Knowledge Distillation for Object Detection

Parul Negi 7048016
Ismail Shah 7047226

Abstract

In this work, we explore knowledge distillation techniques for the object detection task, utilizing two major teacher-student model pairs. The first pair was trained using the Pascal VOC 2012 dataset, while the second pair was trained on the COCO dataset.

For the first teacher model, we initially employed a DINO-based backbone with a Faster R-CNN detection head. However, this configuration yielded suboptimal performance, prompting an investigation into potential issues with the spatial dimensions. To address this, we switched to a Vision Transformer (ViT)-based backbone, which significantly improved the mean Average Precision (mAP). The distilled student model, utilizing MobileNetV2 as the backbone, further increased its mAP after applying distillation losses, highlighting the effectiveness of knowledge distillation in enhancing the performance of lightweight models.

For the second teacher-student model pair, we used the original DETR architecture for the teacher model. In the case of the student model, the backbone was replaced, allowing for a comparative analysis of distillation effectiveness across different architectures and datasets.

1. Introduction

Knowledge distillation (KD) is a powerful technique for transferring the learned features and representations from a larger, complex model (the teacher) to a smaller, efficient model (the student). This process enables the deployment of high-performing models in resource-constrained environments. In this work, we focus on applying KD to the object detection task, using the Pascal VOC 2012 [1] dataset and COCO dataset [7]. We investigate the effectiveness of different backbone networks and distillation strategies in improving the performance of the student model.

2. Related Work

Knowledge distillation was first introduced by Hinton et al. [2], where the soft predictions of a large teacher model are used to train a smaller student model. Over the years, this technique has been widely applied in various tasks, including object detection. Methods like CrossKD [3] have improved distillation by transferring intermediate features of the student's detection head to the teacher's detection head. For object detection, recent works have explored different backbone networks and distillation losses to enhance the detection accuracy of student models.

3. Method

3.1. DINO-ViT

3.1.1 Teacher Model

We initially experimented with a DINO-based backbone [5] integrated with the Region Proposal Network (RPN) and Region of Interest (ROI) heads from Faster R-CNN [6]. The model was trained on the Pascal VOC 2012 dataset, achieving a mean Average Precision (mAP) of 4%. This poor performance was attributed to the lack of spatial dimensionality in the DINO backbone's output, which necessitated replicating the feature vector across height and width dimensions before feeding it to the RPN head.

3.1.2 ViT-based Teacher Model

To address the spatial dimension issue, we replaced the DINO backbone with a Vision Transformer (ViT) [9] while retaining the RPN and ROI heads. The ViT backbone, maintaining spatial dimensions, resulted in a significant performance boost, with the model achieving an mAP of 50% on the Pascal VOC 2012 dataset. This model, known as ViTDet [9], contains 106M parameters.

3.1.3 Student Model

The student model employs MobileNetV2 [4], a lightweight network with 19M parameters. Initially, this model was trained from scratch without any distillation loss, achieving an mAP of 29% after 10 epochs. Subsequently, we applied knowledge distillation, using Mean Squared Error (MSE) loss on the bounding boxes and Kullback-Leibler (KL) divergence loss on the classifier outputs. After distillation, the model's mAP improved to 40%, demonstrating a 10% increase due to the distillation process.

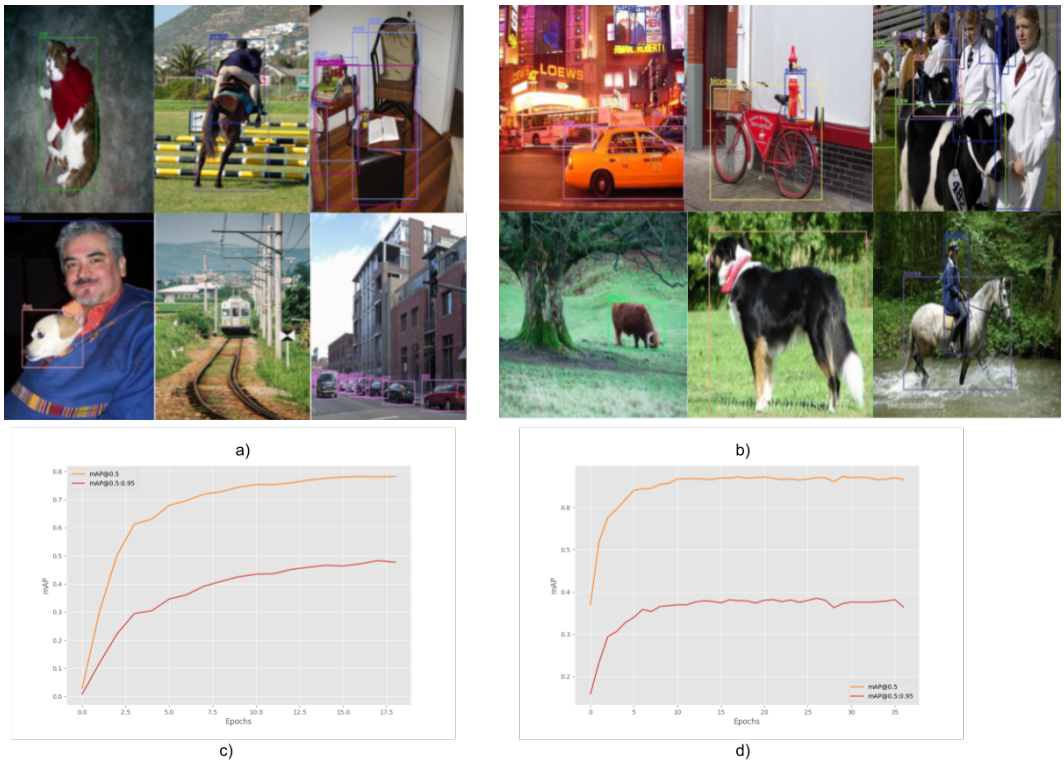


Figure 1. **(a)** Results from the Teacher Model **(b)** Results from the Student Model **(c)** Training curve of the Teacher Model **(d)** Training curve of the Student Model with Distillation Loss.

3.2. DETR - Detection Transformer

3.2.1 Teacher Model

In our project, we utilized the original architecture of the DEtection TRansformer (DETR) model, as outlined in the seminal paper [8]. This architecture incorporates a ResNet-50 backbone as the Convolutional Neural Network (CNN) component, along with a Transformer module composed of 6 encoding and decoding layers, each equipped with 8 multi-head attention mechanisms. Given the extensive training time required for the COCO dataset, which exceeded 7 days, we opted to leverage pre-trained weights for the teacher model. This decision was made to ensure the timely completion of the project while maintaining the model's performance integrity. It consisted of approximately 41 million parameters.

3.2.2 Student Model

In the student DETR model, we replaced the ResNet-50 backbone with the more compact ResNet-18. This substitution resulted in a reduction of the feature map's channel dimensionality from 2048 to 512 channels. Attempts to reduce the number of encoder and decoder layers in the Transformer did not yield satisfactory results, so we retained the original configurations of the Transformer module.

Due to time constraints, the student model was trained on only two classes, specifically 'person' and 'dog'. We employed logit distillation as the primary technique for training the student model, although alternative methods such as contrastive learning and feature map distillation were considered but not explored in this project. The model was trained for 10 epochs, achieving a mean Average Precision (mAP) score of approximately 46%.

The modifications led to a significant reduction in the total number of parameters, bringing the model size down to approximately 18.3 million parameters.

4. Experimental Results and Analyses

Model	Parameters (M)	mAP (0.5-0.95)
DINO Backbone + Faster R-CNN	110	4%
ViT Backbone + Faster R-CNN	106	50%
MobileNetV2 (Baseline)	19	29%
MobileNetV2 (KD)	19	40%

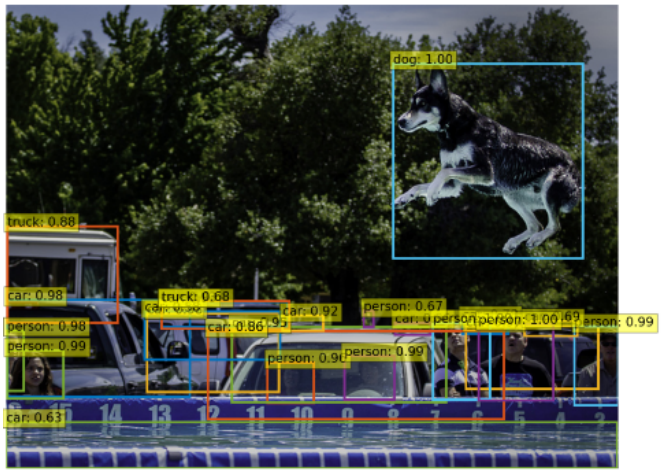
Table 1. Performance comparison of different models on Pascal VOC 2012 dataset.

Our experimental results, summarized in Tab. 1 and Fig. 1, highlight the impact of backbone choice and knowledge distillation on object detection performance. The ViT-

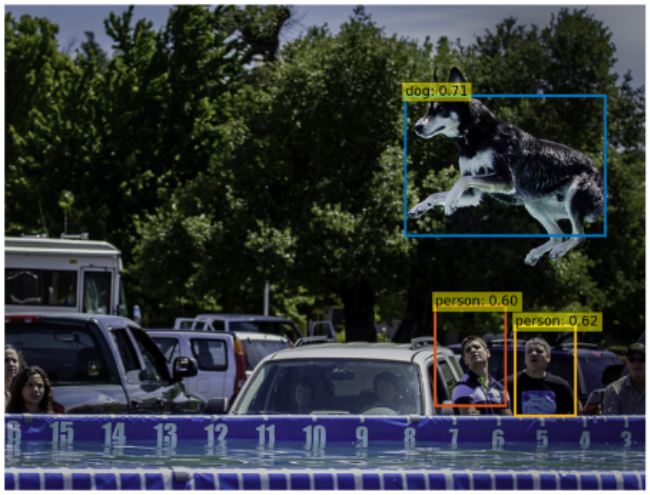
based teacher model significantly outperformed the initial DINO-based model, and the application of KD to the MobileNetV2 student model resulted in a substantial performance improvement.

Model	Parameters (M)	mAP (0.5-0.95)	F1-Score
Teacher Model-ResNet 50	41	62%	0.78
Student Model-ResNet 18	18.3	46%	0.53

Table 2. Performance comparison of DETR teacher-student model on COCO dataset.



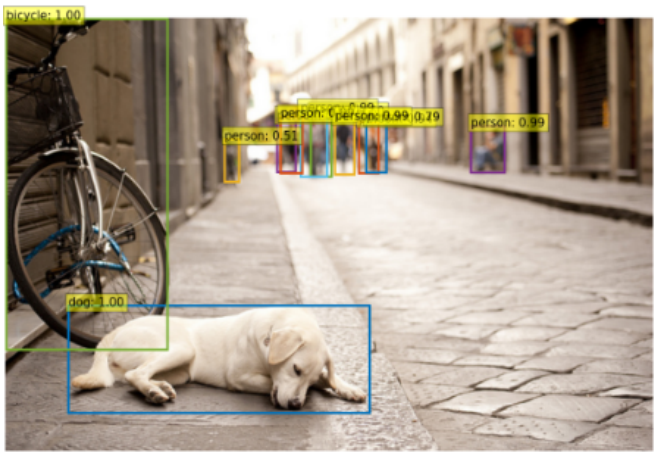
(a) DETR Teacher Model trained on all classes



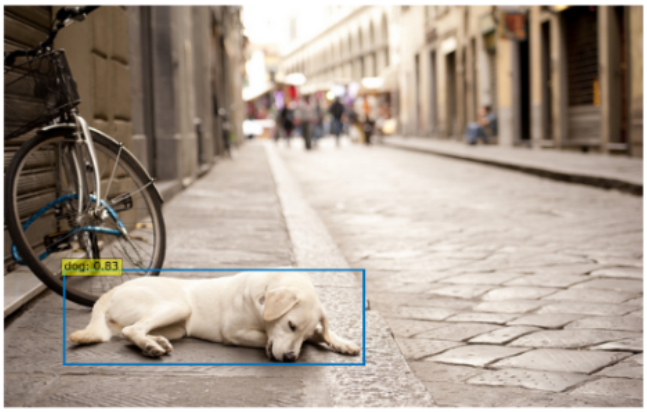
(b) DETR Student Model trained for person and dog class

Figure 2. Comparison of DETR Models

The experimental results illustrated in Fig. 2 demonstrate that the teacher model was able to detect all people in the scene, including those partially obscured behind the car mirror and those located at the edges of the image. In contrast, the student model struggled with these challenging cases, only successfully detecting a subset of the individuals who were not occluded in any way however for that



(a) DETR Teacher Model trained on all classes



(b) DETR Student Model trained for person and dog class

Figure 3

too the bounding box coordinates did not cover the object completely in some cases.

Additionally, Fig. 3 highlights that the student model failed to detect people in the background who were blurred, whereas the teacher model had no such difficulties. On the other hand, the student model was able to accurately detect the dog in the scene, performing comparably to the teacher model in this specific task.

5. Conclusion

This work highlights the potential of knowledge distillation in boosting the performance of object detection models. By strategically choosing the teacher model's backbone and applying targeted distillation losses, we achieved notable improvements in the student model's accuracy. Future research could explore additional distillation strategies and experiment with various backbone architectures to further enhance detection performance. Moreover, further investigation into adapting DINO for object detection could yield valuable insights and potentially lead to even greater advancements in this field.

References

[1] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 1

[2] Jeff Dean Geoffrey Hinton, Oriol Vinyals. Distilling the knowledge in a neural network. In *NIPS 2014 Deep Learning Workshop*, 2014. 1

[3] Zhaohui Zheng Xiang Li Ming-Ming Cheng Qibin Hou Jiabao Wang, Yuming Chen. Crosskd: Cross-head knowledge distillation for object detection. *arXiv:2306.11369*, 2023. 1

[4] Menglong Zhu Andrey Zhmoginov Liang-Chieh Chen Mark Sandler, Andrew Howard. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 1

[5] Ishan Misra Hervé Jégou Julien Mairal Piotr Bojanowski Armand Joulin Mathilde Caron, Hugo Touvron. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 1

[6] Ross Girshick Jian Sun Shaoqing Ren, Kaiming He. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS Proceedings*, 2015. 1

[7] Serge Belongie James Hays Pietro Perona Deva Ramanan Piotr Dollár C. Lawrence Zitnick Tsung-Yi Lin, Michael Maire. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[8] Lewei Lu Bin Li Xiaogang Wang Jifeng Dai Xizhou Zhu, Weijie Su. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 2

[9] Ross Girshick Kaiming He Yanghao Li, Hanzi Mao. Exploring plain vision transformer backbones for object detection. In *ECCV 2022*, 2022. 1