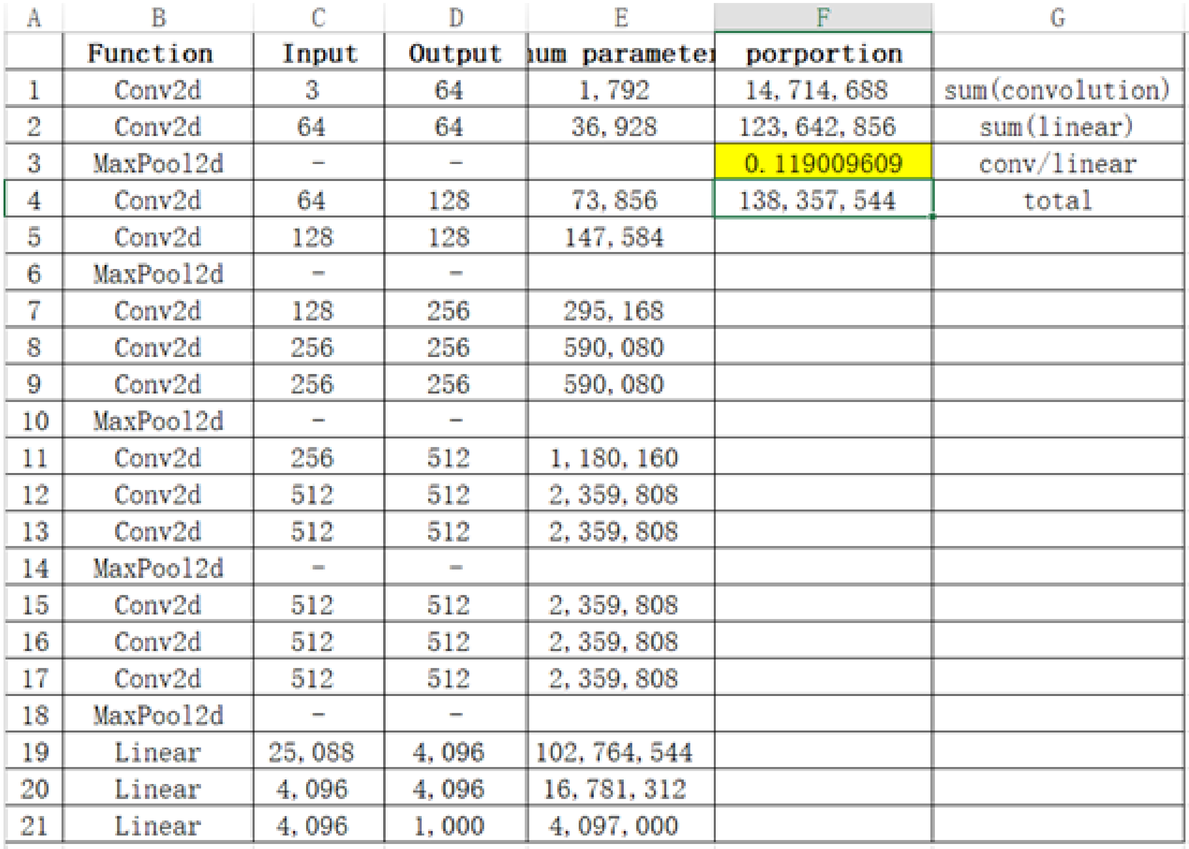
Report to Deep Learning Applications

Lecheng WANG, Guanyu CHEN
M2A, Sorbonne University

Contents

2-a: Transfer Learning

★Question 1



A	B	C	D	E	F	G
	Function	Input	Output	num parameter	porportion	
1	Conv2d	3	64	1,792	14,714,688	sum(convolution)
2	Conv2d	64	64	36,928	123,642,856	sum(linear)
3	MaxPool2d	-	-		0.119009609	conv/linear
4	Conv2d	64	128	73,856	138,357,544	total
5	Conv2d	128	128	147,584		
6	MaxPool2d	-	-			
7	Conv2d	128	256	295,168		
8	Conv2d	256	256	590,080		
9	Conv2d	256	256	590,080		
10	MaxPool2d	-	-			
11	Conv2d	256	512	1,180,160		
12	Conv2d	512	512	2,359,808		
13	Conv2d	512	512	2,359,808		
14	MaxPool2d	-	-			
15	Conv2d	512	512	2,359,808		
16	Conv2d	512	512	2,359,808		
17	Conv2d	512	512	2,359,808		
18	MaxPool2d	-	-			
19	Linear	25,088	4,096	102,764,544		
20	Linear	4,096	4,096	16,781,312		
21	Linear	4,096	1,000	4,097,000		

Figure 1: A table to count the parameters in each layer.

This table shows nearly 90% of parameters are produced by fully connected layers.

★Question 2

Output size of last layer of VGG16 is (1,1000), each element in the vector represents the probability (or score) of the input image belonging to one of the 1000 classes.

★Question 3

Role of the ImageNet normalization:

1) Stabilize the training procedure. Normalization reduces the chance of vanishing or exploding gradients, thus, leads to a faster and more stable convergence.

2) Improve generalization ability. Normalization helps to center the data and scale it, reducing the effects of, for example, varying light conditions. This allows the model to focus on more important features like shapes and textures thus, leads to a better performance.

Setting the model to evaluation mode ensures that it remains unchanged, preventing accidental modifications. Additionally, since gradient calculations are not required, computation becomes faster. In evaluation mode, networks with dropout layers will also behave differently compared to training mode.

Question 4

We can see the different output channels have focus on different features. We use (i,j) to refer the picture in i th row and j th column. Picture (4,2) and (4,4) extract features from large, distinctly varied color blocks. Such feature extraction helps the network differentiate between the background and the main subject of the image. Picture (4,3) can be regarded as the extraction of textures in the finer details. Picture (2,4) high lights the brighter areas within the image.

It is difficult for the human brain to fully understand the meaning of each channel, but there is no doubt that each of these channels extracts some type of feature of the image, and these features are useful for later operations.

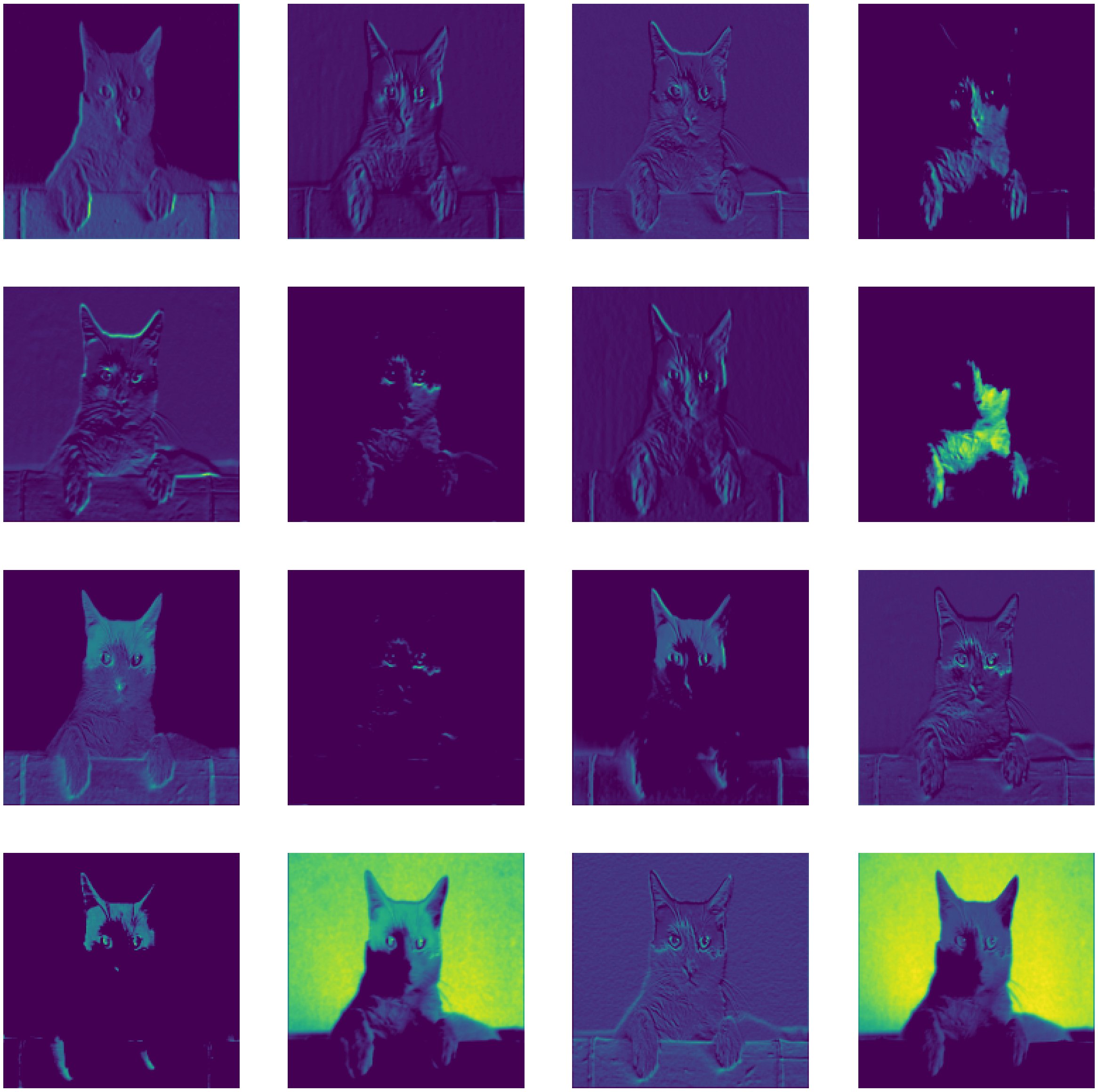


Figure 2: Visualize the output of first convolution.

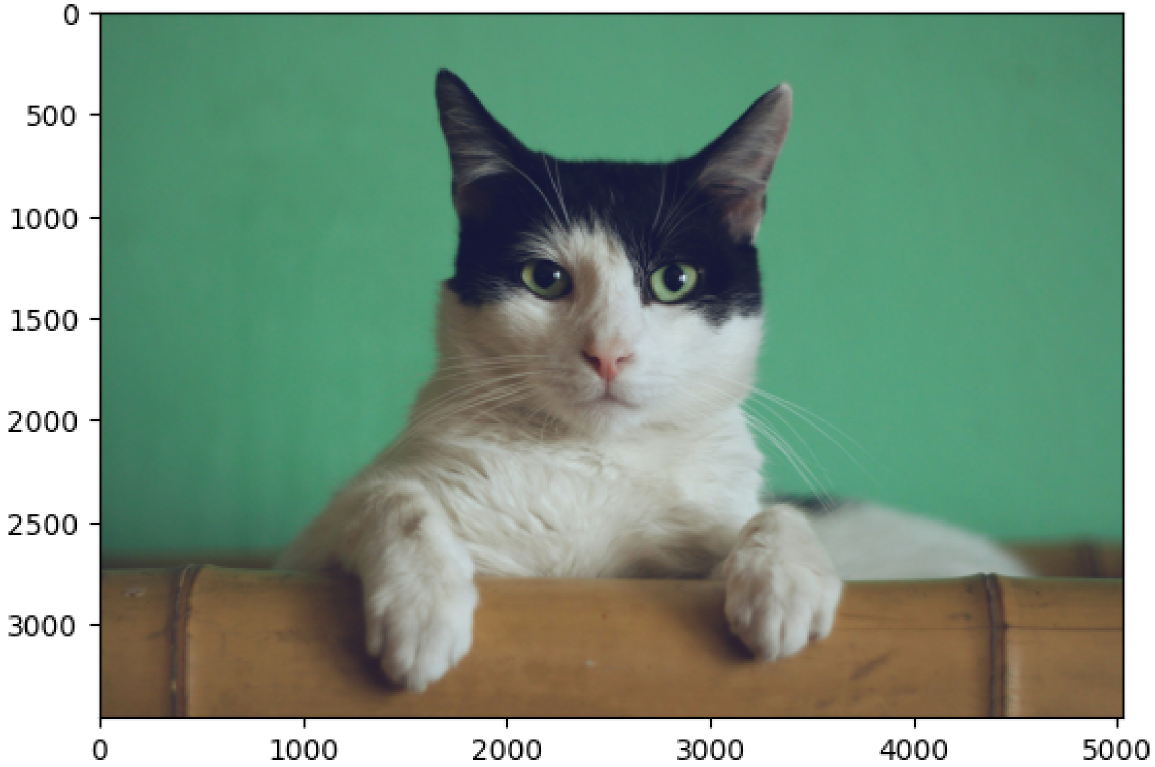


Figure 3: Origin input image of a cat.

★Question 5

Because VGG16 has too many parameters and 15 Scene is a relatively small dataset. If we directly train VGG16 on 15 Scene, it can cause significant overfitting.

★Question 6

Pre-training on ImageNet provides a strong baseline model that has already learned to recognize a wide range of features, such as edges, textures, shapes, and other complex patterns. This ability serves as a "universal" foundation for feature extraction, reducing training time and helping to prevent overfitting (as noted in question 5). This makes it a more effective approach for building a model that performs well on the 15 Scene dataset.

Question 7

First, the training dataset (ImageNet) may differ from the test dataset (15 Scene), so the features learned and the normalization parameters may not fully align. Second, since the task has changed, there might be a more suitable network architecture for extracting features specifically tailored for this classification task.

Question 8

In the early layers, the model primarily extracts basic graphical features like edges, textures, and light and shade. As more convolutional layers are added, the extracted features grow more complex, capturing shapes and partial structures of objects. In the final layers, the model can interpret complete information about the entire image, such as the object's class.

Question 9

Since the image is black and white, there is only one channel. We can make three copies of this channel after normalizing it and use it as the input to the RGB three channels

Question 10

A fully connected layer can certainly replace the SVM classifier. However, this approach has some drawbacks. For instance, it brings more parameters to be learned, and if the dataset is small, the large number of parameters can lead to overfitting.

Question 11

going further