

Figure 5: Architecture of the text detector DB. This figure comes from the paper of DB (Liao et al. 2020). The red and gray rectangles show the backbone and head of the text detector separately.

tion, regularization parameters, learning rate warm-up, light head, pre-trained model and PACT quantization. Finally, the model size of the text recognizer is only 1.6M for Chinese and English recognition and 900KB for alphanumeric sym- bols recognition.

bols recognition. In order to implement a practical OCR system, we con- struct a large-scale dataset for Chinese and English recog. nition as an example. Specifically, text detection dataset has 97K images. Direction classification dataset has 600k im ages. Text recognition dataset has 17.9M images. A small amount of the data are selected to conduct ablation exper- iments quickly and choose the appropriate strategies. We make a lot of ablation experiments to show the effects of different strategies in Figure2 Besides, we also verify the proposed PP-OCR system for other languages recognition which including alphanumeric symbols, French, Korean, Japanese and German.

Japanese and German. The rest of the paper is organized as follows. In section 2, we present the bag of model enhancement or slimming strategies. Experimental results are discussed in section 3 and conclusion is conducted in section 4.

# 2Enhancement or Slimming Strategies

In this section, the details of six strategies for enhancing the model ability or reducing the model size of a text detector will be introduced. Figure 5 shows the architecture of the text detector DB.

text detector DB. Light Backbone The size of backbone has dominant effect on the model size of a text detector. Therefore, light backbones should be selected for building the ultra lightweight models. With the development of image clas. sification, MobileNetV1, MobileNetV2, MobileNetV3 and ShuffleNetV2 series are often used as the light backbones. Each series has different scale. Thanks to the inference time on CPU and accuracy of more than 20 kinds of back bones are provided by PaddleClas as shown in Figure6

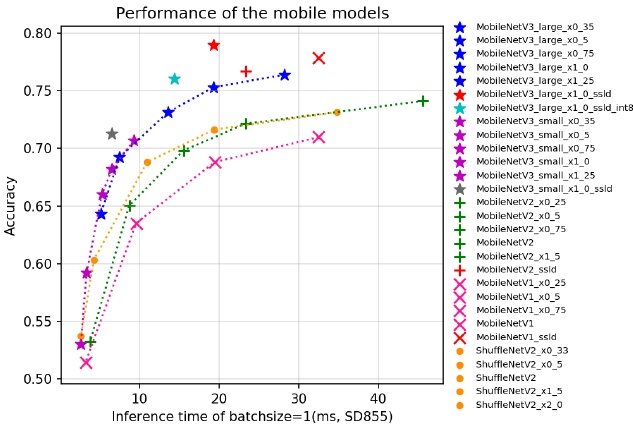


Figure 6: The performance of some light backbones on the ImageNet 1000 classification, including MobileNetV1, Mo- bileNetV2, MobileNetV3 and ShuffleNetV2 series. The in- ference time is tested on Snapdragon 855 (SD855) with the batch size set as 1.

MobileNetV3 can achieve higher accuracy when the pre dict time are same. As for the choice of scale, we adopt MobileNetV3\_large\_x0.5 to balance accuracy and efficiency empirically. Incidentally, PaddleClas provides a total of up to 24 series of image classification network structures and training configurations, 122 models' pretrained weights and their evaluation metrics, such as ResNet, ResNet\_vd, SERes- NeXt, Res2Net, Res2Net\_vd, DPN, DenseNet, EfficientNet, Xception, HRNet, etc.

Xception, HRNet, etc. Light Head The head of the text detector is similar as the FPN (Lin et al.2017) architecture in object detection and fuse the feature maps of the different scales to im- prove the effect for the small text regions detection. For con- venience of merging the different resolution feature maps, 1 1 convolution is often used to reduce the feature maps to the same number of channel (we use inner channels for