基于钻石数据集的 FATE 实验报告

蔡剑平

本实验考虑了如下的纵向联邦学习场景。在该场景中包含了两个参与方，分别是提供专业影像数据（图片）的医疗机构 A 以及通过表格形式记录患者信息（如基本信息或体检信息）的诊所 B。为了让合作诊所 B 能够利用机构 A 的影像资料库建立联邦诊断模型，并且实现联合诊断的同时不泄露机构 A 中所包含的专业影像资料，本文建立了如下联邦学习模型。

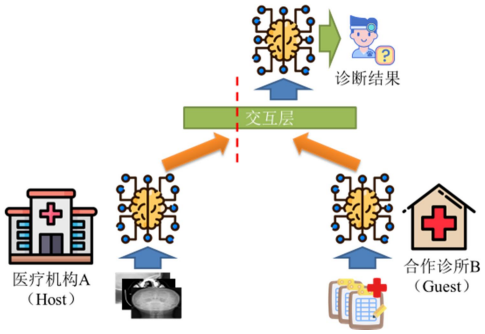

图 1 医疗图像机构与合作诊所的联合诊断

本实验采用了 HeteroNN 模块实现。不过，由于目前没有找到合适于该医疗场景的联邦学习数据集，本文先采用钻石数据集作为替代进行实验。

1. 数据集介绍

实验采用了一个钻石数据集，来自于 Kaggle 网站，网址为 https://www.kaggle.com/datasets/harshitlakhani/natural-diamonds-prices-images。根据原始数据集，本文构造了子数据集用于建立模型来判断某一些属性的钻石是否为“祖母绿”。实验数据包含了 495 个钻石样本的图片以及 713 个带有属性的钻石信息。

图 2 钻石样本图片示例（上为非祖母绿、下为祖母绿）

在预处理阶段，本文将透明度、色泽、切割、抛光、对称、荧光等几个指标进行数值化处理，按照优->劣（分值高->低）进行给定相应数字。对于图片数据，由于 FATE 框架并不能直接接受图片数据作为输入，因此本文将 30*30 的图片按照 RGB 三颜色处理为 30*30*3（高*宽*颜色通道）向量。这里的一个关键点在于 FATE 中的模型按照行优先原则组织张量的，所以向量排列时应当以颜色通道为优先变化，其次是宽。例如，第 1 个向量对应像素（1,1）

的 R 通道，第 2 个向量对应像素（1,1）的 G 通道，第 3 个向量对应像素（1,1）的 B 通道，第 4 个向量对应像素（1,2）的 R 通道，依次类推。

2. 单纯的 CNN 实验

本文首先采用如图 3 所示的 CNN 模型进行训练。实现代码见“zs_cnn.ipynb”。

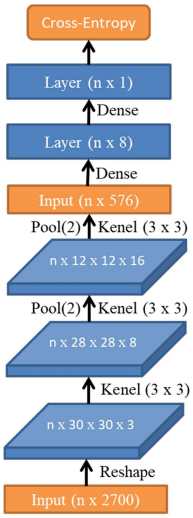

图 3 神经网络模型

实验采用交叉熵作为 Loss 函数，训练结果为 Loss 值 0.53，预测正确率为 0.87。

3. 基于钻石数据的联邦学习

根据上述应用场景，本文建立了如下结构的联邦神经网络模型进行联合学习。实验代码见 py 文件“zs_nn_pic_attr.py”。

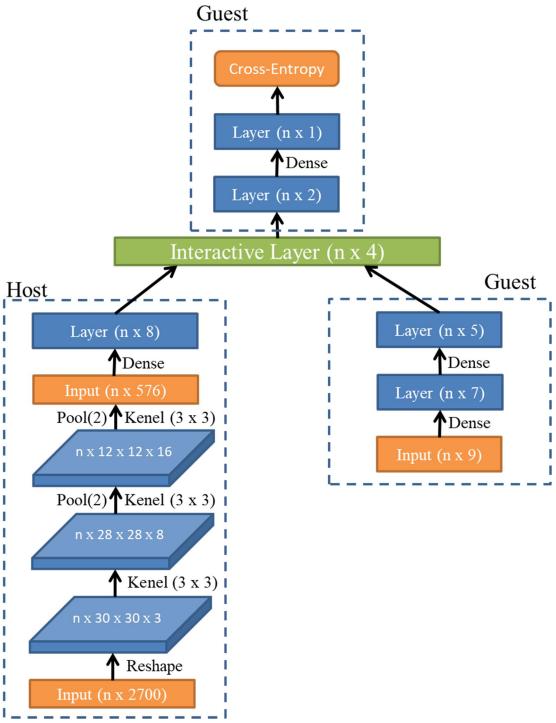

图 4 联邦神经网络模型

本实验采用了 FATE 中的 HeteroNN 模块进行，实验代码见 py 文件“sbt_vmd.py”，实验编号为 202204271143397547850。实验设置的迭代次数为 50 次。训练时长为 13 分 02 秒，训练过程的 Loss 变化如下图所示：

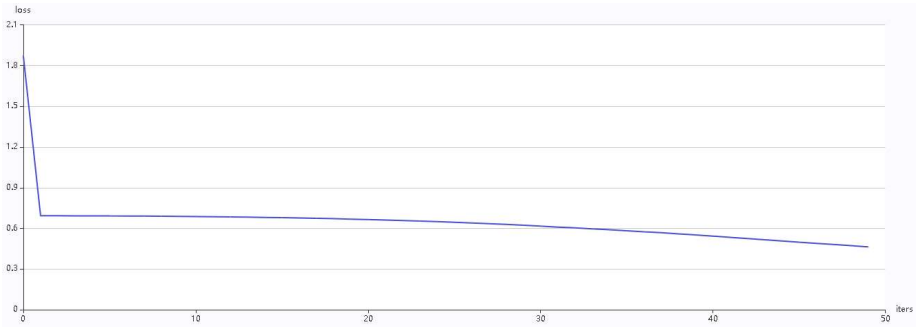

图 5 Loss 变化图

Loss 采用交叉熵函数计算。从 Loss 变化来看，Loss 值能够相对稳定地收敛。而其评价结果如下图所示：

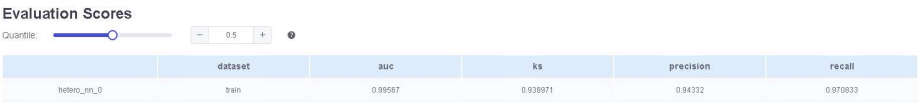

图 2.2 整体评价

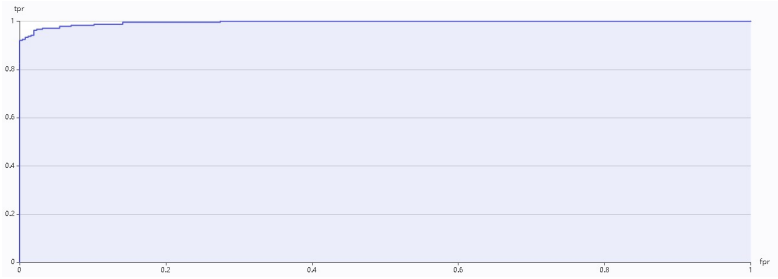

图 6 ROC 图

从上述实验结果来看，该实验获得了优秀的实验结果，AUC 值高达 0.995，精度为 0.94，召回率为 0.97，并且 ROC 曲线几乎达到饱满。上述结果初步说明，本文所构建的 CNN+Dense 联邦学习网络能够较好地胜任钻石分类学习的任务。相比于无联邦学习的算法来说，联邦学习模型具有更好预测效果。不过，在预测阈值方面，在阈值 0.5 下的预测准确率为 91.7%；在最佳阈值 0.38 下的预测准确率为 97%。虽然，该结果看起来挺不错的，但观察最佳阈值可以看出阈值 0.38 相比于 0.5 存在较大的偏差。为此，我们观察了 K-S 曲线，可以看出 K-S 曲线并不是一个完美的“橄榄”形状。也就是说，目前训练的模型存在划分点敏感的问题，若划分点选择不恰当，则所训练的模型将无法完成预测任务。

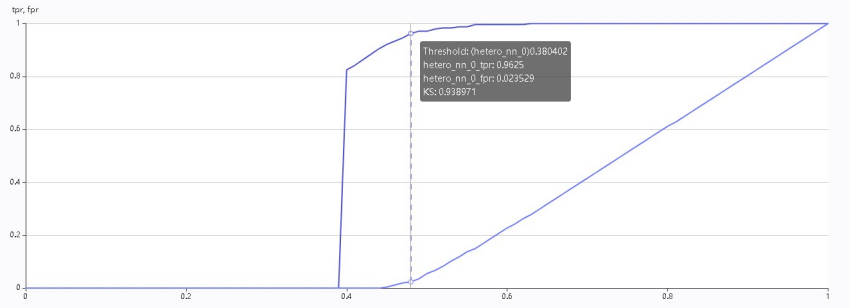

图 7 K-S 曲线

总结

从上述实验结果来看，针对于“钻石数据集”，图片和表格数据结合的形式确实能够在一定程度上提升算法准确性。不过，实验过程也暴露了 HeteroNN 的一些问题

1. 由于 FATE 框架目前并不支持 GPU 训练，因此 HeteroNN 训练的时间开销过大，这导致算法的迭代周期变得很长。
2. 训练过程对于模型结构是敏感的。不合理的模型结构容易出现早停的情况（即模型并未收敛却被 FateFlow 停止）。该问题的产生主要是网络结构不够合理导致的。虽然早停的一个常见原因是学习率设置得过大，但经过多次试验，本文设置了一个很小的学习率依旧会出现早停现象。因而，该问题需要做进一步研究。
3. 虽然模型在不出现早停的情况下几乎能够获得不错的学习效果。但容易出现类别划分点偏差的情况，如 4 所示情况是一个相对理想的实验结果。在其他实验中，常常出现，最佳划分点小于 0.1 的情况，这是不平衡数据集常出现的一个现象。目前，该问题产生的具体原因尚不完全明确。经测试，本文认为可能出现的原因是原始数据集的所有值都是非负的，将数据预处理映射到[-1,1]区间可以改善这一问题。还有一个改进思路是在 Input 层上叠加一个无激活函数层，也可以解决该问题。