ROMANIAN ACADEMY

School of Advanced Studies of the Romanian Academy

"Simion Stoilow" Institute of Mathematics

PhD Thesis Summary

Learning in Video using the Teacher-Student Paradigm:
From Unsupervised Object Detection to Text-Video Retrieval

Thesis advisor:

Prof. Dr. Marius Leordeanu

PhD Student:

Ioana Croitoru

2022

Contents

Abstract

Computer vision and natural language processing are two fields of high interest in artificial intelligence. The focus of this work is generalized distillation which is studied and applied on tasks related to computer vision and natural language processing. The tasks we consider are object segmentation, object detection and text-video retrieval. We proved that distillation can be used in various settings, such as supervised or unsupervised learning. For the unsupervised learning, we address the task of learning to detect and segment foreground objects in single images. We achieve our goal by training a student pathway, consisting of a deep neural network that learns to predict, from a single input image, the output of a teacher pathway that performs unsupervised object discovery in video. Further we proved that the performance can be boosted over several generations of students and teachers. For the supervised learning we considered the task of text-video retrieval. For this, we are the first to investigate the design of algorithms that learn from multiple pre-trained text emebeddings and propose a novel generalized distillation method, called TEACHTEXT, which leverages complementary cues from multiple text encoders to provide an enhanced supervisory signal to the retrieval model. Moreover, we extend our method to video side modalities and show that we can effectively reduce the number of used modalities at test time without compromising performance. Our approach advances the state of the art on several video retrieval benchmarks by a significant margin and adds no computational overhead at test time.

Chapter 1

Introduction

Computer vision is a track of Artificial Intelligence (AI) which process images, videos and other visual inputs in order to derive meaningful information. Information can mean anything from edge detection to object detection, object segmentation, tracking of an object and so on. Usually, the developed methods in Computer Vision (CV) aim to replicate the capability of human vision. There are a lot of areas of research and a lot of applications, such as: medical imaging [49], motion capture, automotive safety [18] etc.

Natural Language Processing (NLP) is, on the other hand, a field in artificial intelligence which focuses on deriving meaningful information from texts. While the purpose of computer vision is to enable computers to understand the visual input similar to the capability of human vision, the purpose of NLP is to enable computers to understand the spoken words the same way human beings can. Some of the areas of research in NLP are: machine translation [56], text summarization [32] and so on.

Needles to say, there is a lot of interest also in tasks where computer vision intersects with natural language processing, such as: text-video retrieval [11], video captioning [3], image captioning [58] and so on.

Nowadays, deep learning [17] is at the core of most of the state of the art computer vision methods, as well, as natural language processing methods. Since 2014, deep learning boosted the performance on various benchmarks on both fields. In order for a deep network to have competitive results, huge amounts of data is required. To achieve good results, the data should be labelled. But labeled data is hard to collect since it involves that several

persons should annotate the image, text or video. So, there is also a huge interest in the unsupervised, semi-supervised and self-supervised area of both computer vision and natural language processing. There are huge amounts of videos, text and images on the web, which made this area popular.

Moreover, in order to boost the performance of a deep network without requiring any additional labeled data, the idea of learning from other methods arises. Generalized distillation [33] is a framework that unifies knowledge distillation [20] and priviledge information [55], two methods that enable machines to learn from other machines. Generalized distillation involves a machine, which is usually called student, that is able to learn from another trained machine, usually called teacher. So, using generalized distillation recent works [10, 11] proved that the performance of a deep network student can be boosted by using the information provided from the teacher. Naturally, the combination of two or multiple different methods that are trained to solve the same task and are combined together into an ensemble proved to achieve higher scores than a single model [2].

1.1 Contributions

In this work we plan to study the behaviour of learning through generalized distillation from an ensemble of teachers, both in the unsupervised scenario and supervised scenario. The tasks we used for generalized distillation in the unsupervised scenario is object segmentation and object detection, while the task in the supervised case is text-video retrieval. In this way, we prove the effectiveness of learning from an ensemble of teachers in several scenarios: in the unsupervised case and in the supervised case as well as in a computer vision application and in an application that is at the intersection of computer vision and natural language processing, namely text-video retrieval.

Object segmentation. Object detection is the task were you need to localize and classify each object in an image or video. Object segmentation is the task were each pixel in an image should be labeled with the corresponding class.

For object segmentation we focused on unsupervised learning from visual data since it has an immense practical value, as huge quantities of unlabeled videos can be collected at low cost. We achieve our goal by training a student pathway, consisting of a deep neural network that

learns to predict, from a single input image, the output of a teacher pathway that performs unsupervised object discovery in video. This has a dual benefit: firstly, it allows, in principle, unlimited generalization possibilities during training, while remaining fast at testing. Secondly, the student not only becomes able to detect in single images significantly better than its unsupervised video discovery teacher, but it also achieves state of the art results on two current benchmarks, YouTube Objects and Object Discovery datasets. At test time, our system is two orders of magnitude faster than other previous methods.

Furthermore, we improved the proposed system by designing it to learn over several generations of teachers and students. At every generation the teacher performs unsupervised object discovery in videos or collections of images and an automatic selection module picks up good frame segmentations and passes them to the student pathway for training. At every generation multiple students are trained, with different deep network architectures to ensure a better diversity. The students at one iteration help in training a better selection module, forming together a more powerful teacher pathway at the next iteration.

Text-video retrieval. In order to prove the effectiveness of generalized distillation we also considered a supervised scenario for a very popular task which is at the intersection of computer vision and natural language processing, naming text-video retrieval. Given a natural language sentence and a collection of videos, the goal is to design a system that is able to retrieve the video that is best described by the query. In recent years, considerable progress on the task of text-video retrieval has been achieved by leveraging large-scale pretraining on visual and audio datasets to construct powerful video encoders. By contrast, despite the natural symmetry, the design of effective algorithms for exploiting large-scale language pretraining remains under-explored. In this work, we are the first to investigate the design of such algorithms and propose a novel generalized distillation method, TEACHTEXT, which leverages complementary cues from multiple text encoders to provide an enhanced supervisory signal to the retrieval model.

Chapter 2

Unsupervised learning from video to detect foreground objects in single images

Unsupervised learning from visual data is one of the most difficult challenges in computer vision. It is essential for understanding how visual recognition works. Learning from unsupervised input has an immense practical value, as huge quantities of unlabeled videos can be collected at low cost. Here we address the task of unsupervised learning to detect and segment foreground objects in single images. We achieve our goal by training a student pathway, consisting of a deep neural network that learns to predict, from a single input image, the output of a teacher pathway that performs unsupervised object discovery in video. Our approach is different from the published methods that perform unsupervised discovery in videos or in collections of images at test time. We move the unsupervised discovery phase during the training stage, while at test time we apply the standard feedforward processing along the student pathway. This has a dual benefit: firstly, it allows, in principle, unlimited generalization possibilities during training, while remaining fast at testing. Secondly, the student not only becomes able to detect in single images significantly better than its unsupervised video discovery teacher, but it also achieves state of the art results on two current benchmarks, YouTube Objects and Object Discovery datasets. At test time, our system is two orders of magnitude faster than other previous methods.

This chapter is based on the paper "Unsupervised learn-ing from video to detect foreground objects in single images." Croitoru, Ioana, Simion-Vlad Bogolin, and Marius Leordeanu. International Conference on Computer Vision. 2017.

2.1 Introduction

Unsupervised learning is one of the most difficult and intriguing problems in computer vision and machine learning today. Researchers believe that unsupervised learning from video could help decode hard questions regarding the nature of intelligence and learning. As unlabeled videos are easy to collect at low cost, solving this task would bring a great practical value in vision and robotics.

Our system is presented in Figure 1. We have an unsupervised training stage, in which a student deep neural network learns frame by frame from an unsupervised teacher, which performs object segmentation in videos, to produce similar object masks in single images. The teacher method takes advantage of the consistency in appearance, shape and motion manifested by objects in video. In this way, it discovers objects in the video and produces a foreground segmentation for each individual frame. Then, the student network tries to imitate for each frame the output of the teacher, while having as input only a single image - the current frame. The teacher pathway is much simpler in structure, but it has access to information over time. In contrast, the student is much deeper in structure, but has access only to one image. Thus, the information discovered by the teacher in time is captured by the student in depth, over neural layers of abstraction. In experiments, we show a very encouraging fact: the student easily learns to outperform its teacher and discovers by itself general knowledge about the shape and appearance properties of objects, well beyond the abilities of the teacher. Since there are available methods for video discovery with good performance, the training task becomes immediately feasible. In this work we chose the VideoPCA algorithm introduced as part of the system in [52] because it is very fast (50-100 fps), uses very simple features (pixel colors) and it is unsupervised - with no usage of supervised pre-trained features.

2.2 Scientific context

Recent unsupervised methods follow two directions. One is to learn powerful features in an unsupervised way and then use them in a classic supervised learning scheme in combination with different classifiers, such as SVMs or CNNs [41]. In very recent work [39], developed

independently from ours, a deep network learns, from an unsupervised system using motion cues in video, image features that are applied to several transfer learning tasks. The second approach to unsupervised learning is to discover, at test time, common patterns in unlabeled data using clustering or data mining formulations [21]. Unsupervised learning in video is also related to co-segmentation [22] and weakly supervised localization [12]. Earlier methods are based on local feature matching and detection of their co-occurrence patterns [52], while more recent ones [24] discover object tubes by linking candidate bounding boxes between frames with or without refining their location. Traditionally, the task of unsupervised learning from image sequences, has been formulated as a feature matching or data clustering optimization problem, which is computationally very expensive.

2.3 System architecture

In our experiments, the student indeed outperforms its teacher. Moreover, it achieves state of the art results on two different benchmarks. The success of this unsupervised learning paradigm is due to the fact that the student is forced to capture from appearance only visual features that are good predictors for the presence of objects. The overview of our system is presented in Figure 1.

Teacher path: unsupervised discovery in video. We used the VideoPCA algorithm, which is a part of the whole system introduced in [52]. It runs at $50-100$ fps and at this speed we can produce one million unsupervised soft segmentations in a reasonable time.

Student path: single-image segmentation. It consists of a deep convolutional network, with ten layers (seven convolutional, two pooling and one fully connected layer). We treat foreground object segmentation as a regression problem, where the soft mask given by the unsupervised video segmentation system acts as the desired output.

Unsupervised soft masks selection. We used a simple measure of masks quality based on the following observation: when masks are close to the ground truth, the mean of their nonzero values is usually high. Thus, when the discoverer is confident is more likely to be right. The mean value of non-zero pixels in the soft mask is then used as a score indicator for each segmented frame.

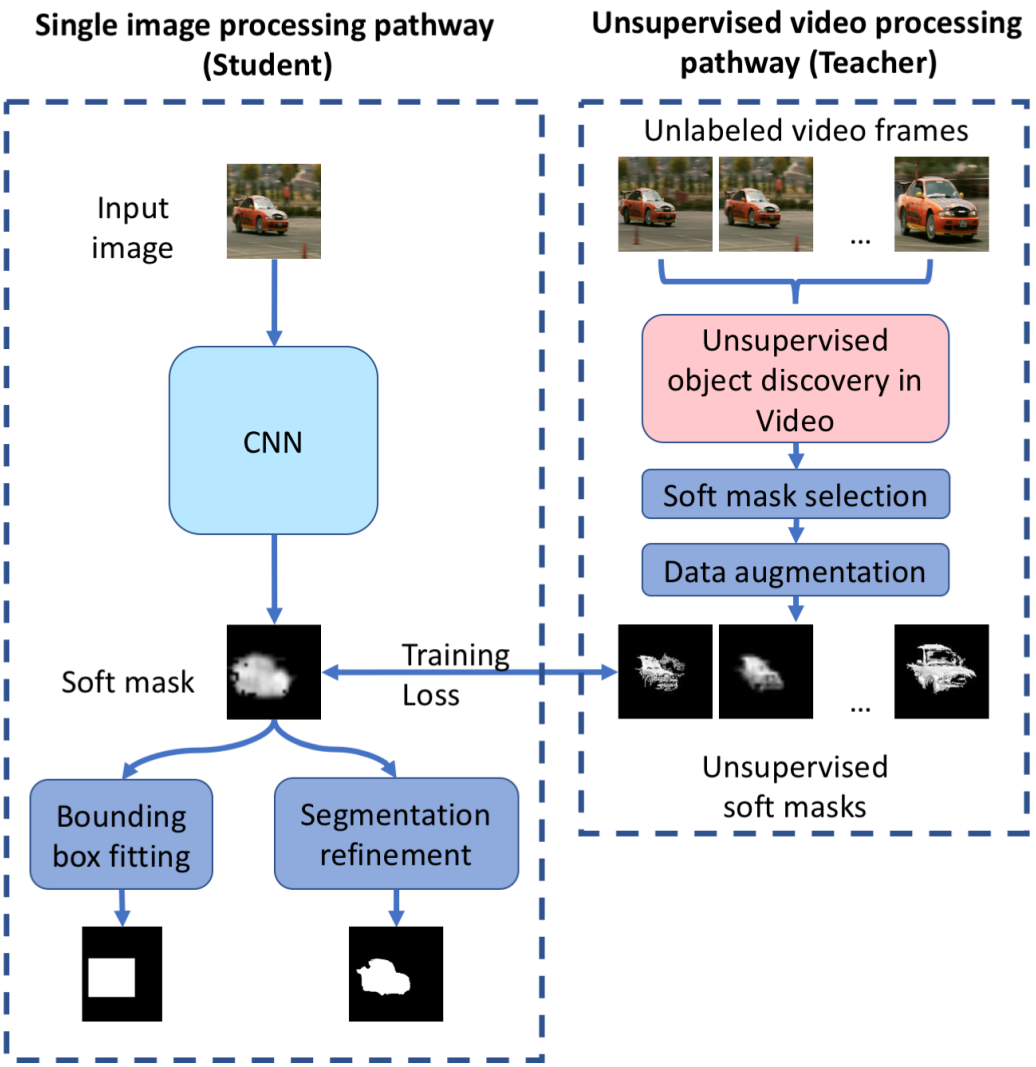

Figure 1: **The dual student-teacher system proposed for unsupervised learning to detect foreground objects in images.** It has two pathways: the teacher, on the right, discovers in an unsupervised fashion foreground objects in video. It outputs soft masks for each frame. The resulting masks, are then filtered based on a simple and effective unsupervised quality metric. The set of selected segmentations is then augmented in a relatively simple manner, automatically. The resulting final set of pairs - input image (a video frame) and soft mask (the mask for that particular frame which acts as an unsupervised label) - are used for training the student CNN pathway.

	Airplane		Car		Horse	
	P	J	P	J	P	J
[27]	80.20	7.90	68.85	0.04	75.12	6.43
[22]	49.25	15.36	58.70	37.15	63.84	30.16
[23]	47.48	11.72	59.20	35.15	64.22	29.53
[47]	88.04	55.81	85.38	64.42	82.81	51.65
[8]	90.25	40.33	**87.65**	64.86	86.16	33.39
Ours_1	90.92	**62.76**	85.15	66.39	**87.11**	54.59
Ours_2	**91.41**	61.37	86.59	**70.52**	87.07	**55.09**

Table 1: **Results on the Object Discovery in Internet images [47] dataset (P, J metric).** Ours_1 represents our network trained using the VID dataset (with 10% selection), while Ours_2 represents our network trained on VID and YTO datasets (with 10% selection). We observe that Ours_2 has better results with mean P of **88.36** and mean J of **62.33** compared to Ours_1 (mean P: 87.73, mean J: 61.25).

2.4 Experimental analysis

Comparisons with other methods. We compare our unsupervised system with state of the art methods designed for the task of object discovery in collections of images, that might contain one or a few main object categories of interest. A representative current benchmark in this sense is the Object Discovery in Internet Images dataset. Different from other methods, we do not need a collection of images during testing, since each image is processed independently by our system, at test time. Therefore, our performance is not affected by the structure of the image collection or the number of classes of interest being present in the collection.

We tested our system on the task of fine foreground object segmentation and compared to the best performers in the literature on the Object Discovery dataset in Table 1. We evaluate based on the same P, J evaluation metric as described by Rubinstein *et al.* [47] - the higher P and J, the better. P refers to the per pixel precision, while J is the Jaccard similarity (the intersection over union of the result and ground truth segmentations). In Figure 2 we present some qualitative samples.

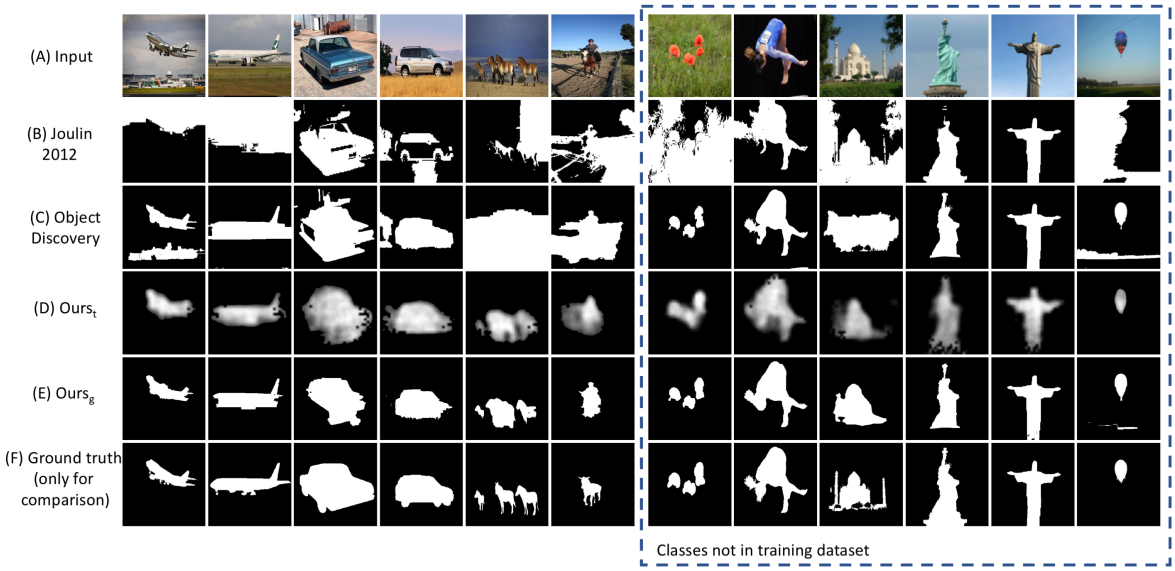

Figure 2: **Visual results on the Object Discovery dataset.** A: input image, B: segmentation obtained by [23], C: segmentation obtained by [47], D: thresholded soft mask produced by our network, E: segmentation mask produced after refining the soft output of our network with GrabCut [46], F: ground truth segmentation. More details and results: https://sites.google.com/view/unsupervisedlearningfromvideo.

2.5 Conclusions

We have shown in extensive experiments that it is possible to use a relatively simple method for unsupervised object discovery in video to train a powerful deep neural network for detection and segmentation of objects in single images. The result is interesting and encouraging and shows how a system could learn, in an unsupervised fashion, general visual characteristics that predict well the presence and shape of objects in images. The network essentially discovers appearance object features from single images, at different levels of abstraction, that are strongly correlated with the spatiotemporal consistency of objects in video.

The student network, during the unsupervised training phase, is thus able to learn general "objectness" characteristics that are well beyond the capabilities of its teacher. These characteristics include good form, closure, smooth contours, as well as contrast with its background. What the simpler teacher discovers over time, the deep, complex student is able to learn across several layers of image features at different levels of abstraction. Therefore, our unsupervised learning model, tested in extensive experiments, brings a valuable contribution to the unsupervised learning problem in vision research.

Chapter 3

Unsupervised learning of foreground object segmentation

Unsupervised learning represents one of the most interesting challenges in computer vision today. The task has an immense practical value with many applications in artificial intelligence and emerging technologies, as large quantities of unlabeled images and videos can be collected at low cost. In this paper, we address the unsupervised learning problem in the context of segmenting the main foreground objects in single images. We propose an unsupervised learning system, which has two pathways, the teacher and the student, respectively. The system is designed to learn over several generations of teachers and students. At every generation the teacher performs unsupervised object discovery in videos or collections of images and an automatic selection module picks up good frame segmentations and passes them to the student pathway for training. At every generation multiple students are trained, with different deep network architectures to ensure a better diversity. The students at one iteration help in training a better selection module, forming together a more powerful teacher pathway at the next iteration. Our method achieves top results on three current datasets for object discovery in video, unsupervised image segmentation and saliency detection.

This chapter is based on the paper "Unsupervised learn-ing of foreground object segmentation." Croitoru, Ioana, Simion-Vlad Bogolin, and Marius Leordeanu. International Journal of Computer Vision 127.9 (2019): 1279-1302.

3.1 Introduction

In this thesis, we propose a novel approach to unsupervised learning that successfully tackles many of the challenges associated with this task. We present a system that is composed of two main pathways, one that performs unsupervised object discovery in videos or large image collections - the teacher branch, and the other - the student branch, which learns from the teacher to segment foreground objects in single images. The unsupervised learning process could continue over several generations of students and teachers. The key aspects of our approach, which ensure improvement in performance from one generation to the next, are: 1) the existence of an unsupervised selection module that is able to pick up good quality masks generated by the teacher and pass them for training to the next generation students; 2) training of multiple students with different architectures, able through their diversity to help train a better selection module for the next iteration and form together with the selection a more powerful teacher pathway at the next iteration and 3) access to larger quantities of, and potentially more complex, unlabeled data, which becomes more useful as the generations become stronger.

In Figure 3 we present a graphic overview of our full system. In the unsupervised training stage the student network (module A) learns, frame by frame, from an unsupervised teacher pathway (modules B and C) to produce similar object masks in single images. Module B discovers objects in images or videos, while module C selects which masks produced by module B are sufficiently good to be passed to module A for training. Thus, the student branch tries to imitate the output of module B for the frames selected by module C, having as input only a single image - the current frame, while the teacher can have access to an entire video sequence.

3.2 Scientific context

The literature on unsupervised learning follows two directions. 1) One is to learn powerful features in an unsupervised way and then use them for transfer learning, within a supervised scheme and in combination with different classifiers, such as SVMs or CNNs ([41]). 2) The second direction is to discover, at test time, common patterns in unlabeled data, using

clustering, feature matching or data mining formulations ([21]).

Belonging to the first category and closely related to our work, the approach in [39] proposes a system in which a deep neural network learns to produce soft object masks from an unsupervised module that uses optical flow cues in video. Recently, researchers have started to use the natural, spatial and temporal structure in images and videos as supervisory signals in unsupervised learning approaches that are considered to follow a *self-supervised learning* paradigm ([44]). Methods that fall into this category include those that learn to estimate the relative patch positions in images ([13]) and predict color channels ([29]).

The second approach to unsupervised learning includes methods for image co-segmentation ([22]) and weakly supervised localization ([12]). Earlier methods are based on local features matching and detection of their co-occurrence patterns ([52]), while more recent ones ([24]) discover object tubes by linking candidate bounding boxes between frames with or without refining their location. Traditionally, the task of unsupervised learning from image sequences has been formulated as a feature matching or data clustering optimization problem, which is computationally very expensive due to its combinatorial nature. There are also other papers ([30]) that tackle unsupervised learning tasks but are not fully unsupervised, using powerful features that are pre-trained in supervised fashion on large datasets, such as ImageNet ([48]) or VOC2012 ([15]).

With respect to the end goal, our work is more related to the second research direction, on unsupervised discovery in video. Unlike that research, we do not discover objects at test time, but during the unsupervised training process, when the student pathway learns to detect foreground objects.

3.3 System architecture

We propose a genuine unsupervised learning algorithm for foreground object segmentation that offers the possibility to improve over several iterations. Our method combines in complementary ways multiple modules that are well suited for this task.

Student path (Module A): single-image segmentation. The student pathway (module A in Figure 3) consists of a deep convolutional network. We test different network architectures, some of which are commonly used in the recent literature on semantic image segmentation.

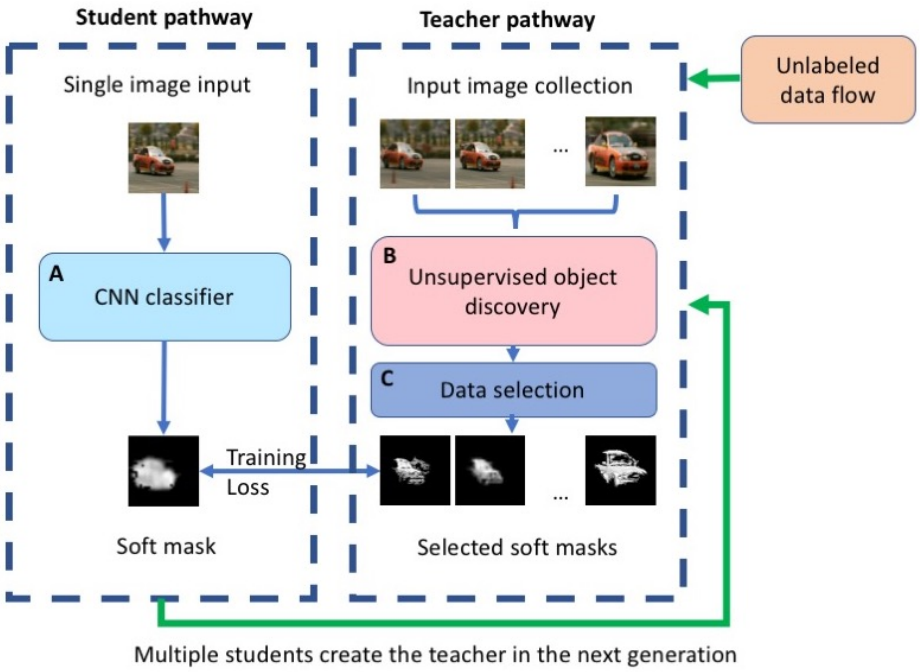

Figure 3: **The dual student-teacher system proposed for unsupervised learning to segment foreground objects in images.** It has two pathways: along the teacher branch, an object discoverer in videos or large image collections (module B) detects foreground objects. The resulting soft masks are then filtered based on an unsupervised data selection procedure (module C). The resulting final set of pairs - input image (or video frame) and soft mask for that particular frame (which acts as an unsupervised label) - are used to train the student pathway (module A). The whole process can be repeated over several generations. At each generation several student CNNs are trained, then they collectively contribute to train a more powerful selection module C (modeled by a deep neural network) and form an overall more powerful teacher pathway at the next iteration of the overall algorithm.

We create a small pool of relatively diverse architectures. In total we use 5 different architectures: LowRes-Net (which produces low resolution segmentation output), FConv (a fully convolutional network and three variations of U-Nets [45].

Combining several student nets. The student networks with different architectures produce varied results that differ qualitatively. While the bounding boxes computed from their soft-masks have similar accuracy, the actual soft-segmentation output looks different. They have different strengths, while making different kinds of mistakes. Their diversity will be the basis for creating the teacher pathway at the next generation.

We experimented with the idea of using several student networks, by combining them to form an ensemble or by letting them produce separate independent segmentations for each image. In our final system we preferred the latter approach, which is more practical, easier to implement and gives the freedom of having the students run independently, in parallel with no need to synchronize their outputs.

When forming an actual ensemble, which we term Multi-Net, the final output is the one obtained by multiplying pixel-wise the soft-masks produced by each individual student net. Thus, only positive pixels, on which all nets agree, survive to the final segmentation. The students at the second iteration are all trained directly on outputs from individual students at the first iteration, filtered with EvalSeg-Net (the selection module for the second iteration). Multi-Net is used only to train the unsupervised selection network, EvalSeg-Net.

Teacher (Module B): unsupervised object discovery. For module B in Figure 3 at first iteration, we use the VideoPCA algorithm, which is a part of the whole system introduced in [52]. For the second iteration we considered all the outputs from all five students to form the training labels.

Unsupervised soft masks selection (Module C) The average value of non-zero pixels in the soft mask is used as a score indicator for each segmented frame at the first iteration.
At the next iterations, we propose an unsupervised way for learning the EvalSeg-Net to estimate segmentation quality. Multi-Net provides masks of higher quality as it cancels errors from individual student nets. Thus, we use the cosine similarity between a given individual segmentation and the ensemble Multi-Net mask, as a cost for ”goodness” of segmentation. Having this unsupervised segmentation cost we train the EvalSeg-Net deep neural net.

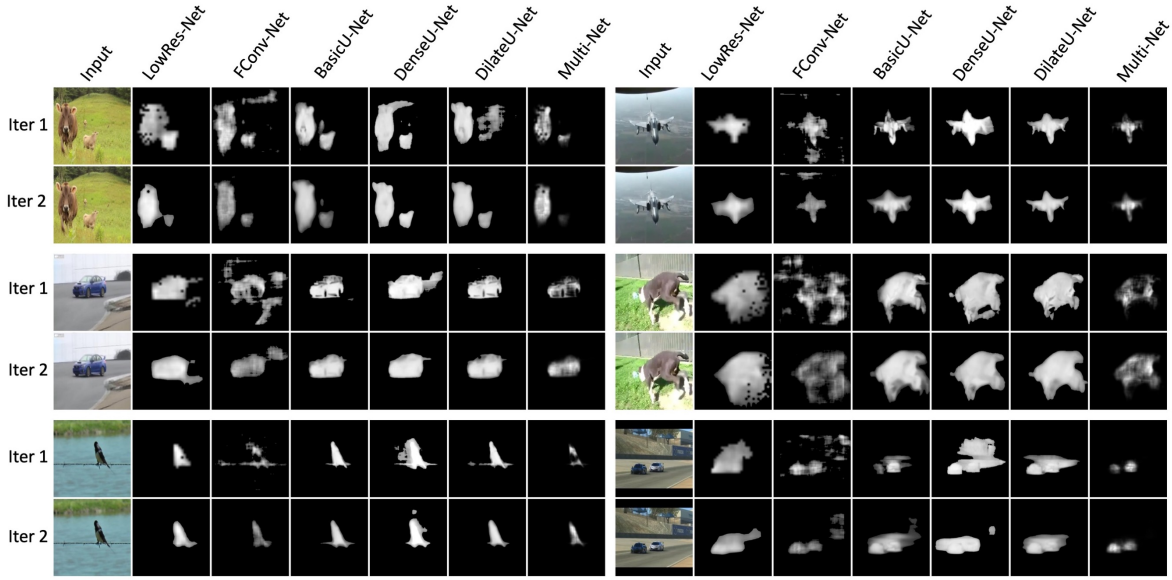

Figure 4: **Visual comparison between models at each iteration (generation).** The Multi-Net, shown for comparison, represents the pixel-wise multiplication between the five models. Note the superior masks at the second generation students, with better shapes, fewer holes and sharper edges. Also note the relatively poorer recall of the ensemble Multi-Net, which produces smaller, eroded masks.

In Figure 4 we present a visual comparison between students and iterations. We can observe that at the second iteration the quality of the segmentation improves.

3.4 Experimental analysis

Comparison with state of the art methods

We first performed comparisons with methods specifically designed for object discovery in video. For that, we choose the YouTube Objects dataset and compare it to the best methods on this dataset in the literature (Table 2). Evaluations are conducted on both versions of YouTube Objects dataset, YTOv1 ([40]) and YTOv2.2 ([26]).

3.5 Conclusions

In this thesis, we present a novel and effective approach to learning from large collections of images and videos, in an unsupervised fashion, to segment foreground objects in single

Method	Aero	Bird	Boat	Car	Cat	Cow	Dog	Horse	Mbike	Train	Avg	Time	Version
[40]	51.7	17.5	34.4	34.7	22.3	17.9	13.5	26.7	41.2	25.0	28.5	N/A	v1
[37]	65.4	67.3	38.9	65.2	46.3	40.2	65.3	48.4	39.0	25.0	50.1	4s	
[25]	64.3	63.2	73.3	68.9	44.4	62.5	71.4	52.3	**78.6**	23.1	60.2	N/A	
[19]	76.3	**71.4**	65.0	58.9	**68.0**	55.9	70.6	33.3	69.7	42.4	61.1	0.35s	
LowRes-Net$_{\text{iter1}}$	77.0	67.5	**77.2**	68.4	54.5	68.3	72.0	**56.7**	44.1	34.9	62.1	0.02s	
LowRes-Net$_{\text{iter2}}$	**83.3**	**71.4**	74.3	69.6	57.4	**80.0**	77.3	**56.7**	50.0	37.2	65.7	0.02s	
DilateU-Net$_{\text{iter2}}$	**83.3**	66.2	**77.2**	**70.9**	63.4	75.0	**80.0**	53.3	50.0	**44.2**	**66.4**	0.02s	
Multi-Net$_{\text{iter2}}$(ensemble)	*87.4*	*72.7*	*77.2*	64.6	62.4	75.0	*82.7*	*56.7*	52.9	39.5	*67.1*	0.15s	
[19]	76.3	**68.5**	**54.5**	50.4	**59.8**	42.4	53.5	30.0	**53.5**	**60.7**	54.9	0.35s	v2.2
LowRes-Net$_{\text{iter1}}$	75.7	56.0	52.7	57.3	46.9	57.0	48.9	44.0	27.2	56.2	52.2	0.02s	
LowRes-Net$_{\text{iter2}}$	79.0	48.2	51.0	**62.1**	46.9	65.7	55.3	**50.6**	36.1	52.4	54.7	0.02s	
DilateU-Net$_{\text{iter2}}$	**84.3**	49.9	52.7	61.4	50.3	**68.8**	**56.4**	47.1	36.1	56.7	**56.4**	0.02s	
Multi-Net$_{\text{iter2}}$(ensemble)	83.1	53.2	54.3	*63.7*	50.6	*69.2*	*61.0*	*51.1*	37.2	48.7	*57.2*	0.15s	

Table 2: **Results on Youtube Objects dataset, versions v1 ([40]) and v2.2 ([26]).** We achieve state of the art results on both versions. Please note that the baseline LowRes-Net already achieves top results on v1, while being close to the best on v2.2. We present results of the top individual models and the ensemble and also keep the baseline LowRes-Net at both iterations, for reference. For each column we highlight with bold the best model and in blue italic the cases where the ensemble is better or equal.

images. We present a relatively general algorithm for this task, which offers the possibility of learning several generations of students and teachers. We demonstrate in practice that the system improves its performance over the course of two generations. Our system is one of the first in the literature that learns to detect and segment foreground objects in images in an unsupervised fashion, with no pre-trained features given or manual labeling, while requiring only a single image at test time.

The convolutional networks trained along the student pathway are able to learn general "objectness" characteristics, which include good form, closure, smooth contours, as well as contrast with the background. What the simpler initial VideoPCA teacher discovers over time, the deep, complex student is able to learn across several layers of image features at different levels of abstraction.

Chapter 4

TEACHTEXT: CrossModal Generalized Distillation for Text-Video Retrieval

In recent years, considerable progress on the task of text-video retrieval has been achieved by leveraging largescale pretraining on visual and audio datasets to construct powerful video encoders. By contrast, despite the natural symmetry, the design of effective algorithms for exploiting large-scale language pretraining remains under-explored. In this work, we are the first to investigate the design of such algorithms and propose a novel generalized distillation method, TEACHTEXT, which leverages complementary cues from multiple text encoders to provide an enhanced supervisory signal to the retrieval model. Moreover, we extend our method to video side modalities and show that we can effectively reduce the number of used modalities at test time without compromising performance. Our approach advances the state of the art on several video retrieval benchmarks by a significant margin and adds no computational overhead at test time. Last but not least, we show an effective application of our method for eliminating noise from retrieval datasets. Code and data can be found at https://www.robots.ox.ac.uk/˜vgg/ research/teachtext/.

This chapter is based on "TEACHTEXT: CrossModal Generalized Distil-lation for Text-Video Retrieval." Croitoru, Ioana, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie and Yang Liu. International Conference on Computer Vision. 2021.

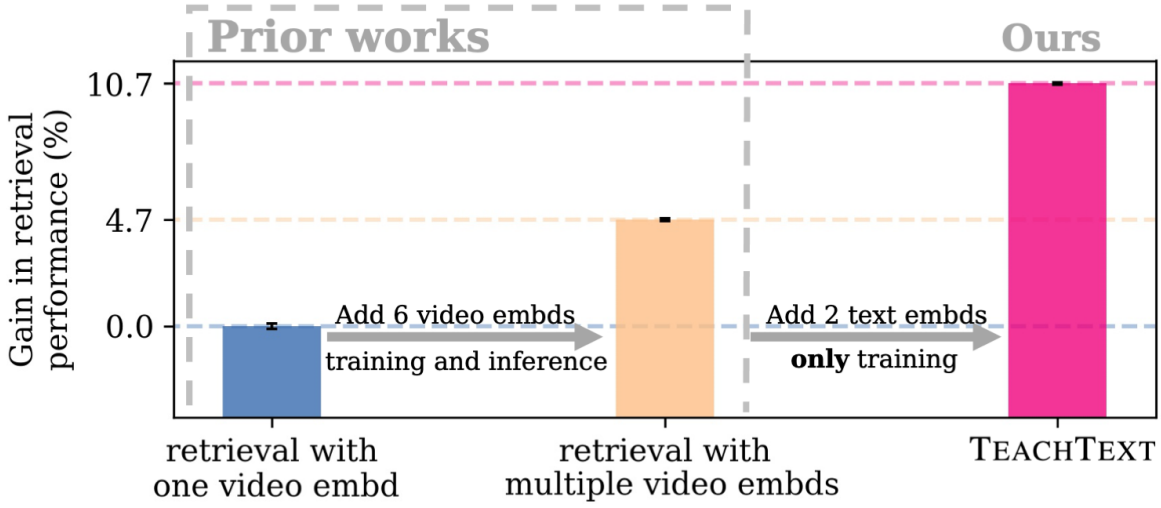

Figure 5: **Distilling the knowledge from multiple text encoders for stronger text-video retrieval.** Prior works [34, 31, 16] have shown the considerable benefit of transitioning from video encoders that ingest a single modality (*left*) to multi-modal video encoders (*centre*). In this work, we show that retrieval performance can be further significantly enhanced by learning from multiple text encoders through the TEACHTEXT algorithm which imposes no additional cost during inference. Text-to-video retrieval performance gain (geometric mean of R1-R5-R10) is reported for a [31] model as well as for our method on the MSR-VTT [57] dataset.

4.1 Introduction

The focus of this work is *text-video retrieval*—the task of identifying which video among a pool of candidates best matches a natural language query describing its content. Video search has a broad range of applications across domains such as wildlife monitoring, security, industrial process monitoring and entertainment. Moreover, as humanity continues to produce video at ever-increasing scale, the ability to perform such searches effectively and efficiently takes on critical commercial significance to video hosting platforms such as YouTube.

A central theme of recently proposed retrieval methods has been the investigation of how to best use multiple video modalities to improve performance. In particular, architectures based on mixtures-of-experts [34, 31] and multi-modal transformers [16] have shown the benefit of making use of diverse sets of pre-trained models for related tasks (such as image classification, action recognition and ambient sound classification) as a basis for video encoding during training and testing.

In this work, we explore whether commensurate gains could be achieved by leveraging mul-

tiple text embeddings learned on large-scale written corpora. Different from video embeddings using multiple modalities and pretraining tasks, it is less obvious that there is sufficient diversity among collections of text embeddings to achieve a meaningful boost in performance. In fact, our inspiration stems from a careful investigation of the performance of different text embeddings across a range of retrieval benchmarks. Strikingly, we observe not only that there is considerable variance in performance across text embeddings, but also that *their ranking is not consistent*, strongly supporting the idea of using multiple text embeddings.

Motivated by this finding, we propose a simple algorithm, TEACHTEXT, to effectively exploit the knowledge captured by collections of text embeddings. Our approach requires a "student" model to learn from a single or multiple "teacher" retrieval models with access to different text embeddings by distilling their text-video similarity matrices into an enhanced supervisory signal. As shown in Fig. 5, TEACHTEXT is capable of delivering a significant performance gain. Moreover, this gain is complementary to that of adding more video modalities to the video encoder but importantly, unlike the addition of video modalities, does not incur additional computational cost during inference.

In Fig. 5 we highlight that the gain for a model that uses multiple text embeddings (last bar) is comparable with the gain of a model that uses multiple video modalities (middle bar). The first bar represents the CE [31] model trained with one video embedding, namely **Obj(IG)**. The second bar represents a CE model using 7 video modalities both for inference and training. In the third and final bar of the chart we present the performance of using three different text embeddings with TEACHTEXT at training, while using only one text embedding at inference time.

Our main contributions are: (1) We propose the TEACHTEXT algorithm, which leverages the additional information given by the use of multiple text encoders; (2) We show that directly learning the retrieval similarity matrix between the joint query video embeddings is an effective generalized distillation technique for this task (and we compare our approach to alternatives among prior work such as uni-modal relationship distillation [38]); (3) We show an application of our approach in eliminating noise from modern training datasets for the text-video retrieval task; (4) We demonstrate the effectiveness of our approach empirically, achieving state of the art performance on six text-video retrieval benchmarks.

4.2 Related Work

Video retrieval methods. The task of indexing video content to enable retrieval has a rich history in computer vision—sophisticated systems have been developed to find specific objects [50], actions [28] and near-duplicates [9]. In this work, we focus on the task of retrieving content that matches a given natural language description. For this particular task, there has been considerable interest in developing cross-modal methods that employ a joint-embedding space for text queries and video content [1]. These joint video-text embeddings, which aim to map videos and text descriptions into a common space such that matching video and text pairs are close together, form an attractive computational model for tackling this problem, since they allow for efficient indexing (although hierarchical embeddings have also been investigated [7]). Recently, two key themes have emerged towards improving the quality of these embeddings. First, large-scale weakly supervised pretraining methods [35] have sought to expand their training data by exploiting the speech contained in the videos themselves as a supervisory signal. Second, the integration of multiple modalities (which has long been considered important for semantic indexing [51]) has been shown to yield significant gains in performance [34, 31]. We focus on candidates from this latter theme as a basis for investigating our approach.

Knowledge Distillation/Privileged Information. The purpose of knowledge distillation is to transfer knowledge from one model (teacher) to another model (student). This idea was originally introduced in the context of decision tree simplification [4] and model compression [5], and later extended by [20] who formalised this knowledge transfer as the temperature-parameterised process of *knowledge distillation*. The concept was further generalised in the unifying framework of *generalized distillation* [33] for learning with privileged information [54] (via *similarity control* and *knowledge transfer* [53]), together with knowledge distillation [20]. Our approach distills knowledge of the similarities between video and text samples into the student and therefore represents a form of generalized distillation.

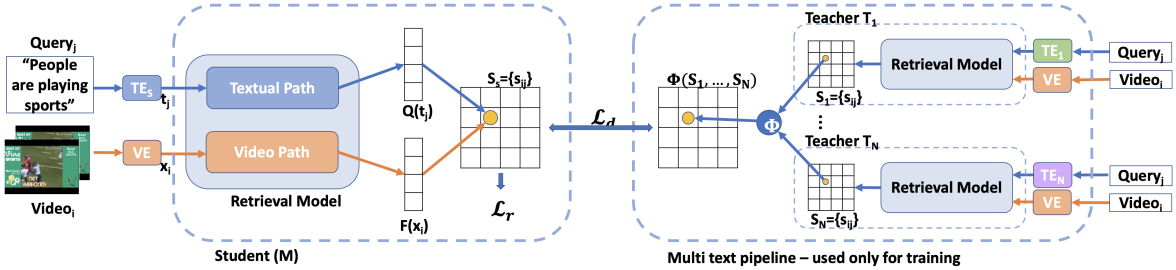

Figure 6: **TEACHTEXT teacher-student framework overview.** Given a batch of input videos and queries in natural language during training, the student model, M (left) and teacher models $T_1,\ldots,T_N$ (right) each produce similarity matrices (visualised as square grids). The similarity matrix produced by M is encouraged to match the aggregated matrices of the teachers through the distillation loss $\mathscr{L}_d$ in addition to the retrieval loss $\mathscr{L}_r$. Note that both the student and teachers ingest the same video embeddings (VE), but employ different text embeddings (TE_S for the student, $\text{TE}_1,\ldots,\text{TE}_N$ for the teachers). At test time, the teacher models are discarded.

4.3 Method

TEACHTEXT algorithm. We propose the TEACHTEXT algorithm which seeks to exploit cues from multiple text embeddings. An overview of our approach is provided in Fig. 6. In the initial phase of the algorithm, we train a collection of teacher models $\{T_k : k \in \{1,\ldots,N\}\}$ for the text-video retrieval task. The teachers share the same architecture but each model T_k uses a different text embedding as input (extracted using a pre-trained text encoder TE_k). In the second phase the parameters of the teachers are frozen. We then proceed by sampling a batch of B pairs of videos and captions and computing a corresponding similarity matrix $S_k \in \mathbb{R}^{B\times B}$ for each teacher T_k (Fig. 6 right). These N similarity matrices are then combined with an aggregation function, $\Phi : \mathbb{R}^{N\times B\times B} \rightarrow \mathbb{R}^{B\times B}$, to form a single supervisory similarity matrix (Fig. 6, centre-right). Concurrently, the batch of videos and captions are likewise processed by the student model, M, which produces a further similarity matrix, $S_s \in \mathbb{R}^{B\times B}$. Finally, in addition to the standard retrieval loss, a distillation loss, $\mathscr{L}_d$, encourages the S_s to lie close to the aggregate $\Phi(S_1,\ldots,S_N)$. During inference, the teacher models are discarded and the student model M requires only a single text embedding.

Student model. A key advantage of our approach is that it is agnostic to the architectural form of the student and teachers, and thus the student (and teachers) can employ any method from the current literature.

Model	$R@1 \uparrow$	$R@5 \uparrow$	$R@10 \uparrow$	$MdR \downarrow$
Dual[14]	7.7	22.0	31.8	32.0
HGR[7]	9.2	26.2	36.5	24.0
MoEE[34]	$11.1_{\pm 0.1}$	$30.7_{\pm 0.1}$	$42.9_{\pm 0.1}$	$15.0_{\pm 0.0}$
CE[31]	$11.0_{\pm 0.0}$	$30.8_{\pm 0.1}$	$43.3_{\pm 0.3}$	$15.0_{\pm 0.0}$
TT-CE	$11.8_{\pm 0.1}$	$32.7_{\pm 0.1}$	$45.3_{\pm 0.1}$	$13.0_{\pm 0.0}$
TT-CE+	$\mathbf{15.0}_{\pm 0.1}$	$\mathbf{38.5}_{\pm 0.1}$	$\mathbf{51.7}_{\pm 0.1}$	$\mathbf{10.0}_{\pm 0.0}$

Table 3: **MSR-VTT full split: Comparison to state of the art.**

Teacher models The teacher models use the same architecture as the student model. Concretely, we create a pool of multiple teachers, each using a different pre-trained text embedding as input. The candidate text embeddings we consider in this work are: mt_grovle [6], openai-gpt [42], gpt2-large [43], gpt2-xl [43], w2v [36]. In this way, we obtain a set of up to five models that form the teachers T_k, $k = 1..5$ used by TEACHTEXT.

4.4 Experimental setup

Comparison to prior work. As it can be seen in Tab.3 our approach is effective and achieves state of the art results. All methods are trained for the retrieval task using only the samples from the target datasets. Moreover, in order to be as fair as possible, in each comparison we included the results of our TEACHTEXT (abbreviated TT in the table) applied also to the best existing method for that dataset.

4.5 Conclusion

In this work, we present a novel algorithm TEACHTEXT for the text-video retrieval task. We use a teacher-student paradigm where a student learns to leverage the additional information given by one or multiple teachers, sharing the architecture, but each using a different pre-trained text embedding. In this way, we achieve state of the art results on six benchmarks.

Chapter 5

Conclusions

In this thesis we have shown how generalized distillation can be applied on different task such as object segmentation, object detection and text-video retrieval. For the unsupervised scenario we considered the tasks of object segmentation and object detection and shown that the distillation technique can successfully be applied to learn from an unsupervised teacher which automatically generates labels. Moreover, we proved that with a simple selection module the noisy labels can be filtered which improved the results. Furthermore, we shown that we can form ensembles of multiple students to improve the results. These ensembles can form a new teacher in the second iteration to boost the performance of new trained students. We demonstrate in practice that the system improves its performance over the course of two generations. The presented system is one of the first in the literature that learns to segment foreground objects in images in an unsupervised way where no manually annotated labels where used.

For the supervised scenario we considered the task of text video retrieval and shown that using generalized distillation you can learn from teachers trained with different pre-trained text embedding and improve the performance. Using different pre-trained text embeddings, the performance varies drastically, suggesting the presence of complementary information. We took advantage of this complementary information through generalized distillation and trained students that learn from one or multiple teachers trained with various pre-trained text embeddings.

Bibliography

[1] Aytar, Y., Shah, M., and Luo, J. (2008). Utilizing semantic word similarity measures for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[2] Beluch, W. H., Genewein, T., Nürnberger, A., and Köhler, J. M. (2018). The power of ensembles for active learning in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[3] Bogolin, S.-V., Croitoru, I., and Leordeanu, M. (2020). A hierarchical approach to vision-based language generation: from simple sentences to complex natural language. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

[4] Breiman, L. and Shang, N. (1996). Born again trees. *University of California, Berkeley, Berkeley, CA, Technical Report*, 1:2.

[5] Buciluǎ, C., Caruana, R., and Niculescu-Mizil, A. (2006). Model compression. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[6] Burns, A., Tan, R., Saenko, K., Sclaroff, S., and Plummer, B. A. (2019). Language features matter: Effective language representations for vision-language tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[7] Chen, S., Zhao, Y., Jin, Q., and Wu, Q. (2020). Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[8] Chen, X., Shrivastava, A., and Gupta, A. (2014). Enriching visual knowledge bases

via object discovery and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[9] Chum, O., Philbin, J., Isard, M., and Zisserman, A. (2007). Scalable near identical image and shot detection. In *Proceedings of the International Conference on Image and Video Retrieval (CIVR)*.

[10] Croitoru, I., Bogolin, S.-V., and Leordeanu, M. (2017). Unsupervised learning from video to detect foreground objects in single images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[11] Croitoru, I., Bogolin, S.-V., Leordeanu, M., Jin, H., Zisserman, A., Albanie, S., and Liu, Y. (2021). Teachtext: Crossmodal generalized distillation for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[12] Deselaers, T., Alexe, B., and Ferrari, V. (2012). Weakly supervised localization and learning with generic knowledge. *International Journal of Computer Vision (IJCV)*, 100(3).

[13] Doersch, C., Gupta, A., and Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[14] Dong, J., Li, X., Xu, C., Ji, S., and Wang, X. (2019). Dual dense encoding for zero-example video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[15] Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision (IJCV)*, 111(1):98–136.

[16] Gabeur, V., Sun, C., Alahari, K., and Schmid, C. (2020). Multi-modal transformer for video retrieval. *Proceedings of the European Conference on Computer Vision (ECCV)*.

[17] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

[18] Grigorescu, S., Trasnea, B., Cocias, T., and Macesanu, G. (2020). A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386.

[19] Haller, E. and Leordeanu, M. (2017). Unsupervised object segmentation in video by efficient selection of highly probable positive features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[20] Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

[21] Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *Proceedings of the ACM Computing Surveys*, 31(3):264–323.

[22] Joulin, A., Bach, F., and Ponce, J. (2010). Discriminative clustering for image co-segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[23] Joulin, A., Bach, F., and Ponce, J. (2012). Multi-class cosegmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[24] Joulin, A., Tang, K., and Fei-Fei, L. (2014). Efficient image and video co-localization with Frank-Wolfe algorithm. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[25] Jun Koh, Y., Jang, W.-D., and Kim, C.-S. (2016). Pod: Discovering primary objects in videos based on evolutionary refinement of object recurrence, background, and primary object models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[26] Kalogeiton, V., Ferrari, V., and Schmid, C. (2016). Analysing domain shift factors between videos and images for object detection. *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 38(11).

[27] Kim, G., Xing, E., Fei-Fei, L., and Kanade, T. (2011). Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[28] Laptev, I. and Pérez, P. (2007). Retrieving actions in movies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[29] Larsson, G., Maire, M., and Shakhnarovich, G. (2016). Learning representations for automatic colorization. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[30] Lee, Y. J., Kim, J., and Grauman, K. (2011). Key-segments for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[31] Liu, Y., Albanie, S., Nagrani, A., and Zisserman, A. (2019). Use what you have: Video retrieval using representations from collaborative experts. *The British Machine Vision Conference (BMVC)*.

[32] Liu, Y. and Lapata, M. (2019). Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

[33] Lopez-Paz, D., Bottou, L., Schölkopf, B., and Vapnik, V. (2016). Unifying distillation and privileged information. *Proceedings of the International Conference on Learning Representations (ICLR)*.

[34] Miech, A., Laptev, I., and Sivic, J. (2018). Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*.

[35] Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., and Sivic, J. (2019). Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[36] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[37] Papazoglou, A. and Ferrari, V. (2013). Fast object segmentation in unconstrained video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[38] Park, W., Kim, D., Lu, Y., and Cho, M. (2019). Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[39] Pathak, D., Girshick, R., Dollar, P., Darrell, T., and Hariharan, B. (2017). Learning features by watching objects move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[40] Prest, A., Leistner, C., Civera, J., Schmid, C., and Ferrari, V. (2012). Learning object class detectors from weakly annotated video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[41] Radenović, F., Tolias, G., and Chum, O. (2016). Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[42] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

[43] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *preprint*, 1(8):9.

[44] Raina, R., Battle, A., Lee, H., Packer, B., and Ng, A. Y. (2007). Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[45] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

[46] Rother, C., Kolmogorov, V., and Blake, A. (2004). Grabcut: Interactive foreground extraction using iterated graph cuts. In *Proceedings of the ACM Transactions on Graphics*, volume 23, pages 309–314.

[47] Rubinstein, M., Joulin, A., Kopf, J., and Liu, C. (2013). Unsupervised joint object discovery and segmentation in internet images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[48] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3).

[49] Shen, D., Wu, G., and Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19:221–248.

[50] Sivic, J. and Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[51] Snoek, C. G. and Worring, M. (2005). Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35.

[52] Stretcu, O. and Leordeanu, M. (2015). Multiple frames matching for object discovery in video. In *The British Machine Vision Conference (BMVC)*.

[53] Vapnik, V. and Izmailov, R. (2015). Learning using privileged information: similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16(1):2023–2049.

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

[55] Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999.

[56] Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A. N., Gouws, S., Jones, L., Kaiser, Ł., Kalchbrenner, N., Parmar, N., et al. (2018). Tensor2tensor for neural machine translation. *arXiv preprint arXiv:1803.07416*.

[57] Xu, J., Mei, T., Yao, T., and Rui, Y. (2016). Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[58] You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J. (2016). Image captioning with semantic attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.