University of Colorado Boulder

The Bigger, the better? Resnet50 vs Naive CNN on CIFAR-10

Sijia Ge Guiye Li

Problem Space

To improve the performance of machine learning models on specific problems, we are building increasingly complicated and larger models nowadays. Most of them are computationally expensive and data-heavy.
However, do we really need such a large model? Garbage in, garbage out. We probably need to pay more attention to the data quality itself since the ability of the model might be limited by the quality of the data. Or the job of a large classification model is actually the combination of data improvement and simple classification? We will conduct comparative experiments to see if adequate preprocessing that aims to improve the quality of the image can help simpler classification methods reach the SOTA accuracy achieved by huge models.

Dataset

CIFAR-10 is our dataset, the total number of samples is $60,000$ and it has been split by the builders based on the ratio of $5:1$ with a training set and test set. Each sample is a matrix with the shape of $(32, 32, 3)$, means the size is 32×32 and the number channel is 3 (RGB) representing red, green, and blue respectively. Each pixel value in the matrix is between $0 - 255$. We use the training set only for the current steps, and split them based on the ratio of $8:2$ into a 'train set' and a 'validation set'.

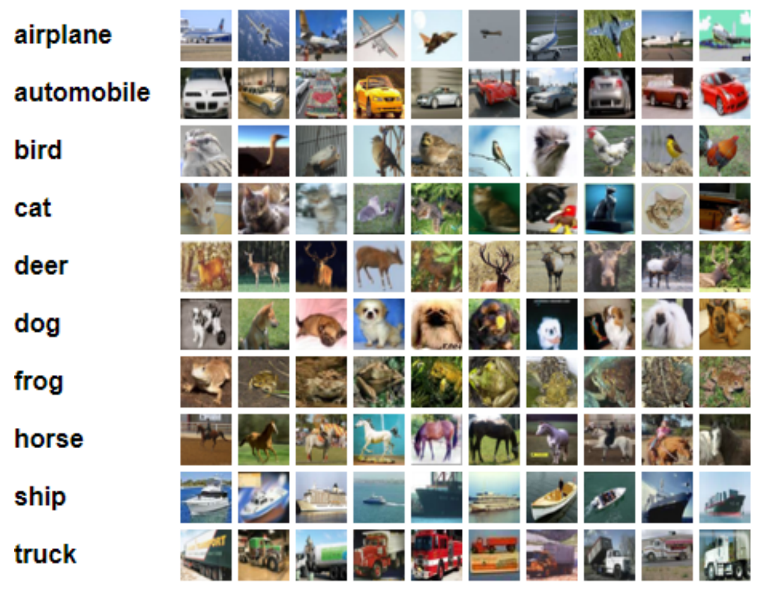
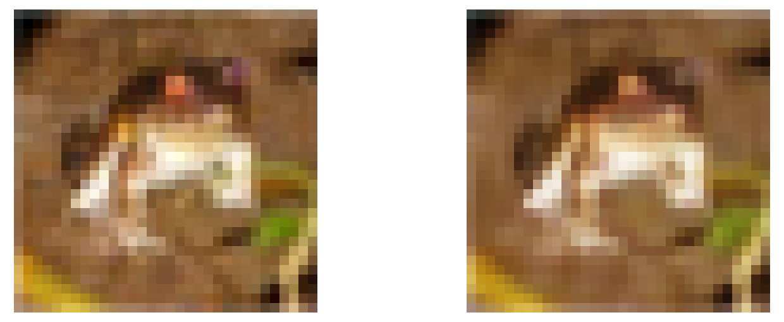


Figure 1. Cifar-10 samples



(a) Ground Truth

(b) Denoising by DIP

Figure 2. A denoising sample generated by DIP.

Dataset pre-processing

Denoising

We have tried using some deep-learning-based denoising models such as Deep-Image-Prior (DIP) on the official CIFAR-10 dataset. Unfortunately, there is no noticeable effect since there is not enough noise in the image.

Fig 2 shows a random sample from the official dataset and the corresponding denoising result by DIP. There is almost no difference between the two pictures.

Manually adding noise

Therefore, we decided to manually add Gaussian distribution noise to experiment with the effect of noise on classification accuracy. Fig 3 shows a sample image randomly chosen from the training set. The image from left to right is the original version, adding noise $z \sim \mathcal{N}(0, 0.01)$ and the image adding noise $z \sim \mathcal{N}(0, 0.04)$. The original version dataset will perform the role of the cleaned dataset and we are trying to analysis the influence to model performance by adding different levels of random noise. Our experiment shows that the DIP can achieve the result that denoising the noise data into the original one after tuning (not included here).

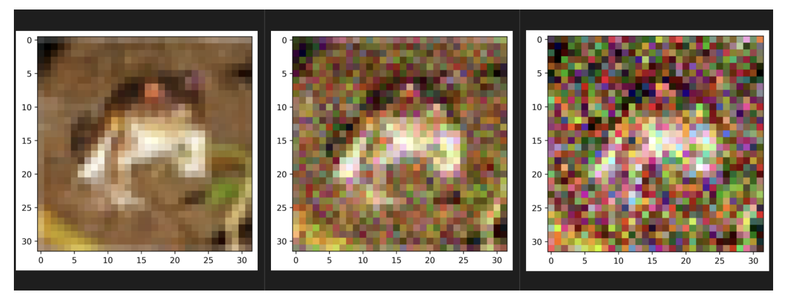
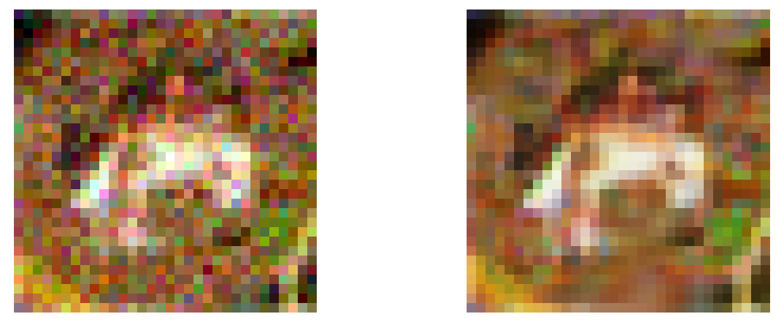


Figure 3. A sample image randomly chosen from the training dataset of the cleaned version and noisy version datasets.



(a) Image with noise

(b) Denoising result

Figure 4. DIP denoising example.

Classification models

In our experiments, we used two different models, one is a simpler Convolutional neural network model built from scratch, the model architecture shown as Fig 5, and the total number of parameters is $111,114$; The other model leverages the Resnet-50 with pre-trained weights on the ImageNet dataset and fine-tuned it on our dataset, in order to classify, we removed the last layer and added some extra layers at the end. The number of parameters is $23,859,978$. The architecture shown as Fig 6.

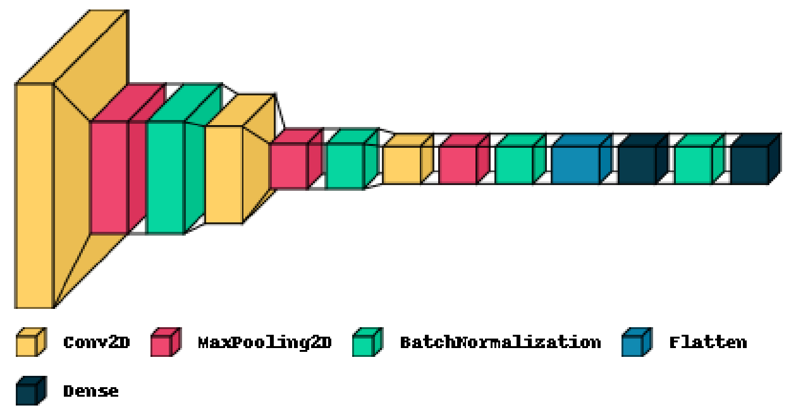


Figure 5. simple CNN model architecture

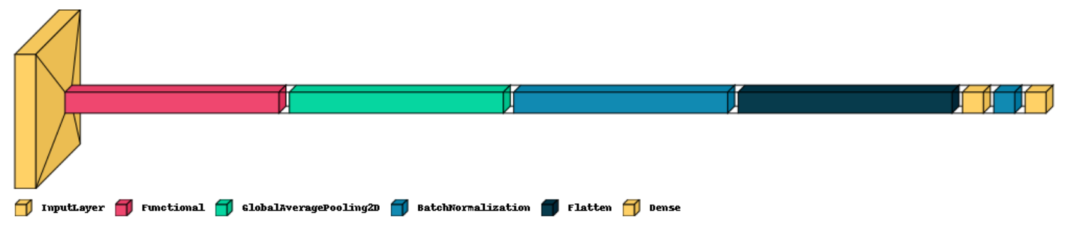


Figure 6. complex model architecture including ResNet50 pre-trained model

We conducted 6 groups experiments, we utilized two models to train on three types of data with different levels of noise. we also recorded the training time, and for the complex model, we recorded the result without pre-trained weight as well. The results show as as we can see, the simple CNN model can run on the original data to achieve the result of the complex model trained on the noise data, and the training time is less than 7 times of that on the complex model, which signifies that we might need to pay more attention to the preprocessing such as denoising.
Besides, we conduct the further experiment with different hyperparameters, one variable is the optimizer. The above experiment was conducted with the SGD optimizer, while if we turn into the Adam optimizer, we find out that, with the Adam optimizer, the result on the complex model does not change a lot while the performance on the simple model optimizes a bit, making the difference of the two models is smaller and smaller, we might consider that Adam requires more hyper-parameters to tune, and the default settings impacts the training ability of the complex model.
Furthermore, we also find out that the size of the data impacts differently on the two types of models. As we upsampling data by nearest, the performance on the complex model improves a lot while making no big difference on the simple models, which signifies the ResNet might be more sensitive to the size of the data.

Classification accuracy comparison

Table 1. Classification accuracy comparison between different classifiers and datasets using optimizer **SGD**.

Datasets	**Simple** classifier	**Complex** classifier(w,n.w weights)
Cleaned dataset	$0.649/1.24$ mins	$(0.782, 0.576)/(9.44, 9.44$ mins)
Noisy dataset $z \sim \mathcal{N}(0, 0.01)$	$0.59/0.94$ mins	$(0.725, 0.552)/(7.05, 7.45$ mins)
Noisy dataset $z \sim \mathcal{N}(0, 0.04)$	$0.53/0.97$ mins	$(0.651, 0.514)/(7.45, 7.44$ mins)

Discussion

Our experiments compare the performance of different 'complexity' models with different levels of noisy data. our conclusion mirrors that cleaner data on a much simpler model can achieve the performance on a complex model which is trained with noisy data. The more complex model is capable of modeling more complex data space, however, the data quality also impacts the performance. Compared to training complex models at a cost of longer time and computing resources, we might also turn into the initial and original process for data cleaning and preprocessing, this work might inspire researchers to focus not only on the model design but also on the data. Furthermore, we also find out that factors such as optimizer and data size would impact differently on the different models, which might spark the tuning and training process for researchers.
In future work, we might explore further the explainability such as analyzing the feature maps generated by different models and figuring out the secret of the 'Black Box'.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[2] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3): 379–423, 1948. doi:10.1002/j.1538-7305.1948.tb01338.x.

[3] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018.