Proceedings of the 4th International Conference on Computing and Data Science (CONF-CDS 2022)
DOI: 10.54254/2755-2721/2/20220530

Integrated Knowledge Distillation for Efficient One-stage Object Detection Network

Zixu Cheng

University College London, London WC1E 6BT, UK

zixu.cheng.21@ucl.ac.uk

Abstract. Due to the low latency requirements in object detection, numbers of one-stage methods like YOLO and SSD adopt a shared head for both classification and localisation tasks. While the decoupled head used to decouple the subtasks into different heads are getting more popular in one-stage detection because they improve accuracy. In contrast, the computational complexity caused by the decoupled head can't be ignored. To solve these problems, we propose an integrated knowledge distillation framework for transferring the representation ability of the decoupled head to the original coupled head and contributing to efficient one-stage object detection. It solves the problem that the coupled head is insufficient in handling the conflict of subtasks and avoids the time delay introduced by the coupling head and the increase of network parameters.

Keywords: object detection, knowledge distillation, decoupled head.

1. Introduction

One-Stage detection is one of the most well-known mainstream object detection methods in the computer vision field because of its fairish trade-off between accuracy and latency. Contrary to two stage detection, it stops the regional recommendation stage and makes the network run faster, and at the same time, obtains competitive accuracy. As a state-of-the-art model of one-stage detection, YOLOv5 [1] inherits the structural style of the YOLO series [1-5] and introduces plenty of optimised structures and tricks to its backbone and neck parts, which makes it achieve outstanding accuracy and inference speed and can be well applied on edge computing devices.

However, few studies have focused on the detection head used by YOLOv5 and optimised this structure to solve potential self-problems. For example, its detection head might have insufficient feature expression and detection capabilities because of the conflict problem between bounding boxes classification and localisation tasks. Towards that, decoupled head techniques were adopted to decouple the classification and localisation issues, which is an efficient and helpful strategy to solve the conflict problem. Even so, it would still bring more parameters, longer latency, and more computational complexity to the model. YOLOX's research [6] has shown that the accuracy of the vanilla YOLO model increases quite a lot after changing to the decoupled head, which is also used in FCOS [7-8], especially when applying it to End-to-End YOLO. This result indicates that the coupled detection head, also used in YOLOv5, contains indistinguishable and confusing information but the classification and localisation of bounding boxes prediction and may be harmful to the model accuracy. However, the model's increased parameters, latency, and computational complexity are inevitable after utilising the

DOI: 10.54254/2755-2721/2/20220530

decoupled head. It is challenging to solve the defect of the coupled head without greatly increasing the calculation of the equipment

Pruning, quantification and knowledge distillation are three commonly used methods of model compression in the deep learning community. However, pruning and quantification are both methods to damage the model structure, such as model weights or network structures, but knowledge distillation does not. Knowledge distillation was first proposed by Hilton. As a new network slimming solution, it was widely used in deep learning in the classification task. Knowledge distillation refines the knowledge from the teacher model and exchanges the information between the teacher and the student models based on their outputs without changing the structure of the student model, which usually is a lighter model, and promoting the performance of the student model simultaneously. Without modifying the model structure, knowledge extraction is becoming a practical solution to improve the accuracy of the model without increasing the amount of computation.

In this paper, a response-based knowledge extraction framework is proposed, which transfers the capability of decoupler in the primary object detection network to the original models with only one coupler. The framework reduces the computation complexity and extra memory consumption of the supplement neck in decoupled structures. By applying response-based knowledge distillation as a teacher model between outputs of the coupled common YOLOv5 model, pre-training on MS COCO 2017 dataset [9], and taking the decoupled YOLOv5 model as the student model, it can be used to solve the shortcomings of the detector in YOLOv5. As shown in Fig 1., the method adopts offline distillation to train the vanilla YOLOv5 model to acquire more semantic information from the decoupled teacher model, using L2 Loss and KL divergence as the distillation loss functions. Furthermore, in order to optimise the distillation process, the framework also introduces weighted distillation according to the difficulties of subtasks and achieves efficient and reasonable use of distilled knowledge.

The contributions of our paper are summarised as follows:

- We propose an integrated distillation framework for one-stage detection head improvements. We transfer the representation abilities into decoupled heads and integrate them into a coupled head by response-based knowledge distillation.
- We introduce task weight to our distillation framework to balance the knowledge from classification and localisation and make the framework more robust and adaptive.
- Our framework provides a method to explore the representation ability of the decoupled head by using different types of decoupled heads and visualising the heatmap of the feature map in future work.

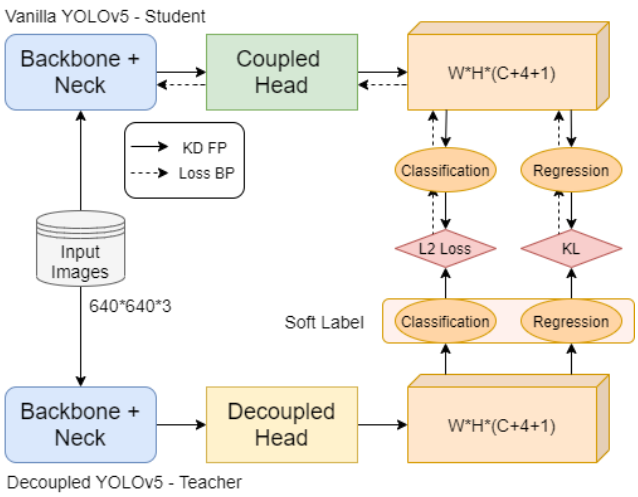

Figure 1. The flowchart of integrated Distillation for Efficient One-stage Object Detection.

DOI: 10.54254/2755-2721/2/20220530

2. Related Work

2.1. One-Stage Object Detection

Since SSD [10] and YOLO [2] first proposed a one-stage object detection method, the conflict between classification and location has been widely discussed. SSD applied the convolutional head in the detection head part, while YOLO adopted the fully connected head for the classification and localisation subtasks. However, the subsequent YOLOv2 to YOLOv5 models introduced the convolution head similar to SSD in the later YOLO series version. Nevertheless, all of them use the coupled head for multiple subtasks in object detection. Many studies on optimisations and improvement of YOLO and SSD series mainly focus on the fusion of multi-level feature scales and the refinement of anchor boxes in the backbone and neck network. Aiming at the multi-scale problem, M2MET [11] introduced MLFPN and built a more efficient feature pyramid network based on the SSD detection framework. RefineDet [12] designed an ARM for anchor refinement like RPN in two-stage networks and promoted the accuracy of the SSD to a higher level like the two-stage network, and maintained its speed at the same time. Gradually, more and more researchers found that the combined classification and localisation tasks are too complex to learn for a shared detection head in the detection model. Therefore, it may lead to the result that the precision of bounding box prediction is not satisfactory. Song et al. [13] pointed out that the shared head (or the sibling head) raised in Fast-RCNN would hurt the accuracy during the training process, and they proposed TSD to decouple the combined tasks.

2.2. Decoupled Head

The decoupled head is a widely used solution to the lack of understanding of two tasks information in a coupled detection head. Mask-RCNN [14] proposed a new head for the instance segmentation task and found that it could slightly improve the performance of the detection head, which indicates that an extra head for another task could get a promotion in model accuracy. RetinaNet [15] introduced two subnets for classification and box regression from the same feature map and used independent parameters in subnets. CenterNet [16] came up with a link structure to decouple the heads into different subtasks such as keypoint heatmap, 3D detection and human pose estimation. FCOS [7] also put forward decoupled designs like RetinaNet and proposed centre-ness for more accurate detection. The Double-Head [17] is a new method raised to optimise the structure of the decoupled head. It integrates the FC-head and Convhead as a Double-Head to decouple the classification and localisation tasks, gaining obvious AP promotion when applying to FPN on the MS COCO dataset. YOLOX-darknet53 [6] introduced decoupled head and found that decoupled head leads to faster convergence and a better result.

2.3. Knowledge Distillation

There are many pieces of research on the application of knowledge extraction in object detection. Generally speaking, according to the types of knowledge extracted in object detection, knowledge extraction can be divided into three types: response-based extraction (or output-based extraction), feature-based extraction and relationship-based extraction [18]. Response-based distillation mainly focuses on the output of the object detection network. Chen et al. [19] first proposed a general framework of output-based knowledge distillation in object detection, and the framework has been widely used in the object detection community. Featured-based knowledge distillation usually applies to the backbone network focusing on the semantic feature information of input images. Liu et al. [20] raised two new loss functions to introduce instance relation and spatial information to solve the lousy performance in feature-based knowledge distillation. SemCKD [21] was proposed by Chen et al. for solving the problem of semantic mismatch between teachers and students via attention to let each layer in the student model learn the multi-layer knowledge instead of being confined to a particular layer artificially designated in the teacher model. Tree-like Decision Distillation [22] gives the student model the same problem-solving mechanism as the teacher model and makes the student model converge more quickly and achieve higher final accuracy. Zhu et al. [23] considered the inter-sample relation and raised CRCD to build anchor-student and anchor-teacher pairs and distil the anchor-student relation supervised by

DOI: 10.54254/2755-2721/2/20220530

corresponding anchor-teacher relation. In order to pay attention to the representation knowledge brought by the decoupling header, response-based knowledge extraction is selected in the framework, and the decoupled representation is extracted through the outputs.

3. Methods

This paper adopts the vanilla YOLOv5 as the student model and the decoupled head YOLOv5 as the teacher model in the response-based knowledge distillation framework. We introduce decoupled head structure to the YOLOv5 to enhance the representative ability of its detection head. The response-based knowledge distillation is utilised to transfer decoupled head ability to coupled student model and fix the increasing computational complexity problem.

3.1 Decoupled Head

The structure of decoupled head introduces more parameters than coupled head to models because one head in decoupled structure only focuses on one subtask in the detect stage. We assume that each decoupled head contains clear semantic information needed for different tasks in object detection. In contrast, the knowledge of classification and localisation are mixed up indistinguishably in the vanilla coupled detection head in the YOLOv5. Therefore, we change the YOLOv5 coupled head to the decoupled design used in YOLOX to endow the detection head with more representation ability of the object information. The detailed construction of decoupled head used in the YOLOv5 is shown in Fig 2.

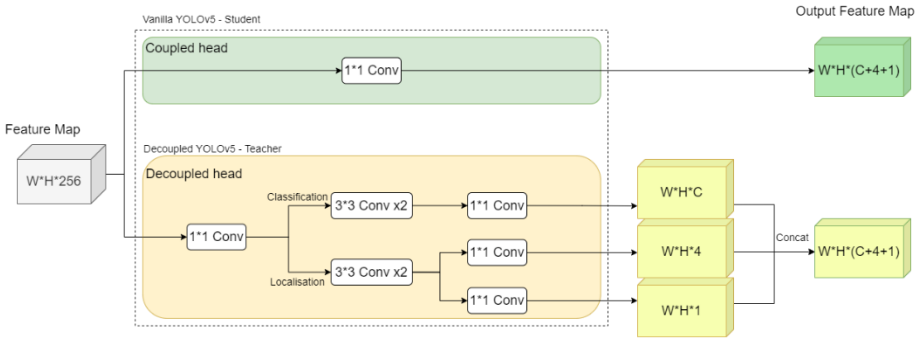

Figure 2. The construction of decoupled head in our framework.

3.2 Response-Based Knowledge Distillation

For the problem of increased computation, we introduce response-based knowledge distillation to settle this problem. We used the vanilla YOLOv5 model as the student model and YOLOv5 with the decoupled head as the teacher model, applied L2 Loss as the loss function of classification and KL divergence as the Loss function of localisation for distillation. The inputs of the distillation are the outputs of the two kinds of detection heads from the vanilla YOLOv5 model and decoupled YOLOv5 model. Then we get soft labels from the outputs of the decoupled model through the SoftMax function. Finally, we use soft labels and the output predictions from the vanilla model to compute the distillation loss and combine it with the training loss. We expect the knowledge distillation between the student and teacher models could transfer the semantic information and representation ability in the decoupled YOLOv5 to the vanilla YOLOv5 to improve performance and achieve model slimming simultaneously. The loss function is computed as follows:

$$L_{cls} = L2\left(X_{Coupled}^{cls} \parallel X_{Decoupled}^{cls}\right) \tag{1}$$

$$L_{reg} = L2\left(X_{Coupled}^{reg} \parallel X_{Decoupled}^{reg}\right) \tag{2}$$

$$L_{total} = L_{cls} + L_{reg} \tag{3}$$

DOI: 10.54254/2755-2721/2/20220530

3.3 Task Weight

Motivated by Double-Head design, we consider the difficulty and complexity of the subtasks in object detection different. The knowledge distillation should not pay equal attention to all the subtask heads. Therefore, we introduced task weight λ to the distillation to balance the framework's Loss of classification and localisation. The overall knowledge distillation Loss function is computed as follows:

$$L_{total} = \lambda \cdot L_{cls} + (1 - \lambda) L_{reg} \tag{4}$$

Overall, the calculation flow of our framework can be represented by the following pseudocode:

Algorithm 1 Integrated Knowledge Distillation

Input: Training Set X, Label Set Y, learning rate γ

Training:

repeat
  for each Batch **do**
    compute prediction p_s of student network;
    compute prediction p_t of teacher network;
    get soft label p_{st} from p_t;
    Compute KD Loss: $L = \lambda \cdot L2(p_s^{cls}, p_{st}^{cls}) + (1 - \lambda) \cdot KL(p_s^{reg} || p_{st}^{reg})$
    BP to student network and update weights: $W \leftarrow W + \gamma \frac{\partial L_W}{\partial W}$
  end for
until Convergence

4. Experiment

We perform our experiments on MS COCO 2017 object detection dataset as the benchmark. We use the vanilla ultralytics-YOLOv5 model as the student model and the decoupled YOLOv5 as the teacher model for our framework baseline. We use the same augmentation methods as vanilla YOLOv5 and the pre-train backbone weight to initialise the models. We set the learning rate to 0.001 and apply SGD as the optimiser in the distillation.

We compare our results with the vanilla ultralytics-YOLOv5 model and the YOLOX model. As shown in Table 1., our framework can promote the performance of the vanilla YOLOv5 model on the MS COCO dataset from mAP 36.7 to 37.4 without any increase in parameter and computational complexity.

Table 1. Our results compared with ultralytics-YOLOv5.

Method	Teacher	FLOPS	Head FLOPS	Total Parameter	Head Parameter	COCO mAP
YOLOv5-s	/	17.1G	0.55G	7.3M	0.19M	36.7
Decoupled YOLOv5-s	/	26.1G	9.57G	7.8M	0.68M	37.8
YOLOv5-s	Decoupled YOLOv5-s	17.1G	0.55G	7.3M	0.19M	37.4

5. Conclusion

Our framework introduces decoupled heads to settle the conflict problem of classification and localisation in one-stage detection networks, raises integrated knowledge distillation to transfer the representation ability from the decoupled teacher model to the original student model, and figures out the increasing computational complexity problem. We use L2 Loss for classification and KL divergence for localisation as loss functions of our distillation framework. We also introduce task weight into our framework to balance the knowledge learned from subtasks according to their difficulty levels.

DOI: 10.54254/2755-2721/2/20220530

In the future, we will adopt more distillation experiments of various types of networks and decoupled heads to explore the differences by visualising the heatmap and further research the interpretability of the detection head.

References

[1] glenn jocher et al. yolov5. https://github.com/ultralytics/yolov5, 2021.

[2] Redmon J, Divvala S, Girshick R, et al.: You Only Look Once: Unified, Real-Time Object Detection[J]. IEEE, 2016.

[3] Redmon J, Farhadi A.: YOLO9000: Better, Faster, Stronger[J]. IEEE Conference on Computer Vision & Pattern Recognition, 2017:6517-6525.

[4] Redmon J, Farhadi A.: YOLOv3: An Incremental Improvement[J]. arXiv e-prints, 2018.

[5] Bochkovskiy A, Wang C Y, Liao H.: YOLOv4: Optimal Speed and Accuracy of Object Detection[J]. 2020.

[6] Ge Z, Liu S, Wang F, et al.: Yolox: Exceeding yolo series in 2021[J]. arXiv preprint arXiv:2107.08430, 2021.

[7] Tian Z, Shen C, Chen H, et al.: Fcos: Fully convolutional one-stage object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 9627-9636.

[8] Wang J, Song L, Li Z, et al.: End-to-end object detection with fully convolutional network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 15849-15858.

[9] Lin T Y, Maire M, Belongie S, et al.: Microsoft coco: Common objects in context[C]//European conference on computer vision. Springer, Cham, 2014: 740-755.

[10] Liu W, Anguelov D, Erhan D, et al.: SSD: Single Shot MultiBox Detector[C]// European Conference on Computer Vision. Springer, Cham, 2016.

[11] Zhao Q, Sheng T, Wang Y, et al. M2det: A single-shot object detector based on multi-level feature pyramid network[C]//Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 9259-9266.

[12] Zhang S, Wen L, Bian X, et al.: Single-shot refinement neural network for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4203-4212.

[13] Song G, Liu Y, Wang X.: Revisiting the sibling head in object detector[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11563-11572.

[14] He K, Gkioxari G, Dollár P, et al.: Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[15] Lin T Y, Goyal P, Girshick R, et al.: Focal Loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[16] Zhou X, Wang D, Krähenbühl P.: Objects as points[J]. arXiv preprint arXiv:1904.07850, 2019.

[17] Wu Y, Chen Y, Yuan L, et al.: Rethinking classification and localisation for object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10186-10195.

[18] Gou J, Yu B, Maybank S J, et al.: Knowledge distillation: A survey[J]. International Journal of Computer Vision, 2021, 129(6): 1789-1819.

[19] Chen G, Choi W, Yu X, et al.: Learning efficient object detection models with knowledge distillation[J]. Advances in neural information processing systems, 2017, 30.

[20] Liu Y, Cao J, Li B, et al.: Knowledge distillation via instance relationship graph[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 7096-7104.

[21] Chen D, Mei J P, Zhang Y, et al.: Cross-layer distillation with semantic calibration[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(8):

DOI: 10.54254/2755-2721/2/20220530

7028-7036.

[22] Song J, Zhang H, Wang X, et al.: Tree-Like Decision Distillation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 13488-13497.

[23] Zhu J, Tang S, Chen D, et al.: Complementary relation contrastive distillation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 9260-9269.