A Training Strategy to Optimize the Performance of Target Detection Model

Ninghui He[1*], Pei Zhang [1], Xiu Zhou [1], Shitao Liu [1], Jinpeng Hao [1], Rui Liu[2]
[1.] State Grid Ningxia Electric Power Co. Ltd, Yinchuan 750000, China
[2] Beijing Smartchip Microelectronics Technology Company Limited, BeiJing City, 100000, China
**Corresponding Author.*

Abstract

Although neural networks have made tremendous progress in feature extraction, each year new models refresh the accuracy of previous models on major data sets, which also shows that there are still a considerable number of effective features in the data set. This paper presents a model and a training strategy to fully exploit the effective features in the data set. The main feature of the model is the use of focal loss based on gradient cropping; the main steps of the strategy include: (1) build a model that contains all necessary optimization techniques, and debug the model capacity to the highest accuracy, get an initial model; (2) corrects the missing and wrong labels of the data set based on the initial model to obtain the training set 1, training the model, and obtaining the model 1; (3) extracting the FN and FP as the training set 2, training the model, and obtaining the model 2; (4) repeating (3) ; (5) NMS is used to merge the prediction results of filtering model 1 to model n. Experiments show that, based on the model obtained by this strategy, MAP increases by 2.7%.

Keywords: *Deep Learning; Neural Network; Computer vision; Object Detection; Focal Loss; Hard Sample Mining*

I. Introduction

It is an indisputable fact that neural networks show excellent feature extraction capabilities in the field of machine vision. Since the birth of AlexNet, the constantly improved neural network has also continued to refresh the accuracy of neural networks in the fields of image recognition, target detection, and image segmentation. What's interesting is that the latest neural networks can always improve the accuracy, which shows that there are still valid features in datasets such as ImageNet that have not been mined.

The current neural network model is developed based on statistical theory. Although we have reason to believe that such a neural network is not the most effective in processing features, it is likely that there is a better structure that uses fewer samples to achieve the purpose of learning. But with the continuous increase of datasets, existing neural networks can also rely on training on large amounts of data to achieve good learning results. Moreover, there is no experiment to prove that there are any characteristics that the current neural network cannot learn based on a large amount of data. In the field of machine vision, we cannot directly know the distribution of features in the dataset as in the field of data mining. However, we can assume that all natural datasets always obey the same distribution, that is, the normal distribution. We assume that the number of effective features in the image dataset is normally distributed, and the neural network also extracts features based on the number of such distributed features.

II. Related Work and Existing Problems

When using deep learning algorithms, we sample the training set, and then select parameters to reduce the training set error. Use the trained model to verify the test set to get the generalization (test) error. In this process, the generalization error expectation will be greater than or equal to the training error expectation. We should work towards the following to improve the accuracy of the algorithm:

Reduce training error;

Reduce the gap between training error and generalization error.

These two factors correspond to the two main challenges of machine learning in figure 1: underfitting and overfitting. Underfitting means that the model cannot obtain a sufficiently low error on the training set, while overfitting means that the gap between the training error and the test error is too large.

By adjusting the capacity of the model, you can control whether the model is biased towards over-fitting and under-fitting. More simply, the capacity of which is the ability of the model to fit a variety of functions. Models with low capacity are difficult to fit the training set, models with high capacity may overfit [1].

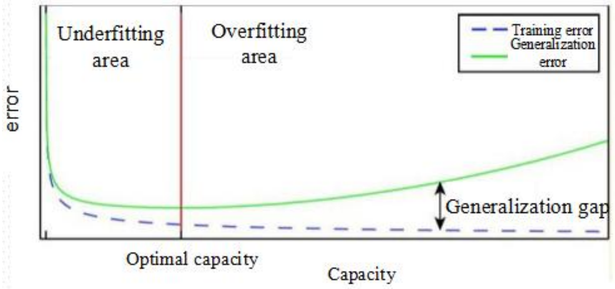


Figure 1 The relationship between model capacity and generalization error

However, based on our previous assumptions, the number of effective features in the dataset is normally distributed, that is, the imbalance of feature distribution is common in the dataset. As the model capacity increases, the problem of feature imbalance cannot be solved.

To deal with the problem of category imbalance, Jiankang Deng et al. proposed focalloss to solve this problem [2]. In fact, the imbalance of the category, in essence, also belongs to the imbalance in the number of features. In theory, we can use focalloss to deal with feature imbalance.

The number of features is inversely related to loss. The loss distribution that focalloss expects to see is as shown in the figure 2. In other words, the expected result of focalloss is that as long as the number of features can be balanced by focalloss, all features can be learned by the network [3].

This is not the case.,neural networks need to be trained based on large amounts of data. Based on the assumption of normal distribution of features, there are some features in the dataset, and their number is not enough to support the training of the neural network, and the model cannot normally converge on this part of the data [4].

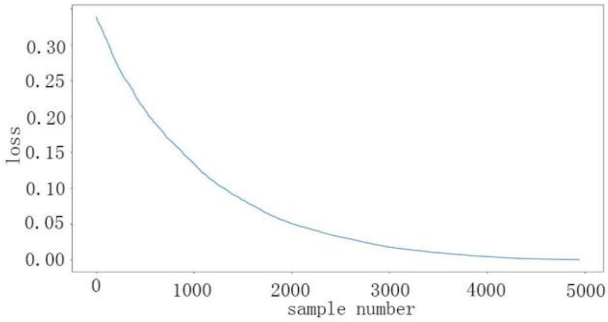


Figure 2 The relationship between the number of features expected by the modulus focalloss and loss

Figure 3 shows the loss performance of the model on a dataset. The larger the loss, the smaller the number of samples corresponding to the loss value, that is, the smaller the number of samples corresponding to the feature.

Theoretically, there is a critical point for each feature. When the number of features is less than the critical value, the model will not be able to effectively fit the features, which manifests as an overfitting of the sample.

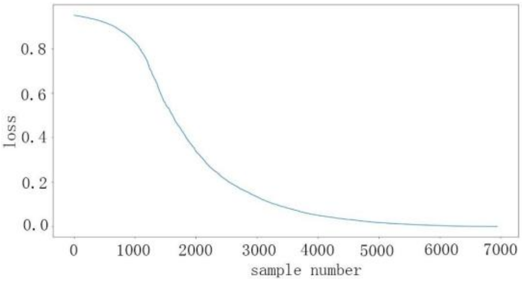


Figure 3 The relationship between the number of features and loss in practice

If the proportion of this part of the data set cannot be ignored, paying too much attention to this part of the unlearnable samples will cause the model to become unstable on the entire dataset. In fact, we found through experiments that when the γ value of focalloss is set to be large in fugure 4, large fluctuations in model loss do occur, and the model does not converge.

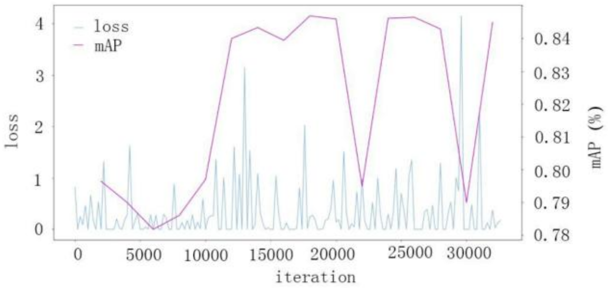


Figure 4 The characteristics of the loss curve in the case of a large γ value

Therefore, in solving the problem of feature imbalance, focalloss is not perfect. If there are abnormal points in the sample, or the number of difficult-to-separate sample features is not enough to make the model converge, the effect of focalloss's strategy of focusing on difficult-to-separate samples will be greatly reduced [5]. In fact, when we remove the part of the sample (about 15% of the total number of samples) where the number of features is not enough to train the model, the accuracy of the model is not affected. This further shows that the direct use of focalloss on the actual dataset is undesirable.

III. Our Work

As mentioned earlier, we assume that the number of features in the natural dataset is normally distributed. This dataset has the following characteristics:
(1) The number of features is normally distributed, and there is an obvious feature imbalance.

(2) There is part of the feature, which each feature corresponds to the number of samples is very small, the model cannot learn this part features correctly, while the total number of samples corresponding to these features and a lot of it.

(3) Due to the influence of subjective and objective factors such as manual labeling, it is inevitable that there are abnormal samples such as incorrect labeling in the dataset.
Aiming at this characteristic of the dataset, we try to give a reasonable model and a training strategy to fully mine the effective features in the dataset.

3.1 Model building

The use of focallloss on actual datasets is limited, but the problem of feature imbalance is widespread in the dataset [6]. Unlike focallloss which only pays attention to the loss value of the feature, we believe that we should also pay attention to whether the number of features has reached the level of learning. Based on these constraints, we believe that a model suitable for actual datasets should be able to deal with the problem of feature imbalance and the interference of unlearnable features.

In the training process, the loss of the corresponding sample of the learnable features tends to 0 with the iteration of the training, and for the non-learnable features, the model must overfit the sample to reduce the loss of the sample, and when the amount of data is large under the circumstances, the model capacity is difficult to meet this requirement. Therefore, the sample loss corresponding to this part of the feature shows that the fluctuation does not converge. When the sample corresponding to the learnable feature tends to 0, this part of the loss will interfere with the convergence of the model [7].

In order to deal with this situation, theoretically speaking, it is necessary to limit the value of this part of the loss. That is, gradient clipping is performed on the model. Therefore, we adopt gradient clipping on the basis of focalloss [8], Among them, focalloss is used to effectively solve the problem caused by feature imbalance, and at the same time, the gradient of the backpropagation is cropped, so that the model does not pay too much attention to difficult samples. Theoretically, the loss of the samples corresponding to this part of the unlearnable features will be prominent in the later stage of training [8]. In order not to affect the convergence speed, it is recommended to perform gradient clipping when the training is close to convergence.

The above are our requirements for the model, and other techniques that do not destroy the effect of the above combination are also allowed to be added. In the model we used, more verified techniques have been added to improve the accuracy of the model, including label smoothing and CIOU loss. The main structure of the model uses Yolov4 [9].

3.2 Model

As mentioned earlier, the latest neural networks can always improve accuracy, which shows that there are still valid features in datasets such as ImageNet that have not been mined. And there is no experiment to prove that there are any characteristics that the current neural network cannot learn based on large amounts of data. This paper presents a boosting-based training strategy for fully mining useful information in big datasets, so as to obtain a better accuracy on the basis of the SOTA model [10]. The strategy mainly includes the following steps:

3.2.1 Model capacity determination

The influence of model capacity on training results is fundamental and crucial. Too small a capacity will lead to under-fitting, otherwise it will lead to over-fitting. The capacity of the model can be modified by adjusting the number of layers of the model and the parameters of each layer.

3.2.2 Perform missing and mislabeled corrections on the dataset to obtain training set 1, and train the model to obtain model1

Data labeling is ultimately done through manual labeling. Missing and mislabeling are often inevitable. While adjusting the model capacity in step (1), we recommend setting appropriate confidence thresholds and region of interest thresholds to detect the entire dataset. And from the false negative samples and false positive samples to extract mislabeled and missed samples for labeling and correction. As the accuracy of the model continues to improve, new samples of missed and mislabeled samples are often mined. In fact, we suggest to correct the missing and mislabeled datasets after each next step of training.

Perform finetune based on the dataset that removes the missing and mislabeled data, and get the first model.

3.2.3 Extract false positive samples and false negative samples as training set 2, train the model, and get model2
Use model 1 to detect the training set, and use false positive samples and false negative samples as a new training set for training to obtain model2.

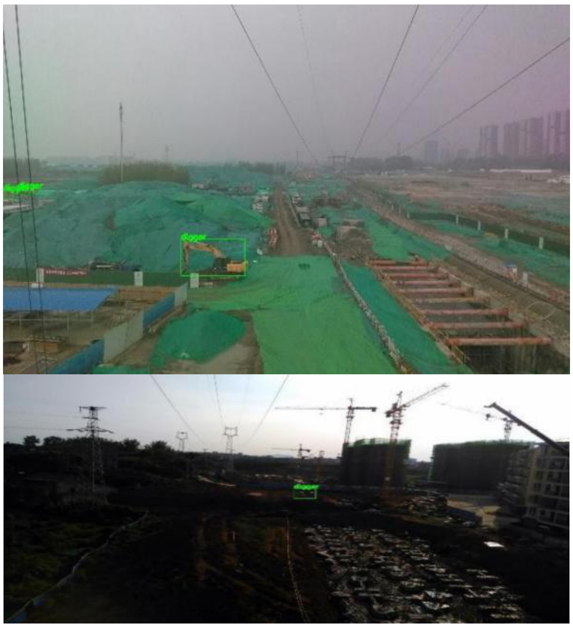
The method used in this step is hard case mining, and there is a problem with simple hard case mining. According to our experience, when finetune is performed for difficult cases, it will inevitably lead to a certain degree of degradation of mAP that is easy to sample detection results.

3.2.4 Repeat step
Use model2 to detect training set 2 and obtain new false positive samples and false negative samples, which are used as training set 3. Model3 is trained based on this dataset.

3.2.5 Use NMS to merge and filter the prediction results from model1 to model.

IV. Experimental Dataset

We extract the excavator category from the transmission line channel detection dataset as the experimental dataset. As shown in the figure 5, the total number of targets in the dataset is about 27k, most of which are small targets. The experimental dataset is randomly sampled and divided into training set, validation set and test set at a ratio of 0.8:0.1:0.1.



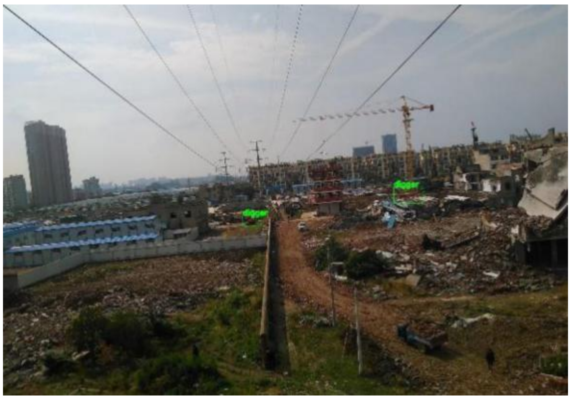


Figure 5 Example of training data set

V. Result

We adopted the model and strategy defined above, and we conducted experiments on a single category digger dataset. The following table 1 shows the superimposition effect of the three models. Some previous papers have proposed difficult mining methods to improve mAP. From the table 1, we can see that the mAP from model1 to model3 is gradually improved. The mAP of model3 even exceeds the mAP of model1+model2.This proves the effectiveness of hard case mining. We further merge the model results and we can see that mAP will continue to improve. Compared with model1, mAP has increased by nearly 2.7%.

Table 1 The MAP value of the model trained in each step

Model	mAP（%）
Model1	73.54
Model2	74.13
Model3	75.02
Model1+model2	74.83
Model1+model2+model3	76.23

The following figure shows the loss performance of model1, model2, and model3 on data set 3. Among them, -1 means missed detection.

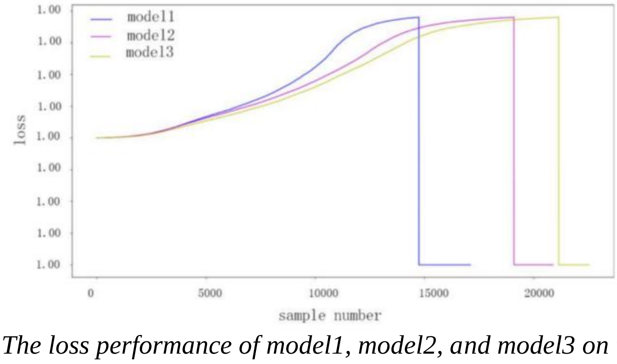


Figure 6 The loss performance of model1, model2, and model3 on the dataset

It can be seen from the figure 6 that as the training steps increase, the number of missed targets decreases. This is consistent with our expectations. On the other hand, we can see that the total number of targets is also increasing. The newly added targets are roughly composed of three parts, one is the corrected missed target, the other is the newly detected missed target, and the other is the misdetected target.

Comparing the number of missed detections (-1) and the number of false detections of the three models, we can see that when the number of false detections increases significantly, the decrease of the number of missed detections is relatively small. This part of the missed target is what we think is the unlearnable sample. That is to say, there are some specific number of features, and the number of corresponding samples is not enough to support network training. Therefore, there is a critical point. When the number of features in the data set is lower than the critical point, the model no longer has the ability to learn this feature. When the number of such features in the data set is large, it is very unfavorable to the convergence of the focalloss-based model, just like the loss fluctuation mentioned in the previous figure 6.

In fact, when we remove this part of the sample from the training set, the model training based on focalloss reduces the mAP fluctuation a lot in the later stage of training.

VI. Conclusion

This paper combines the characteristics of general data set feature imbalance, mislabeling and missing labeling, and unlearnable features, and gives a usable model and training strategy to fully mine the effective features in the data set, which improves mAP by nearly 2.7%. In addition, our experiments also show that because of the ubiquity of abnormal data, the focalloss that requires relatively high quality in the data cannot often cope with the accuracy drop caused by the imbalance of the category. In fact, it only plays a role in mitigating.

References

[1] S. Bell, C.L. Zitnick, K. Bala, R. Girshick, “Insideoutside net: Detecting objects in context with skip pooling and recurrent neural networks,” In CVPR, July, 2016.

[2] S.R. Bulo, G. Neuhold, P. Kontschieder, “Loss maxpooling for semantic image segmentation,” In CVPR, vol. 3, 2017.

[3] J. Dai, Y. Li, K. He, J. Sun, “R-FCN: Object detection via region-based fully convolutional network,” In NIPS, vol. 1, 2016.

[4] N. Dalal, B. Triggs, “Histograms of oriented gradients for human detection,” In CVPR, vol. 2, 2005.

[5] P. Dollar, Z. Tu, P. Perona, S. Belongie, “Integral channel ´features,” In BMVC, vol. 2, pp. 3, 2009.

[6] D. Erhan, C. Szegedy, A. Toshev, D. Anguelov, “Scalable object detection using deep neural network,” In CVPR, vol. 2, 2014.

[7] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, “The PASCAL visual object classes (VOC) challenge. IJCV, vol. 2, 2010.

[8] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, “Cascade object detection with deformable part models,” In CVPR, vol. 2, pp. 3, 2010.

[9] C.Y. Fu, W. Liu, A. Ranga, A. Tyagi, A.C. Berg, “DSSD: Deconvolutional single shot detector,” arXiv:1701.06659, vol. 1, no. 2, pp. 8, 2016.

[10] R. Girshick, “Fast R-CNN,” In ICCV, pp. 1, 2, 4, 6, 8, 2015.