课程大作业

2021 亚太眼科大数据竞赛-技术报告

迟慧璇[1*]，朱赫[1]，黄琳焱[2]

1. 中科院计算所，学号：202128013229082，202128013229031

2. 厦门大学信息学院

* E-mail: chihuixuan21@mails.ucas.ac.cn

摘要 2021 亚太眼科学会大数据竞赛是亚太眼科学会 (Asia Pacific Tele-Ophthalmology Society, APTOS) 主办的全球人工智能挑战赛，其主题是预测糖尿病性黄斑水肿 (DME) 患者的 Anti-VEGF 抗血管内皮生长因子 (简称 Anti-VEGF) 治疗转归。首先，我们对数据集进行了预处理操作，得到了模型可用的干净数据。接着，我们设计了包含上游主网络、上游辅助网络以及下游网络这三部分的神经网络模型，并进行了训练和推理。最终，我们队 (UCAS_DM_GTL) 在初赛中的排名为 **85/10006***，进入了复赛。代码地址：`https://github.com/ytchx1999/UCAS_DM_GTL_Tianchi`[†].

关键词 APTOS, Data Mining, Tianchi-Competition

1 问题描述

1.1 问题背景

抗血管内皮生长因子（Anti-VEGF）药物可用于治疗若干导致眼后视网膜黄斑区下新生血管生长或肿胀的眼部疾病。但是，据不同研究报告，有大约 10% 到 50% 的患者每月都接受 Anti-VEGF 注射治疗，却仍然没有达到充分的治疗效果。本次项目的目的为建立机器学习模型来预测糖尿病性黄斑水肿 (DME) 患者接受 6 个月 Anti-VEGF 治疗后的反应，以此来协助眼科医生在治疗前确定无反应患者，并为其定制个性化治疗方案。比赛时间轴如表1所示。

*由于初赛的时间延长（11.19 初赛结束），时间关系，我们并没有参与复赛，此处的排名为初赛（赛季 1）的排名，特此说明。

†代码仓库共包含 5 个分支，其中 `tianchi_v4` 版本是最终版本，我们已经将其合并到了 `main` 分支。

表 1 时间轴

Date	Activity
2021-08-28	开始报名
2021-09-01～2021-11-19	**初赛（赛季 1）**
2021-11-23～2021-12-24	复赛

1.2 数据集

本项目数据集分为 CSV 文件和 OCT 图像两部分，CSV 文件如图1所示，包含了患者的基本信息、诊断结果、治疗方案以及治疗前后 OCT 图像的一系列标注数据。

patient ID	gender	age	diagnosis	preVA	anti-VEGF	preCST	preIRF	preSRF	prePED	preHRF	VA	continue injection	CST	IRF	SRF	PED	HRF
			1=wet														
	1=male		AMD		1=bevacizumab		1=present	1=present	1=present	1=present		1=yes		1=present	1=present	1=present	1=present
	2=female		2=PCV		2=ranibizumab		0=absent	0=absent	0=absent	0=absent		0=no		0=absent	0=absent	0=absent	0=absent
			3=DME		3=aflibercept												
			4=RVO		4=conbercept												
			5=CME		0=not receiving anti-VEGF												
			6=fellow eye(not receive anti-VEGF)														
			9=other diagnosis														

图 1 csv 文件内容

OCT 图像包含患者治疗前后的眼球扫描图像，如图2所示，左半部分为扫描的横截面位置，右半部分为对应的横截面图像。

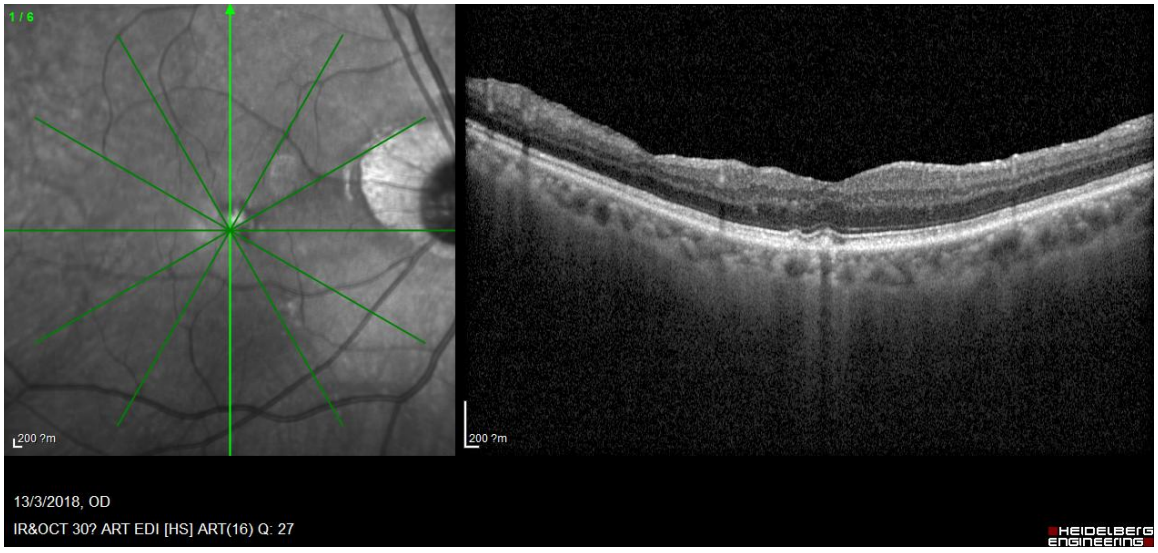

图 2 OCT 图像

1.3 问题定义

本项目所要解决的问题涉及医学领域内的图像分割和图像处理，通过结合 CSV 文件中的患者信息和图像标注数据来学习患者眼球 CT 图像中的信息，以此解决相应的分类问题（continue，IRF，

SRF，HRF，PED）和回归问题（CST，VA）。

2 数据预处理

2.1 数据分析

CSV 文件：首先对训练集中 CSV 文件内的各个属性进行初步分析，包括属性的变量类型、值域、缺值信息、均值方差等，掌握数据的大致结构。训练集中提供了每个患者的个人信息（gender，age）、疾病种类（diagnosis）、治疗方案（Anti-VEGF）、治疗前后视力（preVA，VA）、治疗前后中心视网膜厚度（preCST，CST）、治疗前后的四种症状（IRF，preIRF，SRF，preSRF，PED，prePED，HRF，preHRF）。

我们计算这些属性间的 Pearson 相关系数，并绘制了如图3所示的 Heatmap 图像。其中的正相关性较强（r>0.6）的有 preVA 与 VA、prePED 与 PED、preIRF 与 IRF、preHRF 与 HRF；负相关性较强（r<-0.4）的有 preSRF 与 diagnosis、prePED 与 diagnosis、continue injection 与 diagnosis、PED 与 diagnosis。

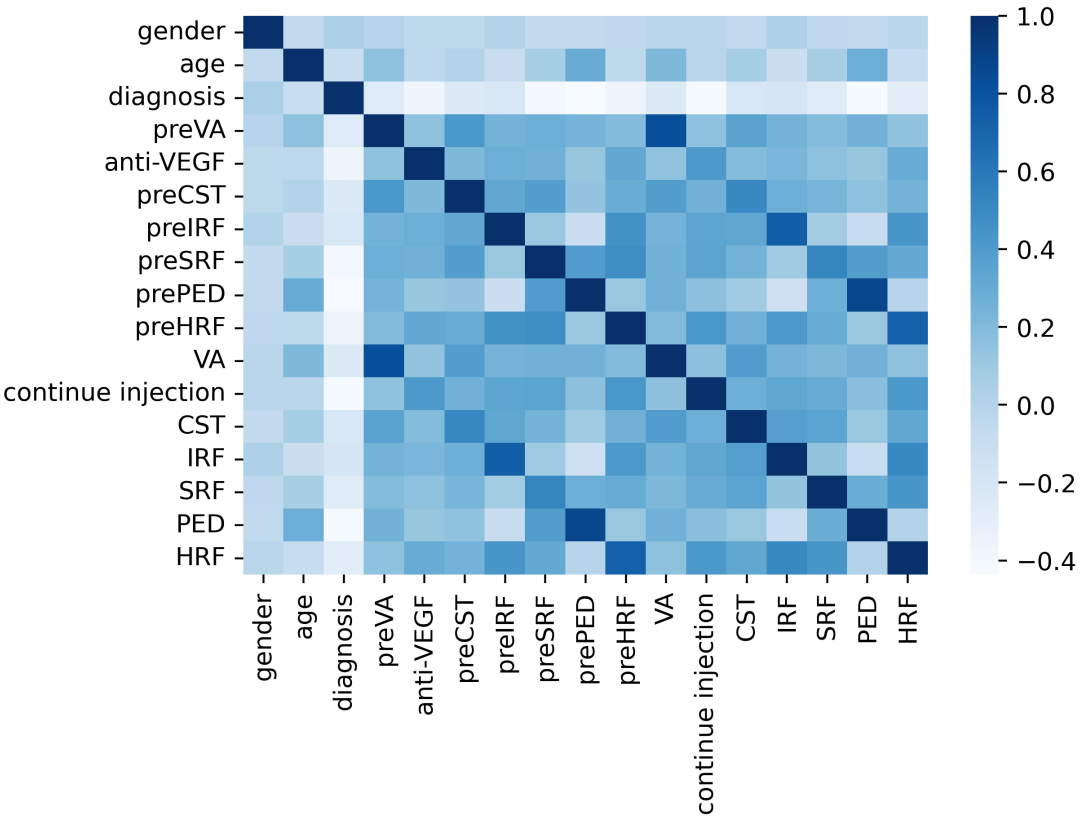

图 3 Heatmap 图像

在测试集的 CSV 文件中，只提供了患者 gender、age、diagnosis、preVA 和 anti-VEGF 五个

属性，剩下的属性均需要根据患者的 OCT 图像进行预测。

OCT 图像：在提供的图像数据集中，每个子文件夹包含一个患者的所有眼球扫描图像，根据官方提供的命名规则来区分左右眼球和治疗前后。通过抽样分析我们可以得出，OCT 图像数据质量较差，存在以下几种问题：有些患者只包含单只眼数据；有些患者治疗前后图像数量不一致，存在图像缺失；OCT 图像格式不统一，眼球横截面有水平和垂直两种方式。

2.2 数据预处理

CSV 文件：首先处理数据中的缺失信息。因为缺失的属性值多为患者某类症状的判别信息，不容易补全，并且缺失属性值的患者数量较少，所以我们选择直接剔除包含 NaN 的患者信息。

接下来，我们对训练集中的属性值进行预处理。首先我们对 gender、age 和 preVA 三个属性值做 L2 归一化，和 diagnosis、anti-VEGF 拼接后作为模型的输入向量。然后我们将 preCST、CST 和 VA 进行 Z-Score 标准化后和其余属性拼接作为模型的输出向量。测试集的 CSV 文件和训练集的输入向量处理方法相同，我们将处理后的训练集和测试集数据分别存储，等待下一步模型调用。

OCT 图像：我们首先对图像进行裁剪，只保留右半部分的横截面图像。因为数据集的命名不规范，我们需要在遍历文件夹时匹配前面的编号字段。考虑到图像数据存在缺失、格式不统一等问题，我们将每个患者根据左右眼、治疗前后共分成四类，如果每一类下有图片则进行平均化处理。如图4所示，患者 0 的左眼治疗前有 5 张 OCT 图像，经过裁剪和累加后我们可以得到左下角的融合后的图像。

得到融合的数据集后，我们同样对每个眼睛治疗前后的图片中随机抽出一张组成补充数据集，以进一步提高模型的学习精度。在将上述图片读入模型前，我们还需要对训练集的图像做随机裁剪和旋转，最终完成图像的预处理工作。

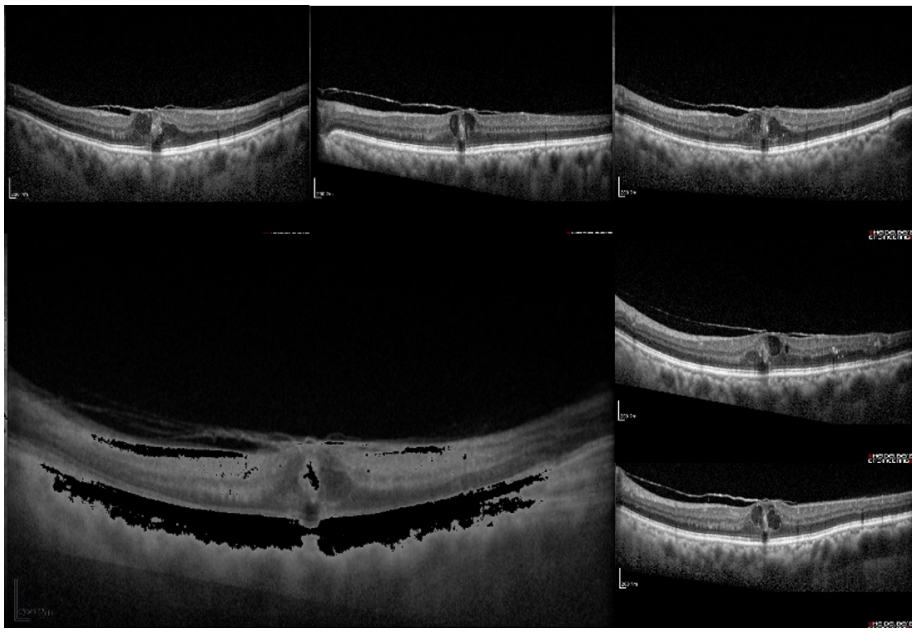

图 4 预处理后的图像

2.3 最终数据集

在对 CSV 文件和 OCT 图像均进行预处理后，我们得到最终的目录树结构如图5所示：

```
├── LICENSE
├── README.md
├── data
│   ├── APTOS\ Big\ Data\ Competition\ Preliminary\ Dataset.rtf
│   ├── Flowchart.png
│   ├── PreliminaryValidationSet_Info.csv
│   ├── TrainingAnnotation.csv
│   ├── data_train.ipynb
│   ├── resnet18-5c106cde.pth
│   ├── submit.csv
│   ├── test_data.pk
│   ├── train_data.pk
│   └── tree.txt
├── dataset
|   ├── split
│   ├── mix_train
│   │   ├── 0000-0000
│   │   │   ├── 0000-0000L_1.jpg
│   │   │   ├── 0000-0000L_2.jpg
│   │   │   ├── ...
│   │   └── ...
│   |── mix_test
|   ├── pkl_train
|   └── pkl_test
└── src
    ├── __init__.py
    ├── main.py
    ├── models
    │   ├── __init__.py
    │   └── model.py
    └── utils
        ├── __init__.py
        ├── corr_csv.py
        ├── dataset.py
        ├── mix_images.py
        ├── preprocess_csv.py
        ├── transfer_img.py
        └── preprocess_img.py
```

图 5 目录树

- 分割图像数据集 (https://drive.google.com/u/0/uc?id=1i9tnBbfM3tkO5GyMtytsZcUIxU1GytOW&export=download)
- 融合图像数据集 (https://drive.google.com/u/0/uc?id=1Wc0CmqeZg_gJkiiqoB1EZTOS0seB1MF4&export=download)
- 数据集补充特征 (https://drive.google.com/u/0/uc?id=1h2aHyAxEaVbM23YGP6dr_1pdwWIbG2NT&export=download)

3 模型结构与代码实现

3.1 模型结构

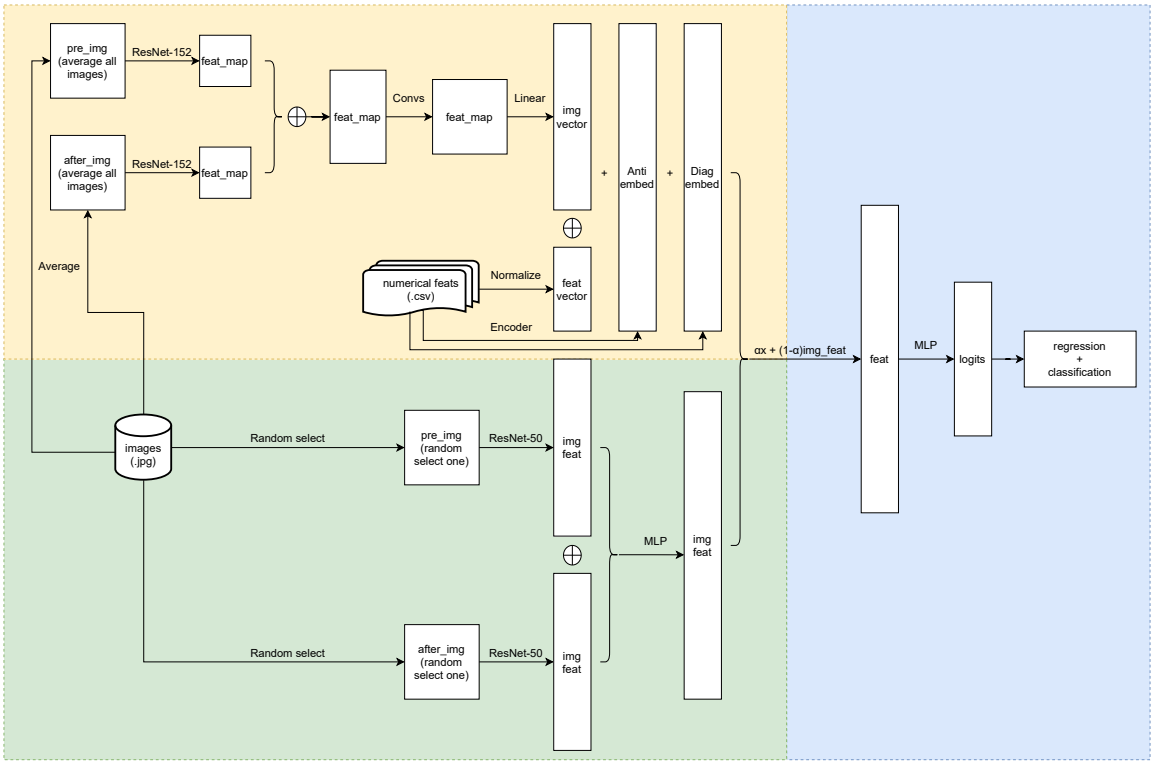

图 6 模型整体结构

如图6所示，模型（`model.py`）总共分为三个部分：黄色区域代表的上游主网络部分、绿色区域代表的上游辅助网络部分以及蓝色区域代表的下游网络部分。

3.1.1 上游主网络

在这个部分，我们使用已经预处理好的治疗前（`pre_img`）和治疗后（`after_img`）的图片作为输入。我们将同一个人的左眼和右眼视为不同的样例输入，同一只眼睛的治疗前和治疗后图片视为同一个样例的输入。考虑到训练集和测试集中每个人的图片数量不一，我们采用了一种简单的方法——对同一只眼睛的所有图片进行平均处理，使得每只眼睛均得到统一数量的输入。

首先，我们将输入数据通过在 ImageNet 中训练得到的 ResNet-152 模型的前半部分，分别得到 `feature map`。对治疗前和治疗后的图片进行 concat 后，进行一个两层卷积混合，最后再作为两个分支进行拼接得到 `img_vector`。

与此同时，我们也使用训练集和测试集 csv 文件中共有的一些 feature，对模型进行辅助训练。我们果断舍弃了没有信息的 gender 和 age 维度；对于整型特征 diagnosis 和 anti-VEGF，我们使用 Encoder 对这两个属性进行编码得到可学习的 `anti embedding` 和 `diag embedding`；对于浮点型

特征 preVA，我们对其进行 Normalization 之后得到 `feat_vector`。

将 `img_vector` 和 `feat_vector` 进行拼接后，与两个 embedding 相加得到上游主网络的输出特征 `x`。

3.1.2 上游辅助网络

为了减少简单平均操作对图片信息产生的损失，我们还设计了一个辅助网络。对于每一只眼睛，我们分别从治疗前个治疗后的图片集合中随机选择一张原始图片，使用预训练的 ResNet-50 分别得到 `img_feat` 并进行拼接。随后，经过 MLP 降维后得到一个维度较低的 `img_feat`。

3.1.3 下游网络

在这一部分，我们首先对上游的两个网络得到的输出特征进行加权求和，接着通过一个 MLP 得到最终输出的 logits，然后分别进行 regression 和 classification。对于 classification，我们选取最大值所指示的 index 作为预测类；对于 regression，我们根据原始数据的均值和方差，将最终得到的 logits 进行逆变换，得到最终的预测值。

3.2 训练和推理的整体架构

训练和推理的代码位于 `main.py` 中。

- 数据载入，数据 Transform
- Minibatch 进行训练
- Inference 推理并保存结果到 csv 文件中

3.3 亮点和技巧展示

- 图片特征增强：对训练集图片进行随机裁剪和翻转，测试集图片保持不变
- 学习率调整：每 10 个 epoch，学习率衰减 0.1
- 舍弃无用的特征属性（gender、age）
- 对整形离散特征（diagnosis、anti-VEGF）进行 encoding，得到低维的 embedding
- 预训练模型 Resnet 并冻结卷积层参数进行 Fintune

4 实验结果及分析

4.1 实验结果和初赛排名

队伍：UCAS_DM_GTL，初赛排名：**85 / 10006**，进入复赛，如图7所示。

提交的最好成绩历史如表2所示。

硬件设备：

- **CPU:** skylake - Intel(R) Xeon(R) Gold 5118 CPU @ 2.30GHz
- **GPU:** Tesla V100 (32GB)

图 7 初赛排名

表 2 最好成绩历史

Date	Score
2021-10-08	**0.4104**
2021-10-07	0.3973
2021-10-02	0.3807
2021-10-02	0.3716
2021-10-01	0.2853

4.2 结果分析

本次比赛基本达到了预定的目标，仅使用最传统的 ResNet 模型进行模型搭建和设计，就进入了复赛。虽然如此，仍然有很多的不足之处：

- 没有使用 k-fold cross validation 和 ensemble 方法
- 没有用到医学成像以及图像分割的 domain knowledge 和 SOTA

5 总结与展望

首先，我们对数据集进行了预处理操作，由于数据集比较脏，所以这一部分占用了我们大量的时间。接着，我们设计了包含上游主网络、上游辅助网络以及下游网络这三部分的神经网络模型，在这一部分，我们仅使用了传统的模型并使用了一些技巧来进行实现。最终，我们在初赛排名为 85/10006，并成功进入了复赛。当然，我们还有很多的不足，比起很多队伍，我们并没有做模型融合以及使用

专业知识，这也是在今后的比赛中我们需要去注意的地方。

致谢

由于我和朱赫同学都来自中科院计算所高通量中心，都不是研究 CV 方向的（这个比赛是更加专业的医学图像分割问题），这对于我们来说是一个不小的挑战。厦门大学 MAC 实验室的黄琳焱同学作为外援，负责了前期数据预处理的部分工作，也给了我们很多的理论和技术支持，这给我们 baseline 的搭建打下了一个良好的基础，特此感谢。从头搭建 baseline 的过程，让我逐渐熟悉了从数据处理到模型搭建的全过程，使我受益匪浅。——— 迟慧璇 (2021.11.21)