Distillation Multiple Choice Learning for Multimodal Action Recognition Supplementary Material

Nuno Cruz Garcia[1,2], Sarah Adel Bargal[3], Vitaly Ablavsky[4]
Pietro Morerio[6,7] , Vittorio Murino[5,6,7], Stan Sclaroff[3]

[1]Faculdade de Ciências, Universidade de Lisboa, Portugal [2]Copelabs, ULHT, Portugal
[3]Boston University [4]University of Washington [5]Dipartimento di Informatica, University of Verona, Italy
[6] Istituto Italiano di Tecnologia [7]Ireland Research Center, Huawei Technologies Co. Ltd., Dublin, Ireland

nrgarcia@ciencias.ul.pt, sbargal@bu.edu, vxa@uw.edu
{pietro.morerio,vittorio.murino}@iit.it, sclaroff@bu.edu

1. Hyperparameters

As referred in the paper, we present more details on the hyperparameters so that our experiments can be reproducible. Our results were obtained using $\lambda = 1$, in Equation 3. For some runs, we obtained slightly better results using $\lambda = 0.5$. We observed that the temperature $T = 2$ generally works best for all datasets, although results using $T = 5$ are similar in some cases. We trained our networks for 220 epochs using using SGD optimizer with Momentum 0.9, and an initial learning rate of 10_3, and decay of 10_1 at epoch [100,150,180,200]. We used 5% of the training data for validation and early stopping, but observed that the accuracy stabilizes at the maximum value.

2. Comparing MCL Methods.

As mentioned in the main paper, we provide more comprehensive results for comparison against other MCL methods and independently trained modality networks. We compare the performance of SMCL and CMCL with our proposed DMCL. We also compare against independently trained modality networks. For each method we present the accuracy of the RGB, Depth, and Optical Flow modality networks, the sum of all modality network predictions ($\sum$), and the oracle accuracy (Φ). Please note that SMCL [2] does not give a single prediction, and following [2] we sum all network predictions in the presented results here. We compare against both versions of CMCL, CMCL_0 and CMCL_1 proposed by Lee *et al*. [1]. Table 1, 2, 3, and 4 are extensions of Table 1 of the main paper.

We note our method is significantly effective for three out of four datasets. As mentioned in the paper, the distillation effect is less effective for the largest dataset. One possibility may be related to hyperparameter tuning. While for the other smaller datasets, we were able to test temperatures in the range of 2 to 10, this is not practical for the largest dataset.

Using our method, the RGB network is improved $\sim$5% in comparison to the baseline for the UWA3DII and NWUCLA. This result hold across other combinations of training and test views, as showed in the next section.

3. Results on NWUCLA and UWA3DII

In the main paper, we presented results for the most commonly used view setting for these datasets. As referred in the paper, we present here the results on the remaining views, for NWCULA on table 5, and for UWA3DII on table 6.

The last column of each table shows the increase in performance that our method gives to the modality networks, using only one modality at test time. We confirm that the results are consistent across views, what shows the effectiveness of our method.

	Ind.	**SMCL**	**CMCL_0**	**CMCL_1**	**DMCL**
RGB	87.53	24.83	12.23	11.13	**93.64**
Depth	80.30	24.46	15.41	13.30	**83.29**
Flow	89.58	50.68	73.16	84.60	**91.07**
$\sum$	**93.79**	49.00	83.08	84.73	93.28
Φ	**97.86**	86.79	88.82	89.65	97.64

Table 1. **Northwestern-UCLA dataset.View$_3^{1,2}$**

	Ind.	**SMCL**	**CMCL_0**	**CMCL_1**	**DMCL**
RGB	73.74	25.19	3.03	22.28	**78.39**
Depth	77.09	24.70	46.86	21.65	**81.87**
Flow	**89.66**	38.60	52.01	45.49	88.26
$\sum$	**89.75**	60.70	85.53	31.90	89.50
Φ	**95.52**	88.51	90.25	83.89	94.96

Table 2. **UWA3DII dataset. View$_{1,3}^{2,4}$**

	Ind.	**SMCL**	**CMCL_1**	**DMCL**
RGB	79.66	26.67	29.61	**81.25**
Depth	77.97	30.41	32.27	**78.98**
Flow	84.19	33.30	32.69	**84.45**
$\sum$	**86.57**	62.22	05.28	86.23
Φ	**92.11**	86.19	86.29	91.71

Table 3. **NTU120^{mini} dataset.**

	Ind.	**SMCL**	**CMCL_0**	**CMCL_1**	**DMCL**
RGB	**84.86**	22.31	24.42	22.37	84.31
Depth	**83.31**	25.77	29.45	25.77	82.29
Flow	**86.72**	32.82	38.78	32.82	86.44
$\sum$	**89.74**	5.54	85.44	5.06	88.46
Φ	**94.36**	79.81	92.17	85.20	93.21

Table 4. **NTU120 dataset.**

Method	**Training Modality**	**Testing Modality**	**View**$_1^{2,3}$	**View**$_2^{1,3}$	**View***$_3^{1,2}$	**Avg.**	Avg. δ
Independent	RGB	RGB	54.73	55.13	87.52	65.79	-
Independent	Depth	Depth	47.32	29.59	80.30	52.40	-
Independent	Flow	Flow	74.08	**78.05**	89.58	80.57	-
Ours	RGB, Depth, Flow	RGB	**59.95**	**62.01**	**93.64**	**71.86**	**+6.12**
Ours	RGB, Depth, Flow	Depth	**49.79**	**32.54**	**83.29**	**55.20**	**+2.99**
Ours	RGB, Depth, Flow	Flow	**74.59**	77.85	**91.07**	**81.17**	**+1.48**

Table 5. **NWUCLA dataset.** This table shows results for all the combinations of views for the cross-view protocol defined in the original paper [4]. The superscript refers to the training views, and subscript to test. Each result is the average of 3 runs. For each column, *i.e.* for each view, results in bold represent the best result per modality, with each colour representing a test modality. The last column shows that, on average, our method increases significantly the performance for all networks with respect to the baseline.

Method	**Training Modality**	**Testing Modality**	**View**$_{3,4}^{1,2}$	**View**$_{2,4}^{1,3}$	**View**$_{2,3}^{1,4}$	**View***$_{1,3}^{2,4}$	**View**$_{1,4}^{2,3}$	**View**$_{1,2}^{3,4}$	**Avg.**	Avg. δ
Independent	RGB	RGB	63.42	62.16	71.03	73.74	59.39	78.18	67.99	-
Independent	Depth	Depth	68.15	67.85	69.22	77.09	71.95	80.99	72.54	-
Independent	Flow	Flow	**87.94**	83.05	79.65	**89.66**	84.82	**86.76**	85.31	-
Ours	RGB, Depth, Flow	RGB	**66.49**	**64.78**	**73.28**	**78.39**	**63.92**	**81.93**	71.47	**+3.48**
Ours	RGB, Depth, Flow	Depth	**72.80**	**73.04**	**71.96**	**81.87**	**74.31**	**82.05**	76.00	**+3.45**
Ours	RGB, Depth, Flow	Flow	87.09	**84.42**	**81.08**	88.26	**85.01**	86.47	85.38	**+0.07**

Table 6. **UWA3DII dataset.** This table shows results for all the combinations of views for the cross-view protocol defined in the original paper [3]. The superscript refers to training views, and subscript to test. View* is the one presented in the main paper. Each result is the average of 3 runs. For each column, *i.e.* for each view, results in bold represent the best result per modality, with each colour representing a test modality. The last column shows that, on average, our method increases the performance of RGB and Depth networks in respect to the baselines in $\sim 3.4\%$. The results are not so visible in the Optical Flow network.

References

[1] Kimin Lee, Changho Hwang, Kyoung Soo Park, and Jinwoo Shin. Confident multiple choice learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2014–2023. JMLR. org, 2017.

[2] Stefan Lee, Senthil Purushwalkam Shiva Prakash, Michael Cogswell, Viresh Ranjan, David Crandall, and Dhruv Batra. Stochastic multiple choice learning for training diverse deep ensembles. In *Advances in Neural Information Processing Systems*, pages 2119–2127, 2016.

[3] Hossein Rahmani, Arif Mahmood, Du Huynh, and Ajmal Mian. Histogram of oriented principal components for cross-view action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 38(12):2430–2443, 2016.

[4] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2649–2656, 2014.