. \tag{2}$$

A temperature value of $\tau > 1$ produces a softer probability distribution over the classes and has been proposed to avoid overfitting [10].

3.1.1. Loss Function

In our experimental evaluation, we show that the loss function (2) is not optimal for cross-modal knowledge transfer. In particular, finding an optimal τ is difficult and it strongly depends on the student network. Instead, we propose to use the cross-entropy loss

$$\mathcal{CE}(\mathrm{P}_\mathrm{S}, \mathrm{P}_\mathrm{T}) = -\log\left(\mathrm{P}_\mathrm{S}(\hat{c}_T)\right) \tag{3}$$

where $\hat{c}_T = \mathrm{argmax}_c \, \mathrm{P}_\mathrm{T}(c)$. This means that the teacher makes a hard decision and we use the class label estimated by the teacher as supervision for the student network.

3.2. Mutual Learning

In the context of fully supervised image classification, [22] proposed a deep mutual learning strategy. Instead of learning a single network with full supervision, an ensemble of networks is learned collaboratively and the network teach each other throughout the training process.

We show that mutual learning is also useful for cross-modal knowledge transfer. In this case, we train an ensemble of K student networks together such that each network learns to mimic the probability distribution of the teacher network, as well as to match the probability estimate of its peers. Our approach for cross-modal knowledge transfer with mutual learning is shown in Fig. 1(b) for $K = 2$.

Since the students are applied to the same modality, we can apply the KL-loss with softened temperature τ (1). The loss functions L_{Θ_1} and L_{Θ_2} for the student networks with parameters Θ_1 and Θ_2, respectively, are then given by

$$L_{\Theta_1} = \mathcal{CE}(\mathrm{P}_1, \mathrm{P}_\mathrm{T}) + \mathcal{KL}(\mathrm{P}_1^\tau, \mathrm{P}_2^\tau) \tag{4}$$

and

$$L_{\Theta_2} = \mathcal{CE}(\mathrm{P}_2, \mathrm{P}_\mathrm{T}) + \mathcal{KL}(\mathrm{P}_2^\tau, \mathrm{P}_1^\tau). \tag{5}$$

Noise%	0	5	10	14	20	25
Acc	78.50	73.20	72.58	71.51	69.70	68.01

Table 1: Impact of noisy labels during training on the classification accuracy of the ST-GCN model. The *Student-Train* set is used for training and the *Test* set for evaluation.

τ	1	2	5	10	20	50
Acc	51.05	52.00	70.80	71.17	68.90	64.00

Table 2: Accuracy of the ST-GCN student network on the *Test* set using the KL-loss with different values for the softmax temperature τ.

#Students	Method	Accuracy	Accuracy
(K)	(supervision)	(Max)	(Average)
1	Full Supervision	-	78.50
1	Teacher-Student	-	71.17
2	Ensemble without mutual	71.93	72.32
2	Mutual Learning	73.20	73.60
3	Mutual Learning	73.60	74.22
4	Mutual Learning	73.30	73.50

Table 3: Impact of mutual learning and the number of student networks K. In case of multiple student networks, we combine the predictions of the student networks during inference either by averaging the class probabilities or taking the maximum probability of each class. For the experiments, the KL-loss with $\tau = 10$ is used.

The proposed approach can be extended to more student networks. For K students, the loss function for optimizing the k-th student network is given by

$$L_{\Theta_k} = \mathcal{CE}(\mathrm{P_k}, \mathrm{P_T}) + \frac{1}{K-1}\sum_{l\neq k} \mathcal{KL}(\mathrm{P}_\mathrm{k}^{\tau}, \mathrm{P}_l^{\tau}). \qquad (6)$$

4. EXPERIMENTS

We evaluate our approach on the large scale multi-modal action recognition dataset NTU RGB+D [12] which contains more than 56 thousand video samples. The videos are collected from 40 distinct subjects and contain 60 different action classes. We use the RGB videos as source modality for the teacher network and the skeleton data as target modality. We adapt the cross subject evaluation protocol with 40,320 samples from 20 subjects for training and 16,560 samples from the remaining 20 subjects for testing. To evaluate the knowledge transfer, we divide the 20 training subjects into two groups of 10 subjects each, resulting in the *Teacher-Train* set for training the teacher network and the *Student-Train* set for training the student networks. While the RGB videos with class labels are used for the *Teacher-Train* set, the *Student-Train* comprises pairs of RGB videos and sequences of 3D human poses, but no class labels. We evaluate the accuracy of the student networks on the pose data of the *Test* set. We use Temporal Segment Networks [2] (TSN) as our teacher network and use optical flow as the teacher modality. We use the same hyper-parameters as in [2]. For the student networks, we use the Spatio Temporal Graph Convolution Networks (ST-GCN) [6] and the Hierarchical Co-occurrence Network (HCN) [7] which both use the skeleton modality as their input data. We train the ST-GCN model using two GPUs with a batch size of 16 for a total of 200 epochs. All other hyper-parameters are the same as in [6] and [7].

In order to analyze how much knowledge we can extract from the teacher network, we evaluate the action recognition accuracy of the teacher network, which has been trained on the *Teacher-Train* set. On the *Student-Train* set, we obtain an accuracy of 86%, i.e., the teacher network will produce around 14% wrong labels during the knowledge transfer to the student networks.

Next, we study the effect of noisy labels on the performance of the ST-GCN network. To conduct this experiment, we assign randomly wrong labels to a percentage of the training videos in the *Student-Train* set. We then train the ST-GCN model on *Student-Train* in a fully supervised manner using this noisy labeled data as ground-truth. Table 1 reports the action recognition accuracy on the *Test* set for different percentages of noisy labels during training. 78.5% is the upper bound that can be achieved by cross-modal knowledge transfer if the teacher network is perfect since it corresponds to training the student network with full supervision. The accuracy drops from 78.5% to 73.2% if 5% of the videos are wrongly labelled. Given that the teacher network misclassifies 14% of the videos on *Student-Train*, we can expect to achieve 71.51% accuracy using cross-modal knowledge transfer.

Given some bounds for the accuracy that we can expect, we now analyse the impact of the loss functions for the task of cross-modal knowledge transfer. For the rest of the experiments, we train the student networks on *Student-Train* using the teacher network as supervision and evaluate the action recognition accuracy of the student network on *Test*. We first analyze the impact of the temperature τ for the KL-loss (1). Table 2 shows that for $\tau \leq 2$, the accuracy is very low since ST-GCN overfits on the *Student-Train* set.

We still keep the KL-loss with $\tau = 10$, but evaluate the benefit of using more than one student network for mutual learning. Table 3 compares the accuracy of mutual learning with multiple student networks (last three rows). In this case, we obtain an ensemble of student networks where the predictions are combined by averaging the class probabilities (average). We also report the results if for each class the highest probability among all student networks is taken (max). The results show that averaging performs better than taking the maximum. $K = 3$ gives the best accuracy and mutual learning increases the accuracy compared to a teacher-student

Loss	# of students	Accuracy
Full supervision	-	78.50
KL	1	71.17
Cross-entropy	1	74.91
KL + Mutual	2	73.60
Cross-entropy + Mutual	2	**77.83**

Table 4: Results for the cross-entropy loss. For mutual learning, we average over the student networks. The cross-entropy loss outperforms the KL loss reported in Table 3.

Loss	# of students	Accuracy
Full supervision	-	80.60
KL ($\tau = 1$)	1	74.40
KL ($\tau = 2$)	1	74.90
Cross-entropy	1	77.40
Cross-entropy + Mutual	2	79.00
Cross-entropy + Mutual	3	**79.50**

Table 5: Accuracy of the HCN student network on the *Test* set using different loss functions and varying number of student networks.

setup as proposed in [10] by $+3\%$. To analyze if the improvement stems from the ensemble model or the mutual learning, we also trained two student networks without the mutual loss (ensemble without mutual). The result shows that 50% of the improvement is due to the ensemble and the rest due to the mutual learning. It is interesting to note that mutual learning already achieves a higher accuracy than training the network with 5% of randomly assigned wrong labels (Table 1).

So far we have only used the KL-loss, but we have not evaluated the proposed approach using the cross-entropy loss (6). We report the results with the cross-entropy loss in Table 4. Compared to the KL-loss, the accuracy increases from 71.17% to 74.91% for one student network and from 73.6% to 77.83% for two student networks. While the second term in (6) uses the KL-loss with $\tau = 10$ for mutual learning, we observed that the accuracy decreases if cross-entropy is used for both terms. Compared to [10], the proposed approach improves the accuracy by $+6.66\%$. Note that the proposed approach nearly achieves the accuracy of ST-GCN trained with full supervision.

In order to demonstrate that the proposed approach is insensitive to the type of student network, we also evaluated the accuracy of cross-modal knowledge transfer if we use HCN [7] as student network. Table 5 reports the results for the HCN model. For the KL-loss (1), we had to adjust the temperature τ. While ST-GCN performs better for a large value of τ as reported in Table 2, it is the other way around for HCN since HCN is a smaller network which suffers less from overfitting. For HCN, $\tau = 2$ performs best and larger values of τ actually decrease the accuracy. This shows that is very difficult to chose the hyper-parameter τ [10] in the context of cross-modal action recognition. If we use the proposed cross-entropy loss, this problem does not occur and it outperforms the KL-loss. If we use mutual learning with two or three students, the action recognition accuracy is improved by $+4.1\%$ or $+4.6\%$, respectively, compared to KL with $\tau = 2$ [10]. Note that we still use $\tau = 10$ in (6) and we found that (6) is not sensitive to the parameter τ since the mutual loss is computed only for the student networks, which have the same network architecture applied to the same modality.

Method	Full Train	*Student-Train*
Skeletal Quads [23]	38.62	
Lie Group [24]	50.08	
HBRNN-L [15]	59.07	
Dynamic Skeletons [25]	60.23	
PA-LSTM [12]	62.90	
STA-LSTM [14]	73.40	
ST-LSTM+TS [26]	69.20	
Temporal Conv [16]	74.30	
VA-LSTM [27]	79.20	
ST-GCN [6]	81.60	78.50
Two-stream CNN [8]	83.20	
HCN [7]	86.50	80.60
Cross-modal ST-GCN		77.83
Cross-modal HCN		79.50

Table 6: Comparison with the state-of-the-art for the cross-subject protocol. Note that the numbers are not directly comparable since the other approaches are trained with full supervision on the entire training set. While our approach is trained only on *Student-Train* and with less supervision.

Finally, we compare our approach with the current state-of-the-art methods for the skeleton modality on the NTU RGB+D dataset in Table 6. Although our student networks are trained on less data and with less supervision, they achieve a higher accuracy than many other approaches that are trained with full supervision on the entire training set.

5. CONCLUSION

We have presented an approach that uses knowledge distillation for cross-modal action recognition. The approach is able to transfer knowledge from one modality to another modality without the need of any additional annotations. Instead, pairs of sequences of both modalities are sufficient for the knowledge transfer. We evaluated our approach on a large-scale multi-modal dataset using two different student networks. For both networks, the accuracy of the networks trained with cross-modal knowledge transfer is very close to the accuracy of the networks if they are trained with full supervision.

6. REFERENCES

[1] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems 27*. 2014.

[2] 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 *ECCV*, 2016.

[3] João Carreira and Andrew Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *CVPR*, 2017.

[4] Ali Diba, Mohsen Fayyaz, Vivek Sharma, M. Mahdi Arzani, Rahman Yousefzadeh, Juergen Gall, and Luc Van Gool, "Spatio-temporal channel correlation networks for action classification," in *ECCV*, 2018.

[5] Yong Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *CVPR*, 2015.

[6] Sijie Yan, Yuanjun Xiong, and Dahua Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, 2018.

[7] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," in *IJCAI*, 2018.

[8] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu, "Skeleton-based action recognition with convolutional neural networks," in *ICME Workshops*, 2017.

[9] Cristian Buciluă, Rich Caruana, and Alexandru Niculescu-Mizil, "Model compression," in *KDD*, 2006.

[10] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[11] Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang, "A survey of model compression and acceleration for deep neural networks.," *CoRR*, vol. abs/1710.09282, 2017.

[12] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in *CVPR*, 2016.

[13] Mao Ye, Qing Zhang, Liang Wang, Jiejie Zhu, Ruigang Yang, and Juergen Gall, "A survey on human motion analysis from depth data.," in *Time-of-Flight and Depth Imaging*. 2013, vol. 8200 of *Lecture Notes in Computer Science*, pp. 149–187, Springer.

[14] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *AAAI*, 2017.

[15] Yong Du, Wei Wang, and Liang Wang, "Hierarchical recurrent neural network for skeleton based action recognition," .

[16] Tae Soo Kim and Austin Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," *in CVPR Workshops*, 2017.

[17] Zelun Luo, Jun-Ting Hsieh, Lu Jiang, Juan Carlos Niebles, and Li Fei-Fei, "Graph distillation for action detection with privileged modalities," in *ECCV*, 2018.

[18] Nuno C. Garcia, Pietro Morerio, and Vittorio Murino, "Modality distillation with multiple stream networks for action recognition," in *ECCV*, 2018.

[19] Saurabh Gupta, Judy Hoffman, and Jitendra Malik, "Cross modal distillation for supervision transfer," in *CVPR*, 2016.

[20] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *ACM Multimedia*, 2018.

[21] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi, "Through-wall human pose estimation using radio signals," in *CVPR*, June 2018.

[22] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu, "Deep mutual learning," in *CVPR*, 2018.

[23] G. Evangelidis, G. Singh, and R. Horaud, "Skeletal quads: Human action recognition using joint quadruples," in *ICPR*, 2014.

[24] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *CVPR*, 2014.

[25] Jian-Fang Hu, Wei-Shi Zheng, Jian-Huang Lai, and Jianguo Zhang, "Jointly learning heterogeneous features for rgb-d activity recognition," in *CVPR*, 2015.

[26] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang, "Spatio-temporal LSTM with trust gates for 3d human action recognition," in *ECCV*, 2016.

[27] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," *in ICCV*.