Better and Faster: Knowledge Transfer from Multiple Self-supervised Learning Tasks via Graph Distillation for Video Classification

Chenrui Zhang and **Yuxin Peng**[*]
Institute of Computer Science and Technology, Peking University, Beijing 100871, China
pengyuxin@pku.edu.cn

Abstract

Video representation learning is a vital problem for classification task. Recently, a promising unsupervised paradigm termed *self-supervised learning* has emerged, which explores inherent supervisory signals implied in massive data for feature learning via solving auxiliary tasks. However, existing methods in this regard suffer from two limitations when extended to video classification. First, they focus only on a single task, whereas ignoring complementarity among different task-specific features and thus resulting in suboptimal video representation. Second, high computational and memory cost hinders their application in real-world scenarios. In this paper, we propose a graph-based distillation framework to address these problems: (1) We propose logits graph and representation graph to transfer knowledge from multiple self-supervised tasks, where the former distills classifier-level knowledge by solving a multi-distribution joint matching problem, and the latter distills internal feature knowledge from pairwise ensembled representations with tackling the challenge of heterogeneity among different features; (2) The proposal that adopts a teacher-student framework can reduce the redundancy of knowledge learned from teachers dramatically, leading to a lighter student model that solves classification task more efficiently. Experimental results on 3 video datasets validate that our proposal not only helps learn better video representation but also compress model for faster inference.

1 Introduction

Video representation learning aims to capture discriminative features from video data and is a critical premise for classification problem. In the last decade, supervised methods have achieved remarkable success in a variety of areas, showing extraordinary performance on representation learning. However, heavy reliance on well-labeled data limits the scalability of these methods as building large-scale labeled datasets is time-consuming and costly. Furthermore, learning from manual annotations is inconsistent with biology, as living organisms develop their visual systems without the requirement for millions of semantic labels. Hence, there is a growing interest in unsupervised learning, and *self-supervised learning*, one of the most promising unsupervised representation learning paradigms, is gaining momentum.

In contrast to supervised methods, self-supervised learning exploits structural information of the raw visual data as supervisory signals to yield transferable representation without manual annotations. Concretely, machine is asked to solve an auxiliary task by leveraging self-supervision rather than labels, and this process can result in useful representation. The core hypothesis behind this idea is that solving these tasks need high-level semantic understanding of data, which forces self-supervised models to learn powerful representation.

Self-supervised learning is especially potential in video processing area since video is an information-intensive media that can provide plentiful contextual supervisory cues by nature. While various types of auxiliary strategies in video [Agrawal *et al.*, 2015; Wang and Gupta, 2015; Jayaraman and Grauman, 2016; Fernando *et al.*, 2017] have shown impressive performance, there are two main *limitations* of these works. *First*, they commonly resort to a single task without accounting for complementarity among different task-specific features. Empirically, solving different tasks in video need different features and these features can complement each other to form a comprehensive understanding of video semantics. *Second*, in order to achieve better performance, researchers tend to adopt deeper and wider models for representation embedding at the expense of high computational and memory cost. As video data in real-world workflows is of huge volume, efficiency issue must be addressed before practical application of classification approaches.

In this paper, we argue that heterogeneous video representations learned from different auxiliary tasks in an ad-hoc fashion are not orthogonal among each other, which can be incorporated into a more robust feature-to-class semantics. In analogy to biological intelligence, humans can improve performance on a task via transferring knowledge learned from other tasks and, intuitively, a general-purpose representation is strong enough to handle tasks in different scenarios. In light of above analysis, we propose to learn video representation for classification problem with a *Divide and Conquer* manner: (1) Instead of designing **one** versatile model for cap-

[*] Corresponding author.

turing discriminating features from different aspects simultaneously, we distill knowledge from **multiple** teacher models that are adept at specific aspects to the student model; (2) The student model is expected to be **lighter** since redundancy of knowledge from teachers is reduced after distillation. To this end, we propose a graph-based distillation framework, which bridges the advance of self-supervised learning and knowledge distillation for both exploiting complementaries among different self-supervised tasks and model compression. The main contributions of this paper are as follows:

(1) We propose logits graph (G_l) to distill softened prediction knowledge of teachers at classifier level, where we formalize logits graph distillation as a multi-distribution joint matching problem, and adopt Earth Mover (EM) distance as criteria to measure complementary gain flowing among the vertices of G_l. (2) We propose representation graph (G_r) to distill internal feature knowledge from pairwise ensembled representations yield by compact bilinear pooling, which tackles the challenge of heterogeneity among different features, as well as performs as an adaptation method to assist logits distillation.

Attributed to the above two distillation graphs, student model can incorporate complementary knowledge from multiple teachers to form a comprehensive video representation. Furthermore, distillation mechanism makes sure that student model works with fewer parameters than teachers, which has lower computational complexity and memory cost.

We conduct comprehensive experiments on 3 widely-used video classification datasets, which validate that the proposed approach not only helps learn better video representation but also improve efficiency for video classification.

2 Related Work

2.1 Self-supervised Learning

Self-supervised learning is a recently introduced unsupervised paradigm, its key contribution is answering the question that how to effectively evaluate the performance of models trained without manual annotations. Typically, works in this area design tasks which are not directly concerned, such "auxiliary" tasks are difficult enough to ensure models can learn high-level representations. Nowadays, various self-supervised methods have been studied in computer vision, owing to image/video can provide rich structural information for developing auxiliary tasks.

Previous auxiliary tasks in single image domain involve asking networks to inpaint images with large missing regions [Pathak *et al.*, 2016], colorize grayscale images [Zhang *et al.*, 2017] and solve jigsaw puzzles by predicting relative position of patches [Noroozi and Favaro, 2016], *etc.* Compared to images, videos contain more abundant spatiotemporal information to formulate self-supervision. For example, temporal continuity among frames can be leveraged to build a sequence order verification or sorting task [Misra *et al.*, 2016; Fernando *et al.*, 2017; Lee *et al.*, 2017]. Wang et al. [2015] finds corresponding patch pairs via visual tracking and use the constraint that similarity of matching pairs is larger than that of random pairs for training guidance. Analogously, Dinesh et al. [2016] proposes temporally close frames should have similar features, as well as features change over time should be smooth. In addition, as shown in [Agrawal *et al.*, 2015], ego-motion can also be utilized as meaningful supervision for visual representation learning. Our work is based on four self-supervised approaches in video (see section 3).

2.2 Knowledge Distillation

Knowledge distillation (KD) [Hinton *et al.*, 2015] aims to utilize the *logits*, i.e., pre-softmax activations of trained classifiers (i.e., teacher models), to form softened probabilities that convey information of intra- and inter-class similarities. These extra supervisions (aka soft targets) can be combined with original one-hot labels (aka hard targets) to guide a lighter model (i.e., student model) to learn a more generalizable representation. After [Hinton *et al.*, 2015], a surge of variants emerge to mine different forms of knowledge implied in teacher models. FitNets [Romero *et al.*, 2015] treats internal feature maps of ConvNet as hints that can provide explanation of how a teacher solves specific problems. Sergey et al. [2017] propose learning to mimic the attention maps of teacher model can be helpful. Besides knowledge that teacher learned from single sample, the relationships across different samples are also valuable for student training [Chen *et al.*, 2017]. In this paper, we distill knowledge of multiple teachers from both logits prediction and internal feature, denoted by *logits distillation* and *representation distillation* respectively.

Lopez-Paz et al. [2016] propose *generalized distillation* (GD) to combine KD with privileged information [Vapnik and Izmailov, 2015]. This technique distills knowledge from privileged information learned by teacher model at training stage, aiming at boosting student model at test stage where supervision from teacher is totally absent. More recently, [Luo *et al.*, 2017] considers multi-modal data as privileged information and extends GD paradigm to a graph-based form. In this paper, we advocate that self-supervised tasks could be powerful privileged information which can supply student model with expertise from multiple views of video semantics, and internal feature distillation should be leveraged to further assist softened logits distillation as an adaptation method.

3 Self-supervised Tasks

Under the principle of maximizing complementarity, we first implement four self-supervised methods in video, including frame sorting [Lee *et al.*, 2017], learning from egomotion [Agrawal *et al.*, 2015], tracking [Wang and Gupta, 2015], and learning from frame predicting that we design in this work inspired by [Lotter *et al.*, 2017]. For better readability, we denote them using abbreviations as $\pi_{\mathcal{S}}$, $\pi_{\mathcal{M}}$, $\pi_{\mathcal{T}}$ and $\pi_{\mathcal{P}}$, respectively. $\pi_{\mathcal{S}}$ formulates the sequence sorting task as a multi-class classification problem. Taking symmetry of some actions (e.g., opening/closing a door) into account, there are $n!/2$ possible permutations for each n-tuple of frames. $\pi_{\mathcal{M}}$ captures camera transformation between two adjacent frames that the agent receives when it moves and trains model to determine whether two given frames are temporally close. $\pi_{\mathcal{T}}$ tracks similar patches in video and measures distance of them in representation space. $\pi_{\mathcal{P}}$ predicts the subsequent frames of video, and we adopt a neuroscience inspired architecture named PredNet [Lotter *et al.*, 2017] as visual encoder.

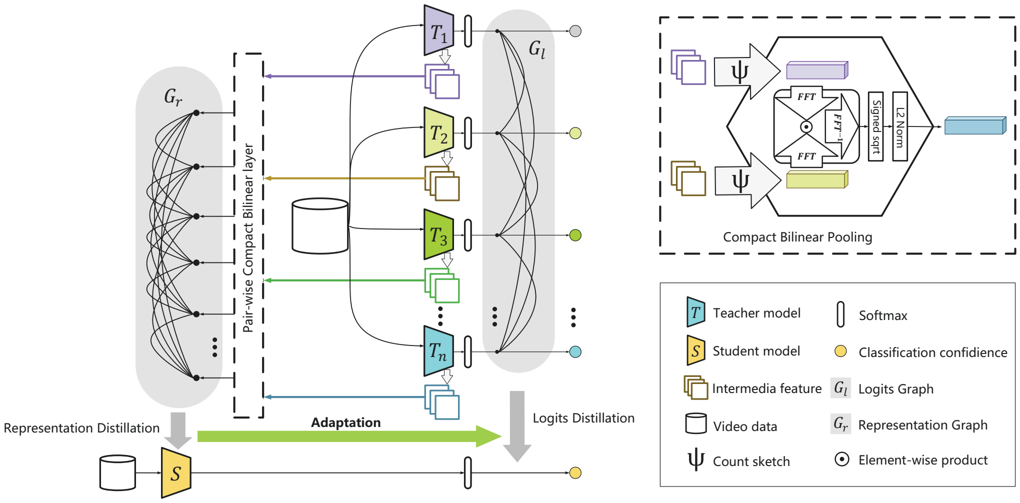


Figure 1: Architecture of the proposed framework.

We give the insights about the complementarity among these tasks as follows: (1) Self-supervised models can be mathematically divided into generative (e.g., $\pi_{\mathcal{P}}$) and discriminative (e.g., other three tasks) categories, and features extracted by them represent two complementary aspects of video understanding, like the imagination and judgment aspects in human understanding. (2) $\pi_{\mathcal{S}}$ and $\pi_{\mathcal{P}}$ force machine to capture temporal information in video, while $\pi_{\mathcal{M}}$ and $\pi_{\mathcal{T}}$ focus on identifying visual elements or their parts, in which spatial feature is more useful. (3) Video contains both local and holistic feature that conveyed by short- and long-term clips, respectively. In our case, $\pi_{\mathcal{M}}$ and $\pi_{\mathcal{T}}$ are adept at utilizing local information, while $\pi_{\mathcal{S}}$ and $\pi_{\mathcal{P}}$ are committed to capture holistic feature in video.

4 Graph Distillation Framework

The proposed framework is illustrated in Figure 1. We distill knowledge of self-supervised teacher models from two perspectives, namely soft probability distribution and internal representation. In this section, we first formalize the problem, then elaborate two components of the proposed approach.

4.1 Problem Formalization

To formalize the problem, assume $\mathcal{D}_{tr} = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}_{tr}|}$ and $\mathcal{D}_{te} = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}_{te}|}$ denote the training set with $|\mathcal{D}_{tr}|$ samples and the test set with $|\mathcal{D}_{te}|$ samples, respectively. $x_i \in \mathbb{R}^d$ is the video clip and $y_i \in [1, c]$ represents the label for c-class video classification problem. $\Pi = \{\pi_{\mathcal{S}}, \pi_{\mathcal{M}}, \pi_{\mathcal{T}}, \pi_{\mathcal{P}}\}$ is the set of self-supervised tasks in section 3 and $\mathcal{F}_t = \{f_t^{(i)}\}_{i=1}^{|\mathcal{F}_t|}$ is the set of teacher functions learned from Π. Our goal is learning a lighter student function $f_s^\star$ with aid of $\mathcal{F}_t$ under the principle of empirical risk minimization (ERM):

$$f_s^\star = \arg\min_{f_s \in \mathcal{F}_s} \frac{1}{|\mathcal{D}_{te}|} \sum_{i=1}^{|\mathcal{D}_{te}|} \mathbb{E}\big[\ell(f_s(x_i), y_i)\big] + \Omega(\| f_s \|) \tag{1}$$

where $\mathcal{F}_s$ is the function class of $f_s : \mathbb{R}^d \to [1, c]$, ℓ is the loss function and Ω is the regularizer. Notably, f_t and f_s in this paper are deep convolutional neural networks unless otherwise specified (see section 5.1).

4.2 Logits Graph Distillation

In vanilla knowledge distillation, the loss in Eq. (1) is composed with two parts:

$$\ell(f_s(x_i), y_i) = (1 - \lambda)\ell_h(f_s(x_i), y_i) + \lambda \ell_s(f_s(x_i), q_i) \tag{2}$$

where ℓ_h denotes the loss stems from true one-hot labels (i.e., hard targets) and ℓ_s denotes the imitation loss that comes from softened predictions (i.e., soft targets), $\lambda \in (0, 1)$ is the hyperparameter which balances ℓ_h and ℓ_s. In practice, ℓ_h is typically the cross entropy loss:

$$\ell_h(f_s(x_i), y_i) = \sum_{k=1}^{c} \mathbb{I}(k = y_i) \log \sigma(f_s(x_i), y_i) \tag{3}$$

where $\mathbb{I}$ is the indicator function and σ is the softmax operation $\sigma(z_i) = \frac{\exp z_i}{\sum_{k=1}^{c} \exp z_k}$. Softened prediction q_i in imitation loss of Eq. (2) is defined as $q_i = \sigma\big(f_t(x_i)/T\big)$ in which $f_t(x_i)$ is the class-probability prediction on x_i produced by teacher f_t, T is the temperature and a higher value for T produces a softer probability distribution over the classes.

Our logits graph G_l extends unidirectional distillation above to distilling logits knowledge from multiple self-supervised tasks dynamically. G_l performs as a directed graph in which each vertex v_m represents a self-supervised teacher and the edge $e_{n\to m} \in [0, 1]$ denotes the information weight from v_n to v_m. Given the video sample x_i, the corresponding total imitation loss in logits graph distillation is calculated by

$$\ell_s(x_i, \mathcal{F}_t^i) = \sum_{v_m \in U} \sum_{v_n \in \mathcal{V}(v_m)} e_{n\to m} \cdot \ell_{n\to m}^{logit}(x_i, \mathcal{F}_t^i) \tag{4}$$

where $\ell_{n\to m}^{logit}$ means the distillation loss flowing on edge $e_{n\to m}$, U is the universe set that contains all vertices of G_l and $\mathcal{V}(v_m)$ is the set of vertices that point to v_m. $\mathcal{F}_t^i$ represents the logits output of teachers on sample x_i. As depicted in Figure 2, we treat distillation in G_l as a multi-distribution matching problem, where teachers dynamically adjust their logits distribution according to complementary gain (CG) received from their neighbors, and CG is based on the imitation feedback of student when it distills knowledge from different

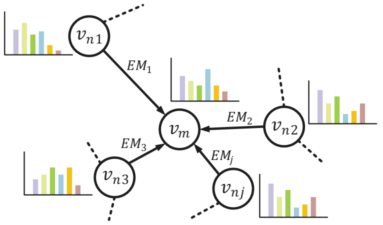


Figure 2: Illustration of logits graph distillation. EM denotes Earth Mover distance and the length of edges represents the information weight between adjacent vertices.

teachers. Take single teacher-student pair for example, we use Earth Mover (EM) distance to measure $\ell^{logit}(x_i, f_t^{(k)})$ as:

$$\ell^{logit}(x_i, f_t^{(k)}) = \inf_{\gamma \sim \Phi(\mathbb{P}_t^{(k)}, \mathbb{P}_s)} \mathbb{E}_{(\mu,\eta)\sim\gamma}\big(\| \mu(x_i) - \eta(x_i) \|\big) \quad (5)$$

$$\mu(x_i) = \sigma(f_t^{(k)}(x_i)/T_k), \eta(x_i) = \sigma(f_s(x_i)) \quad (6)$$

where $\mathbb{P}_t^{(k)}$ and $\mathbb{P}_s$ denote the space of probability belong to $f_t^{(k)}$ and f_s, respectively. $\Phi(\mathbb{P}_t^{(k)}, \mathbb{P}_s)$ represents the set of all joint distributions of $\gamma(\mu, \eta)$ whose marginals are respectively $\mathbb{P}_t^{(k)}$ and $\mathbb{P}_s$. $\mu(x_i)$ distills softened logits from $f_t^{(k)}$ with temperature T_k, where k means applying different temperature for different teachers. $\eta(x_i)$ is output probability of the student network. We implement G_l using adjacency matrix where element $G_l[m][n]$ equals to $e_{m\to n}$, and distillation is trained by minimizing the sum of Eq. (3) and Eq. (4) over all training samples iteratively.

The closest work to our G_l is multi-modal graph distillation (MMGD) [Luo *et al.*, 2017] which distills knowledge from multi-model data, and there are mainly two differences between them. (1) We formalize logits distillation from multiple teachers as a multi-distribution joint matching problem. Instead of cosine distance, we apply EM distance to measure the logits distribution discrepancy between teacher and student. Mathematically, teacher models are devoted to mapping visual characteristics of various video samples in logits distribution supported by low dimensional manifolds, where EM distance is proved to be more useful [Arjovsky *et al.*, 2017]. Besides, EM distance can describe intra-class and inter-class variability more adequately than cosine distance, resulting in more refined knowledge distillation to capture subtle complementarity of teachers. (2) Moreover, we propose a novel feature distillation graph G_r to assist this distribution matching process (section 4.3) and experimental results indicate that our G_l can benefit much more from G_r than MMGD method.

4.3 Representation Graph Distillation

Besides logits distribution, we hypothesize that intermediate feature of different teacher models can also be leveraged via graph-based distillation. The core difficulty stems from the heterogeneous nature of these task-specific features. To tackle this challenge, we propose to pairwise ensemble the original features via compact bilinear pooling [Gao *et al.*, 2016] and adopt the bilinear features as vertices of representation distillation graph G_r, which is the second component in Fig 1. There exist three insights behind this idea: (1) It is reasonable that the internal features of ConvNets reflect certain activation and attention patterns of neurons, and bilinear pooling allows all elements of heterogeneous feature vectors to interact with each other in a multiplicative fashion, which captures the salient information with more complementarity. (2) In the view of feature domain adaptation, bilinear pooling maps original features to a smoother representation space, where their distributions are more homologous and discrepancy among them are easier to measure. (3) Furthermore, since the high level prediction distribution of models is based on their internal representation, aligning at feature level also performs as a powerful adaptation method which assists distillation learning in G_l.

Assume $\mathcal{R} \in \mathbb{R}^{|C\times S|}$ denotes the feature vector with C channels and spatial size S, where $S = H \times W$. Each vertex V_k of G_r is yield by invoking compact bilinear pooling on feature pairs $(\mathcal{R}_m, \mathcal{R}_n)$:

$$\begin{aligned}\Psi(\mathcal{R}_m \otimes \mathcal{R}_n, h, s) &= \Psi(\mathcal{R}_m, h, s) * \Psi(\mathcal{R}_n, h, s) \\ &= FFT^{-1}(FFT(\Psi_m) \odot FFT(\Psi_n))\end{aligned} \quad (7)$$

where $\Psi : \mathbb{R}^{|C\times S|} \to \mathbb{R}^e$ represents Count Sketch projection function [Charikar *et al.*, 2002] and $e \ll |C \times S|$ as Ψ projects the outer product to a lower dimensional space. $\otimes$ and $\odot$ are outer product and element-wise product, respectively. $h \in \{1, \cdots, e\}^{|C\times S|}$ and $s \in \{-1, 1\}^{|C\times S|}$ are two randomly initialized vectors for invocation of Ψ. $*$ means convolution and Eq. (7) utilizes the theorem that element-wise product in the frequency domain equals to convolution in time domain, which can yield a more compact feature vector. Then the bilinear feature Ψ is passed through a signed squareroot ($z = sign(\Psi)\sqrt{|\Psi|}$) and ℓ_2 normalization ($V_k = z/ \| z \|_2$), where $\forall k \in \{1, \mathcal{C}^2_{|\Pi|}\}$ and combinatorics $\mathcal{C}^2_{|\Pi|}$ is the vertex number of G_r. For instance, there will be 6 vertices in G_r after pairwise ensemble of 4 teacher features for each video sample.

Similar to G_l, G_r is defined as a adjacency matrix but the edge of G_r is a vector $E \in \mathbb{R}^{1\times b}$ rather than a scalar, where b is the dimension of vertex V. We distill the bilinear feature with temperature T_k as softened representation distribution:

$$\mathcal{D}_k^{soft}(x_i) = \sigma(V_k(x_i)/T_k) \quad (8)$$

and use Maximum Mean Discrepancy (MMD) to meause the distillation loss:

$$\begin{aligned}\ell_r(\mathcal{R}_s, \mathcal{D}_k^{soft}) = & \frac{1}{C_s^2}\sum_{p=1}^{C_s}\sum_{p'=1}^{C_s}\mathcal{K}\big(\sigma(\mathcal{R}_s^{(p)}), \sigma(\mathcal{R}_s^{(p')})\big) \\ & + \frac{1}{C_k^2}\sum_{q=1}^{C_k}\sum_{q'=1}^{C_k}\mathcal{K}\big(\mathcal{D}_k^{(q)}, \mathcal{D}_k^{(q')}\big) \\ & - \frac{2}{C_s C_k}\sum_{p=1}^{C_s}\sum_{q=1}^{C_k}\mathcal{K}\big(\sigma(\mathcal{R}_s^{(p)}), \mathcal{D}_k^{(q)}\big)\end{aligned} \quad (9)$$

where $\mathcal{R}_s$ denotes the feature map in a certain layer of student model and $\mathcal{R}_s^{(p)}$ is the feature vector at channel p. $\mathcal{D}_k$ is the softened distribution of k-th vertex in G_r, C_s and C_k are channel numbers of student model and $\mathcal{D}_k$, respectively.

$\mathcal{K}$ is kernel function which maps features to higher dimensional Reproducing Kernel Hilbert Space (RKHS), and we use Gaussian Kernel: $\mathcal{K}(\boldsymbol{a}, \boldsymbol{b}) = \exp(-\frac{\|\boldsymbol{a}-\boldsymbol{b}\|_2^2}{2\sigma^2})$ in practice. Note that, softmax operation in Eq. (9) ensures each sample has the same scale and it is unnecessary for $\mathcal{D}_k$ as Eq. (8) have already performed it.

Let $\mathcal{M}_{\ell_r}$ is the matrix whose element is ℓ_r and $\mathcal{T} \in \mathbb{R}^{1\times \mathcal{C}^2_{|\Pi|}}$ is the temperature vector, the total loss of G_r is calculated by

$$\mathcal{L}_r = \sum \sigma(G_r/\mathcal{T}) \odot \mathcal{M}_{\ell_r} \tag{10}$$

The reasons why MMD is used in G_r lie in two aspects: (1) Feature distillation matches activation distribution of teachers, where MMD can be used to measure the distribution discrepancy between teachers and student. In contrast, matching instance-level features directly (without MMD) is hard to make sense as it ignores the statistical dependence of different instances. (2) Internal activations describe different salient patterns of video and MMD loss acts as a regularizer to encourage knowledge transfer in feature selectivity.

5 Experiments

In this work we use conventional transfer learning paradigm for evaluating the utility of student model on video classification. We first train the teacher models for privileged information learning on auxiliary tasks, then transfer the knowledge from teachers to student via the proposed graph distillation framework. Experimental details are described in the sequel.

5.1 Architectures

We first re-implement three self-supervised methods in the literature based on VGG-19 [Simonyan and Zisserman, 2015], for training teacher model from $\pi_\mathcal{S}$, $\pi_\mathcal{M}$ and $\pi_\mathcal{T}$, respectively. Then we introduce PredNet [Lotter *et al.*, 2017] as a self-supervised model for learning from $\pi_\mathcal{P}$. For notation convenience, we use $T_\mathcal{S}$, $T_\mathcal{M}$, $T_\mathcal{T}$, and $T_\mathcal{P}$ to denote the models for these tasks. All models are customized with a slight modification for meeting the demand in our experiments. In particular, we respectively extend channels to 96 and 128 for `conv1` and `conv2` in VGG-19, and change the filter size of `conv1` to 11×11 as better performance have shown in practice. For $\pi_\mathcal{S}$, $\pi_\mathcal{M}$ and $\pi_\mathcal{T}$, siamese-style architectures are conducted for pairwise feature extraction and base visual encoders share parameters. More specifically, $T_\mathcal{S}$ concatenates convolutional feature over pairs of all frames to be sorted in first `fc` layer, followed by a classical `fc` classifier. $T_\mathcal{M}$ is composed with a base-CNN (BCNN) and a top-CNN (TCNN), in which TCNN takes the output pair of BCNN for transform analysis between two frames. $T_\mathcal{T}$ is a siamese-triplet network in order to judge the similarity of patches in a triple. $T_\mathcal{P}$ is a ConvLSTM model where each layer consists of representation module (R), target module (A), prediction module ($\hat{A}$) and error term (E). We conduct a 4-layers version of it and use the output of R in top layer as representation. For the sake of model compression and acceleration, we choose AlexNet [Krizhevsky *et al.*, 2012] as student model S for video classification task.

5.2 Implementation Details

All models are implemented using PyTorch[1]. For the self-supervised model training, we basically follow the settings in original papers and we encourage the reader to check them for more details. For the graph distillation training, we train the model for 350 epochs using Adam optimizer with momentum 0.9. The batch size is set to 128 and we initialize the learning rate to 0.01 and decay it every 50 epochs by a factor of 0.5. Trade-off hyperparameter λ is set to 0.6.

Another vital issue in practice is how to choose the layer of different networks for feature distillation in G_r. In other words, we expect to quantify the transferability of features from each layer of teacher models. Inspired by [Yosinski *et al.*, 2014], we measure 1) the specialization of layer neurons to original self-supervised tasks at the expense of performance on the classification task and 2) the optimization difficulties related to splitting teacher networks between co-adapted neurons, and we finally choose `conv4` of both teachers (VGG-19 and PredNet) and student (AlexNet) as the distillation layer, whose feature maps have 256 output channels.

5.3 Datasets

Datasets in our experiments contain two branches, i.e., auxiliary datasets for self-supervised learning and target datasets for video classification. For the former, we use the same datasets as original works. For the latter, we evaluate our framework on three popular datasets, namely UCF101 [Soomro *et al.*, 2012], HMDB51 [Kuehne *et al.*, 2013] and CCV [Jiang *et al.*, 2011], which contain 13k, 6.7k and 9.3k videos with 101, 51 and 20 categories, respectively.

5.4 Results

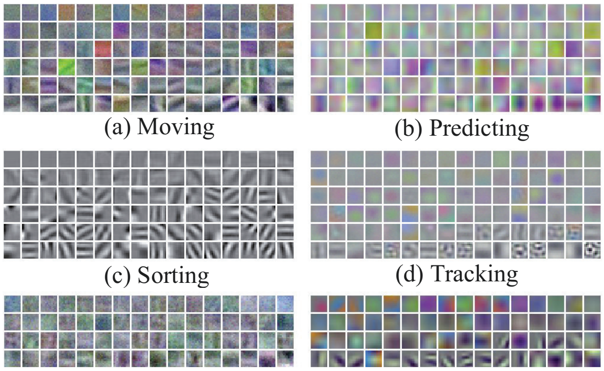
Table 1 compares our framework to several state-of-the-art unsupervised methods for video classification. Obviously, our approach outperforms the previous methods by a large margin, and more importantly, it beats its ImageNet-supervised counterpart ($AlexNet_{pre}$) for the first time.

The results indicate that (1) knowledge learned from different self-supervised tasks is complementary and it can be transferred via graph distillation to form a more comprehensive video semantics. Distillation encourages each teacher to

Methods	UCF101	HMDB51	CCV
$AlexNet_{ran}$	47.8	16.3	39.2
$AlexNet_{pre}$	66.9	28.0	60.8
Wang *et al.* [2015]	40.7	15.6	34.3
Misra *et al.* [2016]	50.9	19.8	44.7
Senthil *et al.* [2016]	55.4	23.6	46.4
Lee *et al.* [2017]	56.3	22.1	48.8
Basura *et al.* [2017]	60.3	32.5	58.1
Ours (G_l+G_r)	**68.9**	**35.1**	**62.6**

Table 1: Comparing with state-of-the-art unsupervised methods for video classification. Average % accuracy for UCF101 and HMDB51, mean average precision for CCV. $AlexNet_{ran}$ means AlexNet trained from scratch with randomly initialized weights.

[1] http://pytorch.org/



(a) Moving (b) Predicting

(c) Sorting (d) Tracking

(e) MMGD (f) Ours

Figure 3: Visualization of `conv1` filters learned from self-supervised tasks and graph distillation. Filters in (a), (b), (c) and (d) are learned from $\pi_\mathcal{M}$, $\pi_\mathcal{P}$, $\pi_\mathcal{S}$ and $\pi_\mathcal{T}$, respectively. The filters of $\pi_\mathcal{S}$ are in grayscale since we use channel splitting for better performance. (e) and (f) show the filters of student trained on UCF101 via MMGD and our proposal, they adopt AlexNet as base model.

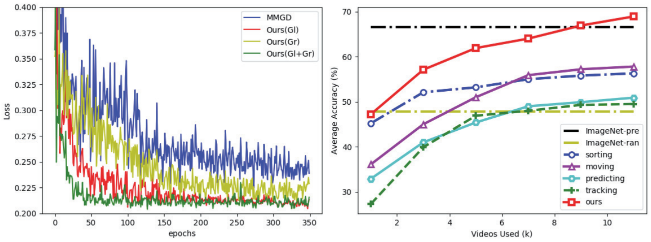
agree with the predictions and patterns other teachers would have learned, which improves the robustness of the student model on video classification. (2) Knowledge from models which trained on ImageNet in a supervised fashion is insufficient for capturing all the details of video, leading to suboptimal performance on classification task.

Representation learned by Student

Qualitatively, we demonstrate the quality of the features learned by student model through visualizing its low-level first layer filters (`conv1`). As exhibited in Figure 3, there are both color detectors and edge detectors in the `conv1` filters of our distilled model. Its filters are sharper and of more varieties than its counterpart learned from MMGD, and tend to be fully utilized when compared to those of the teachers.

5.5 Ablation Studies

We conduct comparison with several baseline methods to evaluate the effectiveness of each component in our proposal. After fine-tuning the self-supervised teacher models for video classification individually, we fuse their confidence scores (model fusion). We also feed the spliced feature of them to a linear classifier (model ensemble). For the conventional knowledge distillation (KD) baseline, we average the confidence scores of teacher models that equipped with KD. G_l and G_r are respectively compared with MMGD and single feature distillation (uniform version of feature distillation). Moreover, in order to verify the mutual promotion between G_l and G_r, we further compare the performance of two combination, i.e., MMGD+G_r and G_l+G_r (ours).



(a) Training process (b) The impact of data volume

Figure 4: Illustration of training process and the impact of video data volume for the classification performance on UCF101.

Methods	UCF101	HMDB51	CCV
Models fusion	57.9	19.4	45.2
Models ensemble	58.3	23.5	47.1
KD (uniform)	60.1	24.3	48.1
Logits graph (MMGD)	62.5	27.2	49.7
Logits graph (G_l)	**65.3**	**30.9**	**52.6**
Feature distill (uniform)	61.7	28.4	56.3
Feature graph distill (G_r)	**66.4**	**33.4**	**58.0**
MMGD+G_r	67.0	34.2	58.9
G_l+G_r	**68.9**	**35.1**	**62.6**

Table 2: Comparison with baseline methods on three vdieo datasets.

As shown in Table 2, graph-based distillation methods outperform the ordinary solutions that in the first three lines. The comparison between row 4 and row 5 suggests that our formalization of logits graph distillation is more effective than MMGD. The comparison between row 6 and row 7 reveals G_r helps model to learn the complementarity of different self-supervised tasks. Results in the last two lines verify that G_l benefits much more from G_r than MMGD since G_r performs as a powerful adaptation method to assist the probability distribution matching. Moreover, the training process of distillation shown in Figure 4 (a) suggests that G_r can boost G_l not only in accuracy but also in training efficiency, which further verify the effectiveness of the two components.

5.6 Comparison with Supervised Methods

Although our approach has not yet yielded video representations as powerful as those learned by supervised methods, it shows advance of integrating complementary video semantics from different self-supervised models. One of the most significant advantages of our approach is that it can utilize massive video data on the web economically, and as displayed in Figure 4 (b), the performance of distilled model increases steadily as the training data for self-supervised tasks grows. We believe learning video representation in such *divide and conquer* manner is potential as it integrates complementary semantic from various models at expert level. Moreover, attributed to the knowledge distillation, our framework dramatically reduces the number of model parameters and inference time, since top performing models for video classification has a much higher computational and memory cost than student.

6 Conclusion

In this paper, we have proposed a graph-based distillation framework for representation learning in video classification. Our framework distills both logits knowledge and internal feature knowledge from teacher models, which utilizes the complementarity of different video semantics from multiple self-supervised tasks. We view logits distillation as a multi-distribution joint matching problem, and pairwise ensemble features via compact bilinear pooling for feature distillation.

Experiments on 3 video datasets verify that our approach can learn better video representation with less capacity.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grant 61771025 and Grant 61532005.

References

[Agrawal *et al.*, 2015] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. Learning to see by moving. In *ICCV*, pages 37–45, 2015.

[Arjovsky *et al.*, 2017] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[Charikar *et al.*, 2002] Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. *ALP*, pages 784–784, 2002.

[Chen *et al.*, 2017] Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Darkrank: Accelerating deep metric learning via cross sample similarities transfer. *arXiv preprint arXiv:1707.01220*, 2017.

[Fernando *et al.*, 2017] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *CVPR*, pages 5729–5738, 2017.

[Gao *et al.*, 2016] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *CVPR*, pages 317–326, 2016.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Workshop*, 2015.

[Jayaraman and Grauman, 2016] Dinesh Jayaraman and Kristen Grauman. Slow and steady feature analysis: higher order temporal coherence in video. In *CVPR*, pages 3852–3861, 2016.

[Jiang *et al.*, 2011] Yu-Gang Jiang, Guangnan Ye, Shih-Fu Chang, Daniel Ellis, and Alexander C Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *ICMR*, page 29, 2011.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[Kuehne *et al.*, 2013] Hilde Kuehne, Hueihan Jhuang, Rainer Stiefelhagen, and Thomas Serre. Hmdb51: A large video database for human motion recognition. In *HPCSE*, pages 571–582. 2013.

[Lee *et al.*, 2017] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *ICCV*, pages 667–676, 2017.

[Lopez-Paz *et al.*, 2016] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *ICLR*, 2016.

[Lotter *et al.*, 2017] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. In *ICLR*, 2017.

[Luo *et al.*, 2017] Zelun Luo, Lu Jiang, Jun-Ting Hsieh, Juan Carlos Niebles, and Li Fei-Fei. Graph distillation for action detection with privileged information. *arXiv preprint arXiv:1712.00108*, 2017.

[Misra *et al.*, 2016] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *ECCV*, pages 527–544, 2016.

[Noroozi and Favaro, 2016] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, pages 69–84, 2016.

[Pathak *et al.*, 2016] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544, 2016.

[Purushwalkam and Gupta, 2016] Senthil Purushwalkam and Abhinav Gupta. Pose from action: Unsupervised learning of pose features based on motion. In *ECCV*, 2016.

[Romero *et al.*, 2015] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2015.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[Soomro *et al.*, 2012] 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.

[Vapnik and Izmailov, 2015] Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: similarity control and knowledge transfer. 16(20232049):55, 2015.

[Wang and Gupta, 2015] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, pages 2794–2802, 2015.

[Yosinski *et al.*, 2014] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *NIPS*, pages 3320–3328, 2014.

[Zagoruyko and Komodakis, 2017] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017.

[Zhang *et al.*, 2017] Richard Zhang, Phillip Isola, and Alexei A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *CVPR*, pages 1057–1068, 2017.