Optimization based on loss function(优化损失函数)

loss优化方法（针对多分类问题）

1 传统方法：Softmax Loss

1.1 理论

softmax loss实际上是由softmax和cross-entropy loss组合而成，两者放一起数值计算更加稳定。

令z是softmax层的输入，f(z)是softmax的输出，则

$$f(z_k) = e^{z_k} / (\sum_j e^{z_j})$$

单个像素i的softmax loss等于cross-entropy error如下:

$$l(y, z) = -\sum_{k=0}^{C} y_c log(f(z_c))$$

展开上式:

$$l((y, z) = log \sum_j e^{z_j} - log e^{z_y}$$

1.2 总结

优化类间的距离非常棒，优化类内距离时比较弱。

2 改进方法

2.1 Weighted Softmax Loss

2.1.1论文

Xie S, Tu Z.Holistically-nested edge detection[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1395-1403.

2.1.2 适用场景

样本数目差距非常之大。比如边缘检测问题，这个时候，明显边缘像素的重要性是比非边缘像素大的，此时可以针对性的对样本进行加权。

2.1.3 理论

$$l(y,z) = -\sum_{k=0}^{C} w_c y_c log(f(z_c))$$

因此 Li 就变成下式：

$$L_i = -\log\left(\frac{e^{\|W_{y_i}\|\|x_i\|\cos(\theta_{y_i})}}{\sum_j e^{\|W_j\|\|x_i\|\cos(\theta_j)}}\right)$$

2.2 Large Margin Softmax Loss

2.2.1 理论

假设一个2分类问题，x属于类别1，那么原来的softmax肯定是希望：

$$W_1^T x > W_2^T x$$

也就是属于类别1的概率大于类别2的概率，这个式子和下式是等效的：

$$\|W_1\|\|x\|\cos(\theta_1) > \|W_2\|\|x\|\cos(\theta_2))$$

那么Large Margin Softmax就是将上面不等式替换成下式：

$$\text{sion margin. So we instead require } \|W_1\|\|x\|\cos(m\theta_1) > \|W_2\|\|x\|\cos(\theta_2) \ (0 \le \theta_1 \le \frac{\pi}{m}) \text{ where } m \text{ is a positive integer. Because the following inequality holds:}$$

因为m是正整数，cos函数在0到π范围又是单调递减的，所以cos(mx)要小于cos(x)。m值越大则学习的难度也越大，这也就是最开始Figure2中那几个图代表不同m值的意思。因此通过这种方式定义损失会逼得模型学到类间距离更大的，类内距离更小的特征。

这样L-softmax loss的Li式子就可以在原来的softmax loss的Li式子上修改得到：

Following the notation in the preliminaries, the L-Softmax loss is defined as

$$L_i = -\log\left(\frac{e^{\|W_{y_i}\|\|x_i\|\psi(\theta_{y_i})}}{e^{\|W_{y_i}\|\|x_i\|\psi(\theta_{y_i})} + \sum_{j\neq y_i} e^{\|W_j\|\|x_i\|\cos(\theta_j)}}\right) \tag{4}$$

in which we generally require

$$\psi(\theta) = \begin{cases} \cos(m\theta), & 0 \le \theta \le \frac{\pi}{m} \\ \mathcal{D}(\theta), & \frac{\pi}{m} < \theta \le \pi \end{cases} \tag{5}$$

where m is a integer that is closely related to the classification margin. With larger m, the classification margin becomes larger and the learning objective also becomes harder. Meanwhile, $\mathcal{D}(\theta)$ is required to be a monotonically decreasing function and $\mathcal{D}(\frac{\pi}{m})$ should equal $\cos(\frac{\pi}{m})$.

下图是从几何角度直观地看两种损失的差别，L-softmax loss学习到的参数可以将两类样本的类间距离加大。通过对比可以看到L-softmax loss最后学到的特征之间的分离程度比原来的要明显得多。

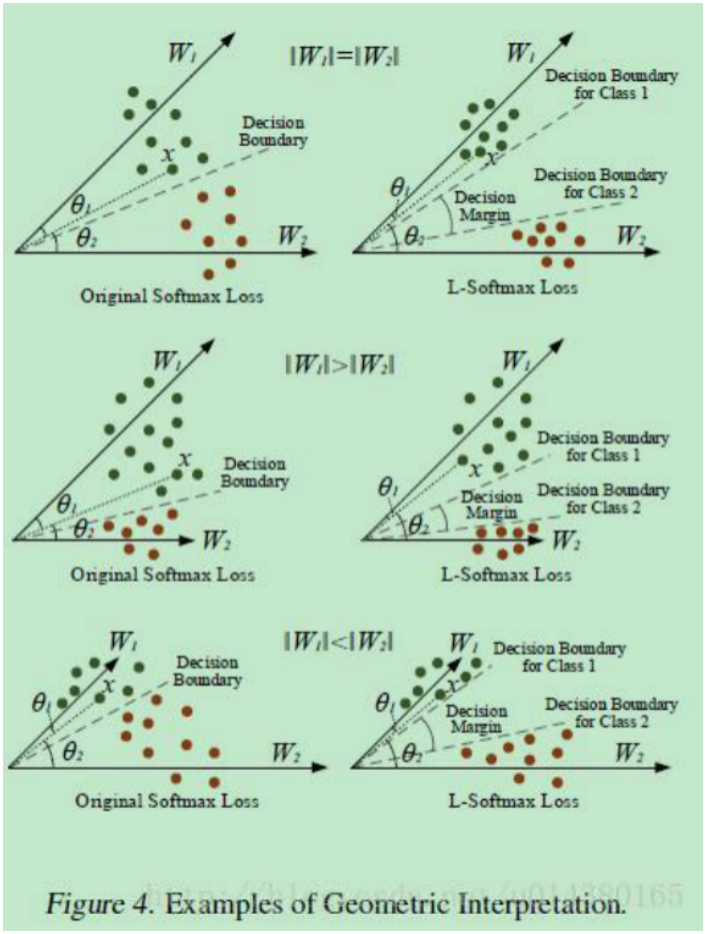

Figure 4. Examples of Geometric Interpretation.

2.2.2 总结

L-softmax loss的思想就是加大了原来softmax loss的学习难度。借用SVM的思想来理解的话，如果原来的softmax loss是只要支持向量和分类面的距离大于h就算分类效果比较好了，那么L-softmax loss就是需要距离达到mh（m是正整数）。

2.2.3 图例

上面一行表示training set，下面一行表示testing set。每一行的第一个都是传统的softmax，后面3个是不同参数的L-softmax，看看类间和类内距离的差距！

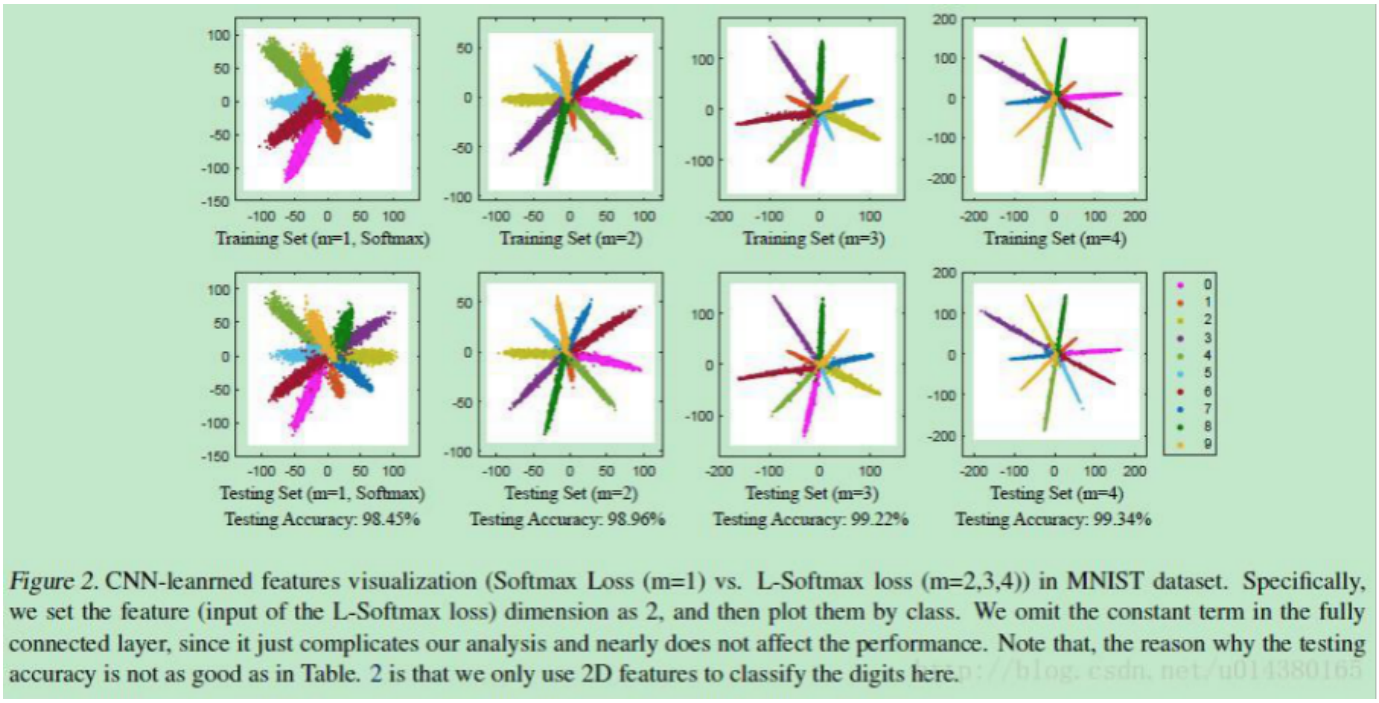

Figure 2. CNN-leanrned features visualization (Softmax Loss (m=1) vs. L-Softmax loss (m=2,3,4)) in MNIST dataset. Specifically, we set the feature (input of the L-Softmax loss) dimension as 2, and then plot them by class. We omit the constant term in the fully connected layer, since it just complicates our analysis and nearly does not affect the performance. Note that, the reason why the testing accuracy is not as good as in Table. 2 is that we only use 2D features to classify the digits here.

2.3 Angular Softmax Loss

2.3.1 论文

Liu W, Wen Y,Yu Z, et al. Sphereface: Deep hypersphere embedding for face recognition[C]//The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017, 1.

2.3.2 理论

angular softmax loss也称A-softmax loss。它就是在large margin softmax loss的基础上添加了两个限制条件||W||=1和b=0，使得预测仅取决于W和x之间的角度 θ 。

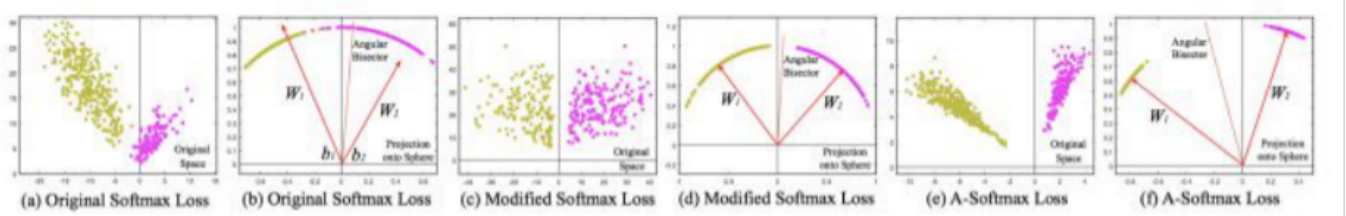

Figure 2: Comparison among softmax loss, modified softmax loss and A-Softmax loss. In this toy experiment, we construct a CNN to learn 2-D features on a subset of the CASIA face dataset. In specific, we set the output dimension of FC1 layer as 2 and visualize the learned features. Yellow dots represent the first class face features, while purple dots represent the second class face features. One can see that features learned by the original softmax loss can not be classified simply via angles, while modified softmax loss can. Our A-Softmax loss can further increase the angular margin of learned features.

上图分别比较了原softmax loss，原softmax loss添加||w||=1约束，以及在L-softmax loss基础上添加||w||=1约束的结果。

为什么要添加|w|=1的约束呢？

作者做了两方面的解释，一个是softmax loss学习到的特征，本来就依据角度有很强的区分度，另一方面，人脸是一个流形，将其特征映射到超平面表面

2.4 L2-constrained Softmax Loss

2.4.1 论文

Ranjan R, Castillo C D, Chellappa R. L2-constrained softmax loss for discriminative face verification[J]. arXiv preprint arXiv:1703.09507, 2017.

2.4.2 理论

将学习的特征x归一化。

文章作者观测到好的正面的脸，特征的L2-norm大，而特征不明显的脸，其对应的特征L2-norm小，因此提出这样的约束来增强特征的区分度。

$$\begin{aligned} \text{minimize} \quad & -\frac{1}{M}\sum_{i=1}^{M}\log\frac{e^{W_{y_i}^T f(\mathbf{x}_i)+b_{y_i}}}{\sum_{j=1}^{C} e^{W_j^T f(\mathbf{x}_i)+b_j}} \qquad (3) \\ \text{subject to} \quad & \|f(\mathbf{x}_i)\|_2 = \alpha, \ \ \forall i = 1, 2, ...M, \end{aligned}$$

2.5 Additive Margin Softmax Loss

2.5.1 论文

Wang F, Liu W,Liu H, et al. Additive Margin Softmax for Face Verification[J]. arXiv preprint arXiv:1801.05599, 2018.

2.5.2 理论

原理：把L-Softmax的乘法改成了减法，同时加上了尺度因子s。

作者这样改变之后前向后向传播变得更加简单。其中W和f都是归一化过的，作者在论文中将m设为0.35。

$$\begin{aligned} \mathcal{L}_{AMS} &= -\frac{1}{n}\sum_{i=1}^{n} log\frac{e^{s\cdot(cos\theta_{y_i}-m)}}{e^{s\cdot(cos\theta_{y_i}-m)}+\sum_{j=1,j\neq y_i}^{c} e^{s\cdot cos\theta_j}} \\ &= -\frac{1}{n}\sum_{i=1}^{n} log\frac{e^{s\cdot(W_{y_i}^T f_i-m)}}{e^{s\cdot(W_{y_i}^T f_i-m)}+\sum_{j=1,j\neq y_i}^{c} e^{sW_j^T f_i}}. \end{aligned} \qquad (6)$$

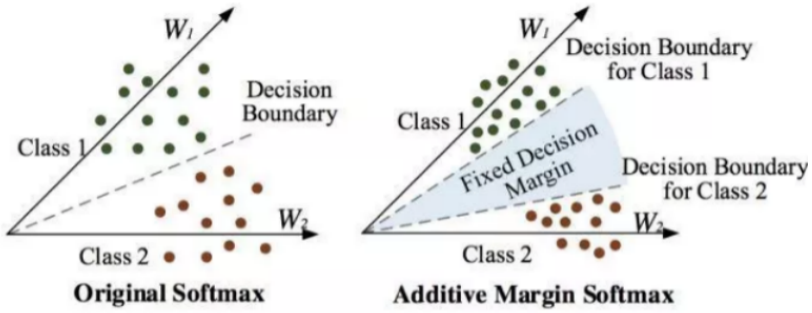

Figure 1. Comparison between the original softmax loss and the additive margin softmax loss. Note that, the angular softmax [9] can only impose unfixed angular margin, while the additive margin softmax incorporates the fixed hard angular margin.

normalization是收敛到好的点的保证，同时，必须加上scale层，scale的尺度在文中被固定设置为30。

什么时候需要normalization什么时候又不需要呢？

这实际上依赖于图片的质量。

公式如下

$$y = \frac{x}{\alpha} \quad \Rightarrow \quad \frac{dy}{dx} = \frac{1}{\alpha}.$$

其中 α 就是向量x的模，它说明模值比较小的，会有更大的梯度反向传播误差系数，这实际上就相当于难样本挖掘了。不过，也要注意那些质量非常差的，模值太小可能会造成梯度爆炸的问题。

2.6 Argface Additive Angular Margin

2.6.1 论文

Deng J, Guo J, Zafeiriou S. ArcFace: Additive Angular Margin Loss for Deep Face Recognition[J]. arXiv preprint arXiv:1801.07698, 2018.

2.6.2 理论

定义：

$$L_7 = -\frac{1}{m}\sum_{i=1}^{m}\log\frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))}+\sum_{j=1,j\neq y_i}^{n}e^{s\cos\theta_j}}, \tag{9}$$

subject to

$$W_j = \frac{W_j}{\|W_j\|}, x_i = \frac{x_i}{\|x_i\|}, \cos\theta_j = W_j^T x_i. \tag{10}$$

3 总结

基于softmax loss及其改进的loss的二分类决策边界条件

Loss Functions	Decision Boundaries
Softmax	$(W_1 - W_2)\,x + b_1 - b_2 = 0$
W-Norm Softmax	$\|x\|\,(\cos\theta_1 - \cos\theta_2) = 0$
SphereFace [23]	$\|x\|\,(\cos m\theta_1 - \cos\theta_2) = 0$
F-Norm SphereFace	$s(\cos m\theta_1 - \cos\theta_2) = 0$
CosineFace [44, 43]	$s(\cos\theta_1 - m - \cos\theta_2) = 0$
ArcFace	$s(\cos(\theta_1 + m) - \cos\theta_2) = 0$

Table 1. Decision boundaries for class 1 under binary classification case. Note that, θ_i is the angle between W_i and x, s is the hypersphere radius, and m is the margin.

参考: https://www.jianshu.com/p/3c8fc9dc5ab1

https://blog.csdn.net/u014380165/article/details/76946358