Adam

Adam 是一个结合了动量法和 RMSProp 的优化算法，其结合了两者的优点。

下面我们来实现以下 adam 算法

```
def adam(parameters, vs, sqrs, lr, t, beta1=0.9, beta2=0.999):
    eps = 1e-8
    for param, v, sqr in zip(parameters, vs, sqrs):
        v[:] = beta1 * v + (1 - beta1) * param.grad.data
        sqr[:] = beta2 * sqr + (1 - beta2) * param.grad.data ** 2
        v_hat = v / (1 - beta1 ** t)
        s_hat = sqr / (1 - beta2 ** t)
        param.data = param.data - lr * v_hat / torch.sqrt(s_hat + eps)
```

```
import numpy as np
import torch
from torchvision.datasets import MNIST # 导入 pytorch 内置的 mnist 数据
from torch.utils.data import DataLoader
from torch import nn
from torch.autograd import Variable
import time
import matplotlib.pyplot as plt
%matplotlib inline

def data_tf(x):
    x = np.array(x, dtype='float32') / 255
    x = (x - 0.5) / 0.5 # 标准化，这个技巧之后会讲到
    x = x.reshape((-1,)) # 拉平
    x = torch.from_numpy(x)
    return x

train_set = MNIST('./data', train=True, transform=data_tf, download=True) # 载入数据
集，申明定义的数据变换
test_set = MNIST('./data', train=False, transform=data_tf, download=True)

# 定义 loss 函数
criterion = nn.CrossEntropyLoss()
```

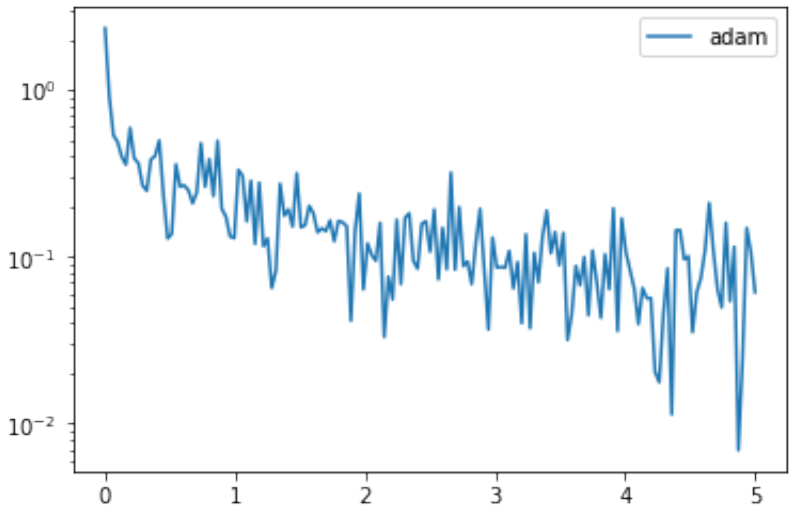
```
train_data = DataLoader(train_set, batch_size=64, shuffle=True)
# 使用 Sequential 定义 3 层神经网络
```

```
net = nn.Sequential(
    nn.Linear(784, 200),
    nn.ReLU(),
    nn.Linear(200, 10),
)

# 初始化梯度平方项和动量项
sqrs = []
vs = []
for param in net.parameters():
    sqrs.append(torch.zeros_like(param.data))
    vs.append(torch.zeros_like(param.data))
t = 1
# 开始训练
losses = []
idx = 0

start = time.time() # 记时开始
for e in range(5):
    train_loss = 0
    for im, label in train_data:
        im = Variable(im)
        label = Variable(label)
        # 前向传播
        out = net(im)
        loss = criterion(out, label)
        # 反向传播
        net.zero_grad()
        loss.backward()
        adam(net.parameters(), vs, sqrs, 1e-3, t) # 学习率设为 0.001
        t += 1
        # 记录误差
        train_loss += loss.data[0]
        if idx % 30 == 0:
            losses.append(loss.data[0])
        idx += 1
    print('epoch: {}, Train Loss: {:.6f}'
          .format(e, train_loss / len(train_data)))
end = time.time() # 计时结束
print('使用时间: {:.5f} s'.format(end - start))
```

```
epoch: 0, Train Loss: 0.372057
epoch: 1, Train Loss: 0.186132
epoch: 2, Train Loss: 0.132870
epoch: 3, Train Loss: 0.107864
epoch: 4, Train Loss: 0.091208
使用时间: 85.96051 s
```

```
x_axis = np.linspace(0, 5, len(losses), endpoint=True)
plt.semilogy(x_axis, losses, label='adam')
plt.legend(loc='best')
```

```
<matplotlib.legend.Legend at 0x10d707908>
```



可以看到使用 adam 算法 loss 能够更快更好地收敛，但是一定要小心学习率的设定，使用自适应的算法一般需要更小的学习率

当然 pytorch 中也内置了 adam 的实现，只需要调用 `torch.optim.Adam()`，下面是例子

```
train_data = DataLoader(train_set, batch_size=64, shuffle=True)
# 使用 Sequential 定义 3 层神经网络
net = nn.Sequential(
    nn.Linear(784, 200),
    nn.ReLU(),
    nn.Linear(200, 10),
)
```

```
optimizer = torch.optim.Adam(net.parameters(), lr=1e-3)

# 开始训练
start = time.time() # 记时开始
for e in range(5):
    train_loss = 0
    for im, label in train_data:
        im = Variable(im)
        label = Variable(label)
        # 前向传播
        out = net(im)
        loss = criterion(out, label)
        # 反向传播
        optimizer.zero_grad()
        loss.backward()
        optimizer.step()
        # 记录误差
        train_loss += loss.data[0]
    print('epoch: {}, Train Loss: {:.6f}'
          .format(e, train_loss / len(train_data)))
end = time.time() # 计时结束
print('使用时间: {:.5f} s'.format(end - start))
```

```
epoch: 0, Train Loss: 0.359934
epoch: 1, Train Loss: 0.173360
epoch: 2, Train Loss: 0.122554
epoch: 3, Train Loss: 0.100869
epoch: 4, Train Loss: 0.085850
使用时间: 93.85302 s
```