Distilling Image Classifiers in Object Detectors

Research paper study and discussion

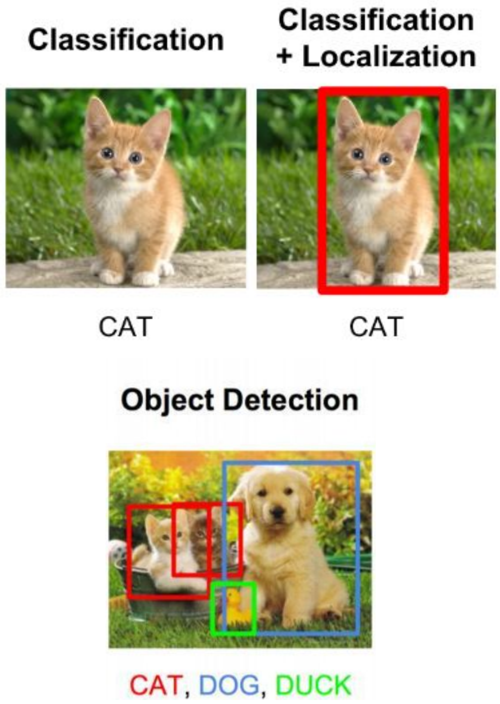
By Dong Yuexi, Chen Mengxuan, Zhang Hongtao, Sun Fan

Agenda

1. Introduction
2. Model Concept
3. Experimental Result Analysis
4. Significance and Impact

Introduction

Object Detection

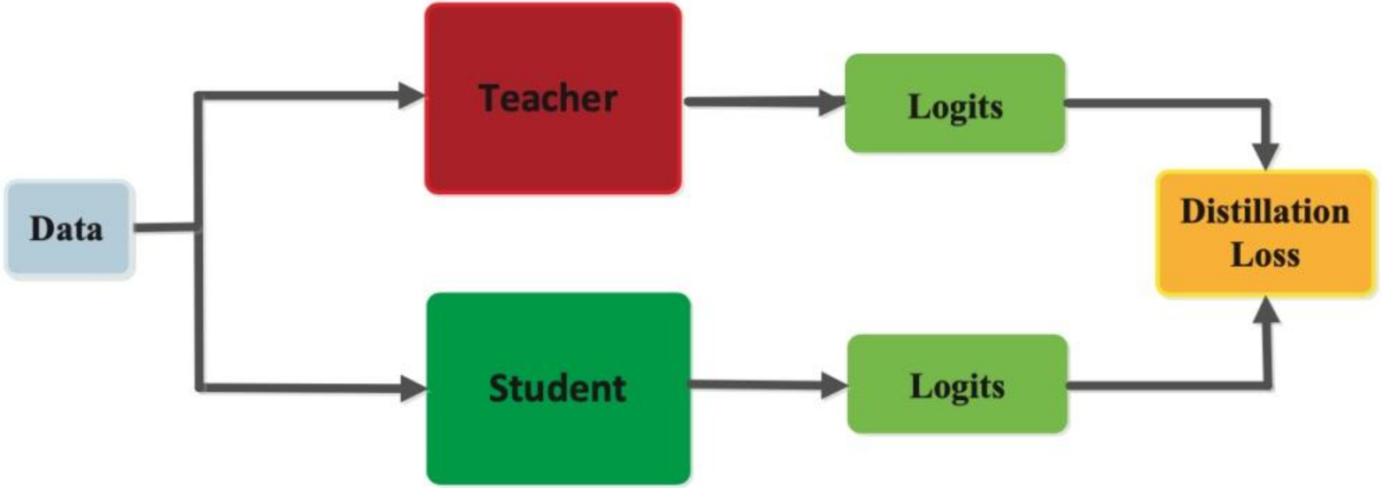


Challenges:
limited memory and computation power

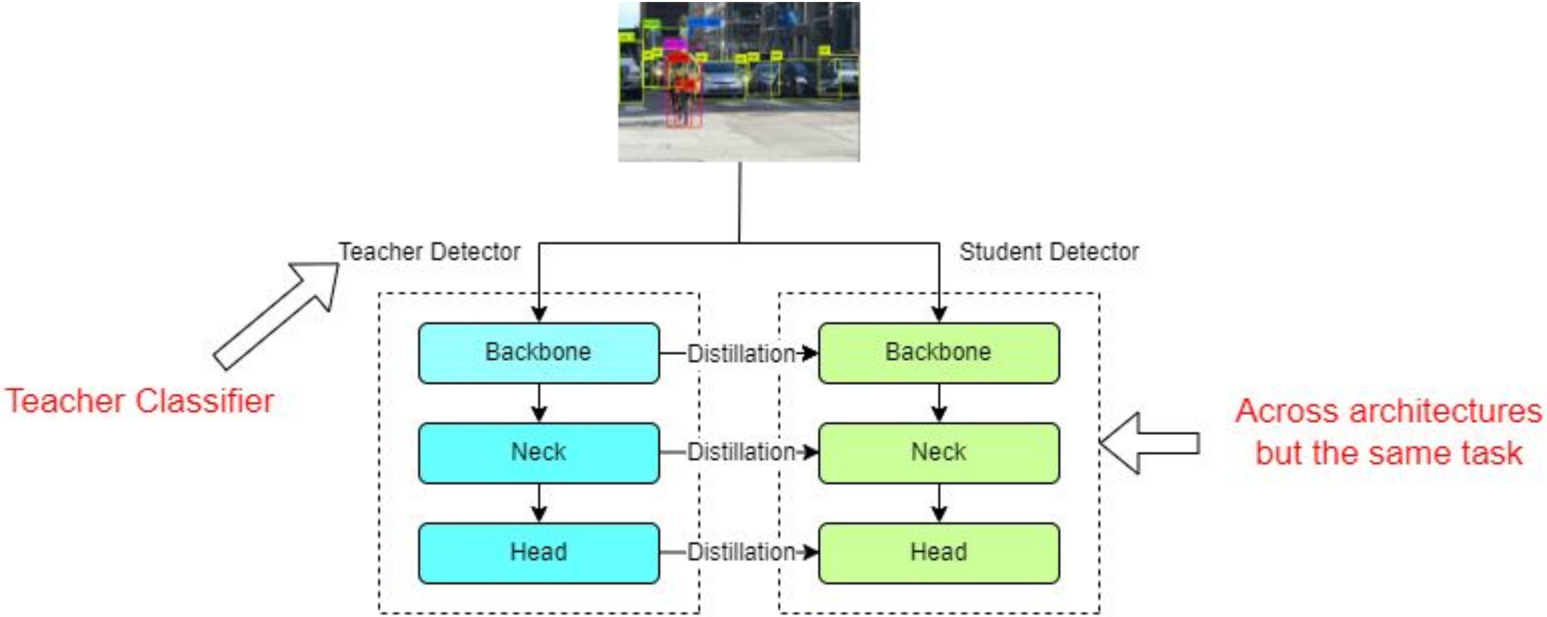
Existing methods:
pruning, quantization, knowledge distillation

This paper:
introduce a knowledge distillation approach for object detection

Knowledge Distillation



Limitation and Goal

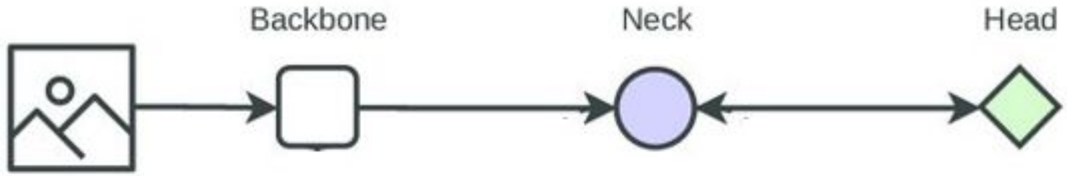


Detecoer-to-Detector Distillation

Model Concept

General Pipeline for a Detector

- Framework of a detector

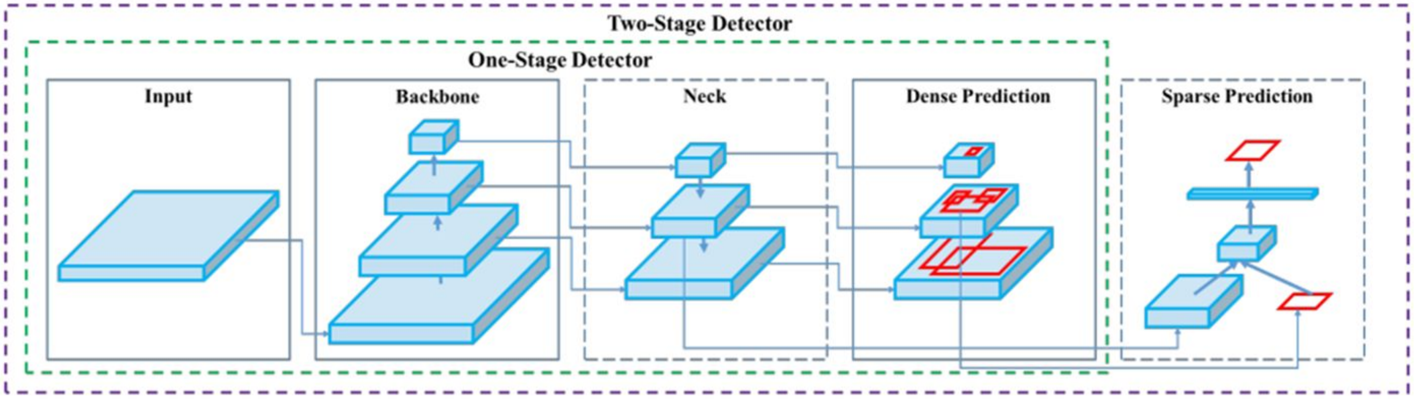


Extract features

Fuse features

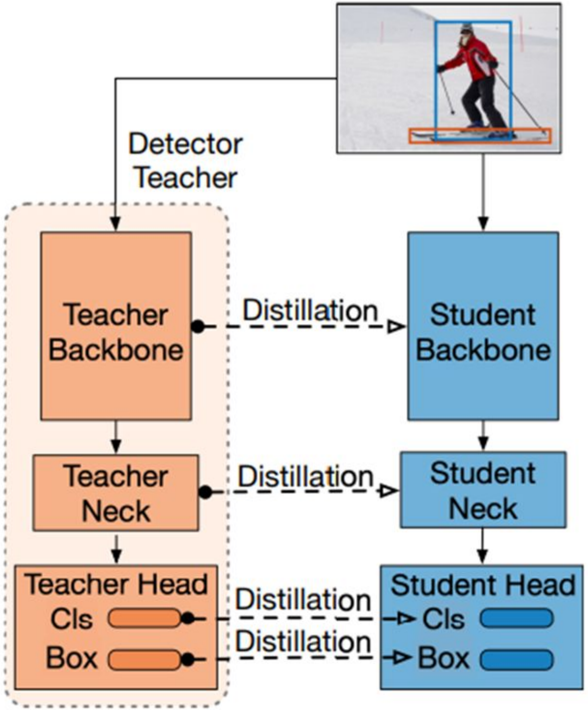
Predict class and bounding box

- Common object detector



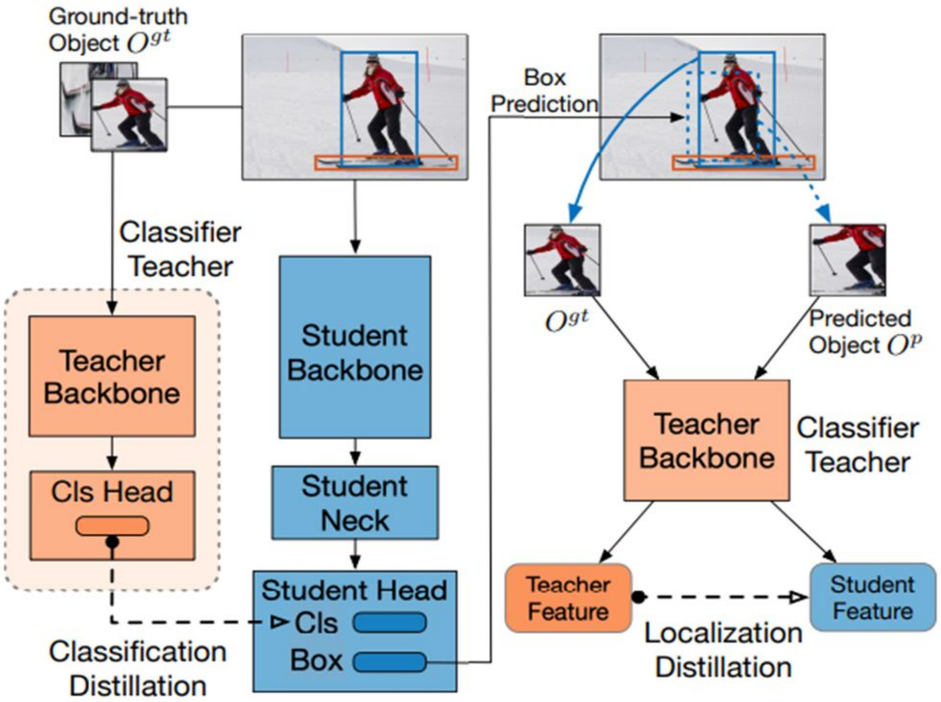
Knowledge Distillation: Detector-to-Detector

- The student and teacher follow the **same kind of detection framework**
- SOTA Method



Knowledge Distillation: Classifier-to-Detector

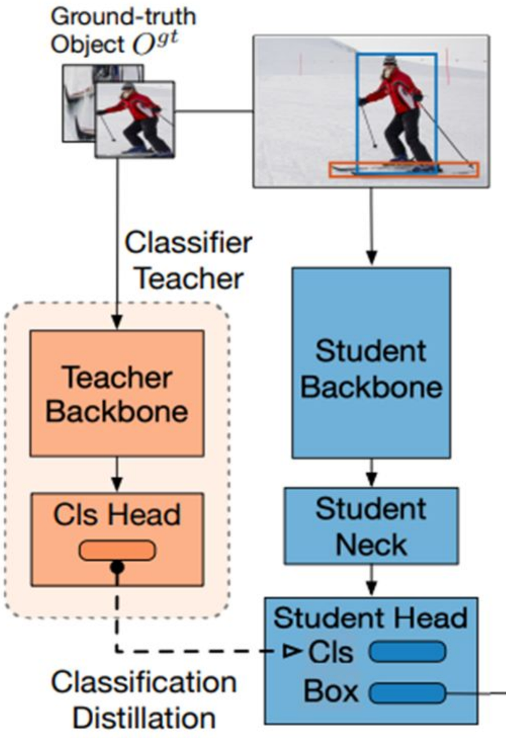
- Knowledge distillation **across tasks**
- From a image **classification teacher** to an **object detection student**



Knowledge Distillation: Classifier-to-Detector

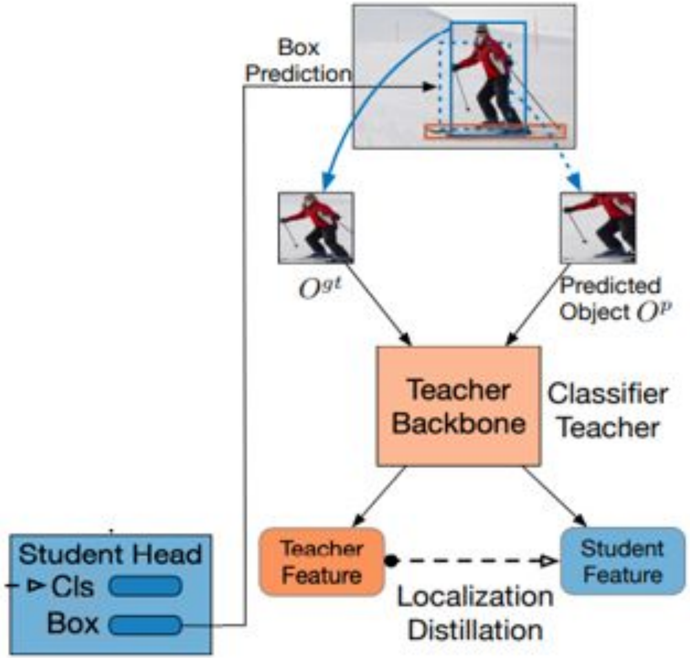
- The classification branch

A loss function defined by the class probabilities of the teacher and the student.



Knowledge Distillation: Classifier-to-Detector

- The localization branch



Feature level distillation strategy: Comparing the features within the predicted bounding box with those within the ground truth.

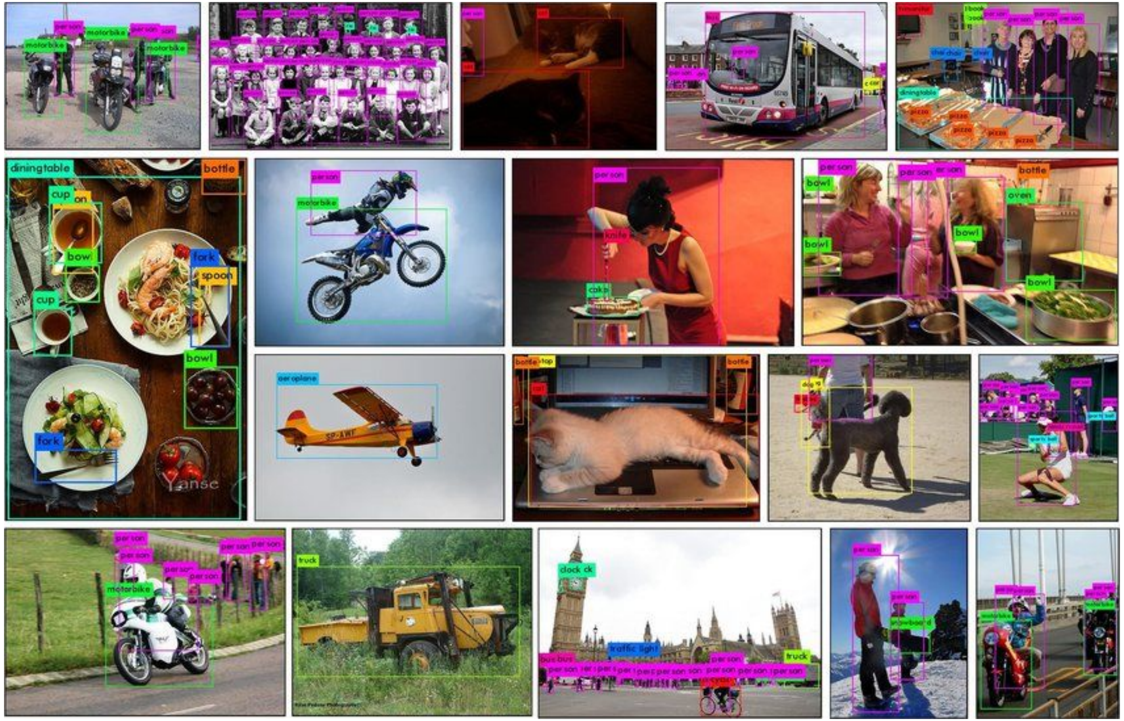
Experimental Result Analysis

Dataset

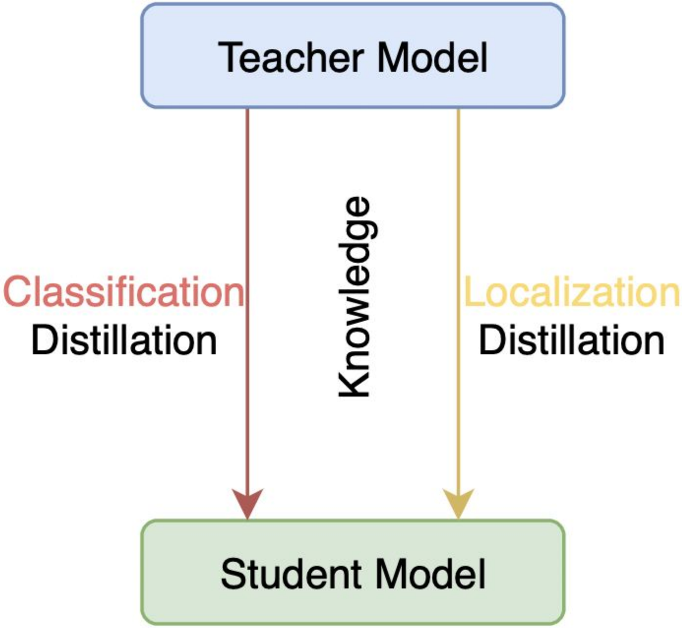
COCO
Common Objects in Context

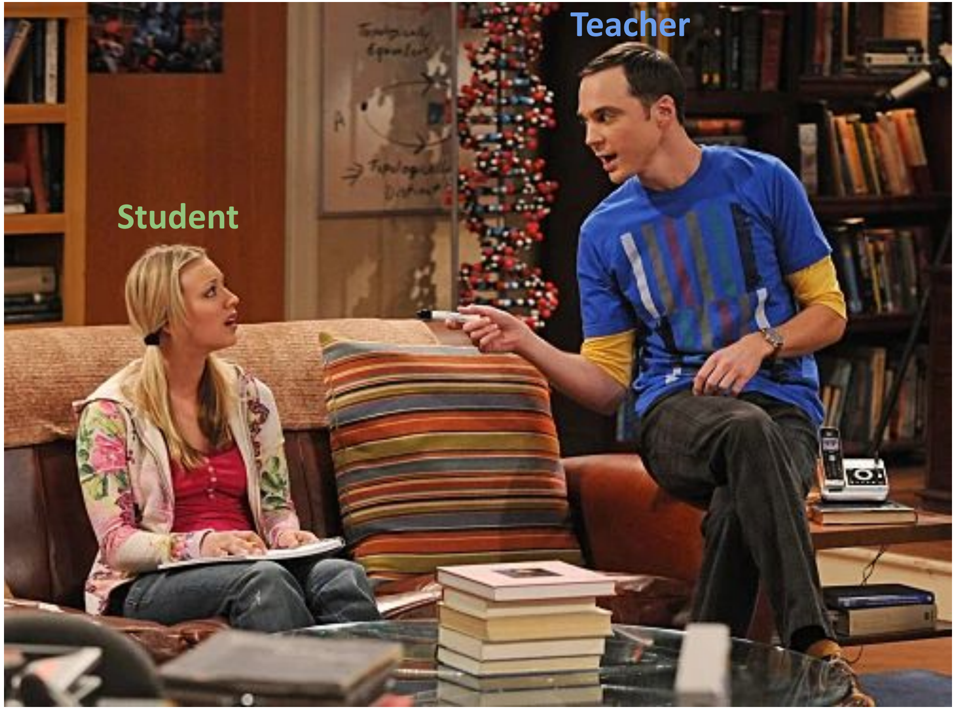
MS COCO 2017(Microsoft) Dataset

- 80 object classes
- 118k images for training
- 5k images for testing



Architecture



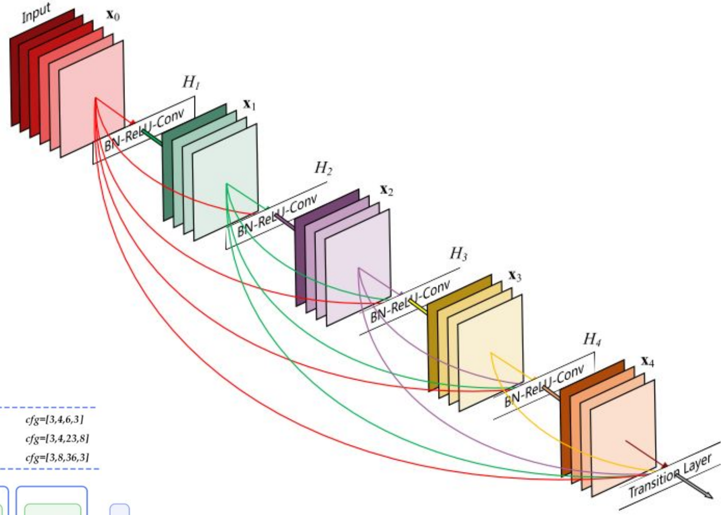


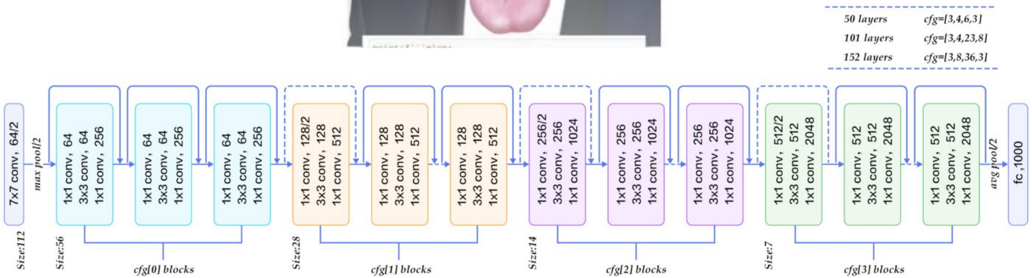
Overall Loss Function

$$\mathcal{L} = \mathcal{L}_{det} + \lambda_{kc}\mathcal{L}_{kd-cls} + \lambda_{kl}\mathcal{L}_{kd-loc}$$

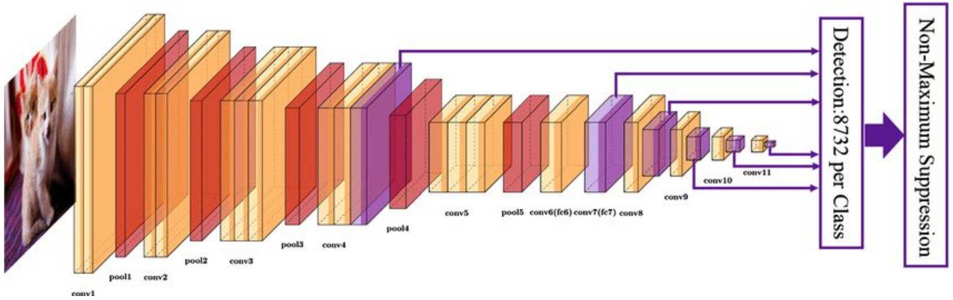
Teacher Model (Classification)

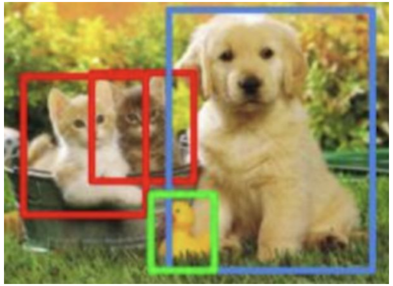
ResNet50
(50 Layers)

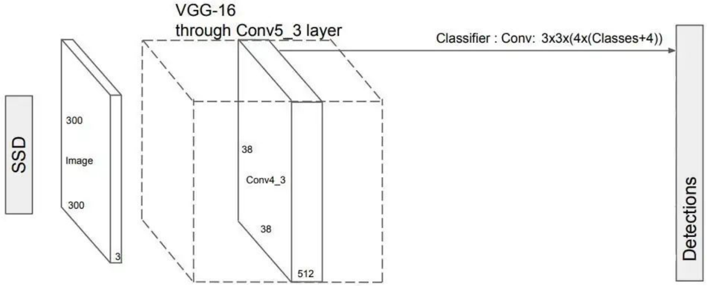




Student Model 1: SSD512(vgg16) (Detection)





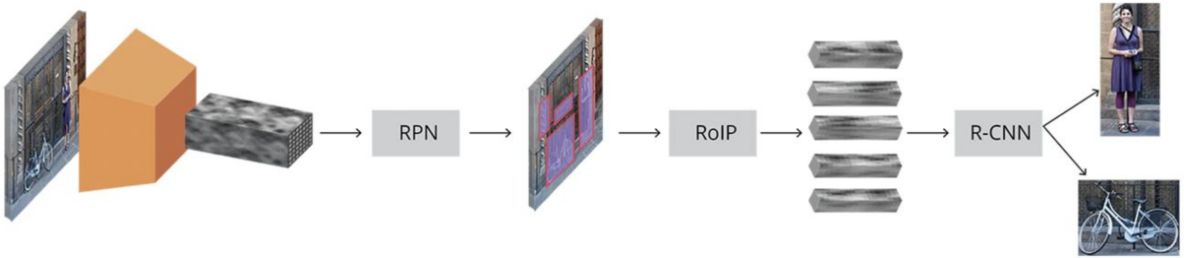


Single Shot Multibox Detector
with VGG16 as backbone

With Distillation:

- 2.7 mAP improvement

Student Model 2: Faster RCNN (Detection)



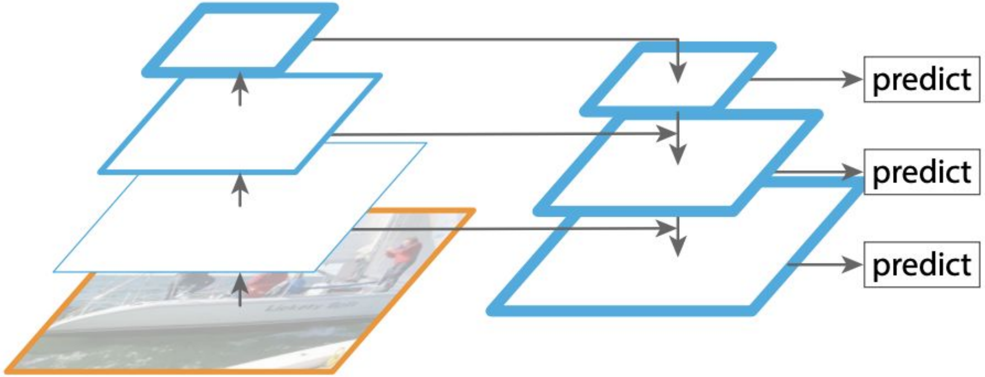
Complete Faster R-CNN architecture

With Distillation:

- 3.9 mAP improvement

Faster RCNN
(Quatered-ResNet50 Backbone)

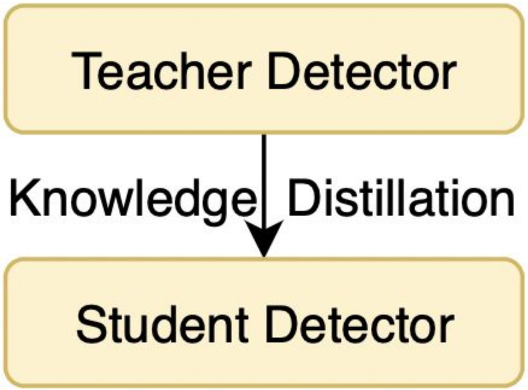
Student Model 3: RetinaNet (Detection)

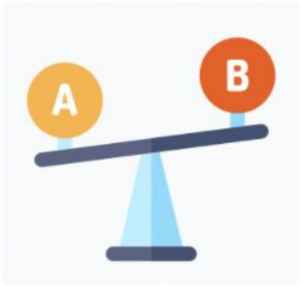


Only used in the comparison experiment

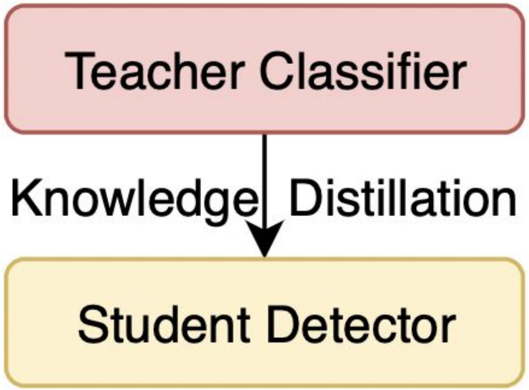
Comparison Experiment

<u>**SOTA**</u> Detector-Detector
Distillation Model: **FKD**

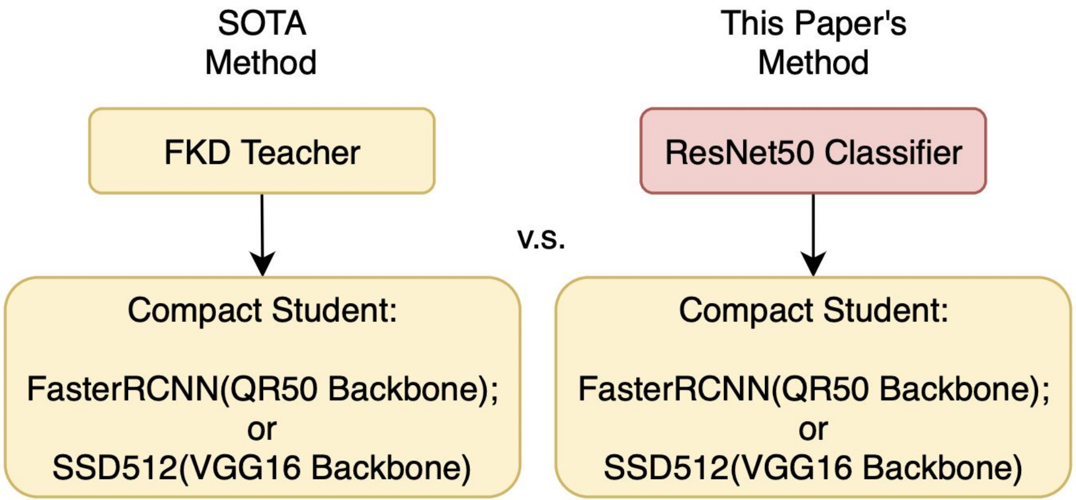




<u>**Novel**</u> Classifier-Detector
Distillation Model



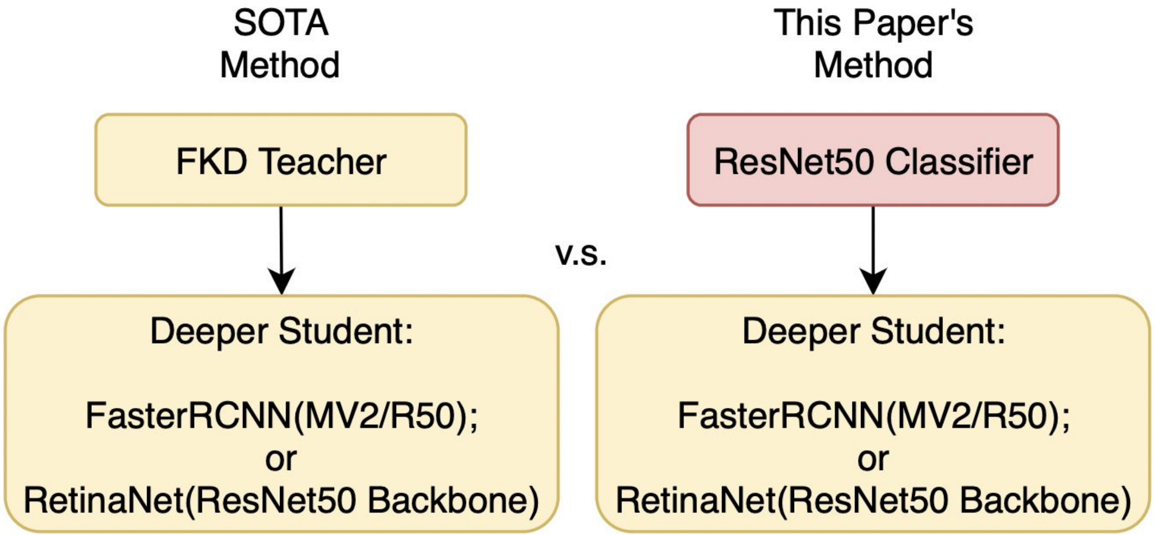
Comparison Using ***Compact*** Student Network



This Paper wins by:

- 1.1 mAP (FasterRCNN)
- 0.9 mAP (SSD512)

Comparison Using ***Deeper*** Student Network



- SOTA Method Wins
- The method of this paper performs better for compact networks

Combination Is Even Better

SOTA FKD + Classifier-Detector Distillation

Table 2: Comparison to detector-to-detector distillation methods on the COCO2017 validation set.

Method	mAP	AP_s	AP_m	AP_l
Faster RCNN-QR50	23.3	13.1	25.0	30.7
+ FKD [43]	26.1	14.6	27.3	35.0
+ Ours	27.2	15.2	29.3	36.2
+ Ours + FKD	**28.0**	**15.4**	**29.8**	**38.5**
SSD512-VGG16	29.4	11.7	34.1	44.9
+ FKD [43]	31.2	12.6	37.4	46.2
+ Ours	32.1	13.3	36.6	47.9
+ Ours + FKD	**32.6**	**13.5**	**37.6**	**48.3**
Faster RCNN-MV2	31.9	18.5	34.4	41.0
+ FKD [43]	33.9	18.3	36.3	45.4
+ Ours	32.7	**19.0**	35.0	42.9
+ Ours + FKD	**34.2**	18.5	**36.3**	**45.9**
Faster RCNN-R50	38.4	21.5	42.1	50.3
+ KD [5]	38.7	22.0	41.9	51.0
+ FGFI [39]	39.1	22.2	42.9	51.1
+ GID [8]	40.2	22.7	44.0	53.2
+ FKD [43]	41.5	23.5	45.0	55.3
+ Ours	38.8	22.5	42.5	50.8
+ Ours + FKD	**41.9**	**23.8**	**45.2**	**56.0**
RetinaNet-R50	37.4	20.0	40.7	49.7
+ FGFI [39]	38.6	21.4	42.5	51.5
+ GID [8]	39.1	22.8	43.1	52.3
+ FKD [43]	39.6	22.7	43.3	52.5
+ Ours	37.9	20.5	41.3	50.5
+ Ours +FKD	**40.7**	**23.1**	**44.7**	**53.8**

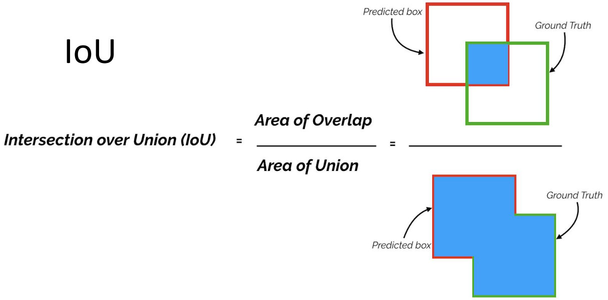
Ablation Study(Component Study)

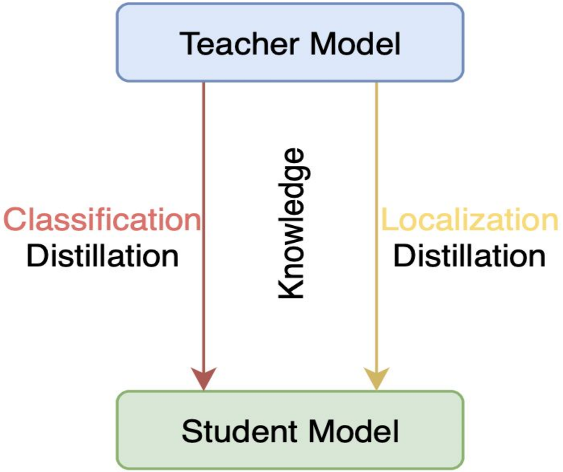
Observations:

- **Classification** distillation yields larger improvement for **smaller** IoU threshold
- **Localization** distillation yields larger improvement for **Larger** IoU threshold

Analysis:

- **Classification** distillation distills category info to student
- **Localization** distillation focuses on precise location





Significance and Impact

Novelty

Old	Detector-Detector Distillation
New	Classifier-Detector Distillation

This research showed:

- Using classification model as teacher in distillation
    - Works!
    - Performs better than baseline
- Improves the SOTA detector-detector distillation

Influenced Paper

1. Knowledge Distillation for 6D Pose Estimation by Keypoint Distribution Alignment

   Contributions:
   - Allowed compact compact network for 6D pose estimation running on embedded devices
   - Powered Robotics Research

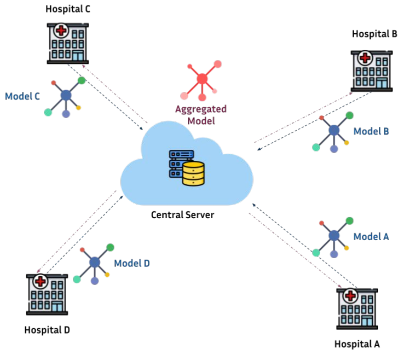
2. Task-Balanced Distillation for Object Detection

   Contributions:
   - Inspired research on task-balanced distillation

Broader Impact

1. Resource-constrained environments:
    a. Improve model performs in mobile devices
    b. Greatly reduced the Training resources

2. Further Research:
    a. Knowledge can be transferred not only across architectures, but also across tasks
    b. Provide new knowledge distillation techniques for federated learning



Q&A

Acknowledgement

- Guo, S., Alvarez, J. M., & Salzmann, M. (2021). Distilling Image Classifiers in Object Detectors. *Advances in Neural Information Processing Systems*, *34*, 1036-1047.

- Guo, S., Hu, Y., Alvarez, J. M., & Salzmann, M. (2022). Knowledge Distillation for 6D Pose Estimation by Keypoint Distribution Alignment. *arXiv preprint arXiv:2205.14971*.

- Tang, R., Liu, Z., Li, Y., Song, Y., Liu, H., Wang, Q., ... & Tan, J. (2022). Task-Balanced Distillation for Object Detection. *arXiv preprint arXiv:2208.03006*.

Thank You