

Figure 11: Illustration of data augmentation, TIA. This fig- ure comes from the paper (Luo et al. 2020)

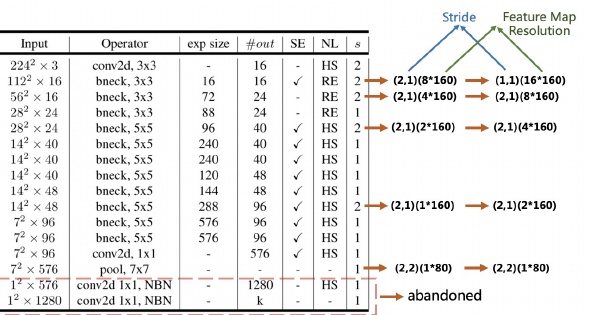


Figure 12: Illustration of the modify of the feature map reso- lution. The table comes from the paper (Howard et al. 2019)

11] at first, a set of fiducial points are initialized on the im age. Then move the points randomly to generate a new im age with the geometric transformation. In PP-OCR, we add BDA and TIA to the training images of the text recognition.

BDA and TIA to the training images of the text recognition. Cosine Learning Rate Decay As mentioned in text de- tection, cosine learning rate decay has become the preferred learning rate reduction method. The experiments show that cosine learning rate decay strategy is also effective to en- hance the model ability for text recognition.

Feature Map Resolution In order to adapt to multilin- gual recognition, particularly in Chinese recognition, in PP- OCR the height and width of the CRNN input are set as 32 and 320. Then, the strides of the original MobileNetV3 is not appropriate for text recognition. As shown in Figure[12 for the sake of keeping more the horizontal information, we modify the stride of the down sampling feature map except the first one from (2,2) to (2,1). In order to keep more verti. cal information, we further modify the stride of the second down sampling feature map from (2,1) to (1,1). Thus, the stride of the second down sampling feature map s2 affects the resolution of the whole feature map and the accuracy of the text recognizer dramaticly. In PP-OCR, s2 is set as (1,1) to achieve the better performance empirically.

to achieve the better performance empirically. Regularization Parameters Overfitting is a common term in machine learning. A simple understanding is that

the model performs well on the training data, but it performs poorly on the test data. To avoid overfitting, many regular ways have been proposed. Among them, weight\_decay is one of the widely used ways to avoid overfitting. After the final loss function, L2 regularization (L2\_decay) is added to the loss function. With the help of L2 regularization, the weight of the network tend to choose a smaller value, and finally the parameters in the entire network tends to O, and the generalization performance of the model is improved ac- cordingly. For text recognition, L2\_decay has a great influ- ence on the accuracy.

ence on the accuracy. Learning Rate Warm-up Similar as the text detection, learning rate warm-up is also helping the text recognition. For text recognition, the experiments show that using this strategy is also effective.

strategy is also effective. Light Head A full connection layer is used to encode the sequence features to the predicted characters in the ordinary. The dimension of the sequence features have an impact on the model size of a text recognizer, especially for Chinese recognition whose characters are more than 6 thousands. Meanwhile, it is not that the higher of the dimension, the stronger of the ability of the sequence features representa- tion. In PP-OCR, the dimension of the sequence features is set to 48 empirically.

set to 48 empirically. Pre-trained Model If the training data is fewer, fine tune the existing networks, which are trained on a large data set such as ImageNet, to achieve fast convergence and better accuracy. The transfer learning in image classification and object detection show the above strategy is effective. In real scenes, the data used for text recognition is often limited. If the models are trained with tens of millions samples, even if they are synthesized ones, the accuracy can be significantly improved with the above models. We demonstrate the effec- tiveness of this strategy through experiments.

PACT Quantization We adopt the similar quantization scheme of the direction classification to reduce the model size of a text recognizer except for skipping the LSTM lay- ers. Those layers will not be quantified at present since the complexity of LSTM quantization.

# 3Experiments

# 3.1Experimental Setup

DataSetsAs shown in Table1 in order to implement a practical OCR system, we construct a large-scale dataset for Chinese and English recognition as an example. For text detection, there are 97k training images and 500

For text detection, there are 97k training images and 500 validation images. Among the training images, 68K im- ages are real scene images, which come from some public datasets and Baidu image search. The public datasets used include LSVT (Sun et al.[2019), RCTW-17 (Shi et al.[2017) MTWI 2018 (He and Yang|2018), CASIA-10K (He et al. 2018), SROIE (Huang et al.2019), MLT 2019 (Nayef et al. 2019), BD1 (Karatzas et al.2011), MSRA-TD500 (Yao et al. 2012) and CCPD 2019 (Xu et al.2018). Most the training images from Baidu image search are document text images. The remaining 29K synthetic images mainly focus on the scenarios for long text, multi direction text and table text. All the validation images come from the real scenes.