Distillation Multiple Choice Learning for Multimodal Action Recognition

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

Abstract

In this work, we address the problem of learning an ensemble of specialist networks using multimodal data, while considering the realistic and challenging scenario of possible missing modalities at test time. Our goal is to leverage the complementary information of multiple modalities to the benefit of the ensemble and each individual network. We introduce a novel Distillation Multiple Choice Learning framework for multimodal data, where different modality networks learn in a cooperative setting from scratch, strengthening one another. The modality networks learned using our method achieve significantly higher accuracy than if trained separately, due to the guidance of other modalities. We evaluate this approach on three video action recognition benchmark datasets. We obtain state-of-the-art results in comparison to other approaches that work with missing modalities at test time. [1]

1. Introduction

Humans perceive the environment by processing a combination of modalities. Such modalities can include audio, touch and sight, with each modality being distinct from and complementary to the others. Deep learning methods may likewise benefit from multimodal data. In this paper, we explore how to leverage the complementary nature of multimodal data at training time, in order to learn a better classifier that takes as input only RGB data for inference.

One popular way to train multimodal deep learning models is to train one network per modality, and mean pool all the network predictions for inference. This is a sub-optimal use of multimodal training data, as modalities do not exchange information while training. For example, considering the task of action recognition, some actions are easier to discriminate using certain modalities over others: the action "open a box" may be confused with "fold paper" when solely relying on the RGB modality, while it is easily classified using depth data [24].

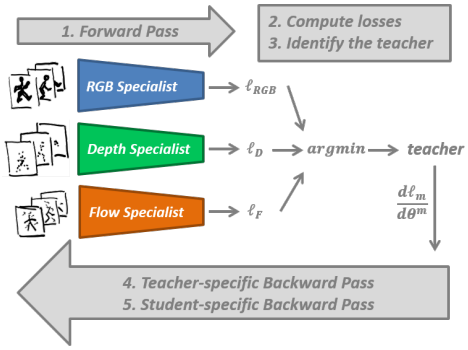

Figure 1: **Distillation Multiple Choice Learning (DMCL)** allows multiple modalities to cooperate and strengthen one another. For each training sample, the modality specialist m that achieves the lowest loss ℓ distills knowledge to strengthen other modality specialists. At test time, any subset of available modalities can be used by DMCL to make predictions.

This suggests that an ensemble of networks could use multimodal data in a more efficient way, *e.g.* by encouraging the network trained with a given modality to focus on the set of classes or samples that maximizes its discriminative power. In this case, each network is referred to as a *specialist network*, as it only sees part of the dataset and specializes in that part of the problem. Assuming that all modalities are available, the ensemble should be able to fuse the specialists' predictions and produce a single output.

The problem of multimodal fusion becomes more challenging when some modalities are not available at test time. This is particularly problematic if the training process encourages the specialization of each modality network of the

[1]Work performed at Boston University. Code available at https://github.com/ncgarcia/DMCL

ensemble. In this case, a missing modality means that the ensemble loses the ability to correctly classify the corresponding part of the task assigned to this specialist.

In this paper, we propose a novel method that is at the intersection of MCL framework and Knowledge Distillation [16, 28], called Distillation Multiple Choice Learning (DMCL). DMCL addresses two practical dimensions of multimodal learning: a) leveraging the complementarity of multiple modalities, and b) being robust to missing modalities at test-time.

We take inspiration from the Multiple Choice Learning (MCL) framework, which is a popular way to train an ensemble of RGB networks [21, 19, 40]. This method chooses the best performing network of the ensemble to backpropagate the task loss. However, extending it to multiple modalities is not straightforward. Networks that are trained using different modalities learn at different speeds. Consequently, the network that learns faster in the beginning of the training dominates the traditional MCL algorithm, and is encouraged to remain dominant throughout the training. We extend MCL to a) address such challenges associated with multimodal data, and b) deal with modalities that may be missing at test time.

The case of a missing modality at test time is related to learning using Privileged Information [42] and Knowledge Distillation [16]. This type of approaches is usually structured as a two-step process: training a teacher network, and then using its knowledge to train a student network. The teacher network has usually a larger capacity, or has access to more data than the student. For example, consider the problem of learning a model for action recognition using a multimodal dataset composed of RGB, depth, and optical flow videos. In practice, it is reasonable to assume that only RGB modality is present for test inference: depth sensors are expensive and optical-flow computation incurs runtime cost that may not meet real-time budget. At the same time, depth and optical flow can provide valuable information on the samples or classes that it perform better, and that could be distilled to the RGB network [37][3].

We build on these ideas to develop a model that learns from multimodal data, exploiting the strength of each modality in a cooperative setting as the training proceeds. This is summarized in Figure 1. Furthermore, our proposed model is able to account for one or more missing modalities at test time. Our main contributions are:

- We conduct a deep evaluation of the MCL framework in the context of multimodal learning and give insights on how multiple modalities behave in such ensemble learning methods.
- We propose DMCL, a MCL framework designed for multimodal data where modalities cooperate to strengthen one another. Moreover, DMCL is able to account for missing modalities at test time.
- We present competitive to or state-of-art results for multimodal action recognition using privileged information on three video action recognition benchmark datasets.

2. Related Work

Generalized Distillation. The Generalized Distillation [28] framework gives a unifying perspective on Knowledge Distillation (KD) [16] and Learning Using Privileged Information (LUPI) [42]. KD was first proposed as a way to transfer knowledge from a large ensemble of networks to a single small capacity network [16]. It uses the smoothed ensemble's probability distribution as a soft target to train the lighter network, in addition to the ground truth target. LUPI refers to the setting where some information available at training time is not be available at test time [42]. The privileged information can be provided by a "teacher" network, for example, a model previously trained on another dataset or modalities. The "student" network leverages the additional information to learn a better model to be used at test time.

These ideas have been applied in many creative ways to a variety of domains such as network compression [2], language tasks [7], defending from adversarial attacks [30], transfer labels across domains [12], unifying classifiers using unlabeled data [43], using distillation without a pre-trained teacher [48][47], and others [8]. Inspired by these ideas, we extend the MCL algorithm for multimodal tasks, allowing knowledge transfer between modalities in a cooperative learning setting via KD.

Video Action Recognition. Video action recognition has a vast body of literature, with deep learning breaking accuracy scores every year [6][22][27][26]. We focus on multimodal deep learning methods in a privileged information setting, *i.e.* using fewer modalities at test time. A more comprehensive review is presented in [45][15][18]. The combination of RGB and Optical Flow is one of the most popular ways to capture appearance and temporal information for video tasks [37]. Some interesting works use modules specifically developed to learn motion features, which are then incorporated in models that use RGB only [39][20][32][49][4][38]. Due to the specificity of these modules, these architectures can be difficult to adapt to incorporate other kind of features or modalities, such as depth. Other methods learn an additional hallucination network to mimic the features of a missing modality [9][10][4]. These works use all data of all modalities indiscriminately, and learning the additional hallucination network requires a pre-trained network. Our method learns by exploring the multimodal data asymmetrically via the MCL algorithm, which leverages the strengths of each modality, without the

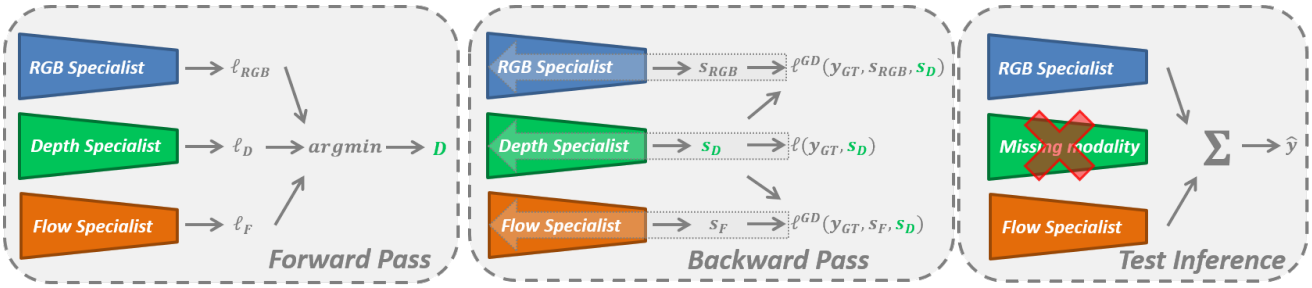

Figure 2: **Distillation Multiple Choice Learning (DMCL)** In the Forward Pass, we calculate the classification cross-entropy losses ℓ for each modality and identify the teacher network - in this case, the Depth network. In the Backward Pass, we compute the soft targets of the teacher, S_D, and use them as an extra supervision signal for the student networks. The loss for the student networks ℓ^{GD} refers to the Generalized Distillation loss, defined on Eq. 3. The loss for the teacher network D uses the normal logits, *i.e.* soft targets with temperature $T = 1$. At test time, we are able to cope with missing modalities. The final prediction is obtained by averaging the predictions of the available modalities.

need of a pre-training step or an additional network at test time. The important work Luo *et al*. [29] uses distillation to transfer knowledge across modalities. This approach consists in learning a graph to mediate the strength of the imitation loss between modalities. The graph is learned after a pre-training stage in which the modality networks are trained separately. Our method moves towards a system that exploits the complementary nature of multimodal data since the beginning of the training process, *i.e*. with randomly initialized networks.

Ensemble Methods. A comprehensive review about ensemble methods is well presented in [35]. The most relevant method to ours is the Multiple Choice Learning (MCL) framework. Guzman-Rivera *et al*. [13] proposed MCL to optimize the oracle accuracy of an ensemble of models. Lee *et al*. [21] proposed Stochastic MCL, an adaptation of MCL to an ensemble of neural networks that learn via stochastic gradient descent. Each network of the ensemble trained via Stochastic MCL produces a set of diverse outputs. The inability to output a single prediction compromises its use in real applications. Lee *et al*. [19] addressed this issue with Confident MCL. The main idea is to avoid confident predictions for the classes not assigned to a given specialist. This allows for the sum of all ensemble's networks outputs to get a single prediction. Tian *et al*. [40] also addressed this issue by training an additional network to estimate the weight of the outputs of each specialist. While [19] and [40] propose ways to get a single prediction out of the ensemble, they do not address how such methods can be used with multimodal data. We draw inspiration on these works to address this issue within the MCL framework.

3. Method: Training Multimodal Specialists

Our goal is to learn an ensemble of multimodal specialists that leverages the specific strengths of each modality to the benefit of the ensemble. This is accomplished by setting a cooperative learning strategy where stronger networks teach weaker networks through knowledge distillation. For a given data point at training time, we identify the best-performing network as a teacher for the remaining networks in the ensemble.

3.1. Distillation Multiple Choice Learning

Algorithm 1 describes our method DMCL. Let $\mathbb{D} = \{(x_i, y_i)\}^N$ be a multimodal dataset having N training samples. Each sample x_i represents the data for the M modalities available, $x_i = \{x_i^1, \ldots, x_i^M\}$, and y_i represents its label.

Our ensemble is composed of a set of M networks f, each using as input a different modality $f^1(x_i^1), \ldots, f^M(x_i^M))$. The MCL algorithm maximizes the ensemble accuracy, often referred to as oracle accuracy. The oracle accuracy assumes that we can choose the correct prediction out of the set of outputs produced by each network. This translates to the minimization of the ensemble loss L, which is defined as the lowest of the individual networks' loss values, calculated for a given data point.

Formally, MCL minimizes the ensemble loss L with respect to a specific task loss $\ell(y_i, \hat{y}_i)$ for each network prediction $\hat{y}_i = f^m(x_i^m)$ for a specific modality m:

$$L(\mathbb{D}) = \sum_{i=1}^{N} \min_{m \in \{1,\ldots,M\}} \ell(y_i, f^m(x_i^m)). \quad (1)$$

In practice, we get all the networks' predictions for each sample of the batch. We calculate the loss $\ell_{criterion}$ for each network and sample (line 5, Algorithm 1). In this case, $\ell_{criterion}$ corresponds to the standard cross-entropy loss. The network with the lowest loss value is designated as the winner network, and the others are set to be loser networks. The loss and gradient updates for a network depend on whether it is a winner or loser network (lines 10-14, Algorithm 1). In our proposed privileged-information formulation, we view the winner network as a teacher, and the loser networks as students.

DMCL function of `update_winner` and `update_losers` of Algorithm 1 define how the teacher

network distills information to the student networks, strengthening them. DMCL updates teachers with respect to the cross-entropy training loss computed using the ground-truth label. The loser networks are updated using a distillation loss, which aims to transfer knowledge from the winner network.

Knowledge Distillation. Matching the students' with the teachers' soft targets is one way to transfer knowledge from one model to another. Soft targets are a smoothed probability distribution than the originally produced by the modality network f^m:

$$s_i^m = \sigma(f_i^m(x_i^m)/T), \tag{2}$$

where σ is the softmax function, f_i^m are the logits, and T is a scalar value. The default temperature T value is set to 1 for models that do not incorporate distillation. Setting T to a higher value produces a smoother probability distribution that reveals valuable information about the relative probabilities between classes, which has shown to improve knowledge transfer and generalization of the new model. In practice, very small probability values become more evident with higher temperatures.

The Generalized Distillation (GD) [28] method consists of three sequential steps: (1) learn the teacher network; (2) fix the teacher and compute the soft target for all samples; (3) use the teacher's soft targets as additional targets to the ground truth to learn student networks. The Generalized Distillation loss is defined as:

$$\begin{aligned}\ell^{GD}(i) = (1-\lambda)\ell(y_i, \sigma(f(x_i))) \\ +\lambda\ell(s_i, \sigma(f(x_i))), \lambda \in [0,1]\end{aligned} \tag{3}$$

In contrast, we use distillation in an online fashion in the context of the MCL framework. The role of teacher / student network is assigned to the winner / loser network respectively, for each sample of the batch. The soft targets are computed using the winner network output, which is used to compute the loss and update the loser networks. We do not pretrain teachers as per conventional distillation, *i.e*. all networks are randomly initialized. In DMCL, teachers and students learn together in a cooperative setting.

This cooperative setting is beneficial in two ways: It gives loser networks the opportunity to build good representations even if they are not the *argmin* chosen network; And it still enables networks to specialize in parts of the problem.

Missing modalities. Our training method encourages each network to learn using ground truth labels for its specialty samples (those obtaining lowest loss), and from the other specialist networks for samples otherwise. By doing so, each specialist incorporates knowledge related to all samples/classes of the task. This enables each network to classify any sample at test time, therefore rendering the ensemble able to account for missing modalities.

Algorithm 1: DMCL

Input: Dataset $\mathbb{D} = \{(x_i, y_i)\}_i^N$, and randomly initialized networks $f^1, \ldots, f^M$ parameterized by $\theta^1, \ldots, \theta^M$
Output: M trained networks $f^1, \ldots, f^M$

1 **for** *step* $\leftarrow 1$ *to convergence* **do**
2 Sample batch $\mathbb{B} \subset \mathbb{D}$
3 **for** $m \leftarrow 1$ *to* M **do**
4 Forward Pass:
5 $\ell^m_{criterion} =$ cross_entropy $(y_i, \hat{y}^m)$
6 **end**
7 **for** $i \leftarrow 1$ *to* $|\mathbb{B}|$ **do**
8 *// Backward Pass:*
9 *// Update winner network* m^*
10 $m^* \leftarrow \underset{m \in \{1,\ldots,M\}}{\arg\min} \{\ell^m_{criterion}\}$
11 $\theta^{m^*} =$ `update_winner` $(\theta^{m^*}, x_i^{m^*}, y_i, f)$
12 *// Update loser networks* m^c
13 $m^c \leftarrow \{1, ..., M\} \setminus \{m^*\}$
14 $\theta^{m^c} =$ `update_losers` $(\theta^{m^c}, x_i^{m^c}, y_i, f)$
15 **end**
16 **end**
17 **return** $f^1, \ldots, f^M$
18 *// **Function Definitions***
19 **Function** `update_winner` $(\theta^{m^*}, x_i^{m^*}, y_i, f)$**:**
20 *// Compute the gradient w.r.t. cross-entropy loss;*
21 $\nabla_{\theta^{m^*}} \ell = \frac{\partial \ell(y_i, f^{m^*}(x_i^{m^*}))}{\partial \theta^{m^*}}$;
22 *// Update parameters of the winner network;*
23 $\theta^{m^*} \leftarrow \theta^{m^*} - \eta \nabla_{\theta^{m^*}} \ell$;
24 **return** θ^{m^*};
25 **Function** `update_losers` $(\theta^{m^c}, x_i^{m^c}, y_i, f)$**:**
26 *// Compute soft targets of* f^{m^*} *using Eq. 2;*
27 $s_i^{m^*} = \sigma(f_i^{m^*}(x_i^{m^*})/T)$;
28 *// Compute soft targets of* f^{m^c} *using Eq. 2;*
29 $s_i^{m^c} = \sigma(f_i^{m^c}(x_i^{m^c})/T)$;
30 *// Compute the gradient w.r.t. GD loss using Eq. 3;*
31 $\nabla_{\theta^{m^c}} \ell^{GD} = \frac{\partial \ell^{GD}(y_i, f^{m^c}, s_i^{m^*}, s_i^{m^c})}{\partial \theta^{m^c}}$;
32 *// Update parameters of the loser networks;*
33 $\theta^{m^c} \leftarrow \theta^{m^c} - \eta \nabla_{\theta^{m^c}} \ell^{GD}$;
34 **return** θ^{m^c};

3.2. Relationship to other MCL methods

The general framework for MCL is described in lines 1-17 of Algorithm 1. The main idea is to enable each of the networks of the ensemble to specialize in different parts of the problem. This algorithm was first devised for RGB ensembles. Two recent instances of MCL are Stochastic MCL (SMCL) [21] and Confident MCL (CMCL) [19]. These methods differentiate from each other and from the general MCL framework in two fundamental ways: 1) the criterion loss used to decide whether a network is a winner or a loser (line 10, Algorithm 1), and 2) how winner and loser models are updated (line 11 and 14, Algorithm 1). In SMCL, $\ell_{criterion}$ corresponds to the task loss, *e.g*. standard cross-entropy for classification. The winner model is updated with respect to that same loss, while the loser models are not updated. This update scheme is also used in [40]. In CMCL, the $\ell_{criterion}$ corresponds to the task loss plus an additional loss that measures how well the other networks

predict the uniform distribution, for the given sample. The winner model is updated as in the SMCL method and the loser models are updated with respect to the KL divergence between its predictions and the uniform distribution.

Neither variations of MCL satisfy our problem statement. SMCL does not result in a single prediction. While CMCL does result in a single prediction by averaging the predictions, it does not account for the idiosyncrasies of multimodal data. The first aspect has to do with heterogeneous training dynamics resulting from having multimodal data as input. Figure 3(right) shows the cross-entropy loss of three networks independently trained for action recognition, using RGB (blue), optical flow (orange), and depth (green). Optical flow learns at a much faster speed than the other modalities. This results in an undesired effect when using CMCL: the optical flow network repeatedly achieves the lowest loss. This behavior is reinforced by the *argmin* operator and the update scheme of CMCL, that does not allow useful gradients to pass to the loser networks. Eventually, the optical flow network ends up winning for all the training samples, which renders the other networks and modalities useless. The second challenge is the probable overfitting. The current training update scheme dictates that only the winner network gets useful gradients to build good representations for the given task, which reduces the data used to train each network. To address this and prevent overfitting, CMCL proposes to share the lower layers of the feature encoders. This is not feasible when the different networks are learning from different modalities as their representations/domains are significantly different.

DMCL addresses these issues for multimodal data by using a cooperative learning setting where the ensemble networks teach each other via Knowledge Distillation. At the same time, DMCL leverages the ensemble learning strategy of the traditional MCL framework, where models specialize depending on their performance with respect to a given input.

4. Experiments

In this section, we present the action recognition benchmark datasets we use to evaluate our approach. We then present the architecture and setup of our experiments. We analyze the performance of our DMCL in comparison to other MCL training strategies. We give insight into why other MCL training strategies fall short for multimodal data. We then demonstrate our privileged information state-of-the-art results and conclude with a discussion of our experimental results.

4.1. Datasets

We test DMCL on three video action recognition datasets that offer RGB and depth data. We augment the three datasets with optical flow frames obtained using the implementation available at [31], based on Liu *et al*. [23].

Northwestern-UCLA (NW-UCLA). This dataset [44] features ten people performing ten actions, captured simultaneously at three different viewpoints. We follow the cross-view protocol suggested by the authors in [44], using two views for training and the remaining for testing.

UWA3DII. This dataset [33] features ten subjects performing thirty actions for four different trials, each trial corresponding to a different viewpoint. As suggested in [33], we follow the cross-view protocol using two views for training and two for testing.

NTU120. The very recent NTU RGB+D 120 dataset [36] is one of the largest multimodal dataset for video action recognition. It consists of a total of 114,480 trimmed video clips of 106 subjects performing 120 classes, including single person and two-person actions, across 155 different viewpoints and 96 background scenes. We follow the cross-subject evaluation protocol proposed in the original paper, using fifty three subjects for training and the remaining for testing. We also create three versions of NTU120, which we refer to as NTU120^{mini}, that contains 50% sampled training data from the 120 classes. We note that NTU120 and NTU120^{mini} share the same test data. When results are reported on NTU120^{mini} they are averaged over the three runs. We also evaluate our method on the smaller less recent version of this dataset, NTU60 [36], that has 60 classes, in order to compare against state-of-the-art reported results.

4.2. Architecture and Setup

Each modality network is implemented as the `R(2+1)D-18` architecture proposed in [41]. This architecture is based on a Resnet-18 network [14], modified such that a 1D temporal convolution is added after every 2D convolution, thus giving the network the ability to learn spatiotemporal features. The factorization of a 3D convolution into a combination of 2D + 1D convolution has shown to be more effective for video classification tasks. The ensemble of modality networks is simultaneously trained following Algorithm 1.

The input of each modality network is a clip of eight frames of the corresponding modality. For each training step, a video is split into eight equal parts and we randomly sample a frame from each of them. Each training input frame is a crop of dimension [224,224,3], cropped around a randomly shifted center, for each video. We also use other data augmentation techniques such as random horizontal flipping and random color distortions. The networks are trained from scratch for all the experiments, using SGD optimizer with Momentum 0.9, and an initial learning rate of $10^{-}3$. At test time, we sample ten clips per video, each clip consisting of eight frames randomly sampled, centered, and

Table 1: **Comparing MCL methods.** We compare the performance of SMCL [21] and CMCL [19] with our proposed DMCL on the NWUCLA, UWA3DII, and NTU120 datasets. We also compare against independently trained modality networks. For each method we present the accuracy of the RGB modality network, the sum of all modality network predictions ($\sum$), and the oracle accuracy (Φ). For each row, corresponding to one dataset, we highlight in bold the best result using RGB only at test time. Using our DMCL methods results in better RGB networks for three out of four datasets.

	Independent			**SMCL [21]**			**CMCL [19]**			**Our DMCL**		
	RGB	$\sum$	Φ	RGB	$\sum$	Φ	RGB	$\sum$	Φ	RGB	$\sum$	Φ
NWUCLA	87.53	93.79	97.86	24.83	49.00	86.79	11.13	84.73	89.65	**93.64**	93.28	97.64
UWA3DII	73.74	89.75	95.52	25.19	60.70	88.51	22.28	31.90	83.89	**78.39**	89.50	94.96
NTU120^{mini}	79.66	86.57	92.11	26.67	62.22	86.19	29.61	5.28	86.29	**81.25**	86.23	91.71
NTU120	**84.86**	89.74	94.36	22.31	5.54	79.81	22.37	5.06	85.20	84.31	88.46	93.21

with no data augmentation techniques. The final prediction for each video is the average of the ten clip predictions. We have experimented with different values of temperature T and hyperparameter λ, and found that T={2,5} and λ={1, 0.5} works best, with little accuracy variations. Further details related to hyperparameters are given in the supplementary material.

4.3. Results

In this section, we demonstrate how DMCL leverages multiple modalities to learn an RGB network that outperforms an independently trained RGB classifier - our baseline, and other MCL training strategies. All MCL strategies are trained using the same training process as our method, including data augmentation techniques, optimizer, and number of steps, and are considered as ablation experiments of our method. We then demonstrate state-of-the-art privileged information results.

Comparison vs. MCL variants. Table 1 shows the action classification performance on the three video action recognition benchmark datasets for MCL variants and independently trained modality networks. We present the classification accuracy using the RGB modality, the sum of predictions of RGB, Flow, and Depth modalities (Σ), and the oracle accuracy (Φ). An oracle Φ is assumed to have the ability to select the modality that gives the best prediction among the ensemble. Our DMCL approach performs better than modalities trained independently, *i.e.* without MCL, and better than SMCL and CMCL variants. While Table 1 focuses on improvement with regard to the RGB modality, we provide similar results for Depth and Optical Flow in the supplementary material. We note that the effect of knowledge distillation is more visible in the three smaller datasets.

Table 1 also shows that combining the predictions of three modalities (Σ) generally improves accuracy. The fact that the oracle accuracy (Φ) is significantly higher than Σ indicates that, for some cases, at least one modality predicted the correct class, however, the sum of predictions (Σ) resulted in an incorrect prediction. However, the gap between Σ and Φ is lower for DMCL compared to the other approaches. This indicates that DMCL combines modality predictions in a more optimal fashion to improve overall accuracy. The low accuracies of SMCL and CMCL are due to artifacts created by the use of multimodal data, which we investigate in the next section. We have checked the implementation of these methods on RGB-only ensembles, which lead to similar results to those reported in the original papers.

KNN accuracy with random features					
Modality	k=1	k=5	k=10	k=50	k=120
RGB	10.53	10.74	11.11	11.32	12.26
Depth	9.72	10.68	10.77	15.37	13.31
Optical Flow	23.23	23.96	25.31	26.35	24.53

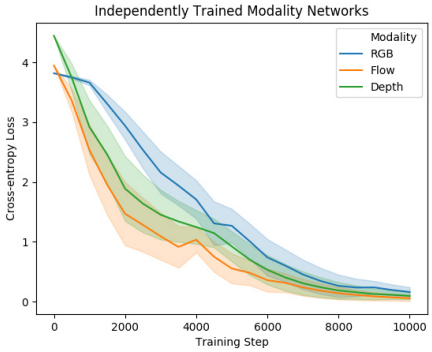

Figure 3: (Left) Accuracy of a KNN classifier with varying k on the NWUCLA dataset. Classified features are computed using randomly initialized networks for each modality. Although all features are randomly generated, optical flow random features tend to achieve a significantly higher accuracy. This helps to explain why optical flow networks learn faster than other modalities. (Right) The cross-entropy loss of three networks independently trained for action recognition on the UWA3DII dataset, using RGB (blue), depth (green), and optical flow (orange). These plots are averaged over three runs. We observe that for the first 10K steps, the training loss of the optical flow network is consistently lower, resulting in a winner-takes-all behavior in traditional MCL algorithms. However, in DMCL, the winner network also teaches the loser networks, strengthening the other modality networks and avoiding this behavior.

Learning speed for different modalities. One of the goals of this paper is to investigate and bring new insights on multimodal learning. In a MCL setting, having a specific modality learn at a faster pace compared to others often leads to an imbalance of the number of data points each modality network is presented with at training time. Networks specializing in different modalities typically do not

Table 2: **Selecting the right teacher network is important.** We present the action recognition classification accuracy on the NWUCLA and UWA3DII datasets for three scenarios, where: modality networks are trained independently; a random teacher is assigned for every sample to guide the other modality networks; and DMCL, where the best-performing teacher (lowest loss) is selected to guide other modality networks. For each column, corresponding to a test modality, we highlight in bold the best result across the three scenarios.

Dataset	NWUCLA					UWA3DII				
Test Modality	RGB	Depth	Flow	Σ	Φ	RGB	Depth	Flow	Σ	Φ
Independent	87.53	80.30	89.58	93.79	97.86	73.74	77.09	**89.66**	89.75	95.52
Random Teacher	89.57	57.81	89.43	86.93	95.71	71.07	79.07	85.03	84.47	92.60
Our DMCL	**93.64**	**83.29**	**91.07**	93.28	97.64	**78.39**	**81.87**	88.26	88.51	94.59

share a backbone of parameters due to the very different nature of the inputs - in contrast to the SMCL and CMCL variants where there is a shared backbone. As a consequence, if a modality network dominates the training process, *i.e.* being the one to consistently achieve the lowest loss for training batches, it will be presented with significantly more training data compared to the other modality networks. We observed that optical flow often dominates the ensemble training process particularly when training using CMCL. This is depicted in Figure 3(right) where the training loss curves of the independently trained networks for Optical Flow, Depth, and RGB are shown over the training steps. Namely, looking at the first steps of the curve we see that Optical Flow curve is consistently lower than Depth, which in turn has lower values than RGB. This is consistent with what we find during training of CMCL, where the RGB network is often ignored, the Depth network learns from a few samples and overfits early, and the Optical Flow network sees the vast majority of the samples.

We further investigate why optical flow dominates the learning process in our action recognition setting. We compute random features extracted from a randomly initialized untrained network for each of the modalities using the same architecture described previously. We then run a kNN classifier using the random features. Figure 3(left) shows results of this experiment on the NWUCLA dataset for $k = 1, 5, 10, 50, 120$. The accuracy of the random features of the optical flow modality is almost twice that achieved using Depth and RGB. The fact that the kNN classifier achieves such good performance compared to the other modalities suggests that Optical Flow data naturally clusters better per class. From the perspective of a deep neural network learning process, this could be interpreted as a better initialization, thus speeding the initial stage of learning.

Leveraging Teacher Strength. In this section, we ablate the mechanism by which the teacher role is determined. The teacher role is assigned to the network that achieves the lowest loss for each sample of the batch, therefore being in the best position to guide/strengthen the other networks. To verify this claim, we train our model with a random assignment of a teacher for each sample of the batch. This can be thought of as a randomized distillation process. We then compare the overall action recognition classification accuracy of both approaches in Table 2. Choosing the right network as teacher consistently achieves better performance compared to a randomly assigned teacher, for every modality. This is in-line with work that combines distillation and graphs, where distillation has a specific direction specified by the direction of the edges [29]. It is interesting to note that random teacher assignment may result in better performance than individual modality networks, *e.g.* for NWUCLA the RGB individual network accuracy is 87.53% *vs.* 89.57% for a random teacher assignment. These may be related to the known regularization effect of knowledge distillation, that has been empirically shown to lead to better performance [16, 7].

State-of-the-art Comparisons. We now compare DMCL to state-of-the-art privileged information methods, and modality baselines, for the task of human action recognition from videos. Table 3 shows results for the UWA3DII and NWUCLA datasets. The top part of the table presents modality baselines for methods that use the same number of modalities in training and testing, including our individually trained modality networks. The bottom part of the table refers to methods that have missing modalities at test time. Our DMCL using RGB only for testing achieves higher accuracy compared to all baselines that use RGB at training and testing, and compared to all state-of-the-art privileged information methods that use RGB at test time, including those that use additional hallucination networks at test time, achieving an absolute improvement of 4.7% for UWA3DII and 6.1% for NWUCLA. Similarly, our DMCL outperforms all baselines when the only available modality is Depth by 4.8% absolute improvement and the state-of-the-art method by 1.3% on UWA3DII.

Table 4 presents results on three versions of the NTU dataset: NTU60, NTU120^{mini}, and the full NTU120. We see that the distillation effect is much more visible in the case of less data. For example, for NTU^{mini}, we achieve an absolute improvement of 1.6% over the baseline for the RGB modality, and of 6% for NTU60. Our best modality network for NTU60 achieves 85.65% compared to the 89.5% of [29] that uses twice the number of modalities we use for training and an additional graph network module.

Table 3: **Accuracy for UWA3DII and NWUCLA dataset.** The first part of the table refers to methods that use unsupervised feature learning (*) or that use the same number of modalities for training and testing. The second part of the table refers to methods that use more modalities for training than for testing. Methods that use RGB$^+$ at test time use an additional network that mimics the missing modality. For each column, corresponding to one dataset, we highlight in colored bold the best result and in normal colored font the second best between our method and the baselines. Each color corresponds to a different test modality. To conduct a fair comparison with baseline methods, this table presents results for the most common view setting for UWA3DII and NWUCLA. Other view settings follow the same trend and results are presented in the supplementary material.

	Method	Training Modalities	Testing Modalities	UWA3DII	NWUCLA
Modality Baselines	R-NKTM [34]	Syn*	RGB	66.3	78.1
	Action Tubes [11]	RGB	RGB	33.7	61.5
	Long-term RCNN [5]	RGB	RGB	74.5	64.7
	Baseline (RGB)	RGB	RGB	73.74	87.52
	MVDI+CNN [46]	D	D	68.3	**84.2**
	Baseline (D)	D	D	77.09	80.30
	Baseline (F)	F	F	**89.66**	89.58
	$\sum$ (RGB, D, F)	RGB, D, F	RGB, D, F	89.75	93.9
Privileged Info.	Hoffman *et al.* [17]	RGB, D	RGB$^+$	66.67	83.30
	Garcia *et al.* [9]	RGB, D	RGB$^+$	73.23	86.72
	ADMD [10]	RGB, D	RGB$^+$	-	91.64
	DMCL	RGB, D, F	RGB	**78.39**	**93.64**
	DMCL	RGB, D, F	D	**81.87**	83.29
	DMCL	RGB, D, F	F	88.26	**91.07**

Table 4: **NTU Datasets.** The test sets for NTU120^{mini} and NTU120 are the same. For each column, corresponding to one dataset, we highlight in bold the best result and in normal colored font the second best between our method and the baselines. Each color corresponds to a different test modality. The approximated values are inferred from a plot in [24]. We note that the effect of the distillation method is more visible on the smaller scale versions NTU60 and NTU120^{mini} of the dataset.

	Method	Training Modalities	Testing Modalities	NTU60	NTU120^{mini}	NTU120
Modality Baselines	ST-LSTM [25]	Skeleton (S)	Skeleton (S)	69.2	$\sim$ 50.0	55.7
	VGG [24]	RGB	RGB	-	$\sim$ 40.0	58.5
	Baseline (RGB)	RGB	RGB	77.59	79.66	**84.86**
	VGG [24]	D	D	-	$\sim$ 20.0	48.7
	Baseline (D)	D	D	78.97	78.67	**83.32**
	Baseline (F)	F	F	81.43	84.21	**86.72**
	VGG [24]	RGB, D	RGB, D	-	-	61.9
	VGG [24]	RGB, D, S	RGB, D, S	-	-	64.0
	$\sum$ (RGB, D, F)	RGB, D, F	RGB, D, F	87.25	86.57	89.74
Privileged Info.	Garcia *et al.* [10]	RGB, D	RGB	73.11	-	-
	ADMD [9]	RGB, Depth	RGB	73.4	-	-
	Luo *et al.* [29]	RGB, F, D, S^{1,2,3}	RGB	**89.5**	-	-
	DMCL	RGB, D, F	RGB	83.61	**81.25**	84.31
	DMCL	RGB, D, F	D	**80.56**	**78.98**	82.22
	DMCL	RGB, D, F	F	**85.65**	**84.45**	86.44

5. Conclusions

MCL is a powerful way for training ensembles of networks, originally proposed for RGB data. We demonstrate undesirable behaviors of this framework when naively applied to multimodal data. We propose DMCL that extends MCL frameworks to leverage the complementary information offered by the multimodal data to the benefit of the ensemble. The cooperative learning is enabled via knowledge distillation that allows the ensemble networks to exchange information and learn from each other. We demonstrate that modality networks trained using our DMCL achieve competitive to or state-of-the-art results compared to the privileged information literature, and significantly higher accuracy compared to independently trained modality networks for human action recognition in videos.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541. ACM, 2006.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[4] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. Mars: Motion-augmented rgb stream for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7882–7891, 2019.

[5] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[6] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 6202–6211, 2019.

[7] Tommaso Furlanello, Zachary C Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. *arXiv preprint arXiv:1805.04770*, 2018.

[8] Chuang Gan, Boqing Gong, Kun Liu, Hao Su, and Leonidas J Guibas. Geometry guided convolutional neural networks for self-supervised video representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5589–5597, 2018.

[9] Nuno C. Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[10] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Learning with privileged information via adversarial discriminative modality distillation. *TPAMI*, 2019.

[11] Georgia Gkioxari and Jitendra Malik. Finding action tubes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 759–768, 2015.

[12] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016.

[13] Abner Guzman-Rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. In *Advances in Neural Information Processing Systems*, pages 1799–1807, 2012.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. Going deeper into action recognition: A survey. *Image and vision computing*, 60:4–21, 2017.

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

[17] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 826–834, 2016.

[18] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *arXiv preprint arXiv:1806.11230*, 2018.

[19] 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.

[20] Myunggi Lee, Seungeui Lee, Sungjoon Son, Gyutae Park, and Nojun Kwak. Motion feature network: Fixed motion filter for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 387–403, 2018.

[21] 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.

[22] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7083–7093, 2019.

[23] Ce Liu et al. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, Massachusetts Institute of Technology, 2009.

[24] Jun Liu, Amir Shahroudy, Mauricio Lisboa Perez, Gang Wang, Ling-Yu Duan, and Alex Kot Chichung. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[25] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European Conference on Computer Vision*, pages 816–833. Springer, 2016.

[26] Xiang Long, Chuang Gan, Gerard De Melo, Xiao Liu, Yandong Li, Fu Li, and Shilei Wen. Multimodal keyless attention fusion for video classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[27] Xiang Long, Chuang Gan, Gerard De Melo, Jiajun Wu, Xiao Liu, and Shilei Wen. Attention clusters: Purely attention based local feature integration for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7834–7843, 2018.

[28] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

[29] Zelun Luo, Jun-Ting Hsieh, Lu Jiang, Juan Carlos Niebles, and Li Fei-Fei. Graph distillation for action detection with privileged modalities. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[30] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE, 2016.

[31] Deepak Pathak. Pyflow - python dense optical flow. `https://github.com/pathak22/pyflow`.

[32] AJ Piergiovanni and Michael S Ryoo. Representation flow for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9945–9953, 2019.

[33] 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.

[34] Hossein Rahmani, Ajmal Mian, and Mubarak Shah. Learning a deep model for human action recognition from novel viewpoints. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):667–681, 2018.

[35] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.

[36] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

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

[38] Jonathan C Stroud, David A Ross, Chen Sun, Jia Deng, and Rahul Sukthankar. D3d: Distilled 3d networks for video action recognition. *WACV 2020*, 2020.

[39] Shuyang Sun, Zhanghui Kuang, Lu Sheng, Wanli Ouyang, and Wei Zhang. Optical flow guided feature: A fast and robust motion representation for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1390–1399, 2018.

[40] Kai Tian, Yi Xu, Shuigeng Zhou, and Jihong Guan. Versatile multiple choice learning and its application to vision computing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6349–6357, 2019.

[41] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[42] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.

[43] Jayakorn Vongkulbhisal, Phongtharin Vinayavekhin, and Marco Visentini-Scarzanella. Unifying heterogeneous classifiers with distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3175–3184, 2019.

[44] 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.

[45] Pichao Wang, Wanqing Li, Philip Ogunbona, Jun Wan, and Sergio Escalera. Rgb-d-based human motion recognition with deep learning: A survey. *Computer Vision and Image Understanding*, 171:118–139, 2018.

[46] Yang Xiao, Jun Chen, Yancheng Wang, Zhiguo Cao, Joey Tianyi Zhou, and Xiang Bai. Action recognition for depth video using multi-view dynamic images. *Information Sciences*, 480:287–304, 2019.

[47] Chenglin Yang, Lingxi Xie, Chi Su, and Alan L Yuille. Snapshot distillation: Teacher-student optimization in one generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2859–2868, 2019.

[48] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018.

[49] Jiaojiao Zhao and Cees GM Snoek. Dance with flow: Two-in-one stream action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9935–9944, 2019.