信息矩阵

$$ds^2 = D_{KL} [P(y|x; \theta) P(y|x; \theta+\delta)] \approx (1)/(2) \delta^ F(\theta)\delta, \{8\}$$

$$F(\theta) = \square_{x\square} P(x), y\square P(y|x) [(\partial P(y|x; \theta))/(\partial \theta) (\partial P(y|x; \theta)^)/(\partial \theta)], \{9\}$$

其中，δ 表示参数的微小变化。上述黎曼度量提供了参数空间的几何视角。通过对黎曼度量的分析，我们可以深入理解归一化方法如何有助于神经网络训练。

5.2.2 归一化广义线性模型的几何特性

我们将几何分析的重点放在广义线性模型（Generalized Linear Model, GLM）上。以下分析结果可以很容易地推广到深度神经网络，其中Fisher信息矩阵采用块对角近似，每个块对应单个神经元的参数。

广义线性模型可以看作是通过权重向量w和偏置标量b对指数族输出分布进行参数化。为与前文保持一致，GLM的对数似然可以用求和输入a表示为：

$$P(y|x; w,b) = ((a+b)y - \eta(a+b))/(\varphi) + c(y,\varphi), \{10\}$$

$$\square[y|x] = f(a+b) = f(w^ x + b), \{Var\}[y|x] = \varphi f'(a+b), \{11\}$$

其中，$f(\cdot)$是传递函数（类比神经网络中的非线性激活函数），$f'(\cdot)$是传递函数的导数，$\eta(\cdot)$是实值函数，$c(\cdot)$是对数配分函数。φ是缩放输出方差的常数。假设使用H个独立的GLM对H维输出向量$y = [y_1,y_2,\cdot s,y_H]$建模，且 $P(y|x; W,b) = \sum_{i=1}^H P(y_i|x; w_i,b_i)$。令W为权重矩阵（其行向量为各GLM的权重向量），b为长度为H的偏置向量，{vec}(·)表示Kronecker向量化算子。多维GLM关于其参数$\theta = [w_1^,b_1,\cdot s,w_H^,b_H]^ =$ {vec}([W,b]^)的Fisher信息矩阵即为数据特征与输出协方差矩阵的期望Kronecker积：

$$F(\theta) = \square_{x\square} P(x) [\{\{Cov\}[y|x]\}\{\varphi^2\} \{bmatrix\} xx^ & x \\\\ x^ & 1 \{bmatrix\}]. \{12\}$$

通过对原始模型中求和输入a应用归一化方法（通过μ和σ），我们得到归一化的GLM。不失一般性，我们用{F}表示带有额外增益参数$\theta =$ {vec}([W,b,g]^)的归一化多维GLM下的Fisher信息矩阵：

$$\{F\}(\theta) = \{bmatrix\} \{F\}_{11} & \cdot s & \\\\ & & \\\\ & \cdot s & \{bmatrix\}$$

归一化多维GLM下的费雪信息矩阵

我们将{F}表示为带有额外增益参数$\theta =$ {vec}([W,b,g]^)的归一化多维GLM下的费雪信息矩阵：

$$ {F}(θ) = {bmatrix} {F}_11 & ·s & {F}_1H \\ & & \\ {F}_H1 & ·s & {F}_HH {bmatrix}, {F}_ij = □_x□ P(x) [{Cov}[(y_i)/(φ), (y_j)/(φ) | x] {bmatrix} (χ_i χ_j^)/(σ_i g_i σ_j g_j) & (χ_j^)/(σ_i g_j) \\ (χ_i)/(σ_j g_i) & (a_i a_j)/(σ_i σ_j) - (a_i μ_j)/(σ_i σ_j) - (a_j μ_i)/(σ_i σ_j) + (μ_i μ_j)/(σ_i σ_j) {bmatrix}] $$

$$ χ_i = x - (∂ μ_i)/(∂ w_i) - (a_i - μ_i)/(σ_i) (∂ σ_i)/(∂ w_i) $$

通过权重向量增长实现的隐式学习率降低

值得注意的是，与标准GLM相比，沿权重向量w_i方向的块{F}_ij受到增益参数和归一化标量σ_i的缩放影响。如果权重向量w_i的范数增长两倍，即使模型输出保持不变，费雪信息矩阵也会不同。

沿w_i方向的曲率将变化(1)/(2)因子，因为σ_i也会变为两倍。因此，在归一化模型中相同的参数更新下，权重向量的范数实际上控制了权重向量的学习率。在学习过程中，具有较大范数的权重向量更难改变其方向。因此，归一化方法通过这种方式隐式地调节了学习率。

实验结果

权重向量学习机制

权重向量的范数（norm）有效地控制了该向量的学习速率。在学习过程中，具有较大范数的权重向量更难改变其方向。因此，归一化方法对权重向量具有隐式的"早停"（early stopping）效应，有助于稳定学习过程直至收敛。

输入权重幅度的学习

在归一化模型中，输入权重的幅度通过增益参数（gain parameters）显式参数化。我们比较了两种情况下模型输出的变化： 1. 在归一化GLM中更新增益参数 2. 在原始参数化下学习过程中更新等效权重的幅度

沿{F}中增益参数的方向捕获了输入权重幅度的几何特性。我们发现标准GLM中沿输入权重幅度的黎曼度量（Riemannian metric）由其输入范数缩放，而批量归一化和层归一化模型对增益参数的学习仅取决于预测误差的幅度。因此，归一化模型中对输入权重重幅度的学习相比标准模型，对输入及其参数的缩放更具鲁棒性。详见附录中的详细推导。

实验设置

我们在6个任务上进行了层归一化实验，重点关注循环神经网络（RNN）应用： 1. 图文排序（image-sentence ranking） 2. 问答系统（question-answering） 3.

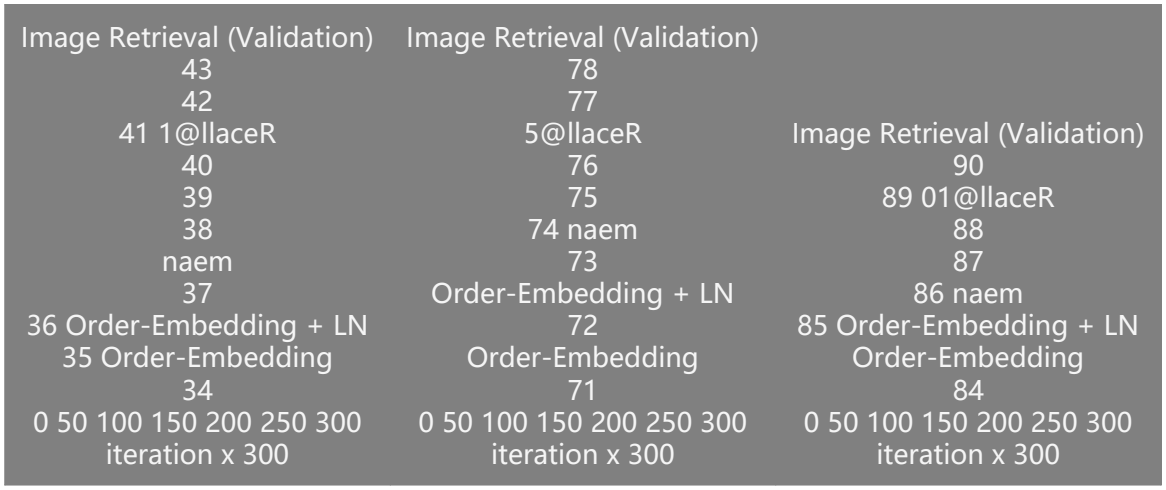
上下文语言建模（contextual language modelling） 4. 生成建模（generative modelling） 5. 手写序列生成（handwriting sequence generation） 6. MNIST分类

除非另有说明，实验中层归一化的默认初始化设置是将自适应增益设为1，偏置设为0。

实验结果图表

图像检索性能

Figure 1: 使用带/不带层归一化的顺序嵌入（order-embeddings）的Recall@K曲线



![图像检索验证集结果] (a) Recall@1 (b) Recall@5 (c) Recall@10

表格数据

Table 2: 在5个测试分割上的图文检索平均结果

| 模型 | 字幕检索 | 图像检索 | |------|---------|---------| | | R@1 | R@5 | R@10 | Meanr | R@1 | R@5 | R@10 | Meanr | | Sym[Vendrov et al., 2016] | 45.4 | 88.7 | - | 5.8 | 36.3 | - | 85.8 | 9.0 | | OE[Vendrov et al., 2016] | 46.7 | 88.9 | - | 5.7 | 37.9 | - | 85.9 | 8.1 | | OE(ours) | 46.6 | 79.3 | 89.1 | 5.2 | 37.8 | 73.6 | 85.7 | 7.9 | | OE+LN | 48.5 | 80.6 | 89.8 | 5.1 | 38.9 | 74.3 | 86.3 | 7.6 |

注： - R@K表示Recall@K（越高越好） - Meanr表示平均排名（越低越好） - Sym对应对称基线方法 - OE表示顺序嵌入方法

6.1 图像与语言的序嵌入

在本实验中，我们将层归一化（layer normalization）应用于Vendrov等人[2016]最近提出的序嵌入模型（order-embeddings mod

el），用于学习图像和句子的联合嵌入空间。我们遵循与Vendrov等人[2016]相同的实验方案，并修改了他们公开可用的代码以加入层归一化[1]，该代码使用了Theano框架[Team等人，2016]。

来自微软COCO数据集[Lin等人，2014]的图像和句子被嵌入到一个共同的向量空间中，其中使用GRU[Cho等人，2014]对句子进行编码，并使用预训练VGG卷积网络[Simonyan和Zisserman，2015]（10-crop）的输出对图像进行编码。序嵌入模型将图像和句子表示为2级偏序关系，并替换了Kiros等人[2014]中使用的余弦相似度评分函数，改用非对称评分函数。

[1] https://github.com/ivendrov/order-embedding