Caffe学习系列(8)：solver优化方法

上文提到，到目前为止，caffe总共提供了六种优化方法：

- Stochastic Gradient Descent (type: "SGD"),
- AdaDelta (type: "AdaDelta"),
- Adaptive Gradient (type: "AdaGrad"),
- Adam (type: "Adam"),
- Nesterov's Accelerated Gradient (type: "Nesterov") and
- RMSprop (type: "RMSProp")

Solver就是用来使loss最小化的优化方法。对于一个数据集D，需要优化的目标函数是整个数据集中所有数据loss的平均值。

$$L(W) = \frac{1}{|D|}\sum_{i}^{|D|} f_W\left(X^{(i)}\right) + \lambda r(W)$$

其中，$f_W(x^{(i)})$计算的是数据$x^{(i)}$上的loss, 先将每个单独的样本x的loss求出来，然后求和，最后求均值。 r(W)是正则项（weight_decay），为了减弱过拟合现象。

如果采用这种Loss 函数，迭代一次需要计算整个数据集，在数据集非常大的这情况下，这种方法的效率很低，这个也是我们熟知的梯度下降采用的方法。

在实际中，通过将整个数据集分成几批（batches), 每一批就是一个mini-batch，其数量（batch_size)为N<<|D|，此时的loss 函数为：

$$L(W) \approx \frac{1}{N}\sum_{i}^{N} f_W\left(X^{(i)}\right) + \lambda r(W)$$

有了loss函数后，就可以迭代的求解loss和梯度来优化这个问题。在神经网络中，用forward pass来求解loss，用backward pass来求解梯度。

在caffe中，默认采用的Stochastic Gradient Descent（SGD）进行优化求解。后面几种方法也是基于梯度的优化方法（like SGD），因此本文只介绍一下SGD。其它的方法，有兴趣的同学，可以去看文献原文。

1、Stochastic gradient descent (SGD)

随机梯度下降（Stochastic gradient descent）是在梯度下降法（gradient descent）的基础上发展起来的，梯度下降法也叫最速下降法，具体原理在网易公开课《机器学习》中，吴恩达教授已经讲解得非常详细。SGD在通过负梯度$\nabla L(W)$和上一次的权重更新值V_t的线性组合来更新W，迭代公式如下：

$$V_{t+1} = \mu V_t - \alpha \nabla L(W_t)$$

$$W_{t+1} = W_t + V_{t+1}$$

其中，α 是负梯度的学习率(base_lr)，μ 是上一次梯度值的权重（momentum），用来加权之前梯度方向对现

昵称：denny402
园龄：5年10个月
粉丝：74
关注：2
+加关注

< 2016年5月 >

日	一	二	三	四	五	六
24	25	26	27	28	29	30
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31	1	2	3	4

搜索
找找看
谷歌搜索

常用链接
我的随笔
我的评论
我的参与
最新评论
我的标签
更多链接

我的标签
python(26)
caffe(25)
opencv3(10)
matlab(9)
mvc(9)
MVC3(8)
ajax(7)
geos(6)
opencv(6)
ml(5)
更多

随笔分类
caffe(26)
GDAL(2)
GEOS(6)
matlab(11)
opencv(19)
Python(25)

随笔档案
2016年1月 (33)
2015年12月 (29)
2015年11月 (10)
2015年7月 (7)
2014年10月 (4)

在梯度下降方向的影响。这两个参数需要通过tuning来得到最好的结果，一般是根据经验设定的。如果你不知道如何设定这些参数，可以参考相关的论文。

在深度学习中使用SGD，比较好的初始化参数的策略是把学习率设为0.01左右（base_lr: 0.01），在训练的过程中，如果loss开始出现稳定水平时，对学习率乘以一个常数因子（gamma），这样的过程重复多次。

对于momentum，一般取值在0.5--0.99之间。通常设为0.9，momentum可以让使用SGD的深度学习方法更加稳定以及快速。

关于更多的momentum，请参看Hinton的《A Practical Guide to Training Restricted Boltzmann Machines》。

实例：

```
base_lr: 0.01
lr_policy: "step"
gamma: 0.1
stepsize: 1000
max_iter: 3500
momentum: 0.9
```

lr_policy设置为step,则学习率的变化规则为 base_lr * gamma ^ (floor(iter / stepsize))

即前1000次迭代，学习率为0.01; 第1001-2000次迭代，学习率为0.001; 第2001-3000次迭代，学习率为0.00001，第3001-3500次迭代，学习率为10^{-5}

上面的设置只能作为一种指导，它们不能保证在任何情况下都能得到最佳的结果，有时候这种方法甚至不work。如果学习的时候出现diverge（比如，你一开始就发现非常大或者NaN或者inf的loss值或者输出），此时你需要降低base_lr的值（比如，0.001），然后重新训练，这样的过程重复几次直到你找到可以work的base_lr。

2、AdaDelta

AdaDelta是一种"鲁棒的学习率方法"，是基于梯度的优化方法（like SGD）。

具体的介绍文献：

M. Zeiler ADADELTA: AN ADAPTIVE LEARNING RATE METHOD. *arXiv preprint*, 2012.

示例：

```
net: "examples/mnist/lenet_train_test.prototxt"
test_iter: 100
test_interval: 500
base_lr: 1.0
lr_policy: "fixed"
momentum: 0.95
weight_decay: 0.0005
display: 100
max_iter: 10000
snapshot: 5000
snapshot_prefix: "examples/mnist/lenet_adadelta"
solver_mode: GPU
type: "AdaDelta"
delta: 1e-6
```

从最后两行可看出，设置solver type为Adadelta时，需要设置delta的值。

3、AdaGrad

自适应梯度（adaptive gradient）是基于梯度的优化方法（like SGD）

具体的介绍文献：

Duchi, E. Hazan, and Y. Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *The Journal of Machine Learning Research*, 2011.

示例：

```
net: "examples/mnist/mnist_autoencoder.prototxt"
test_state: { stage: 'test-on-train' }
test_iter: 500
test_state: { stage: 'test-on-test' }
test_iter: 100
test_interval: 500
test_compute_loss: true
base_lr: 0.01
lr_policy: "fixed"
display: 100
```

2014年7月 (4)
2013年10月 (3)
2013年8月 (5)
2013年7月 (7)
2013年6月 (6)
2011年4月 (4)
2010年6月 (3)

最新评论

1. Re:Caffe学习系列(23)：如何将别人训练好的model用到自己的数据上
您好，看到您的教程学到很多，我没用digst，直接用的命令操作，但是由于电脑原因，我在做图片的lmdb的时候吧图片设置成125~125的，然后运行的时候就出现了错误错误内容是：Check failed.....
--weichang88688
2. Re:Caffe学习系列(4)：激活层（Activation Layers)及参数
给博主赞一个，对入门小白帮助真大！另外可以请问下你，为什么sigmoid层是另建一层，然后将自己输出，relu则本地操作不添加新的层，而后面的TanH,absolute value,power等都是.....
--MaiYatang
3. Re:Caffe学习系列(13)：数据可视化环境（python接口)配置
在哪个路径下Make Clear 呢？Caffe的编译会被清除么？
--TonyFaith
4. Re:Caffe学习系列(13)：数据可视化环境（python接口)配置
@TonyFaith清除以前的编译可以make clear，再重新编译就可以了。缺少python.h我不知道是什么原因...
--denny402
5. Re:Caffe学习系列(17)：模型各层数据和参数可视化
@weichang88688卷积层的输出数据就是net.blobs['conv1'].data[0]，用一个变量保存起来就可以了c1=net.blobs['conv1'].data[0]...
--denny402

阅读排行榜

1. SqlDataReader的关闭问题(9287)
2. 索引超出范围。必须为非负值并小于集合大小。(4655)
3. Caffe学习系列(1)：安装配置ubuntu14.04+cuda7.5+caffe+cudnn(3166)
4. Caffe学习系列(12)：训练和测试自己的图片(2919)
5. Caffe学习系列(2)：数据层及参数(2492)

评论排行榜

1. Caffe学习系列(12)：训练和测试自己的图片(38)
2. SqlDataReader的关闭问题(22)
3. caffe windows 学习第一步：编译和安装（vs2012+win 64)(15)
4. Caffe学习系列(23)：如何将别人训练好的model用到自己的数据上(15)
5. Caffe学习系列(3)：视觉层（Vision Layers)及参数(11)

推荐排行榜

1. SqlDataReader的关闭问题(5)
2. Caffe学习系列(12)：训练和测试自己的图片(4)
3. Caffe学习系列(11)：图像数据转换成db（leveldb/lmdb)文件(2)

```
max_iter: 65000
weight_decay: 0.0005
snapshot: 10000
snapshot_prefix: "examples/mnist/mnist_autoencoder_adagrad_train"
# solver mode: CPU or GPU
solver_mode: GPU
type: "AdaGrad"
```

4、Adam

是一种基于梯度的优化方法（like SGD）。

具体的介绍文献：

D. Kingma, J. Ba. Adam: A Method for Stochastic Optimization. *International Conference for Learning Representations*, 2015.

5、NAG

Nesterov 的加速梯度法（Nesterov's accelerated gradient）作为凸优化中最理想的方法，其收敛速度非常快。

具体的介绍文献：

I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the Importance of Initialization and Momentum in Deep Learning. *Proceedings of the 30th International Conference on Machine Learning*, 2013.

示例：

```
net: "examples/mnist/mnist_autoencoder.prototxt"
test_state: { stage: 'test-on-train' }
test_iter: 500
test_state: { stage: 'test-on-test' }
test_iter: 100
test_interval: 500
test_compute_loss: true
base_lr: 0.01
lr_policy: "step"
gamma: 0.1
stepsize: 10000
display: 100
max_iter: 65000
weight_decay: 0.0005
snapshot: 10000
snapshot_prefix: "examples/mnist/mnist_autoencoder_nesterov_train"
momentum: 0.95
# solver mode: CPU or GPU
solver_mode: GPU
type: "Nesterov"
```

6、RMSprop

RMSprop是Tieleman在一次 Coursera课程演讲中提出来的，也是一种基于梯度的优化方法（like SGD）

具体的介绍文献：

T. Tieleman, and G. Hinton. RMSProp: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning.Technical report*, 2012.

示例：

```
net: "examples/mnist/lenet_train_test.prototxt"
test_iter: 100
test_interval: 500
base_lr: 1.0
lr_policy: "fixed"
momentum: 0.95
weight_decay: 0.0005
display: 100
max_iter: 10000
snapshot: 5000
snapshot_prefix: "examples/mnist/lenet_adadelta"
solver_mode: GPU
type: "RMSProp"
rms_decay: 0.98
```

4. DWZ中刷新dialog的方案解决(2)
5. MVC3学习：利用mvc3+ajax结合MVCPager实现分页(2)

最后两行，需要设置rms_decay值。

分类: caffe

标签: caffe

denny402
关注 - 2
粉丝 - 74
+加关注

« 上一篇：Caffe学习系列(7)：solver及其配置
» 下一篇：Caffe学习系列(9)：运行caffe自带的两个简单例子

posted @ 2015-12-24 20:25 denny402 阅读(1711) 评论(0) 编辑 收藏