信息矩阵

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

$$F(\theta) = \mathbb{E}_{x\sim P(x),\, y\sim P(y|x)}\left[\frac{\partial \log P(y|x;\theta)}{\partial \theta}\frac{\partial \log P(y|x;\theta)^\top}{\partial \theta}\right], \quad (9)$$

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的训练。

5.2.2 归一化广义线性模型的几何性质

我们的几何分析聚焦于广义线性模型（Generalized Linear Model, GLM）。以下分析结果可以轻松推广到具有块对角近似Fisher信息矩阵的深度神经网络，其中每个块对应单个神经元的参数。

广义线性模型可以视为使用权重向量 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 表示如下：

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

$$\mathbb{E}[y|x] = f(a+b) = f(w^\top x + b), \quad \mathrm{Var}[y|x] = \varphi f'(a+b), \quad (11)$$

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 $y = [y_1, y_2, \cdots, y_H]$ 建模，且 $\log P(y|x;W,b) = \sum_{i=1}^{H} \log P(y_i|x;w_i,b_i)$。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，$\mathrm{vec}(\cdot)$ 表示Kronecker向量化算子。多维GLM关于其参数 $\theta = [w_1^\top, b_1, \cdots, w_H^\top, b_H]^\top = \mathrm{vec}([W,b]^\top)$ 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：

$$F(\theta) = \mathbb{E}_{x\sim P(x)}\left[\frac{\mathrm{Cov}[y|x]}{\varphi^2} \otimes \begin{bmatrix} xx^\top & x \\ x^\top & 1 \end{bmatrix}\right]. \quad (12)$$

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

$$\bar{F}(\theta) = \begin{bmatrix} \bar{F}_{11} & \cdots & \vdots \\ \vdots & \ddots & \vdots \\ \vdots & \cdots & \vdots \end{bmatrix}$$

归一化多维GLM中的Fisher信息矩阵

矩阵结构

我们将归一化多维GLM（Generalized Linear Model）中带有额外增益参数$\theta = \mathrm{vec}([W,b,g]^\top)$的Fisher信息矩阵表示为$\bar{F}$：

$$\bar{F}(\theta) = \begin{bmatrix} \bar{F}_{11} & \cdots & \bar{F}_{1H} \\ \vdots & \ddots & \vdots \\ \bar{F}_{H1} & \cdots & \bar{F}_{HH} \end{bmatrix}, \quad \bar{F}_{ij} = \mathbb{E}_{x\sim P(x)} \begin{bmatrix} \mathrm{Cov}[y_i, y_j|x] & \begin{bmatrix} (\chi_i \chi_j^{\mathsf{T}})/(g_i g_j) & (\chi_j)/(g_i \sigma_j) \\ (\chi_i^{\mathsf{T}})/(g_j \sigma_i) & (a_i a_j)/(\sigma_i \sigma_j) - (a_i \mu_j)/(\sigma_i \sigma_j) - (\mu_i a_j)/(\sigma_i \sigma_j) + (\mu_i \mu_j)/(\sigma_i \sigma_j) \end{bmatrix} \end{bmatrix}$$

其中：

$$\chi_i = x - (\partial \mu_i)/(\partial w_i) - (a_i - \mu_i)/(\sigma_i)\,(\partial \sigma_i)/(\partial w_i)$$

权重向量增长带来的隐式学习率降低

与标准GLM相比，沿权重向量w_i方向的块$\bar{F}_{ij}$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，Fisher信息矩阵也会发生变化。

沿w_i方向的曲率将改变(1)/(2)因子，因为σ_i也会变为两倍。因此，对于归一化模型中相同的参数更新，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...

42 41 38 37 36 35 34 0 50 100 150 200 250 300 迭代次数 x 300 1@llaceR naem 图像检索（验证集） Order-Embedding + LN 72 Order-Embedding 71 0 50 100 150 200 250 300 迭代次数 x 300 (a) Recall@1 5@llaceR naem 图像检索（验证集） Order-Embedding + LN 85 Order-Embedding 84 0 50 100 150 200 250 300 迭代次数 x 300 (b) Recall@5 01@llaceR naem 图像检索（验证集） Order-Embedding + LN Order-Embedding (c) Recall@10
图1：使用层归一化与未使用层归一化的order-embeddings模型的Recall@K曲线。

MSCOCO 标题检索 图像检索 模型 R@1 R@5 R@10 Meanr R@1 R@5 R@10 Meanr 对称基线[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(本工作) 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 表2：标题和图像检索在5个测试集上的平均结果。R@K表示Recall@K（数值越高越好），Meanr表示平均排名（数值越低越好）。Sym代表对称基线模型，OE表示order-embeddings模型。

对权重向量具有隐式的"早停"效应，有助于稳定学习过程直至收敛。

学习输入权重的幅度：在归一化模型中，输入权重的幅度通过增益参数显式参数化。我们比较了在归一化GLM中更新增益参数与原始参数化下学习过程中更新等效权重幅度的模型输出变化。bar{F}中沿增益参数的方向捕捉了输入权重幅度的几何特性。我们证明标准GLM沿输入权重幅度的黎曼度量由其输入范数缩放，而批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的幅度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。

6 实验结果

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：图像-句子排序、问答、上下文语言建模、生成建模、手写序列生成和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。

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 model），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，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，GRU[Cho等人，2014]用于编码句子，预训练的VGG卷积网络[Simonyan和Zisserman，2015]（10-crop）的输出用于编码图像。序嵌入模型将图像和句子表示为2级偏序关系，并用非对称评分函数取代了Kiros等人[2014]中使用的余弦相似度评分函数。

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