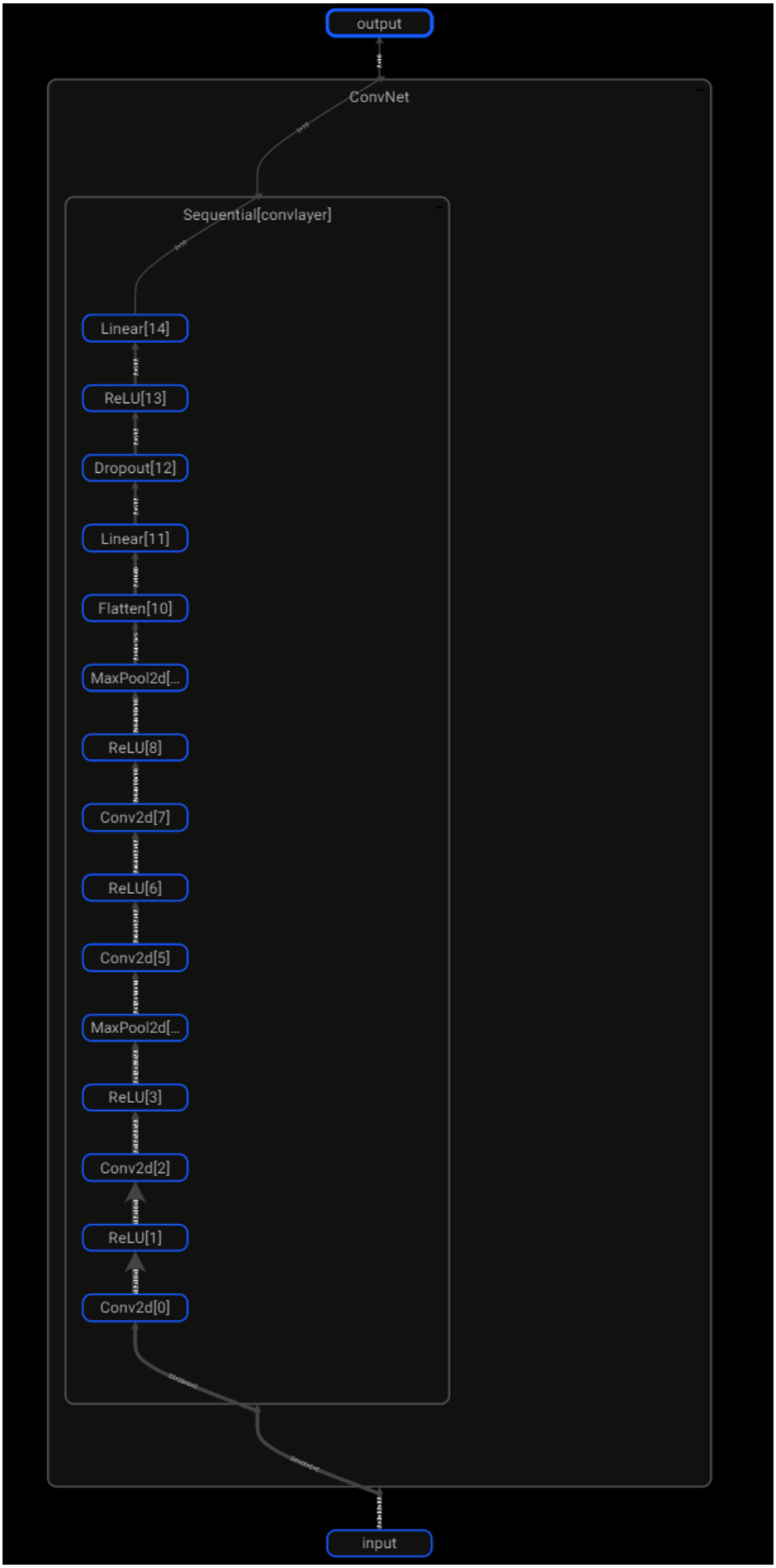
Adversarial Attack and Model Compression: Attempts on Slightly Larger Models

Haoran Geng 2000012975

A Convolutional Neural Network with ReLU Activation Layers for CIFAR10 Classification

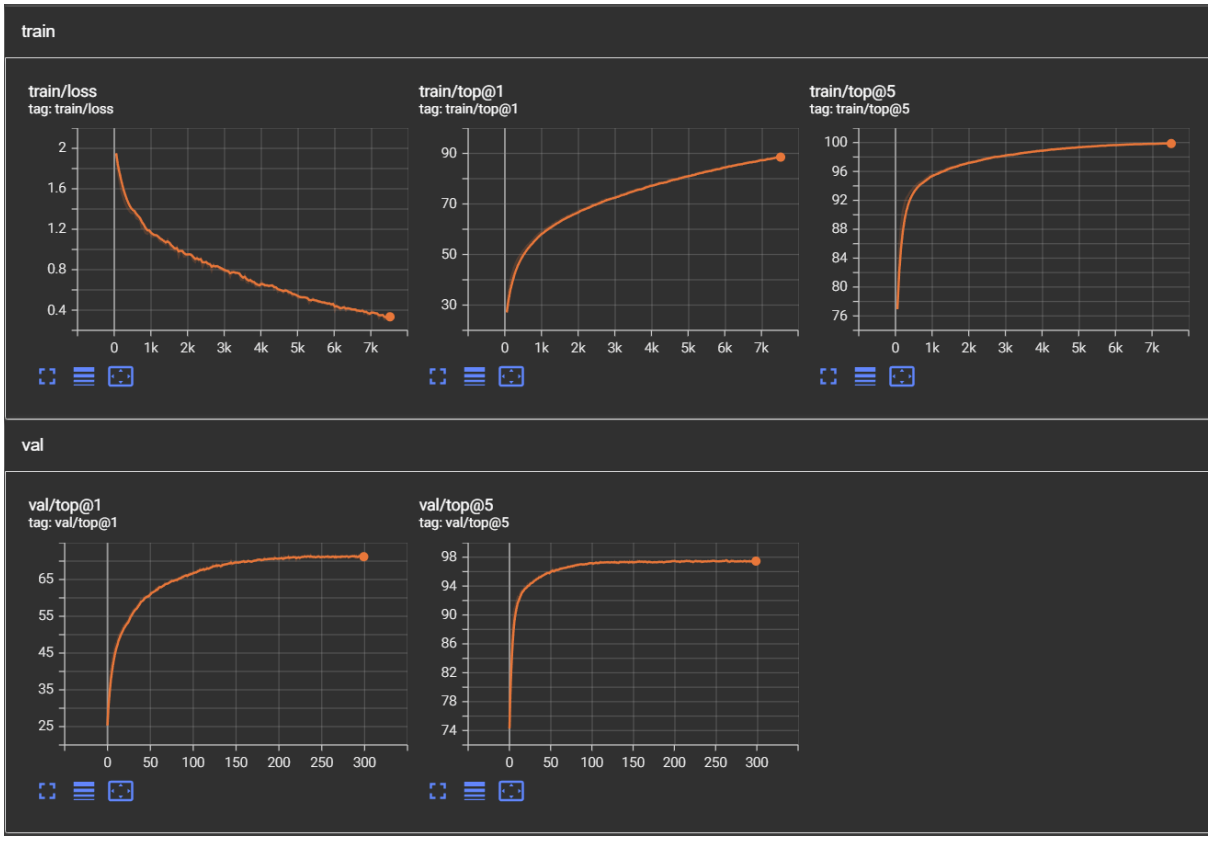
1. Model

```
self.convlayer = nn.Sequential(
  nn.Conv2d(3, 32, 3),
  nn.ReLU(inplace=True),
  nn.Conv2d(32, 32, 3),
  nn.ReLU(inplace=True),
  nn.MaxPool2d(2),
  nn.Conv2d(32, 64, 3),
  nn.ReLU(inplace=True),
  nn.Conv2d(64, 64, 3),
  nn.ReLU(inplace=True),
  nn.MaxPool2d(2),
  nn.Flatten(),
  nn.Linear(64*25, 512),
  nn.Dropout(p=0.5),
  nn.ReLU(inplace=True),
  nn.Linear(512, num_class)
)
```


output
ConvNet
Sequential[convlayer]
Linear[14]
ReLU[13]
Dropout[12]
Linear[11]
Flatten[10]
MaxPool2d[...
ReLU[8]
Conv2d[7]
ReLU[6]
Conv2d[5]
MaxPool2d[...
ReLU[3]
Conv2d[2]
ReLU[1]
Conv2d[0]
input

Model Graph

2. Training Curve



Training Curve

- Evaluation Metrics: **Top@1** Accuracy and **Top@5** Accuracy

3. PGD Attack on Raw Model

```
Attacked !!!
100%| 500/500 [00:01<00:00, 396.92it/s]
 Val Acc@1 36.180
 Val Acc@5 86.540
Not Attacked
100%| 500/500 [00:01<00:00, 375.17it/s]
 Val Acc@1 71.040
 Val Acc@5 97.340
```

`Top@1` 71.040% → 36.180%

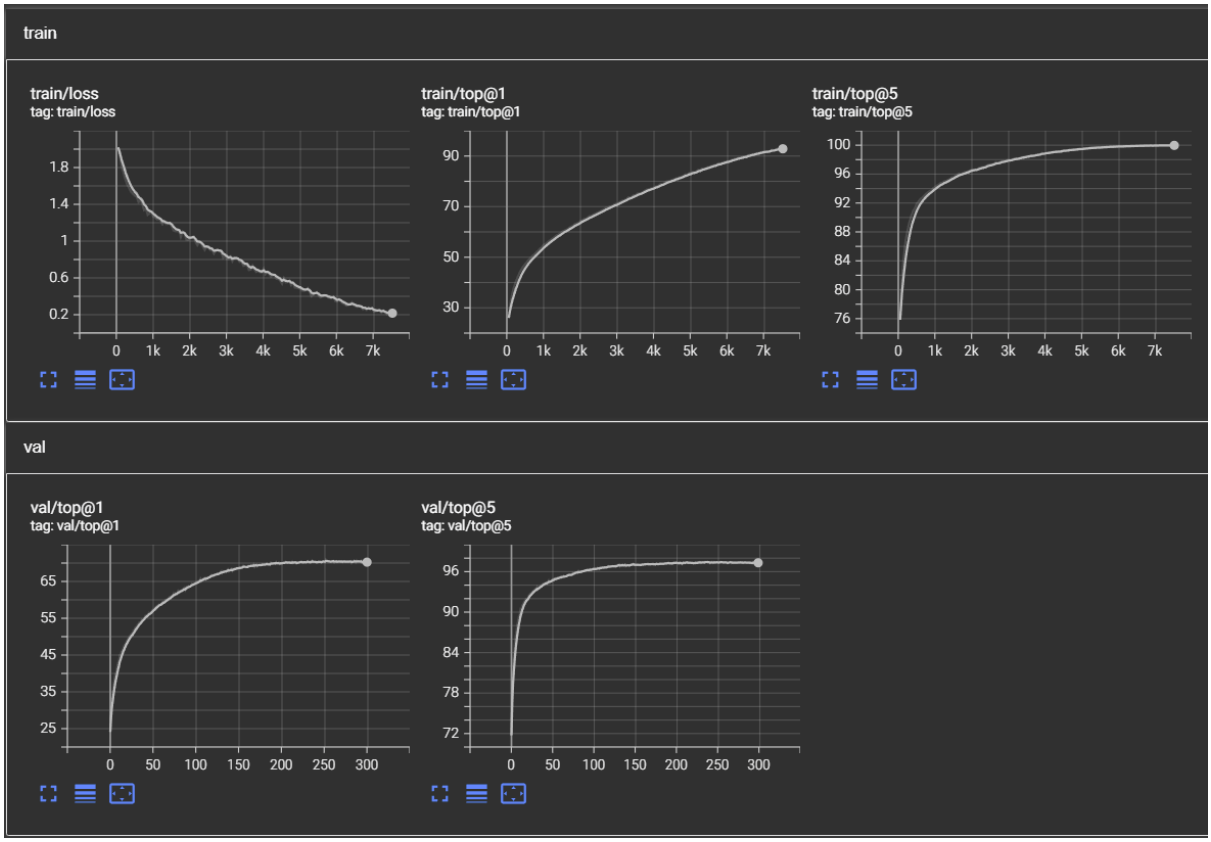
`Top@5` 97.340% → 86.540%

- We found that the impact of PGD Attack on the **Top@1** accuracy is very high and due to the attack mechanism, the impact on **Top@5** accuracy is not that high.

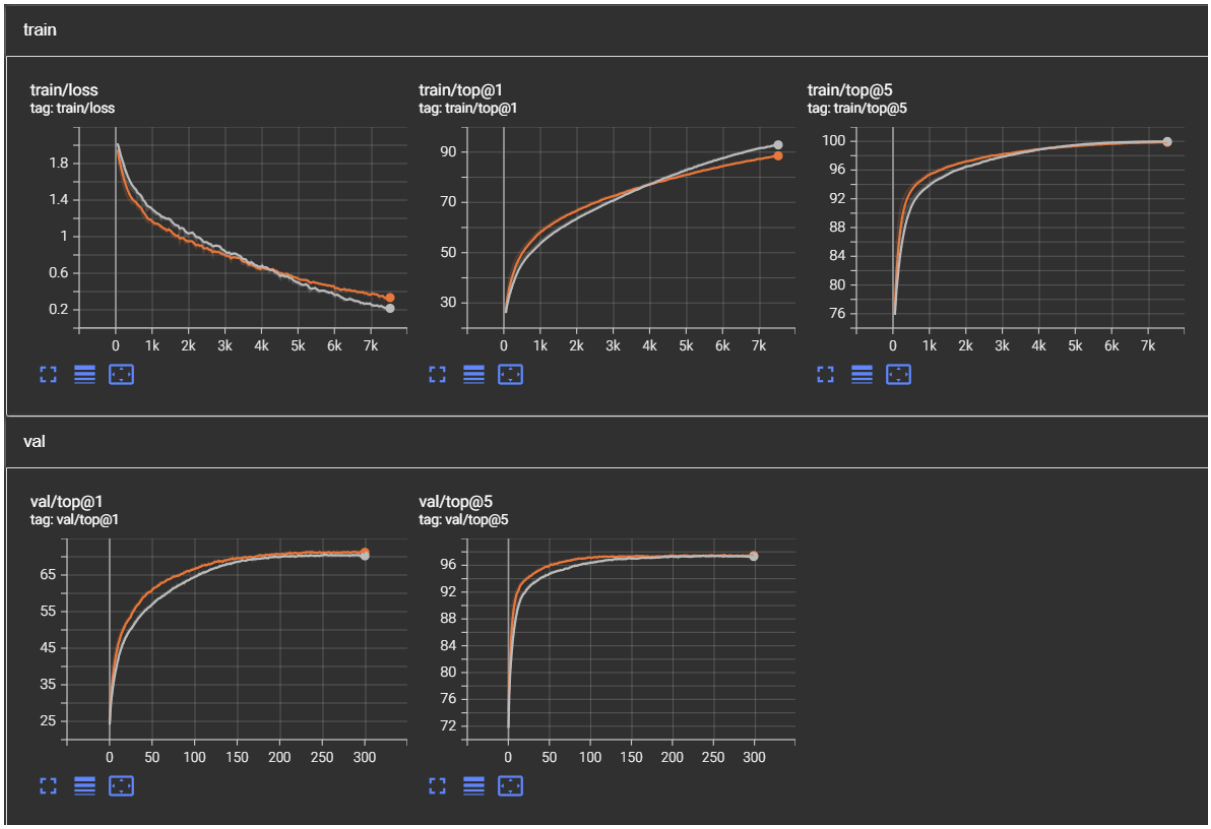
4. PGD-Adversarial Training on this model

Hyperparameter:

$\epsilon = 8/255,\ \alpha = 2/255, Step = 4$



The Contrast between the two training strategies



Orange color is the raw training strategy and grey color is the adversarial training strategy

5. PGD Attack on PGD-Adversarial Trained Model

```
Attacked !!!
100%|                                                                                | 500/500 [00:01<00:00, 408.37it/s]
 Val Acc@1 40.160
 Val Acc@5 87.910
Not Attacked
100%|                                                                                | 500/500 [00:01<00:00, 389.60it/s]
 Val Acc@1 70.590
 Val Acc@5 97.270
```

`Top@1` *70.590% → 40.160%*

`Top@5` *97.270% → 87.910%*

6. Contrast between the two training strategies

- ***Raw Training***

`Top@1` *71.040% → 36.180%*

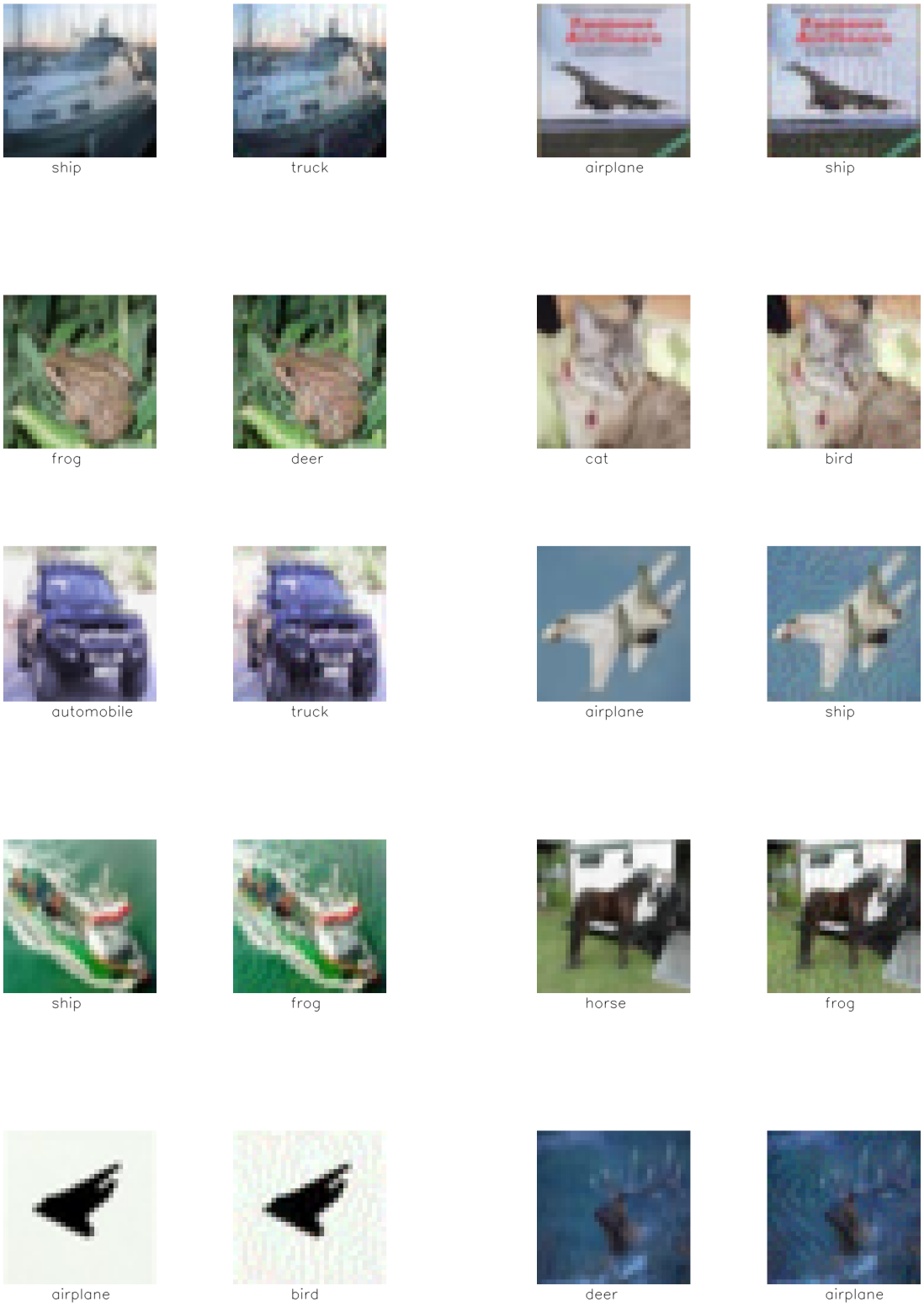
`Top@5` *97.340% → 86.540%*

- ***Adversarial Training***

`Top@1` *70.590% → 40.160%*

`Top@5` *97.270% → 87.910%*

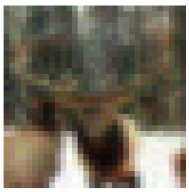
7. Result Visualization



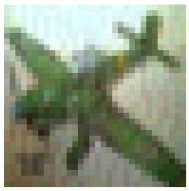

bird


deer

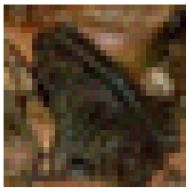

deer


bird


airplane


deer


frog


bird


dog


bird


deer


bird


dog


cat


automobile

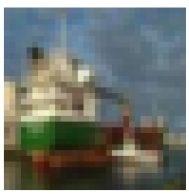

truck

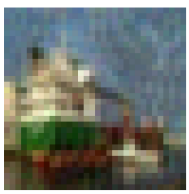

horse

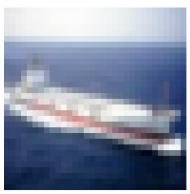

deer

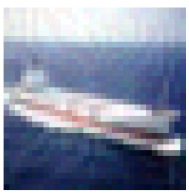

frog


bird


ship


deer

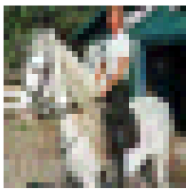

ship


airplane


ship


airplane


horse


dog


frog


deer


bird


frog


bird


airplane


horse


automobile


airplane


truck


ship


deer

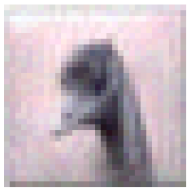

ship

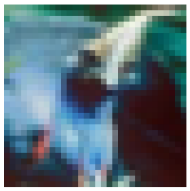

automobile

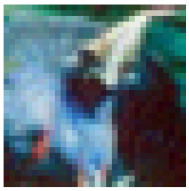

horse


deer


bird


cat


bird


deer


airplane


ship


cat


frog


frog


cat


airplane


deer