Supplementary Material of Dist^2: Distribution-Guided Distillation for Object Detection

Tianchu Guo[1]
tianchu.gtc@alibaba-inc.com
Pengyu Li[1]
lipengyu007@gmail.com
Wei Liu[2]
ustclwwx@gmail.com
Bin Luo[1]
luwu.lb@alibaba-inc.com
Biao Wang[2]
wangbiao225@foxmail.com

[1] Artificial Intelligence Center
DAMO Academy
Alibaba Group

[2] Work was done when
they were employed by Alibaba

1 Details of the Flexible Imitation(FI) Strategy

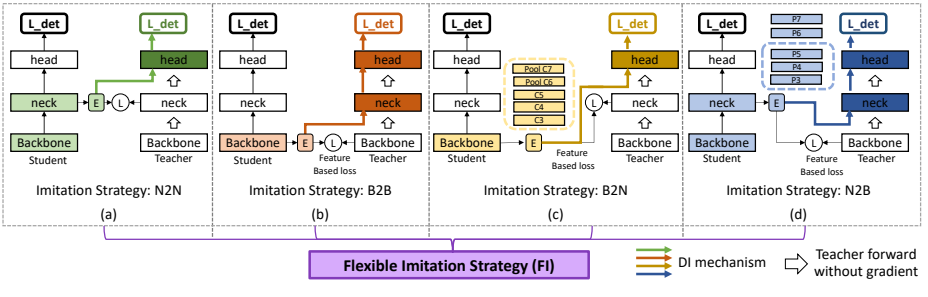

Figure 1: All the imitation strategies of the proposed Flexible Imitation (FI) strategy. The light color parts in the student net denote the feature map generator. The dark color parts in the teacher net denote the detection discriminator. Best viewed in color.

As shown in Fig.1-(a), the backbone and the neck part of the student are treated as the feature generator G^S, denoted as the light green color. The teacher's head is treated as the detection discriminator D^T denoted as the dark green color. The "E" denotes a transferring layer, which is a 1×1 convolution to align the student's channel dimension to the teacher's.

There are four imitation strategies in the FI. Each of the imitation strategies is denoted as "X2Y". It means that the distribution of the X part in the student net imitates the distribution

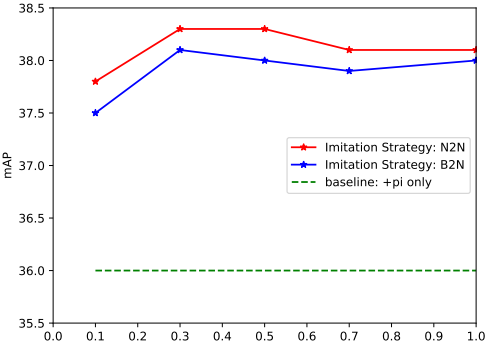

Figure 2: Sensitiveness of the loss weight of the λ_{DI}. It is better to choose the weight at 0.3

of the Y part in the teacher net. The details of all the imitation strategies in FI are shown in the following.

N2N. As shown in Fig. 1-(a). The distribution of student's **N**eck part imitates the distribution of teacher's **N**eck part. The feature generator G^S is the backbone and neck part of the student net. The detection discriminator D^T is the head part of the teacher net. The output of the G^S is equal to $f^{N2N}_{trans}(F^S_{neck}(F^S_{backbone}(x)))$.

B2B. As shown in Fig.1-(b), the distribution of student's **B**ackbone part imitates the distribution of teacher's **B**ackbone part. The feature generator G^S is the backbone part of the student net. The detection discriminator D^T is the neck and head part of the teacher net. The $feat_g$ is equal to $f^{B2B}_{trans}(F^S_{backbone}(x))$.

B2N. As shown in Fig.1-(c), the distribution of student's **B**ackbone part imitates the distribution of teacher's **N**eck part. The feature generator G^S is the backbone part of the student net. The detection discriminator D^T is the head part of the teacher net. The $feat_g$ is equal to $f^{B2N}_{trans}(F^S_{backbone}(x))$. The C6 and C7 are obtained by conducting Pooling on the C5, to fill the position of P6 and P7 in the FPN. The imitation strategy B2N makes high-level semantic information in the teacher's head directly transmit to the student's backbone.

N2B. As shown in Fig.1-(d), the distribution of student's **N**eck part imitates the distribution of teacher's **B**ackbone part. The feature generator G^S is the backbone and the neck part of the student net. The detection discriminator D^T is the neck and head part of the teacher net. The $feat_g$ is equal to $f^{N2B}_{trans}(F^S_{neck}(F^S_{backbone}(x)))$. The P6 and P7 in the FPN output will be removed to match the input of the teacher's neck. The imitate strategy N2B is just similar to reviewing the knowledge.

2 Sensitiveness of the Loss Weight in DI Mechanism

The total loss of our Dist^2 is shown below,

$$\begin{aligned} L =& L_{det}(P,t),(P^{gt},t^{gt})+ \\ & \lambda_{feat}\cdot\Sigma_{is}L_{feat}(feat^S_{is},feat^T_{is})+ \\ & \lambda_{DI}\cdot\Sigma_{is}DI^{is} \end{aligned} \tag{1}$$

where the is is denoted as imitate strategy and $is \in [N2N, B2B, B2N, N2B]$. The λ_{feat} and λ_{DI} are the corresponding loss weights.

In this part, the weight of feature-based loss is fixed, i.e. the λ_{feat} is fixed as 0.1. The influence of the λ_{DI} will be validated. The detection framework is FCOS. The imitation strategies are N2N and B2N. They represent distillation imitation from the same part and cross different parts between student and teacher networks.

As shown in Fig.2, no matter how to choose the λ_{DI} in the range (0,1], the performance is improved compared with the original student net. It can be seen that it's better to choose the weight at 0.3. The performance of the DI mechanism is not sensitive to the choice of the λ_{DI} in the range [0.3,1].