Object detection based on self feature distillation

Zhuguanyu Wu *, a, Zixiao Hu b
School of Informatics, Xiamen University, Xiamen, China.

* a Corresponding author e-mail: 22920182204320@stu.xmu.edu.com; b 22920182204188@stu.xmu.edu.com

Abstract. For object detection, traditional feature distillation methods need to select features from the teacher model to supervise student model training. But what kind of teacher model to choose and what kind of feature transformation to do are two significant problems. In this work, we put forward a self-supervised target detection approach, a simple and effective teacher-free feature distillation method. Specifically, we utilize more semantic features in-depth to distill shallow features, and in the channel dimension, we utilize salient features to monitor the remaining features. We use \( L^2 \) loss to achieve feature minimization, with complex feature transformation and loss definition, because there is a smaller semantic gap between features from the same model and features from the same model. We prove theoretically that such self feature distillation is equivalent to adding feature perturbation and self feature of different subnetworks. For the first time, Self Feature Distillation (SFD) extends feature distillation from teacher model avatar into a general feature regularization method. Our experiments show that SFD is superior to the current feature regularization method. At the same time, self achieves 1.5 times of the training acceleration of SOTA's teacher-based distillation method because it does not need additional training parameters as the teacher model.

Key words: self feature distillation, object detection.

1. Introduction

Recently, as deep learning matures, there exist many significant breakthroughs in various fields. However, the most advanced deep neural network currently calls for a large quantity of computation and memory, limiting its application in the automatic driving vehicle and mobile phone edge devices. A large number of technologies are proposed to solve this problem, including pruning, digitization, model compression, and knowledge distillation. Knowledge distillation, called the teacher-student model, aims to transfer a teacher’s knowledge with high parameterization to students with a lower weight. Because students are trained to imitate the logic or characteristics of teachers, students can inherit knowledge from teachers, so they can often achieve higher accuracy. Because of its simple and practical characteristics, knowledge distillation has become a famous model compression and improve model precision technology. As one of the most critical challenges in computer vision, target detection is an urgent need for accurate and effective models. Unfortunately, most of the existing knowledge
distillation methods in this vision are designed for image classification, but the improvement of target detection has little effect.

In the image to be detected, there are usually more background pixels than foreground objects. However, in the previous knowledge distillation, students are constantly trained to imitate all pixels' characteristics to have the same priority. Therefore, most of the students' attention is focused on learning background pixel features, which inhibits the students' learning of foreground object features. Because foreground pixels are more important in detection, this imbalance seriously affects the performance of knowledge distillation.

Based on the above, we propose a novel and straightforward method, allowing the detection network to enhance its representation learning without additional tags and external supervision. Besides, the method does not increase the reasoning time of the basic model. Our approach is called self feature distillation (SFD). As the name suggests, SFD allows a network to utilize feature mapping from its own layer as its underlying distillation target. Furthermore, this feature extraction mechanism is used to supply the usual segmentation-based super supervised learning.

The motivation of sadness comes from an engaging observation: when we trained the detection network reasonably, attention graphs from different layers are going to capture various rich contextual information. Even if the shallow features imitate the deep features, for example, the second layer imitates the third layer, and the third layer imitates the fourth layer, the network can learn to enhance its expressiveness: (1) the low-level attention map is refined, and the visual attention catches more abundant scene context. (2) Better performance is learned at the lower level, which benefits the more profound level. For example, layer 4 features are more expressive, even though they don't learn anything from any distillation goals. In contrast, the different levels' visual attention of the same network almost did not improve without sadness, although the continuous training of up to 60K episodes.

In addition to the expensive deployment of existing technologies such as multi-task learning and message passing, SFD provides a new choice for training precise planar detection networks. It enables us to train small networks and have incredible visual attention equivalent to intense networks. Our experiments successfully demonstrate the potency of SFD on some famous lightweight models, such as Retinanet-50.

In short, our contribution has two aspects: (1) we propose a new feature extraction method, SFD, to enhance the representation learning of the the CNN-based planar detection model. SFD is only used in the training phase, and there is no computational overhead in the deployment process. What we do is to use the attention graph of the network itself as the distillation target. (2) We verify the effectiveness of SFD in improving the performance of the the general target detection network.

2. Related Work

Object Detection Object detection is a long-standing fundamental problem in computer vision, and it has been a promising research field for decades. The goal of target detection is to determine whether there are target instances of a given category (such as people, cars, bicycles, dogs, and cats) in a given image; If it exists, it returns the spatial location and coverage of each target instance (for example, it returns a bounding box [6]). As the cornerstone of image understanding and computer vision, target detection is the basis for solving more complex and higher-level vision tasks, such as segmentation, scene understanding, target tracking, image description, event detection, and activity recognition. In addition, object detection is widely used in many fields of artificial intelligence and information technology, including robotic vision, consumer electronics, security, autonomous driving, human-computer interaction, content-based image retrieval, intelligent video surveillance, and augmented reality.

Recently, deep learning has been used to completely omit the features of manual production and learn to extract features in an end-to-end way. These technologies have improved the performance of target detection. These methods usually use dense prediction formula, which means that target detection is regarded as a semantic segmentation task, in which each pixel in the image is assigned a label to indicate whether it belongs to the target. For example, He [1] proposed a dual view CNN (dvcnn) to
process target detection. The previous view and top view images are used as output. For example, Neven[8] and others divide target detection into two stages. Specifically, they first perform binary segmentation to distinguish the target pixel from the background pixel. Then, these pixels are divided into different target instances.

Several schemes have been proposed to supplement network-based monitoring and capture more abundant scene background, such as multitasking learning and information transmission. For example, Zhang [14] and others have established a framework to achieve the goal boundary segmentation and target region segmentation simultaneously. The framework also includes the geometric constraints of the expected target boundary and the typical target area to improve the final performance further. Mohsen [7] and others use the standard target label as additional input and integrate the anti generative network (GAN) into the original framework so that the segment graph is more like the label. Pan [9] performs sequential messaging between top-level outputs to utilize structural information better. Although the above methods do bring additional performance gains, multitasking requires additional comments, and messaging is not valid because it sequentially propagates information. In contrast, the SFD proposed in this paper does not need additional comments and does not increase reasoning time.

Knowledge Distillation Knowledge distillation was initially proposed by Hinton [2], which aims to transfer knowledge from extensive network to a small network. In knowledge distillation, a small student network mimics the middle output and label of an extensive teacher network. In Yim’s paper [10], students and teachers share the same capacity on the network and imitate pairs with the exact dimensions. Hou[3] also studies the knowledge extraction between heterogeneous networks. The recent study [11] extended knowledge extraction to attention extraction. For example, Zagoruyko [13] introduces two types of attention distillation, namely, active attention distillation and gradient-based attention distillation. In both methods, students' network is trained by learning attention map derived from teacher network. The difference between our proposed SFD and Zagoruyko is that our method does not require a teacher network. Distillation is carried out in a hierarchical, top-down manner, and attention knowledge is spread layer by layer. This is new in all the literature. It is worth noting that we are concerned with the possibility of distinguishing hierarchical attention and self-learning. This is different from the existing research using visual attention weight features.

3. Methodology

![Detection model diagram](image-url)
3.1. Self feature distillation

In addition to using the aforementioned semantic segmentation and general target existence prediction loss to train our detection network, our goal is to perform hierarchical and top-down attention distillation to enhance the representation learning process. The proposed sad does not require external supervision or additional labels, because the attention map comes from the network itself.

Generally speaking, attention maps can be divided into two categories: activation-based attention maps and gradient-based attention maps. The activation-based attention map is obtained by processing the activation output of a specific layer, and the gradient-based attention map is obtained by the gradient output of the layer. In experiments, we found that activation-based attentional distillation can achieve considerable performance gains, while traditional attentional distillation has little effect. Therefore, in the following chapters, we will only discuss activation-based tension distillation.

Add SFD to the training. The intuition behind SFD is that the previous layer's attention graph can extract useful text information from the continuous layer's attention graph. Therefore, we also conduct spatial softmax operations. If the size of the original attention map is different from that of the target, the bilinear upsampling is added before the softmax operation. The self-attention distillation proposed here is formed in the network itself.

Adding SFD to the existing network is a direct method, and introducing SFD at different training time points may affect the convergence time. We will give an evaluation in the experimental part. Here, let's assume that there is an ENet half-trained to 40000 episodes. A slight attention generator like AT-AEN is added after the encoding block $E_2, E_3, E_4$ in ENET. AT-GEN is a function. The loss formula of continuous stratified distillation is as follows:
Here is a loss function. is the objective loss function of distillation. In this paper, \( M = 4 \). Note that although we can assign different weights to each SFD path, we did not. Nevertheless, we found that the experimental results were good.

### 3.2. Total Loss Function

The overall loss function consists of four parts:

\[
\mathcal{L} = \mathcal{L}_{\text{seg}}(s, \hat{s}) + \alpha \mathcal{L}_{\text{IoU}}(s, \hat{s}) \\
+ \beta \mathcal{L}_{\text{exist}}(b, \hat{b}) + \gamma \mathcal{L}_{\text{distill}}(A_m, A_{m+1})
\]

Here, the first two terms are the segmentation loss caused by the standard cross-entropy loss \( \mathcal{L}_{\text{seg}} \) and \( \text{IOU} \) loss \( \mathcal{L}_{\text{IoU}} \). IOU loss aims to increase the intersection between predicted target pixels and actual target pixels. \( \mathcal{L}_{\text{exist}} \) is binary cross-entropy loss. Parameters \( \alpha, \beta \) and \( \gamma \) are used to balance segmentation loss, existence loss, and separation loss on the final task.

### 4. Experiments

#### 4.1. Experiments Settings

We evaluated the method of self-feature distillation on MS COCO2017. The dataset is a large-scale dataset containing more than 120K images, covering 80 categories [4]. The benchmark detection network comprises a two-stage detection model RCNN [11] and the one-stage detection model retinanet [5]. All the experiments in this paper are implemented with pytorch [10] and a detection framework. For the two-stage model and the one-stage model, we use the same super parameter settings:

\[
\alpha = 7 \times 10^{-4}, \beta = \gamma = 4 \times 10^{-3}
\]

#### 4.2. Experiments Result

This section shows the experimental results of the baseline detector and our model in Table 1. The P-R curve of our Faster RCNN50 is shown in Figure 3. The results show that a consistent and significant AP enhancement can be observed on all two detectors.

![Figure 3. P-R curve of Faster RCNN50](image-url)
5. Conclusion
We propose a target detection training method based on self-feature distillation. Compared with the traditional methods, we do not need to train the teacher model, which is more general. Therefore, we believe that our method can advance the development of target detector-related fields.

References
[1] He, Bei, Rui Ai, Yang Yan, and Xianpeng Lang, “Accurate and robust lane detection based on dual-view convolutionalneutral network,” in “2016 IEEE Intelligent Vehicles Symposium (IV)” IEEE 2016, pp. 1041–1046.
[2] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean, “Distilling the knowledge in a neural network,” arXiv preprintarXiv:1503.02531, 2015.
[3] Hou, Yuenan, Zheng Ma, Chunxiao Liu, and Chen Change Loy, “Learning lightweight lane detection cnns by self attention distillation,” in “Proceedings of the IEEE/CVF International Conference on Computer Vision” 2019, pp. 1013–1021.
[4] Lin, Tsung-Yi, MichaelMaire, SergeBelongie, James Hays, PietroPerona, DevaRamanan, PiotrDollár, and C LawrenceZitnick, “Microsoft coco: Common objects in context,” in “European conference on computer vision” Springer 2014, pp. 740–755.
[5] Mohsen, G, N Cedric, B Nóra, B Olafand, and H Michael, “ELGAN: Embedding Loss Driven Generative AdversarialNetworks for Lane Detection,” Computer Science-Computer Vision and Pattern Recognition, 2018.
[6] Neven, Davy, Bert De Brabandere, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool, “Towards end-to-end laden detection: an instance segmentation approach,” in “2018 IEEE intelligent vehicles symposium (IV)” IEEE 2018, pp. 286–291.
[7] Pan, Xingang, Jianping Shi, Ping Luo, Xiaogang Wang, and Xiaou Tang, “Spatial as deep: Spatial cnn for traffic scene understanding,” in “Proceedings of the AAAI Conference on...
Artificial Intelligence,” Vol. 32 2018.

[8] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” arXiv preprint arXiv:1506.01497, 2015.

[9] Yim, Junho, Donggyu Joo, Jihoon Bae, and Junmo Kim, “A gift from knowledge distillation: Fast optimization, network minimization and transfer learning,” in “Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition” 2017, pp. 4133–4141.

[10] Zagoruyko, Sergey and Nikos Komodakis, “Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer,” arXiv preprint arXiv:1612.03928, 2016.

[11] Zhang, Jie, Yi Xu, Bingbing Ni, and Zhenyu Duan, “Geometric constrained joint lane segmentation and lane boundary detection,” in “Proceedings of the European Conference on Computer Vi.