Dog and Cat Classification with Deep Residual Network

Yao Yu chen
Software School of North University
China
image_cl@163.com

ABSTRACT

I With the development of artificial intelligence, the deep neural network(DNN) has achieved excellent results in image processing domain such as image classification[1] and objct detection[2].The convolution neural networks(CNN) [3] is a Representative algorithm of DNN which have the representation learning ability . According to its convolutional structure, input information is extracted with translation invariance.Based on the widely used CNN ,there are many efficient models.For image classification there are Lenet-5[4],VGG[5], Resnet[6] and so on.For object detection ,the yolo series[7] is well-known.Also few well known datasets are proposed to measure their performance such as ImageNet and Cifar-10[8]. These data sets are dedicated to the classification of multiple objects in natural scenes.Nowadays ,pets play an increasingly important role in our life,so we built a cat and dog dataset, each of which categories with 12500 samples which is larger then 1260 in Imagenet.For our dataset,we trained an image classification model.We focus on the performance of distinguish dog and cat In different scenes, lighting and noise.Our method achieved an accuracy of 92.7 percent and remained robust under adversarial attack.

CCS concepts

• **Applied computing→Computer forensics→Data recovery**

Keywords

Neural network, image classification, convolution neural networks(CNN),adversarial attack

1. INTRODUCTION

In recent years, the image classification task continues to make progress. From the kernel based SVM[9] to the recent neural network models.The accuracy of image classification has been improved considerably. Convolutional neural networks are usually effective in large-scale image and video processing and is often used to extract features from images and adjust the dimensions. Because convolutional neural networks mimic human visual mechanisms,they are widely used in vision tasks.Convolutional Layer is generally used with Regularization layer and Activation function layer simultaneously.Many articles gain pretty result by constructing models based on image classification on account of CNN, for example, lenet-5 reduces the dimensions of the image with two Convolutional Layer and two Pooling Layer, then adjust it to a ten-dimensional vector using the FC(Fully Connection) layer corresponding 10 categories. VGGnet uses several continuous convolution kernels of 3x3 to replace the larger convolution kernels. It also introduces parallel paths, which guarantee a small cost for a given perceptive field.

Datasets is the measurement of models.The current mainstream image classification dataset like mnist[10], which contains ten categories, 6,000 pictures of handwritten numbers, with 28*28 in size and only a single channel.The imagenet dataset contains one million images with one thousand categories. Most of the models use Imagenet for training and testing.Our purpose is to discern dog and cat in different situations,so we made a special dataset for the classification of dogs and cats, which contains 25,000 colorful pictures of size 224*224, and the number of samples of each type is up to 12,500. Moreover, the data distribution is balanced, including samples of different size and backgrounds. we randomly choose 15 cat and 15 dog breeds from Egypt cats, British shorthair, Ragdoll , Huskie , Golden Retriever , Alaskan Malamute and so on which includes most of domestic pets.

We built a model of dog and cat classification based on convolutional neural network and previously-introduced dataset. We design the model maily composed of stacked residual layers and spatial transformer network(STN)[11].The STN generate Affine transformation matrix by a sub-network for the backbone.Our model train with space-invariance after the transform.The backbone is designed based on residual method,which is a best way to avoid gradient vanish and explode.During traing,we choose cross-entropy[12] as loss function,adamax[13] as optimizer,which proved to be suitable for image classification task.

The accuracy have reached 92.1 percent after ten epoch-trained, which is far higher than that of SVM (75%) and VGG (87.5%). In addition, we take some steps to make the model more strong.First we take image augmentation to have a balance of sample distributions.Second we take batch normalize and leaky relu with each convolutional layer.Third we use adamax as optimizer,which which constrain learning rate according to iterations and gradient,What's more.So,our model has considerable robustness since the accuracy can still maintain at 90% under those attack based on FGSM[14].Also,we test the time cost on different machine.Our model costs the average of 60ms on cpu to classify a image is dog or cat,while 15ms on gpu with NVIDIA1660ti.It is a acceptable time for daily use.

ESSE 2020, November 6–8, 2020, Rome, Italy

ACM ISBN 978-1-4503-7762-1/20/11…$15.00
https://doi.org/10.1145/3393822.3432321

Figure1. Images in our dog and cat dataset

2. MODEL

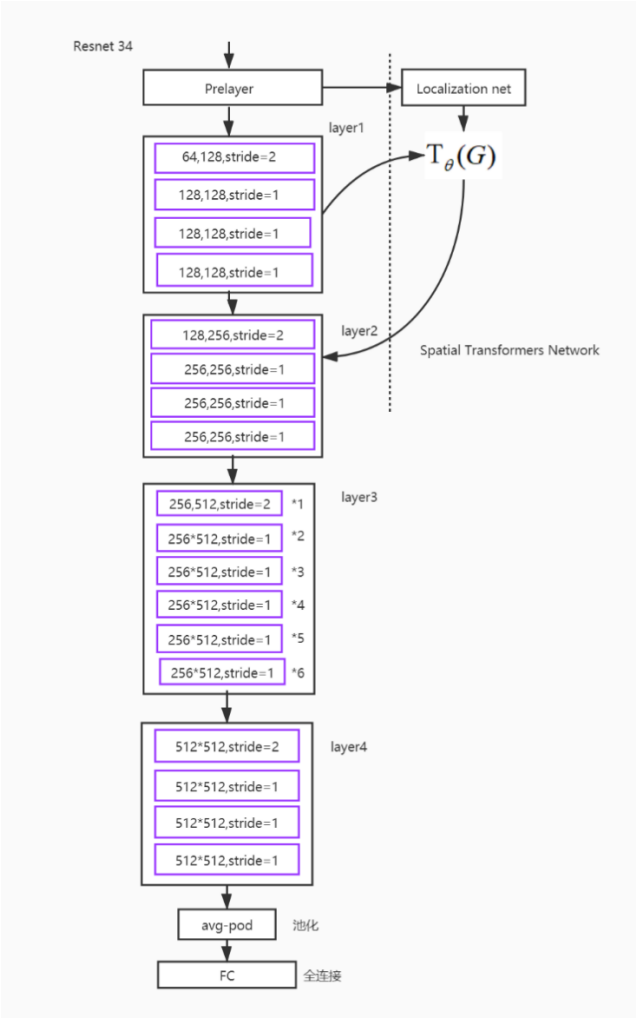

Figure 2. The overall design of the model structure

2.1 Model Design

We designed the model based on residual module and Spatial transforms module.The overall architecture is shown as figure2.The left side is the backbone which is responsible for feature extraction. The right side is called Spatial transforms module which adjust feature map so as to centralizer image.

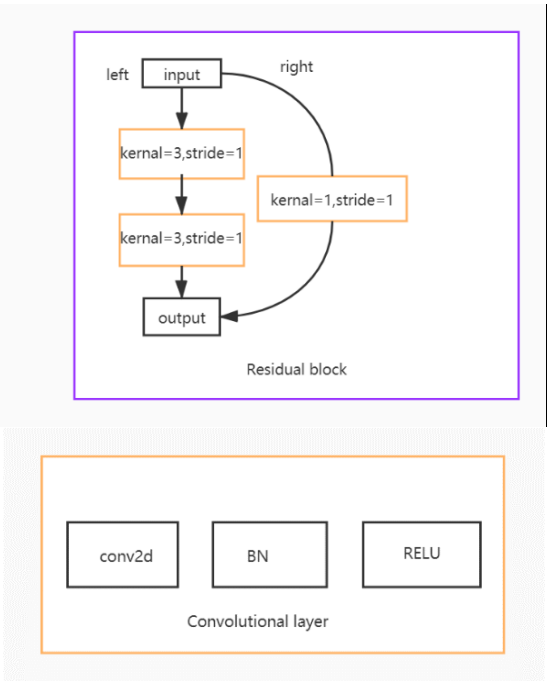

Figure 3. Two important modules in the model

2.2 Backbone

We initially used the pre-layer to increase the number of channels to 64 so as to compressed input images. On the mean time, Localization net caculates the Affine transformation matrix. Then we carried out feature extraction through four layers composed of residual modules. In each layer, the first residuals module is responsible for dimension adjustment, while the following residuals module keeps the dimension unchanged and focuses on the extraction of image information. Since the information of the feature map will be richer as the model goes on, we included more residual modules in the later layers. At the end of the network, a fully connection layer maps the feature map to a two-dimensional vector. We normalized the vector by softmax function, corresponding to two categories of cats and dogs, and selected the larger one as the predicted result.

The backbone are mainly composed four layers which named layer1-4 in figure2.Each of them are composed of few residual blocks with different hyper-parameters. We mark him in blue corresponding to its component in figure 3. There are two paths in a basic residual module such as figure 3 shows, the path on the left is computed by two convolution layers. Also the convolutional layer is marked as orange corresponding to its component in figure 3.The first convolution layer is used to compress the feature map with kernal-size 3, stride 2. The second one is used to extract features without changing the size with kernal-size 3, stride 1, padding 1; the path on the right adds the output to the left of the path directly. It can be summarized as the formula 1.

$$F(x) = x + f(x) \tag{1}$$

The f(x) indicates left path,while indicates right path which keep the dimension of image changeless.During backpropagation.The gradien of residual block can be caculated as formula 2. When the gradient of right side vanish,we can still do backpropagation through left side.

$$\frac{\partial F'(x)}{x} = 1 + f'(x) \tag{2}$$

In figure3,the convolution module consists of convolution layer, regularization layer and activation function layer (Leaky RelU) as shown in figure3, which can enhance the performance of the model and prevent overfitting compared with the simple convolution layer.We choose Leaky Relu as activation function because it soves the Dead Relu problem by applying a tiny coefficient on data when taking a negative value.In ourexperiments,the Leaky can not only improve the model performance,but also make the model more robust.

2.3 Spatial Transform Module

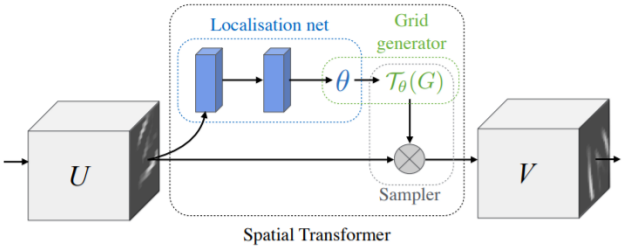

Figure 4. The plug-in module of spatia transform network

The spatital transform module is used to modify the geometric distortion in sample images.First,we suppose the transform can be described within an affine matrix.The module is totally drawn in figure 4, T is the matrix to modify picture U.Localisation net is a smaller network which aim to generate proper matrix T .We design the Localisation net with a four layer fully connection network,which inputs a flatten image,outputs the 2*3 affine matrix.

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \mathcal{T}_\theta(G_i) = \mathtt{A}_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \tag{4}$$

Just as formula(4) shows that adjusted image matrix (x_i^s, ,y_i^s) is caculated by the matrix multiplication of original image matrix(x_i^t, ,y_i^t,1) and transform matrix [θ].The transform matrix can express geometirc transform such as translation,scaling,rotation and shear.For example ,The translation operation Shift the original image along the x and y directions θ_{13} and θ_{23} respectively.That is $x' = x + \theta_{13}$, $y' = y + \theta_{23}$.In this condition,the transform matrix can be expressed as$\begin{pmatrix} 1 & 0 & \theta_{13} \\ 0 & 1 & \theta_{23} \end{pmatrix}$,which is an sub-condition of former .

In our model,we caculate the transform matrix by the input of feature image after pre layer.And the we ues the fine-caculated transorm matrix to adjust fature map between layer1 and layer2.Experiments show that with the additional spatial transform module,the total model is more likely to restrain.Although it is not so helpful to boost the final accuracy of classify, The model can save the time of training,which is a mothod of Tradding space for time.So,there is a question that is th spatial transorm module can boost any manifestion during fature maps?We have done some experiments,which shows that only in shallow layers in network it perform well,that is why we add them between layer1 and layer2,not layer3 or later.

3. EXPERIMENT

3.1 Set Up

We have totally collected 25000 pictures which are divided into cats' and dogs' categories. Training set includes 22,500 pieces and testing test includes 2500 pieces, the division of two sets is according to the ratio of 9:1. When we chose pictures, we respectively selceted different Numbers of samples with various backgrounds and sizes in order to make sure the distribution of samples is reasonable. Meanwhile, the possibility of overfit is avoided, making the model become stronger. We used torchvision to pack the dataset and encapsulate as an iterable object when we read data. We've made all the images 224 by 224 by 3, the length and width are 224 and the number of channels is 3 (RGB). Besides we did random clip and rotation, the diversity of images is guaranteed after data enhancement.

3.2 Train

In the training phase, we aim to make a closer approach to optimize the neural network. As the formula 3 shows, we defined the loss function $J(\theta)$, and optimized it by gradient descent. $h_\theta(x^{(i)})$ indicate the result predicted by our model,while $y^{(i)}$ indicate the true label.During training,we optimize $J(\theta)$ to make the prediction closer to label. We took the parameters in network as low as possible to avoid overfitting. Meanwhile We took the parameters in network as low as possible to avoid overfitting.So we add the l2 norm of parameters θ_j^2 into the loss function.

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right] \tag{3}$$

We selected batchsize 50, learning rate 0.01, with a total of ten rounds of iteration for optimization. We first put the data into the designed network, computed the predicted value via forward propagation and defined the loss function to. measure the gap between the predicted value and the one-hot encoding of the correct label. Because there are two classes, we used one-hot encoding.We introduced L2 regularization[15] and added two norm 0.01 times of all parameters in the loss function with the purpose of improving the performance of the model and avoid overfitting. And we took the gradient of the loss function and then linearly updated the network in the direction of gradient descent aimed to adjust the network. In order to balance the relationship between learning rate, gradient and epoch, Adamax was selected as the optimizer.

3.3 Test

During the test, we first loaded the model with the lowest loss among the previously trained models, that is the model developed in the eighth rotation training. Then the batchsize was selected to be 50, and the prediction vector was obtained by forward propagation. We selected the dimension of the maximum value as the result corresponding to the one-hot coding. At last, iterating

through all the batches to obtain the final accuracy, and the accuracy of this model was 92.1% after testing.

3.4 Result

In order to evaluate our model, we have carried out contrast experiment and adversarial attack experiment.In constract experiment we take SVM and AleexNet as baseline.In adversarial attack experiment,we use FGSM to attack our model so as to verify the robustness and anti-interference capability.

SVM is an image classification method based on hyperplane division, which performs well in the small datasets. We first used PCA to reduce the dimension, and then used SVM classification based on kernel function. the accuracy of the test set reached 76 percent

VGG was second in the 2014 ILSVRC[16] competition which built by Visual Geometry Group in Oxford university.As it is a good image classification model,most articles take it for baseline.We take the same super parameter configuration for VGG and the model in this article. After ten rounds of iteration, the variation of loss is shown in the figure 4. Both experiment showed that during the training process, loss decreased steadily with little rebound phenomenon, and tended to be stable after the seventh round, roughly within the range of 0.15 to 0.2, as shown in the figure 4,which proves the convolutional based neural network is efficient for dog and cat classification.It can be seen that the value of loss decrease faster in our model. And our model became stable after the fifth iteration,while VGG became stable after eighth iteration,which proved that our model was more expressive. VGG's accuracy on the test set reached 87.5%, slightly lower than the 92.1% of the model in this paper.

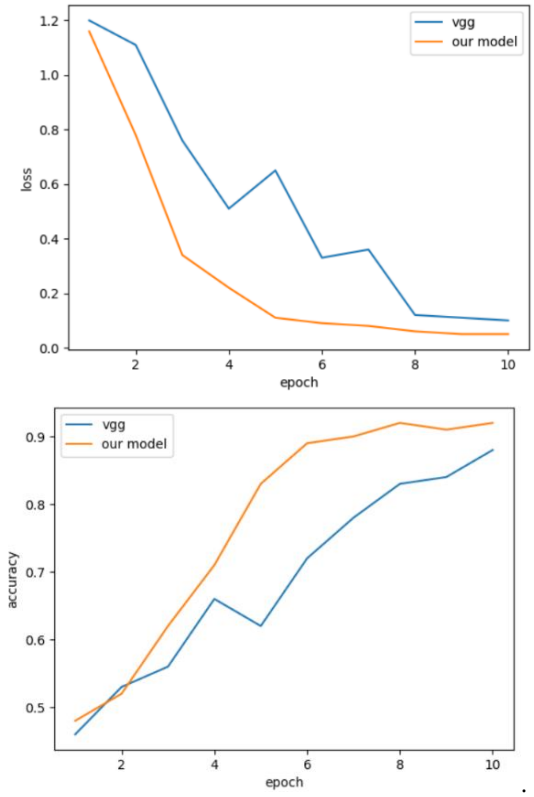

Figure 5. (up)The loss function changes with the epoch in our network and VGG network(down) the accuracy changes with the epoch in our network and VGG network

3.5 Robustness Test

As there are many adversarial sample to attack the neural network,the safety of neural network model is a big problem.So,it is necessary to verify the robustness of our model.the widely used adversial method is FGSM,PGD and so on.

Table 1. Performence under various attack.

model	accuracy	FGSM attack	Time (ms)
Our	0.92	0.88	15
VGG	0.88	0.71	25
SVM	0.75	0.68	72

Fast Gradient Sign Method (FGSM) is a effective white-box one-step attack method to generate adeverarial samples which based on gradient.As formula 5 shows,the original image x is pertubed with step ε in the derections of gradient descend which is hard for human eyes to distinguish.The magnitude of the disturbance is usually controlled by the size of the ε.

$$x' = x + \varepsilon * \mathrm{sign}(\nabla J(x, y_T)) \tag{5}$$

We take ε as 0.5 to control the attack scope so as to measure each models.Our model Accuracy drop to 87.5 percent under the attack of FGSM. Table 1 is an acceptable result which prove proves our model's robustness.While VGG drop to 0.71,SVM drop to 0.68.

Projected Gradient Descent(PGD) is a more powerful multi-step variant of FGSM which enable pertubtions go forward in various derections each step.This method improves the ability of nonlinear interference.In general, the success rate of PGD attack is higher than that of FGSM under the same amplitude disturbance because of its nonlinear perturbations.The method can be formulated as follow.

$$x^{t+1} = \prod_{x+S} (x_t + \varepsilon * \mathrm{sign}(\nabla J(x_t, y_T))) \tag{6}$$

Also we take the same super-parameters ε as 0.5.step as 10.The result is described in table 2.

Table 2. Performance under attack with super-parameters ε=0.5

model	accuracy	PGD attack	Time (ms)
Our	0.92	0.77	15
VGG	0.88	0.35	25
SVM	0.75	0.28	72

As we can see in the two table ,our model shows its excellent ability to resist the adversarial attack.We suppose there are mainly three reasons.First,we take Data balance and data enhancement before training.The number of various category are balanced and randomly raotating are taken to imaged.Second,the spatial transform network works for our model.We also take ablation study that how the network perform whithout spatial transform network.Actually,the sccuracy drop five percent immediately.Third the l2 regularization works.As a result, the

solution of the model is inclined to be with a small norm, which restricts the model space by limiting the size of the norm, thus avoiding overfitting to a certain extent

4. REFERENCES

[1] Krizhevsky, Alex , I. Sutskever , and G. Hinton . "ImageNet Classification with Deep Convolutional Neural Networks." *Advances in neural information processing systems* 25.2(2012)

[2] Everingham, Mark , et al. "The Pascal Visual Object Classes (VOC) Challenge." International *Journal of Computer Vision* 88.2(2010):p.303-338.

[3] Szegedy, Christian , et al. "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning." (2016).

[4] Ahmed El-Sawy, Hazem EL-Bakry, and Mohamed Loey. "CNN for Handwritten Arabic Digits Recognition Based on LeNet-5." (2016).

[5] Simonyan, Karen , and A. Zisserman . "Very Deep Convolutional Networks for Large-Scale Image Recognition." *Computer ence* (2014).

[6] He, Kaiming , et al. "Deep Residual Learning for Image Recognition." *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* IEEE, 2016.

[7] Redmon, Joseph , and A. Farhadi . "YOLOv3: An Incremental Improvement." (2018).

[8] Recht, Benjamin , et al. "Do CIFAR-10 Classifiers Generalize to CIFAR-10." (2018).

[9] Joachims, Thorsten . "Making large-scale SVM learning practical." *Technical Reports* 8.3(1998):499-526.

[10] Deng, L. . "The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]." *Signal Processing Magazine, IEEE* 29.6(2012):p.141-142.

[11] Park, Dongwon , and S. Y. Chun . "Classification based Grasp Detection using Spatial Transformer Network." (2018).

[12] Boer, Pieter Tjerk De , et al. "A Tutorial on the Cross-Entropy Method." *Annals of Operations Research* 134.1(2005):19-67.

[13] Kingma, Diederik P , and J. Ba . "Adam: A Method for Stochastic Optimization." *Computer ence* (2014).

[14] Goodfellow, Ian J, Shlens, Jonathon, and Szegedy, Christian. "Explaining and Harnessing Adversarial Examples." *Computer ence* (2014).

[15] Park, M. . "L1-regularization path algorithm for generalized linear models." *Journal of the Royal Statistical Society* 69.4(2007):659-677.

[16] Yu, Wei , et al. "Visualizing and Comparing Convolutional Neural Networks." *Computer Science* (2014).