计算机科学与技术学院神经网络与深度学习课程实验报告

实验题目：Feature Vis		学号：201918130222
日期：2021/12/8	班级： 智能	姓名： 魏江峰
Email：2257263015@qq.com		

实验目的：

visualizing the features of a pretrained model on ImageNet.

Explore various applications of image gradients, including saliency maps, fooling images, class visualizations

实验软件和硬件环境：

硬件环境：

处理器：Intel core i7 9750-H

电脑：神州 z7m-ct7nk

软件环境：

Pycharm 与 jupyter notebook

实验原理和方法：

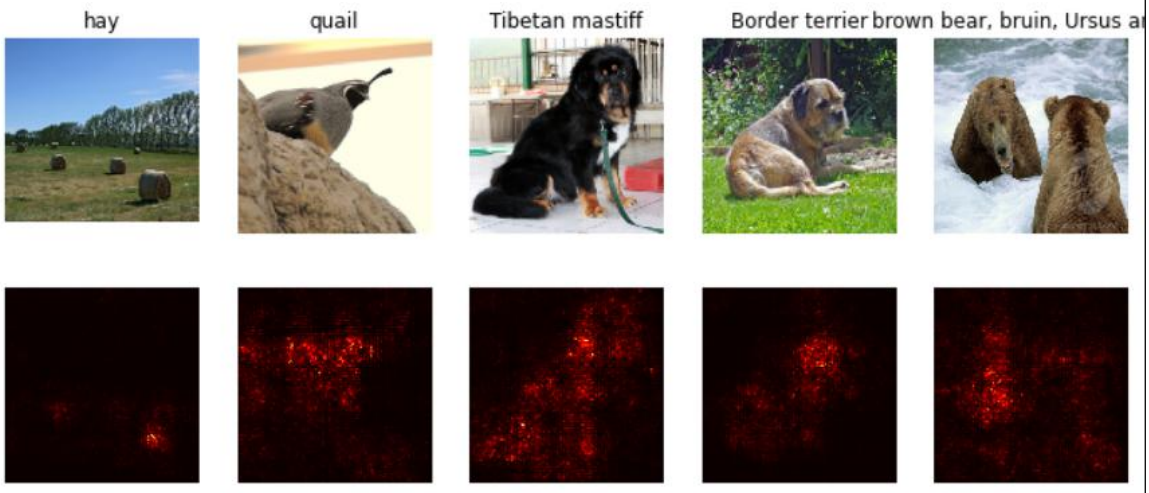
使用 ImageNet 上的一个训练好的模型，利用这个模型定义的 loss 函数，计算图像的梯度，生成一张新的图像。

实验步骤：（不要求罗列完整源代码）

1，Load images:

2，Saliency maps：输出每一个像素在对分类结果的影响力的大小：

```
predict = (model(X).gather(1, y.view(-1, 1)).squeeze())
loss = -torch.log10(predict).sum()
loss.backward()
saliency = torch.abs((X.grad))
```



3，Fooling images：对图像施加一些不显著的变化，以至于人眼无法察觉，但是却会对分类器造成严重的干扰。

```
for i in range(100):
predict = model(X_fooling)
if(predict.argmax() == target_y):
```

```
print("congratulations,you have fooled the model !")
break
scores = predict[:, target_y]
scores.backward()
# print(torch.linalg.matrix_norm(X_fooling.grad).shape)
g = X_fooling.grad
dX = learning_rate * (g / torch.linalg.matrix_norm(g).reshape(3, 1,
1))
X_fooling.data += dX.data
print(scores)
```

可以看出肉眼上两张图片几乎没有区别，但是被处理后的图像却被错误滴分类了。

当将两张图片的差放大 10 倍时，可以看出有一些噪音，这些计算得到的噪音严重地影响了网络的输出。

4，Class Visulization：合成一张可以最大化某一特定的类的分类分数的 image。当它将 image 分成某一个类的时候，它能给我们一些 network 正在寻找的东西的感觉。

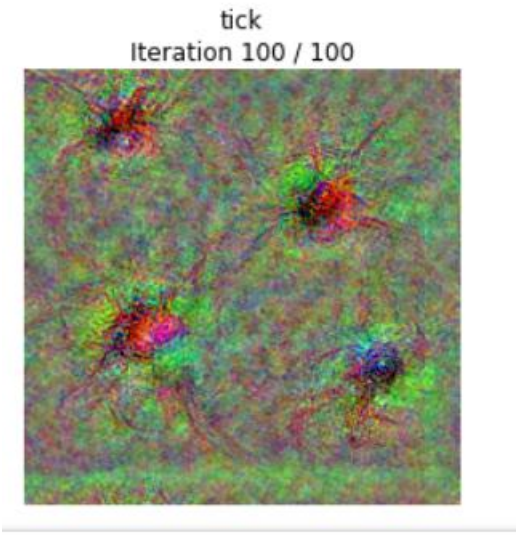
```
loss = model(img)[0, target_y] - l2_reg *((img*img).sum())
        loss.backward()
```

print(torch.linalg.norm(img.grad).shape)

img.data += learning_rate *

(img.grad/torch.linalg.matrix_norm(img.grad).reshape(3, 1, 1))

img.data += learning_rate * (img.grad/torch.linalg.norm(img.grad))



结论分析：

Network visualization 可以将神经网络的一些特征可视化，方便人们对神经网络的观察。

就实验过程中遇到和出现的问题，你是如何解决和处理的，自拟 1 - 3 道问答题：

计算 class visualization 梯度时，没有进行 normalize，结果生成的图像没有任何感性的信息：

loss = model(img)[0, target_y] - l2_reg *((img*img).sum())loss.backward()

这样梯度上升之后，什么图像都没生成。

进行 normalize 之后，就得到了正确的图像。