

The source code is at the end of this document

1 Classification: Convolutional Neural Networks

1.1 Design and Implementation Choices of your Model

Inspired by VGG and ResNet, almost all convolution layers are 3×3 and almost all pooling layers are 2×2 .

The following CNN models are tested.

Model 1

CNN1 is a sample CNN with only one convolutional layer. Dropout layer is used, which is the easiest way to prevent overfitting.

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 26, 26, 32)	320
max_pooling2d (MaxPooling2D)	(None, 13, 13, 32)	0
dropout (Dropout)	(None, 13, 13, 32)	0
flatten (Flatten)	(None, 5408)	0
dense (Dense)	(None, 128)	692352
dense_1 (Dense)	(None, 5)	645
Total params: 693,317		
Trainable params: 693,317		
Non-trainable params: 0		

Model 2

CNN2 is similar to the network in the link given in the assignment. It has four convolutional

layers, each followed by a batch normalization layer, which outputs the previous layer by subtracting the batch mean and dividing it by batch standard deviation. Batch normalization could speed up training and increase the stability of the network.

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 28, 28, 32)	320
batch_normalization (Batch Normalization)	(None, 28, 28, 32)	128
conv2d_2 (Conv2D)	(None, 28, 28, 32)	9248
batch_normalization_1 (Batch Normalization)	(None, 28, 28, 32)	128
dropout_1 (Dropout)	(None, 28, 28, 32)	0
conv2d_3 (Conv2D)	(None, 28, 28, 64)	18496
max_pooling2d_1 (MaxPooling2D)	(None, 14, 14, 64)	0
dropout_2 (Dropout)	(None, 14, 14, 64)	0
conv2d_4 (Conv2D)	(None, 14, 14, 128)	73856
batch_normalization_2 (Batch Normalization)	(None, 14, 14, 128)	512
dropout_3 (Dropout)	(None, 14, 14, 128)	0
flatten_1 (Flatten)	(None, 25088)	0
dense_2 (Dense)	(None, 512)	12845568
batch_normalization_3 (Batch Normalization)	(None, 512)	2048
dropout_4 (Dropout)	(None, 512)	0
dense_3 (Dense)	(None, 128)	65664
batch_normalization_4 (Batch Normalization)	(None, 128)	512
dropout_5 (Dropout)	(None, 128)	0
dense_4 (Dense)	(None, 5)	645
Total params: 13,017,125		
Trainable params: 13,015,461		
Non-trainable params: 1,664		

Model 3

CNN3 is just CNN2 with an extra block(a Conv2D, a Batch normalization and a Dropout layer), because I wanted to check if a deeper network could do a better job.

Layer (type)	Output Shape	Param #
conv2d_5 (Conv2D)	(None, 28, 28, 32)	320
batch_normalization_5 (Batch Normalization)	(None, 28, 28, 32)	128
conv2d_6 (Conv2D)	(None, 28, 28, 32)	9248
batch_normalization_6 (Batch Normalization)	(None, 28, 28, 32)	128
dropout_6 (Dropout)	(None, 28, 28, 32)	0
conv2d_7 (Conv2D)	(None, 28, 28, 64)	18496
max_pooling2d_2 (MaxPooling2D)	(None, 14, 14, 64)	0
dropout_7 (Dropout)	(None, 14, 14, 64)	0
conv2d_8 (Conv2D)	(None, 14, 14, 128)	73856
batch_normalization_7 (Batch Normalization)	(None, 14, 14, 128)	512
dropout_8 (Dropout)	(None, 14, 14, 128)	0
conv2d_9 (Conv2D)	(None, 12, 12, 256)	295168
batch_normalization_8 (Batch Normalization)	(None, 12, 12, 256)	1024
dropout_9 (Dropout)	(None, 12, 12, 256)	0
flatten_2 (Flatten)	(None, 36864)	0
dense_5 (Dense)	(None, 512)	18874880
batch_normalization_9 (Batch Normalization)	(None, 512)	2048
dropout_10 (Dropout)	(None, 512)	0
dense_6 (Dense)	(None, 128)	65664
batch_normalization_10 (Batch Normalization)	(None, 128)	512
dropout_11 (Dropout)	(None, 128)	0
dense_7 (Dense)	(None, 5)	645
Total params: 19,342,629		
Trainable params: 19,340,453		
Non-trainable params: 2,176		

Model 4

CNN4 is a VGG like network (two or three Convolutional layers followed by a pooling layer) given in this link. Batch normalization layer was also applied to speed up computation.

Layer (type)	Output Shape	Param #
conv2d_10 (Conv2D)	(None, 28, 28, 32)	320
conv2d_11 (Conv2D)	(None, 28, 28, 64)	18496
max_pooling2d_3 (MaxPooling2D)	(None, 14, 14, 64)	0
dropout_12 (Dropout)	(None, 14, 14, 64)	0
conv2d_12 (Conv2D)	(None, 14, 14, 128)	73856
conv2d_13 (Conv2D)	(None, 12, 12, 256)	295168
max_pooling2d_4 (MaxPooling2D)	(None, 4, 4, 256)	0
dropout_13 (Dropout)	(None, 4, 4, 256)	0
flatten_3 (Flatten)	(None, 4096)	0
dense_8 (Dense)	(None, 256)	1048832
dropout_14 (Dropout)	(None, 256)	0
dense_9 (Dense)	(None, 256)	65792
dropout_15 (Dropout)	(None, 256)	0
dense_10 (Dense)	(None, 5)	1285
Total params: 1,503,749		
Trainable params: 1,503,749		
Non-trainable params: 0		

VGG19

VGG19 with imagenet pretrained weight. VGG19 contains 16 convolutional layers and 3 fully connected layers. VGG has a relatively simple network structure and it performs well in the image classification. Before using this model, I had to transform the input data into desired form. The origin input format of VGG19 is $224 \times 224 \times 3$, and the minimum input dimension for keras VGG19

model is $32 \times 32 \times 3$. I stacked the data three times and reshaped it to 150×150 , because a too small input would affect the performance of the network and 150×150 is almost the largest size I could get within my computer's capability.

Its performance was really bad(accuracy 65.53%) so I gave it up. Maybe because our labels were twisted and it was not a conventional image classification problem so we should not use the weights directly.

Inception V3

Inception network stacks the result of multiple filters together, therefore, the network can learn small details, middle sized features and almost whole images at the same time. So it will usually achieve a good result.

Since using the pretrained weight directly was not a good idea, I tried transfer learning on Inception V3. We had to adjust the input format to $150 \times 150 \times 3$ like VGG19, and last 10 layers of the model were retrained.

The performance was bad as well(accuracy 40.03%), so I did not try any further improvements.

Inception

The Inception V3 model is too big for us to train from scratch. Following this blog, I built a smaller Inception network. But its accuracy was only 81.12%.



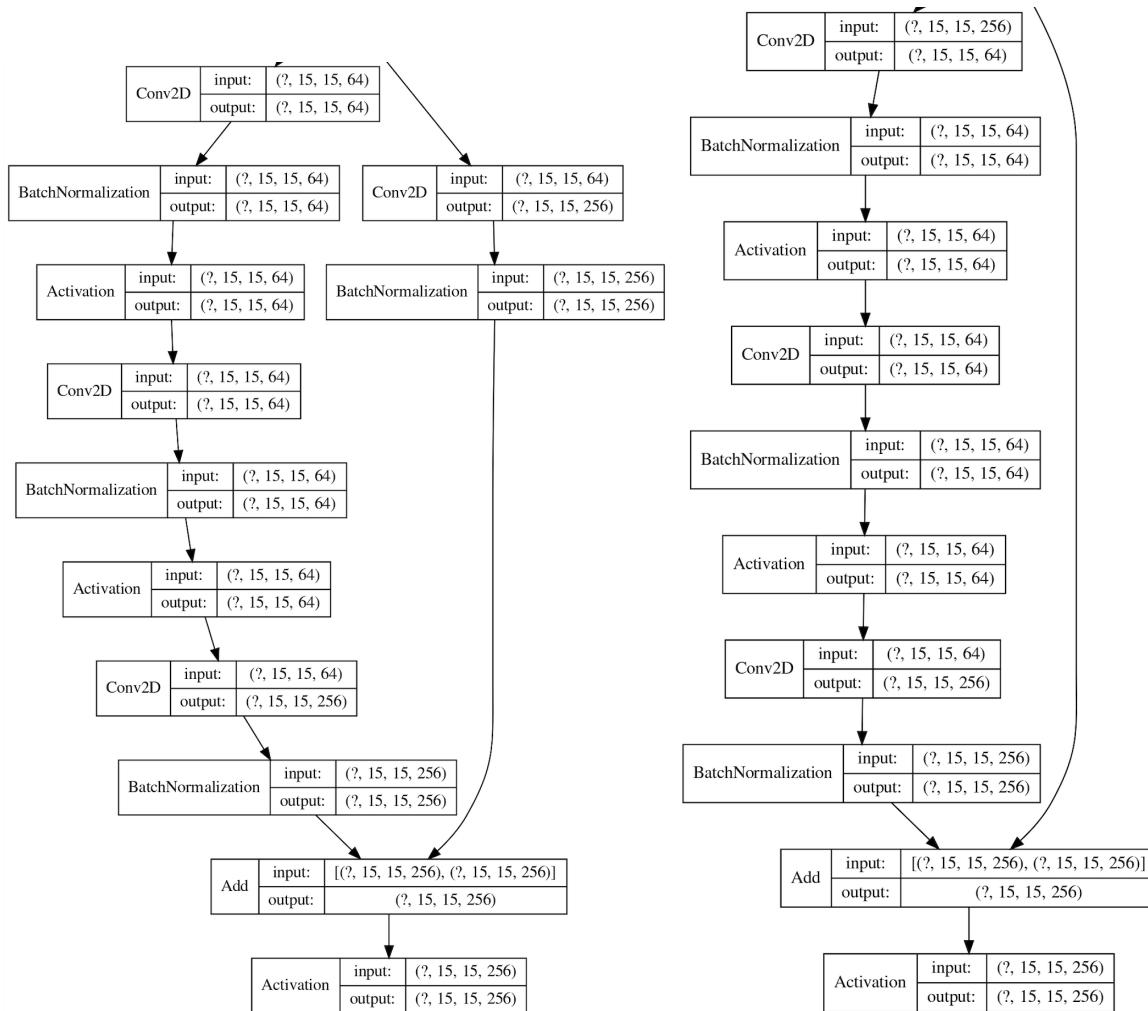
Resnet50

When the network becomes deeper and deeper, the model is prone to overfitting the data. Also, when performing the back-propagation, the gradient may becomes infinitively small or really big,

making it harder to make progress. By skipping one or more layers, that is connecting one layer directly to a layer several depth apart, the model becomes more robust. This shortcut connection makes it possible for us to build deeper networks.

The Resnet50 used here was implemented based on Coursera CNN week 2 programming assignments.

The whole structure is too large to show. Only part of it is shown below. The left one is convolutional block and the right one is identity block. Resnet50 is basically the stacking of these two blocks.



Data Augmentation

ImageDataGenerator was used to generate transformed images for neural networks. Each image was zoomed, rotated, flipped or sheared and then fed to the network. But it did not perform well as expected. The validation accuracy decreased gradually from around 90 percent to 82 percent as more and more data being generated. Hence, data augmentation was not used later.

Hyperparameters, regularization and optimizers will be discussed later.

1.2 Implementation of your Design Choices

1.3 Kaggle Competition Score

0.92240

Achieved using ensemble method. Train Model 1 repeatedly and only keep the predicted result if its testing accuracy is greater than 91 percent. Collect about 30 predicted results and vote like random forest. It took around 36 hours to get 30 qualified results using a single GTX 1080 Ti.

1.4 Results Analysis

1.4.1 Runtime performance for training and testing

The training and predicting time **for one epoch** for each method is shown below:

Model	Training(s)	Testing(s)
1(3 layers)	1.84	1.21
2(7 layers)	5.92	1.84
3(8 layers)	8.01	1.88
4(7 layers)	4.68	1.43
Inception	161.96	14.30
Resnet50	36.01	7.66

For sequential models, the more parameters to train, the slower it will get. But the number of parameters and training time does not grow proportionally. Model 2 has 19,340,453 trainable params and model 4 has 1,503,749, yet their training time for one epoch are similar, which indicates that the structure of the network also affects run time. Inception has only 3,290,597 trainable params, however, it takes way longer to train. Resnet50 has 23,544,837 trainable params but it is 3.5 times faster than Inception.

1.4.2 Comparison of the different algorithms and parameters you tried

Model	Accuracy
1(3 layers)	79.10%-91.36%
2(7 layers)	85.68%-90.15%
3(8 layers)	87.42%-90.47%
4(7 layers)	89.68%-90.15%
Inception	81.12%
Resnet50	88.77%-89.56%

As we can see from the table, generally, the deeper a network, the more stable its performance.

Model 1 only has 3 layers and its accuracy on validation set fluctuate a lot. 79.10% and 91.36% are both very rare cases, they occur only once or twice in hundreds of runs. The accuracy is about 88% on average. Also, we cannot say that model 1 is better than others because its best performance is 91.36%. Because it is the easiest to train(it took only 3s to train a batch with batchsize = 128), I could train it hundreds of times but for other models, it takes tens of minutes to hours to train. So I could only train them a few times. Model 2 and 4 are both 7 layers but model 4 is more stable than model 2. This may due to its VGG like structure.

The performance of Inception and Resnet50 are disappointing. It took 10 times longer to train them but the accuracy is worse than a simple VGG like model. A possible explanation for this is that the training set is not large enough. For the original Inception and Resnet, they are trained by millions of 224×224 colored images. However, we only have 60,000 28×28 grey scale images. The lack of input information made it difficult to build an accurate model.

1.4.3 Explanation of your model

The structure of each model have been shown in section 1.1. This section will mainly focus on hyperparameters, regularization and optimizers.

ReduceLROnPlateau was used as callback function on every model to reduce learning rate when the accuracy has stopped improving.

Only dropout layer as regularisation was used because it is the easiest one.

The epochs was set as 40 initially and later I would check if there was sign of overfitting. If there was, the number would be decreased.

The batch size was set as 128 or 256. These two numbers led to similar performance.

Activation function: softmax for output layer and ReLU for others.

The following hyperparameters are tuned using hyperas and hyperopt following this guide:

- Dropout rate: a random number between 0 and 1.
- Number of units in a Dense layer: 128, 256 or 512.
- Optimizer: Adam, SGD or RMSProp, each with learning rate of 0.001, 0.01 or 0.1.

However, I found hyperas not so efficient and sometimes it generated a set of parameters that led to a really mediocre result after 30-50 evaluation.

1.4.4 Plots

1.4.5 Evaluate your code with other metrics on the training data and argue for the benefit of your approach.