深度神经网络中层特征的隐私保护机制的优化

杨军港

1 背景介绍

最近几年深度学习的领域变得愈发火热，这主要归功于其在数据处理中表现出的异常高效的作用，因此深度学习被应用到越来越多的领域来处理庞大并且复杂的数据。由于移动设备的便携性和普遍性，对于部署深度学习到移动设备处理数据的需求也在日益增多。但是移动端设备的计算能力通常都比较弱，所以运用深度学习处理移动端处理数据时，一般都把大部分神经网络卸载到云端，移动端设备负责收集数据并进行预处理，得到数据的中层特征，再将中层特征上传到云端，进行后续的计算。因为来源非常广泛，所以这些中层特征有可能涉及较多的用户隐私。因此防止用户隐私泄露，保护数据安全成为一个热议的领域。目前存在的方法是，对需要上传的中层特征进行加噪，使得中层特征满足差分隐私的性质，从而达到保护隐私的目的。然而在差分隐私的限定下加噪之后，云端计算得到的模型精度会下降很多。因此为了能够提高后续计算的进度，对于加噪方法的优化显得非常重要。

2 主要的研究思路

这篇文章将按照神经网络分段部署的框架建立模型。根据中层特征的各自的重要程度，增加独特的噪音分布，使得信息能够满足差分隐私的需求，并且最小化扰动成本。

假设

$$F_1(\mathbf{x}) = \mathbf{M} = (m_1, m_2, ..., m_n) \in R^{d\times n}, x \in R^n$$

是移动端的神经网络生成中层特征的过程。F_1 是一个 $R^n \to R^{d\times n}$ 的映射，代表的是移动端的神经网络，而输出 $\mathbf{M} \in R^{d\times n}$ 则为移动端输出的中层特征。

$$F_2(\mathbf{M}) = \mathbf{y}^* \in R^n$$

是云端的神经网络生成最终结果的过程。F_2 是一个 $R^{d\times n} \to R^n$ 的映射，代表的是云端的神经网络，而输出 $\mathbf{y}^* \in R^n$ 则带入为 $\mathbf{x}$ 所得到的精确结果。

为了保护用户隐私，因此在中层特征上传到云端时进行加噪，使其符合差分隐私的需求。

$$P[K(\mathbf{M}) \in O] \le e^{\epsilon} P[K(\mathbf{M'}) \in O]$$

$\mathbf{M} = F_1(\mathbf{x})$，$\mathbf{M'} = F_1(\mathbf{x'})$，$\mathbf{x'}$ 与 $\mathbf{x}$ 为两个正交的数据集，K 为加噪函数，O 为 K 的值域的子集。

在本文中

$$K(\mathbf{M}) = \mathbf{S} = \mathbf{M} + \mathbf{Z} \ \ \in R^{d*n}$$

$\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n) \in R^{d\times n}$ 为噪声矩阵，且 $\mathbf{z}_i, i = 1, 2, ..., n$ 是满足独立分布的随机向量。

因此，将加噪后的中层特征传输到云端进行后续操作。

$$F_2(\mathbf{S}) = \mathbf{y} \in R^n$$

得到结果。

为了使加噪的代价最小，减小其影响。于是，可以确立本文的实验目标:

$$\min \|\mathbf{y}\text{-}\mathbf{y}^*\|$$

即

$$\min \|F_2(K(\mathbf{M})) - F_2(\mathbf{M})\|$$

$$st. \quad P[K(\mathbf{M}) \in O] \le e^{\epsilon} P[K(\mathbf{M'}) \in O]$$

这就是本文所需要解决的带约束条件的优化问题。

首先，F_2 是一个未知函数，并且在大多数的情况下都不相同，因此无法通过分析函数 F_2 来确定最佳噪声分布函数。因此，我们需要对问题进行相应的放缩以及一般化处理。

根据观察，函数 F_2 一般为连续函数或者是分段连续函数。因此，在本问题中，假设 F_2 是连续的，因此函数在闭区间上有界。于是目标函数有如下放缩：

$$\|F_2(M(\mathbf{M})) - F_2(\mathbf{M})\|$$

$$= \|F_2(\mathbf{M} + \mathbf{Z}) - F_2(\mathbf{M})\|$$

$$\leq L\|\mathbf{M} + \mathbf{Z} - \mathbf{M}\|$$

$$= L\|\mathbf{Z}\|$$

于是，将目标函数转化为：

$$\min\|\mathbf{Z}\|$$

$$st. \quad P[K(\mathbf{M}) \in O] \leq e^{\epsilon} P[K(\mathbf{M'}) \in O]$$

现有的很多方法都是得到类似上述的目标函数，但是由于差分隐私的限制条件，噪声的模不得不比较大，使得最后结果的精确度不尽如人意。本文发现，导致这一问题的原因，很大程度上是因为度量函数不够准确导致的。例如，噪声 $\mathbf{z}_1 = (1, 0, ..., 0)$ 和 $\mathbf{z}_2 = (0, 0, ..., 0, 1)$ 加载同一个中层特征上，得到的最终结果大部分情况下是不同的，并且，差距可能还会比较大，但是，$\|\mathbf{z}_1\| = \|\mathbf{z}_2\|$，可以看出，在现在的度量下，中层特征的差距相同并不能表示最后的结果也很接近，因此使用新的度量方法尤为重要。

根据 SVCCA 这篇文章所述，其中提出的全新的度量工具 SVCCA 可以较好得度量层与层之间的距离，因此，本文希望通过 SVCCA 工具，重新定义加噪后的中层特征与未加噪的中层特征之间的差距。

通过了解 SVCCA 的原理，发现其本质是通过比较信息占比较多的部分，比较两层之间的距离。因此，比较方式如下：

$$\mathbf{M} = \mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}}$$

U 和 V 都是正交矩阵，

$$\Sigma = \begin{pmatrix} \sigma_1 & & \\ & ... & \\ & & \sigma_n \\ ... & ... & ... \\ 0 & ... & 0 \end{pmatrix}$$

是与 $\mathbf{M}$ 同样规格的半正定对角矩阵，其中的元素 $\sigma_1 ... \sigma n$ 即代表了 $\mathbf{M}$ 的奇异值。取出占比为 99% 的奇异值组成方阵。即求得最小的 r 使得 $\sum_{i=1}^{r} \sigma_i \geq 0.99 \sum_{i=1}^{n} \sigma_i$，则 $\Sigma' = \Sigma_{r\times r}$

$$\mathbf{M} \approx \mathbf{M'} = \mathrm{U}_{d\times r}\Sigma_{r\times r}\mathrm{V}^{\mathrm{T}}_{\mathrm{r\times n}}$$

每个奇异值的大小，就相当于其在总的信息中所占的比重，因此，在各个奇异值上增加噪声，能体现出所加的噪声在总信息中所占的比重。

$$\mathrm{Z}' = \begin{pmatrix} z'_1 & & \\ & ... & \\ & & z'_r \end{pmatrix}$$

不妨设 z'_i 都是服从高斯分布的噪声，$p(z'_i) = \frac{1}{\sqrt{2\pi}a_i}\exp(-\frac{z'^2}{2a_i^2})$

$$\mathbf{M} + \mathrm{Z} \approx \mathrm{U}_{d\times r}(\Sigma_{r\times r} + \mathrm{Z}')\mathrm{V}^{\mathrm{T}}_{\mathrm{r\times n}}$$

$$\Rightarrow \mathrm{Z} \approx \mathrm{U}_{d\times r}\mathrm{Z}'\mathrm{V}^{\mathrm{T}}_{\mathrm{r\times n}}$$

$$\Rightarrow z_{ij} = \sum_{k=1}^{r} z'_k u_{ki} v_{kj} \quad \mathrm{Z} = \{z_{ij}\}_{d\times n}$$

于是，最小化噪声的影响，即最小化噪声在所传递的信息中的占比即可。

因此，目标函数变为：

$$\min_{a\in R^r} \sum_{i=1}^{r} |(\sigma_i + z'_i) z'_i|$$

$$st. \quad P[K(\mathbf{M}) \in O] \leq e^{\epsilon} P[K(\mathbf{M'}) \in O]$$

由于 z_i 是随机变量，所以，需要对目标函数求期望。目标变为：

$$\min_{a\in R^r}\int_{z\in R^r}\sum_{i=1}^{r}|(\sigma_i+z_i')z_i'|P(dz_1'...dz_r')$$

$$s.t.\quad P[K(\mathbf{M})\in O]\leq e^{\epsilon}P[K(\mathbf{M'})\in O]$$

由 *Differentially-Private Deep Learning from an Optimization Perspective* 这篇文章中所采用的，对约束条件所采用的方法，可以将约束条件变形。首先定义：

$$\Delta=\|x-x'\|,x\in\mathbf{M},x'\in\ \mathbf{M}'$$

$$\alpha=\sup_{\forall\mathbf{M},\mathbf{M}'s.t.d(\mathbf{M},\mathbf{M}')=1}\|x-x'\|$$

则进行如下推导：

$$\Rightarrow P[\mathbf{M}+Z\in O]\leq e^{\epsilon}P[\mathbf{M'}+Z\in O]$$

$$\Rightarrow P[Z\in O-M]\leq e^{\epsilon}P[Z\in O-M']$$

$$\Rightarrow P[Z\in O']\leq e^{\epsilon}P[Z\in O'+M-M']$$

$$\Rightarrow\max_{1\leq i\leq d,1\leq j\leq n}\frac{p(z_{ij})}{p(z_{ij}+\Delta)}\leq e^{\epsilon}$$

$$\Rightarrow\max_{1\leq i\leq d,1\leq j\leq n}\ln\frac{p(z_{ij})}{p(z_{ij}+\Delta)}\leq\epsilon,\forall\|\Delta\|\leq\alpha,\Delta\in R^d$$

因此，我们最后优化的目标为：

$$\min_{a\in R^r}\int_{z\in R^r}\sum_{i=1}^{r}|(\sigma_i+z_i')z_i'|P(dz_1'...dz_r')$$

$$s.t.\max_{1\leq i\leq d,1\leq j\leq n}\ln\frac{p(z_{ij})}{p(z_{ij}+\Delta)}\leq\epsilon,\forall\|\Delta\|\leq\alpha,\Delta\in R^d$$

根据优化的目标，进行相应的分析，最后确定每个 z_i 的确切分布，得到最后的结果。

3 实验目标

中后期，将进行相关的实验，进行如下验证：

1. 模长相同的单位噪声，1 出现的位置不同，会对最后的结果产生不同的影响，并且，在特征值越大的方向上的单位噪声，对结果产生的影响越大。

2. 和之前的加噪方式对比，如果按之前的度量来算，同样模长的噪声，在本文的加噪方式下，得到的最终结果会更加精确。

3. 对于本文所提到的加噪方式，移动端的所需要的计算时间和计算所需要的资源并不会太多。