平安医疗科技疾病问答迁移学习比赛评测报告

国网信通产业集团

福建亿榕信息技术有限公司

团队介绍

- **福建亿榕信息技术有限公司**是**国家电网信息通信产业集团**下属的创新型技术公司，以**非结构化数据的存储、管理、挖掘、行业**应用作为专业技术发展方向；**upside-down团队由亿榕公司研发中心的五个资深工程师组成。**
- 国家规划布局内重点软件企业
- CMMI ML5认证企业
- 全国大数据标准工作组全权成员单位
- 承担全国首个电子文件管理试点项目建设

目录


PART 01
赛题描述与分析

赛题描述及分析

任务说明

本次评测任务的主要目标是针对中文的疾病问答数据，进行病种间的迁移学习。具体而言，给定来自5个不同病种的问句对，要求判定两个句子语义是否相同或者相近。所有语料来自互联网上患者真实的问题，并经过了筛选和人工的意图匹配标注。

示例

问句1:糖尿病吃什么？
问句2:糖尿病的食谱？
label:1
问句1:乙肝小三阳的危害？
问句2:乙肝大三阳的危害？
label:0

数据说明

参赛选手的文件由train.csv、dev.csv、test.csv三个文件构成，train.csv是训练集，包含2万对人工标注好的疾病问答数据，由5个病种构成，其中diabetes10000对，hypertension、hepatitis、aids、breast_cancer各2500对。dev.csv是验证集，包含10000对无label的疾病问答数据，由5个病种构成，其中diabetes，hypertension、hepatitis、aids、breast_cancer各2000对。test.csv是测试集，包含5万对人工标注好的疾病问答数据，其中只有部分数据供验证。
category表示问句对的病种名称，分别对应：diabetes-糖尿病，hypertension-高血压，hepatitis-乙肝，aids-艾滋病，breast_cancer-乳腺癌。label表示问句之间的语义是否相同。若相同，标为1，若不相同，标为0。其中，训练集label已知，验证集和测试集label未知。

总结：这是一个医疗科技疾病领域的二分类评测任务，使用F1作为评测指标

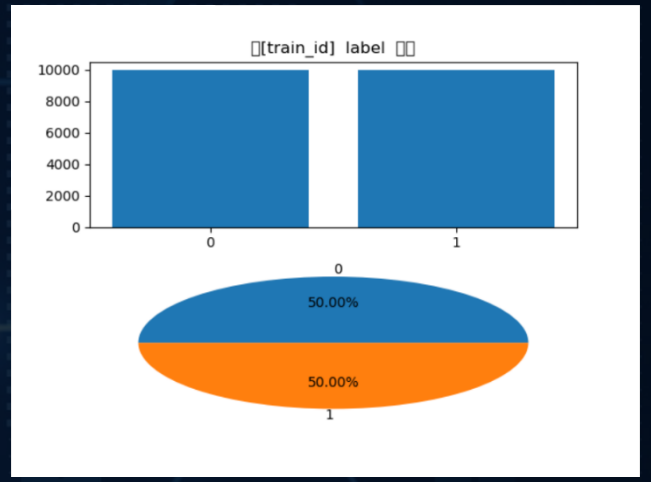
赛题描述及分析

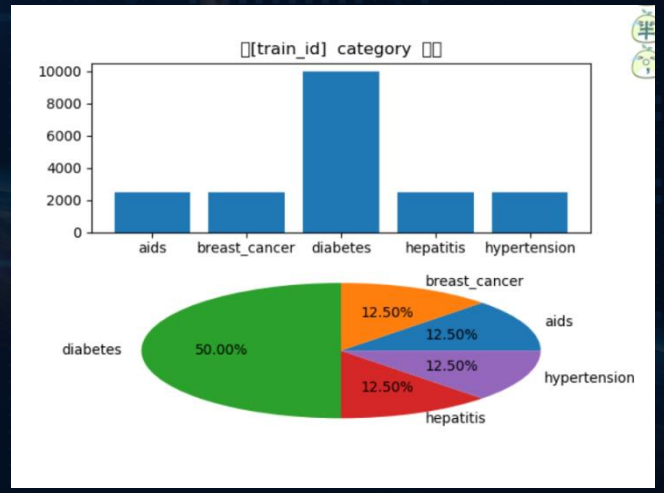
在5万条测试集中，question1有12310条是重复的，question2中有11903条重复。

这5万条数据question1有8443数据与训练集question1完全相同、与训练集question有8421完全相同，其中有458条数据是question1和question2完全相同的。其中训练集与测试集有61条完全相同。

除了以上的数据探索分析当然还可以做分词聚类，分析词语量和词语分布等等。

简单从探索数据可以看出，数据集正负比例分布均匀，数据category数据分布并不平衡，数据存在较大的重复现象，数据分词后词语较为集中。






PART 02
算法流程描述

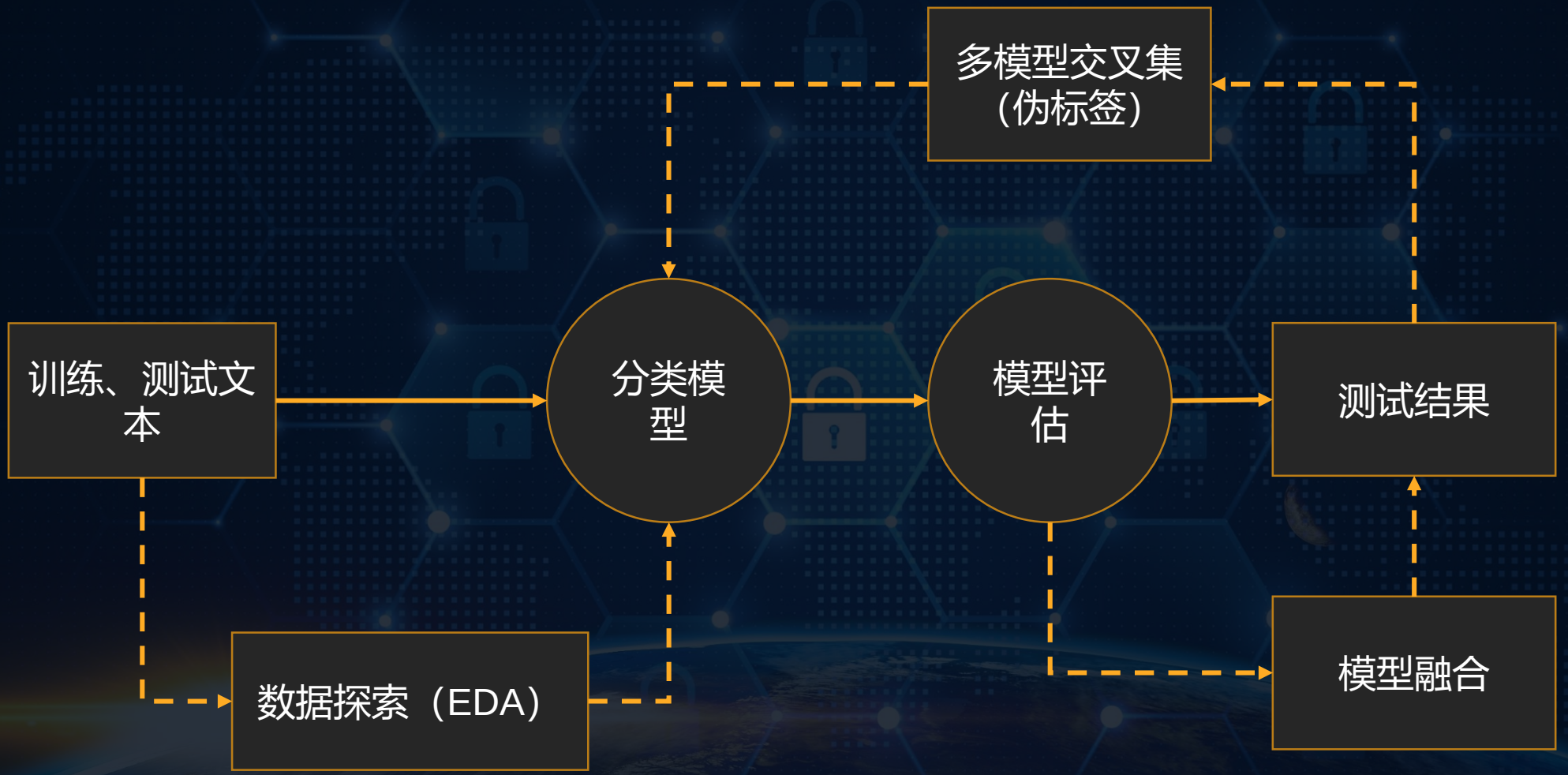
面向工程化的pipeline

利用配置文件控制流水线作业，对于新任务，配置流水线配置文件即可

面向切面的流水线控制作业，模块间可复用，灵活切换

面向持续交付，整个过程按照持续交付的过程设计，便于模型复现

算法方案



落地分析

工程化门槛低

全过程使用流水线配置，自动推荐经验参数，无需ＡＩ开发经验也能训练出适合的模型；

灵活易用

整体框架基于面向切面设计，可以灵活选择适合自己领域的算法进行模型训练，快速实现模型投产；

易于部署和复盘

全过程基于持续交付体系设计，基于docker部署，自动将git repo和commit与流水线中的训练过程连接起来，自动记录模型并在集中的位置创建副本，可以轻松地对模型和初始权重进行共享及复盘。


PART 03
关键技术介绍

伪标签

伪标签学习中使用深度学习网络作为分类器，就是把网络对无标签数据的预测结果，作为无标签数据的伪标签（Pseudo label），用来对网络进行再训练。

方法虽然简单，但是效果很好，比单纯用有标签数据有不少的提升。其主要的贡献在于损失函数的构造：

$$L = \sum_{m=1}^{n} \sum_{i=1}^{C} L(y_i^m, f_i^m) + \alpha(t) \sum_{m=1}^{n'} \sum_{i=1}^{C} L(y'^m_i, f'^m_i)$$

损失函数的第一项是有标签数据的损失，第二项是无标签数据的损失，在无标签数据的损失中， y' 为无标签数据预测得到的伪标签，是直接取网络对无标签数据的预测的最大值为标签。 其中 α(t)决定着无标签数据的代价在网络更新的作用，选择合适的 α(t) 很重要，太大性能退化，太小提升有限。在网络初始时，网络的预测时不太准确的，因此生成的伪标签的准确性也不高。在初始训练时， α(t) 要设为 0，然后再慢慢增加。

本次评测采用多模型预测结果进行交叉取交集并做数据标签平衡参与二次模型训练，单模型性能提高3个百分点。

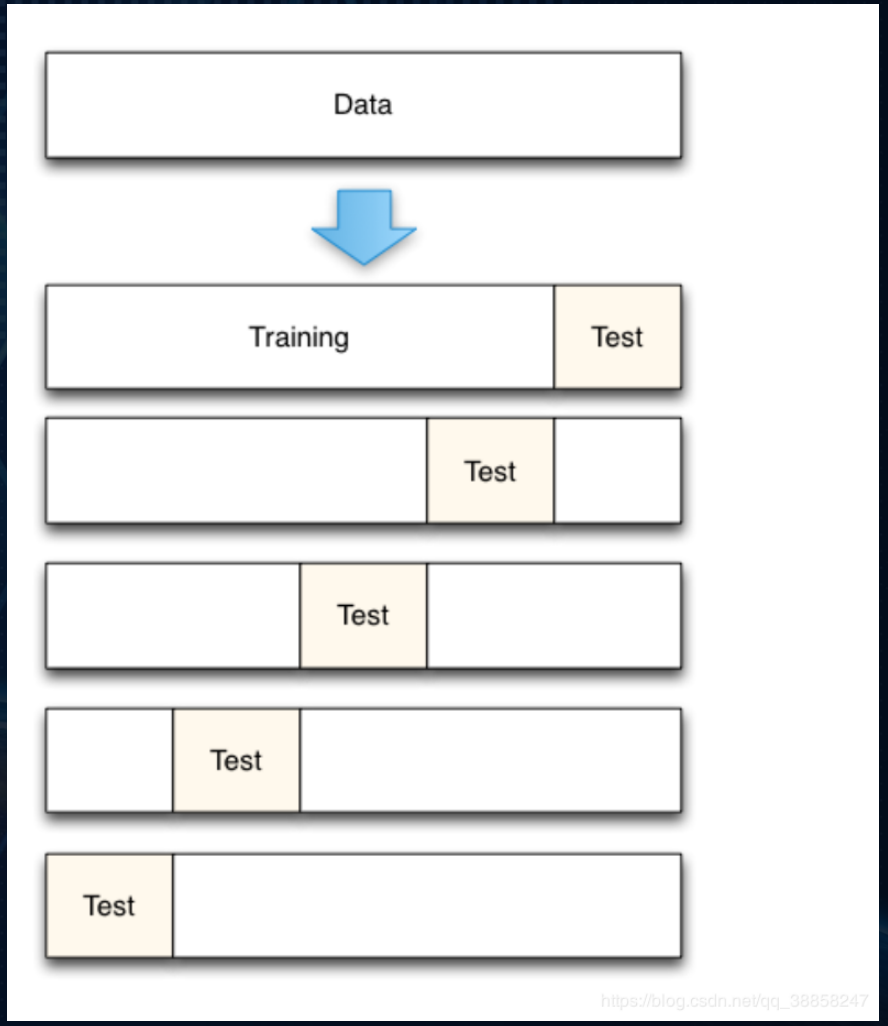
模型融合

在使用训练集对参数进行训练的时候，经常会发现人们通常会将一整个训练集分为三个数据集。一般分为：**训练集**（train_set），**验证集**（valid_set），**测试集**（test_set）这三个部分。这其实是为了保证训练效果而特意设置的。其中测试集很好理解，其实就是完全不参与训练的数据，仅仅用来观测测试效果的数据。而训练集和验证集则牵涉到下面的知识了。

因为在实际的训练中，训练的结果对于训练集的拟合程度通常还是挺好的（初始条件敏感），但是对于训练集之外的数据的拟合程度通常就不那么令人满意了。因此我们通常并不会把所有的数据集都拿来训练，而是分出一部分来（这一部分不参加训练）对训练集生成的参数进行测试，相对客观的判断这些参数对训练集之外的数据的符合程度。这种思想就称为交叉验证

本次评测采用**5折交叉验证**的方法，通多对训练数据进行5次训练，即把数据划分为5等份，5份轮番取1份作为验证数据集，其余4份作为训练集，最终按投票/加权平均的方式对测试集进行推理预测。

该方法能够最大限度使用所有的数据集，防止单次训练模型效果差，通过投票或加权平均的方式能够有效表示该模型的性能。




PART 04
总结与展望

总结

1. 深度学习调试困难来至有许多相互预制的错误来源
2. 为训练bug-free深度学习模型，我们将构建模型看作一个迭代过程
3. 以下步骤可让过程更容易,并且尽可能早捕捉到错误

1.从简单开始 – 尽可能选择最简单的模型和数据

2.实现与调试 – 一旦模型可以运行，过拟合单个批次，并重现已知结果

3.模型评估 – 应用偏差/方差分解方法决定下一步怎么做

4.调试超参 – 使用粗调（coarse-to-fine）随机搜索或贝叶斯优化.

5.提升模型/数据 – 如果欠拟合，则增大模型；如果过拟合，则增大训练数据或正则化

未来展望

1、在现有的pipeline的基础上完成全过程可视化跟踪及复盘

2、实现参数自动化分析及可视化跟踪

3、降低ＡＩ工程化的准入门槛，实现算法在业务领域的快速落地


谢谢