

BankOCRMoE: Bank-scened Optical Character Recognition with Mixture of Experts

摘要：在金融行业中，纸质凭证是记录业务关键信息要素的重要载体。在实际业务场景中，凭证的类别多种多样，有手写、机打等多种类型，并存在错格、模糊、重叠等现象。本次大赛聚焦于金融凭证关键要素的抽取，提供了不同细分场景中的凭证图像切片。然而，现有的开源 OCR 方案在多任务及小样本场景下均表现较差。在本文中，我们设计了中文 TrOCR，并在大规模数据上进行预训练。我们提出了 BankOCRMoE（Bank-scened Optical Character Recognition with Mixture of Experts），使用专家混合来专注不同类型任务，并使用一个管理器模块指导专家的训练和推理。此外，我们观察到了 hard sample 被 easy sample 挤压的现象，为预训练和微调设计了不同的数据增强策略。最后，我们设计了针对 OCR 的异构二级集成策略来进一步提升分数。实验结果表明了我们的方案在金融 OCR 场景下的有效性，方案在 B 榜达到了单模 0.9604，排名第一，最终分数 0.96342，处于绝对领先。

关键词：OCR，预训练，专家混合，数据增强，异构二级集成

目录

1. 大赛背景和任务	1
1.1 大赛背景	1
1.2 大赛任务	1
2. 赛题分析与理解	2
2.1 赛题特点	2
2.1.1 数据特点	2
2.1.2 推理特点	2
2.2 解题思路	2
3. 赛题方案	3
3.1 方案整体流程图	3
3.2 基础模型选型及调整	4
3.2.1 模型选型	4
3.2.2 结构调整	5
3.3 外部数据收集及处理	5
3.4 预训练	6
3.4.1 权重初始化方式	6
3.4.2 预训练阶段数据增强	6
3.5 模型框架设计	6
3.5.1 Feature Extractor	7
3.5.2 Mixture of Experts Module	7
3.5.3 Management Module	8
3.6 微调策略	8
3.6.1 数据预处理	8
3.6.2 微调阶段数据增强	9
3.6.3 Loss	10
3.6.4 Trick	10
3.7 推理策略	11
3.7.1 基础集成策略介绍	11
3.7.2 异构二级集成	12
3.8 后处理	13
3.9 其他尝试	14
4. 实验	16
4.1 预训练实验	16
4.1.1 实验设置	16
4.1.2 实验结果及分析	16
4.2 微调实验	17
4.2.1 实验设置	17
4.2.2 实验结果及分析	18
5. 总结	19
6. 参考文献	19

1. 大赛背景和任务

1.1 大赛背景

在金融业内，纸质凭证是记录业务关键信息要素的重要载体。在实际业务场景中，凭证的类别多种多样，有手写、机打等多种类型，并存在错格、模糊、重叠等现象。利用先进的信息技术自动、高效、准确地抽取凭证内的信息，能够有效提升金融业凭证核验工作的效率，同时减少业务上的操作风险。本届竞赛将从真实场景和实际应用出发，以凭证信息抽取作为赛事主题，期待参赛选手们能在这些任务上相互切磋、共同进步。

1.2 大赛任务

此次竞赛聚焦于金融凭证关键要素的抽取，为选手提供了不同细分场景中的凭证图像切片，具体场景包括：

(1) 交易编码识别：



图像切片		
识别内容	929070	822030

图 3-1 交易编码识别示例

(2) 地址类文字识别：

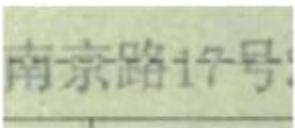
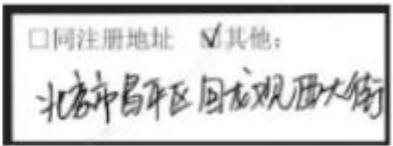
图像切片		
识别内容	南京路 17 号	北京市昌平区回龙观西大街

图 3-2 地址类文字识别示例

(3) 附言文字识别：

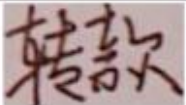

图像切片		
识别内容	转款	港杂费

图 3-3 附言文字识别示例

图像切片内包含手写体和机打体，且存在文字错位、遮挡、模糊、重叠等多种干扰信息的情况，需要参赛队伍通过识别技术，准确识别出图像中的关键要素（包含中英文以及标点符号）。

基于以上场景，赛题分为两个任务，任务一为交易编码识别（包含场景 1），任务二为文字识别（包含场景 2、3），两个任务采用不同的评分标准。本次竞赛允许选手使用开源模型（不可使用 API），允许使用外部数据，且主办方鼓励参赛选手使用单一模型预测不同场景。

2. 赛题分析与理解

2.1 赛题特点

本赛题本质为金融场景下的 OCR，与大多数 OCR 赛题不同的是，该赛题分为两个任务，分别为交易编码识别和文字识别，其各自的验证指标不同，且初赛 A/B 榜可以利用的有标签数据也不同，B 榜提交还有文件大小和推理时长限制。此外，金融场景的 OCR 参考资料较少，开源的 OCR 预训练模型应对难样本能力十分有限。因此，该赛题非常具有挑战性。

2.1.1 数据特点

该赛题分为两个任务，任务一提供 4983 份数字类图像切片作为训练集，A 榜测试集 2534 份数据；任务二提供 88 份文字类图像切片（包含场景 2、3），A 榜测试集 1278 份。数据中许多样本其 label 相同但图像不同。值得注意的是，该赛题 B 榜阶段会放出 A 榜测试集的 label 供选手使用，也就是说，选手更应该注重方案的泛化性和可迁移性，不能将精力完全放在 A 榜的 few-shot 上。

2.1.2 推理特点

B 榜提交时有 2G 的大小限制，因此选手不能使用过大的模型，2 小时的推理时长也限制选手不能融合过多模型或者开设太大的 beam size（如果使用 beam search 作为解码策略）。

2.2 解题思路

根据上述特点及挑战，本团队初步制定了“基础模型选型及结构调整—预训练—适配本赛题的模型框架设计—微调—推理”的大致步骤，拟定了一些用于解决上述挑战的方法，如数据增强、各类损失函数的设计、异构二级集成及后处理等。本团队在实验过程中不断探索与优化，最终达到了 0.9634 的优异成绩，方案详细见第 3 节。

3.2 基础模型选型及调整

3.2.1 模型选型

CNN 卷积神经网络^[1]在计算机视觉任务上起到至关重要的作用,CRNN^[2]是 OCR 任务中主流的算法之一,这里我们尝试使用 CRNN 算法用于 OCR 识别,在交易编码任务上能取得相对不错的效果,但是在文字识别任务中,由于样本量较少,仅 88 条训练样本,该算法在 few shot 情况下效果较差,不能用于实际生产中。考虑到在 NLP 中,对于 few shot 经常使用 Bert^[3]等预训练模型进行知识迁移来提高 few shot 的性能,且随着 transformer 模型 attention 注意力机制进入到视觉任务,同样可以使用 transformer 来进行 OCR 识别任务。通过调研,我们团队了解到了基于 transformer 模型的 OCR 手写文字识别模型 TrOCR^[4] (Transformer OCR), TrOCR 是 Microsoft 发布的一个 OCR 识别模型,其模型架构如下,完全采用了标准的 transformer 模型。

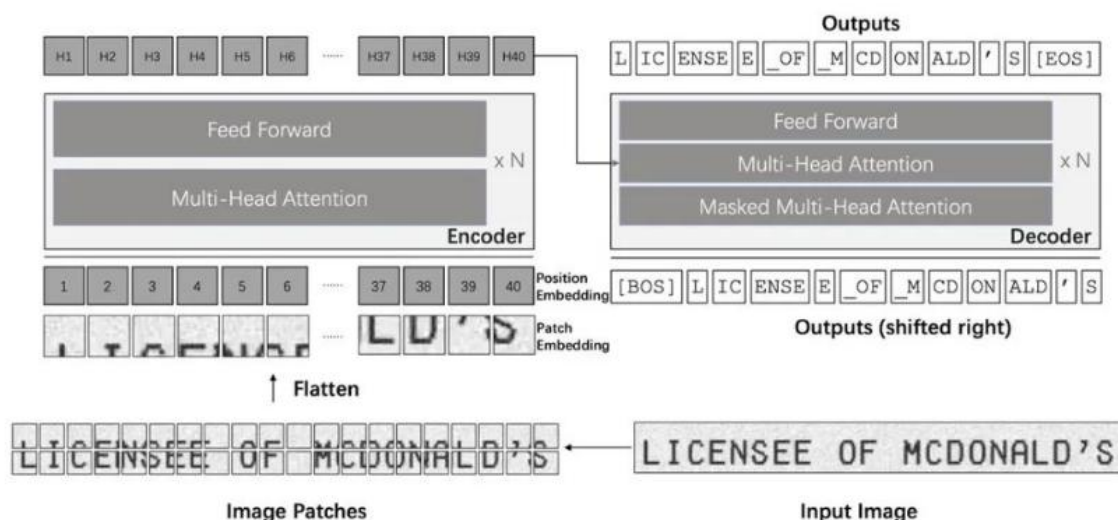


图 3-5 TrOCR 模型结构图

TrOCR 不需要任何复杂的卷积网络作为主干网络,更加易于实现和维护。TrOCR 采用了 Transformer 结构,包括图像 Transformer 和文本 Transformer,分别用于提取视觉特征和建模语言模型,并且采用了标准的 Transformer 编码器-解码器模式。编码器用于获取图像切片的特征;解码器用于生成 wordpiece 序列,同时关注编码器的输出和之前生成的 wordpiece。

相比于传统深度学习 OCR 识别模型,基于预训练的端到端 OCR 识别模型具有以下特点:

(1) TrOCR 使用预先训练好的图像 Transformer 和文本 Transformer 模型,它们利用大规模的未标记数据来进行图像理解和语言建模,而不需要一个外部语言模型。

(2) 在 OCR 数据集上的实验结果表明,TrOCR 可以在印刷、手写和场景文本图像

数据集上实现最先进的结果，而无需任何复杂的前后处理操作步骤。

本团队在对 CRNN、PP-OCRv3^[5]以及 TrOCR 等模型进行大量实验验证后，最终选择了效果相对较好的 TrOCR 作为我们的基础模型，在此基础上进一步调整优化该模型。

3.2.2 结构调整

基于预训练 TrOCR-small 的经验，本团队计划预训练一版 TrOCR-Base，但是考虑到 base 版本模型参数量较大，训练耗时以及模型存储都相对较大以及 B 榜提交模型文件不超过 2G 等限制。我们对 base 版本的 TrOCR 模型的 Decoder 部分参数进行了调整，其调整如下：

表 3-1 Decoder 部分参数调整

模型参数名字	原 TrOCR-base 参数	调整后 TrOCR-basebase 参数
d_model	1024	768
decoder_attention_heads	16	12
decoder_layers	12	8
decoder_ffn_dim	4096	3072

本次参数调整策略是基于本团队在文本生成方面的相关经验来调整的，即强大的 encoder 可以提升后序任务的能力，比如 CPT 模型^[6]相比于 BART^[7]模型，其通过增加 encoder 的层数，减少了 decoder 的层数来提升模型的性能及解码速度。因此我们选择了只对 decoder 侧进行参数调整，保留了 base 版本强大的图像表征能力，这种调整不仅提高了模型性能而且在效果上没有太大的损耗。

3.3 外部数据收集及处理

TrOCR 模型是英文预训练模型，其不能直接应用到本次赛题中，为了使用 TrOCR 解决金融行业 OCR 任务，需要收集一批中文 OCR 任务数据集。本团队在互联网上收集到一批不同场景下的中文 OCR 数据集（<https://github.com/chineseocr/darknet-ocr>），如手写体，印刷体，签名，数字等，其数据样例如下图所示。



图 3-6 预训练数据示例

在得到进行预训练的数据集后，对其进行统计和处理，一共得到约 800w 数据，并将所有的文本提取出来制作用于预训练的词表，该词表共有 12296 个 token，其覆盖了常用的汉字。

3.4 预训练

虽然网上有开源中文版的 TrOCR 预训练模型，但其用于预训练的数据集未知且只有 small 版本，效果达不到预期。因此我们团队在收集到的数据集上进行 small 版本的 TrOCR 预训练，本团队预训练的 TrOCR-small 在该任务是优于开源中文版的 TrOCR-small 的。

3.4.1 权重初始化方式

在预训练的过程中，我们使用英文 TrOCR 模型的 Encoder 权重用于初始化我们的 Encoder 编码器，Decoder 由于参数大小发生改变，则使用了随机初始化。虽然 TrOCR 是英文的，但是其 Encoder 是图像表征过程，用其进行初始化，提高了图像的特征提取能力，对比实验见第 4 节。

3.4.2 预训练阶段数据增强

为了避免预训练模型陷入过拟合以及提升模型的性能，我们采用了数据增强方式来增加训练样本的多样性，提高模型的泛化能力。涉及到的数据增强方式：1) 修改亮度；2) 修改对比度；3) 修改亮度饱和度；4) 转灰度图；5) 仿射变换；6) 高斯变化等。在训练的过程中有 70% 的概率会进行上述 6 种数据增强方式的一种对样本进行增强，来提升模型的性能。

3.5 模型框架设计

因主办方鼓励选手使用单一模型预测两个场景，受启发于专家混合^[8]在各领域的成功应用，我们提出了 BankOCRMoE (Bank-scened Optical Character Recognition with Mixture of Experts)，其包含若干个专家，每个专家有各自专精的领域，在训练时各个专家相互竞争，推理时互相合作。具体来说，BankOCRMoE 包含一个公共 Encoder，三个专家 Decoder，一个管理模块 Manager，三个 Decoder 各自关注不同任务类型，Manager 用来协调各个专家的训练以及推理时分配一个注意力分数，模型结构见图 3-7。

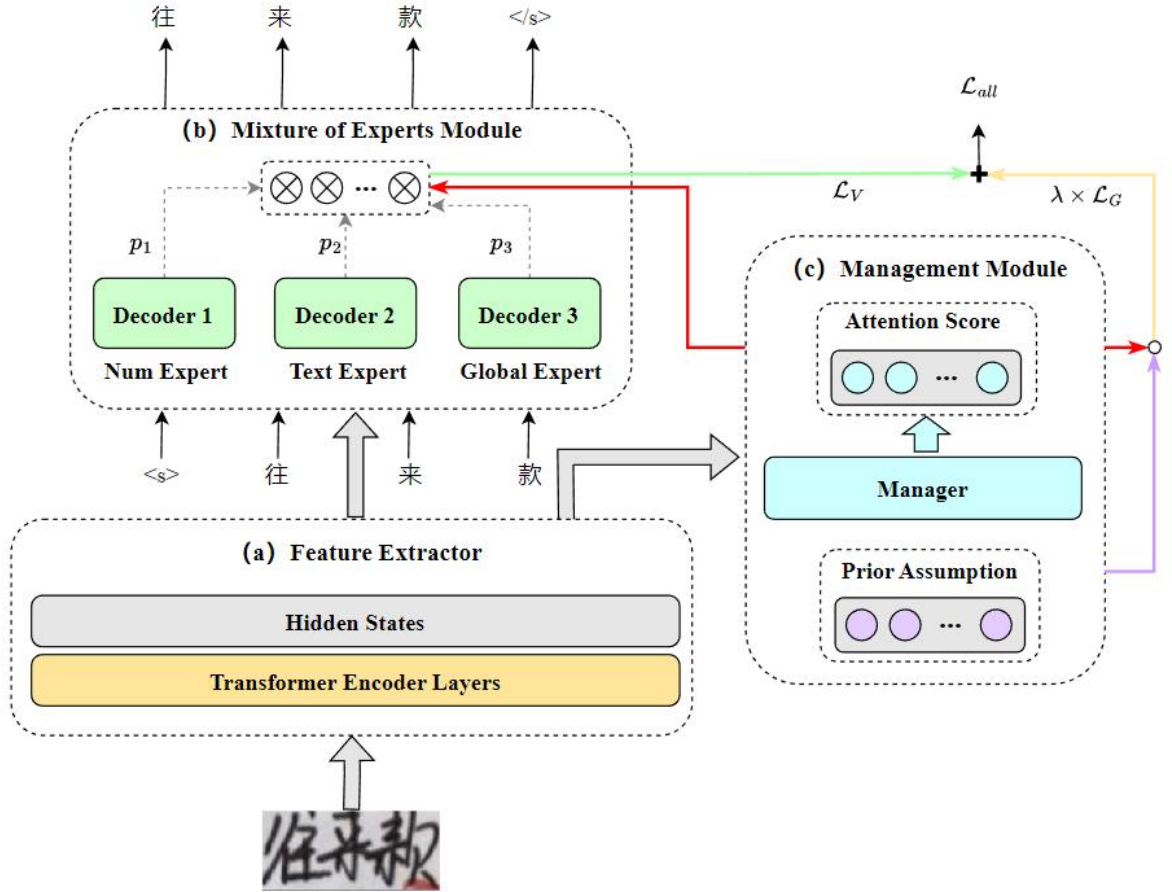


图 3-7 BankOCRMoE 模型结构概述。该模型由三个模块组成：（a）Feature Extractor：用于提取图片特征，作为后续专家和管理器的输入。（b）Mixture of Experts Module：使用三个专家对来自图片的特征解码生成文字，每个专家有各自注重的领域，相互协作。（c）Management Module：我们按经验设置了先验假设来指导 Manager 的分配，然后 Manager 总结专家的全部输出作为最终预测。

3.5.1 Feature Extractor

特征提取器即预训练模型中的 Encoder 部分，用于学习图像的表达，其结构具体细节见上述预训练部分。Encoder 接收经过处理后的图片输入 X ，并输出 hidden states（即图片特征） H ：

$$H = f_{Enc}(X) \quad (1)$$

其中 $X \in \mathbb{R}^{c \times k \times k}$ 为处理后的图片输入， c 为通道数， k 为 size， $H \in \mathbb{R}^{d_1 \times d_2}$ 为 hidden states， d_1 和 d_2 均为维度， f_{Enc} 为 Encoder。

3.5.2 Mixture of Experts Module

在该模块中，专家混合（MoE）根据（1）中提取的图像特征进行文本生成。我们采用三位专家注重不同的任务。具体来说，Num Expert 更侧重于解决任务一，Text expert 更侧重于解决任务二，Global Expert 则对两类任务都进行考虑。然而，专门为各个专家设计不同的框架将限制所提出的框架推广到其他数据集，且上述的预训练效益将大打折扣

扣。受 zhou 等人^[9]的启发，我们使用相同的通用神经架构实现每个专家。具体来说，每个专家都是由调整后的相同 **decoder** 结构组成的，当前时间步的输入由 H 以及上一个时间步的输出组成。

3.5.3 Management Module

MoE 的简单训练可能会存在各种各样的缺陷，"imbalanced experts"是指训练阶段一些专家未能得到充分地训练，且很有可能发展为赢者通吃现象，即只有一个专家起作用，其他专家在推理时分配到的分数接近 0。**Manager** 旨在指导专家的训练并集成所有专家的结果，使其相互竞争，从而提高性能。我们首先人工设计先验假设 a_G ，用于指导 **Manager** 更多地关注两类任务与不同专家输出之间的相关程度，任务一的先验假设为 $[0.4, 0.2, 0.4]$ ，任务二的为 $[0.2, 0.4, 0.4]$ 。之后，我们用 **Manager** 来指导专家的训练，**Manager** 对图片特征 H 进行编码并生成注意力分数 a_M ：

$$h_M = f_{Enc_M}(H) \quad (2)$$

$$a_M = \text{soft max}(\tanh(h_M W_1^M) W_2^M) \quad (3)$$

其中 Enc_M 是管理器的 **encoder**， W_1^M 和 W_2^M 是可训练参数。注意力分数 a_M 数和先验假设 a_G 分别用于指导专家的训练以及教会 **Manager** 分配合理分数，这通过 3.6.2 节中介绍的专门设计的 **loss** 实现的。

3.6 微调策略

3.6.1 数据预处理

通过对任务一的数据集进行分析，我们发现该数据共有以下四种形状，我们规定第一张图像的形状为标准形状。在 **badcase** 分析的时候发现第三类和第四类图像经过处理后变形较为严重，影响识别效果。

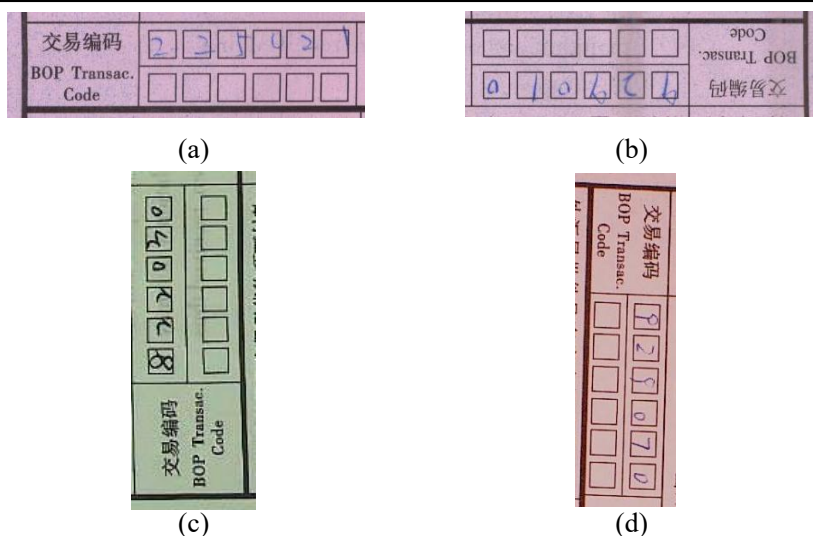


图 3-8 任务一数据四类形状

我们利用图像处理算法，首先识别到要识别的图像的形狀，然后根据其与标准图像得角度进行旋转，将其变化为标准图像，提升了 ocr 的识别效果。

3.6.2 微调阶段数据增强

由于数据中包含大量图像不同而 label 相同的样本，这与预训练阶段时使用的数据增强，即缩放、裁剪、翻转、旋转及变化颜色等操作等价，若在微调时继续采用这些操作则会引入大量噪声，进一步的挤压 hard sample 的空间。因此，我们应设计一种针对此赛题 hard sample 的图像合成策略，来一定程度上抵消数据中 easy sample 带来的“过采样”问题。

通过对 A 榜测试数据的观察，我们发现 text 任务上主要丢分集中在多行长样本上，即上述 hard sample，因此，我们仅针对多行数据进行合成，合成方式为对现有的样本进行纵向拼接，合成限制条件如下：

- (1) 原本就为多行的样本不参与合成；
- (2) 参与合成的样本长度大于 5 小于 18（不能过长也不能过短）；
- (3) 第一行的文本长度要大于第二行（通常人们写满第一行才会换行）；
- (4) 每个样本参与合成的次数不超过 5 次（合成数据不易过多，避免其反向挤压邻居子空间）；
- (5) 拼接的两个样本图像宽度比大于 0.9；

合成图像示例如图 3-9 所示，共合成 1338 数据，将其加入原始样本，然后进行训练。

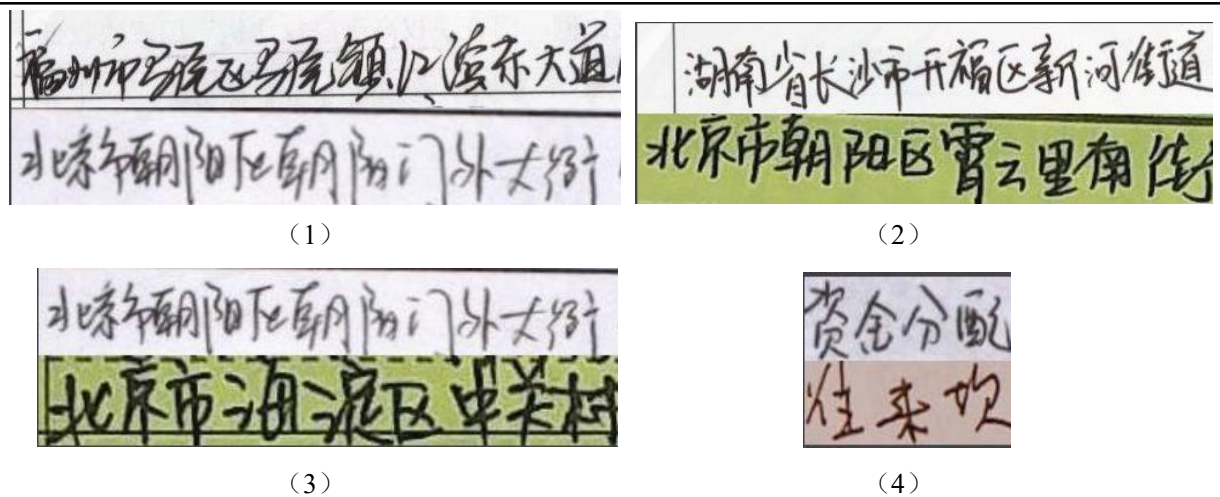


图 3-9 微调阶段合成数据示例

3.6.3 Loss

我们设计了两个 loss functions: 生成损失 \mathcal{L} 以及指导损失 \mathcal{L}_G 。前者是每个专家的生成损失的加权和, 注意力分数由 Manager 分配给专家。后者衡量先验假设和注意力分数之间的差异, 并引导 Manager 为专家分配合理的注意力分数。我们通过最小化这两个项的加权和来联合优化我们的模型:

$$\mathcal{L}_{all} = \mathcal{L} + \lambda \mathcal{L}_G \quad (4)$$

其中 λ 是用于控制 \mathcal{L}_G 比率的超参数, 两个损失函数的细节如下:

(1) 生成损失。我们独立计算每个专家的生成损失, 其中每个专家的生成损失为 decoder 的输出和原文 ground truth 之间的 cross entropy, 然后通过注意力分数 a_M 对各个专家的 cross entropy 进行加权:

$$\mathcal{L}_i = \frac{1}{m} \sum_{j=1}^m \mathcal{L}_{CE}(p_j, t_j) \quad (5)$$

$$\mathcal{L} = \sum_{i=1}^{n_e} (a_M)_i \cdot \mathcal{L}_i \quad (6)$$

其中 \mathcal{L}_i 是第 i 个专家的生成损失, m 为句子长度, \mathcal{L}_{CE} 表示交叉熵损失, p_j, t_j 分别表示第 j 个 token 的预测和真实 token, n_e 是专家数量, $(a_M)_i$ 是 Manager 为第 i 个专家分配的注意力分数。

(2) 指导损失。为了缓解 3.5.3 节提到的专家不平衡现象, 我们使用了另一个损失函数 \mathcal{L}_G , 该函数计算先验假设 a_G 和注意力得分 a_M 之间的 KL 散度:

$$\mathcal{L}_G = D_{KL}(a_G \| a_M) \quad (7)$$

其中 $D_{KL}(\cdot \| \cdot)$ 表示 Kullback–Leibler divergence。通过最小化 \mathcal{L}_G , Manager 学会为专家分配合理的注意力分数。此外, \mathcal{L}_G 还使得专家的训练变得更加平衡。

3.6.4 Trick

本团队没有使用许多 trick, 仅使用了 SWA 平均多个局部最优点。

3.7 推理策略

本团队采用的模型均为 Encoder-Decoder 结构，Encoder 提取图片特征，Decoder 接收图片特征以及上一个时间步输出的 token 作为当前时间步的输入，如图所示。

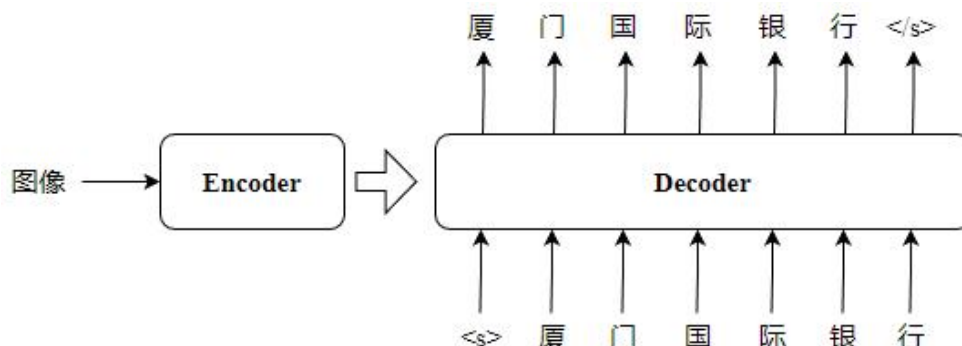


图 3-10 Encoder-Decoder 结构模型处理 OCR 任务时文本生成过程

搜索策略采用 Beam Search，这种算法的主要思想是在每一步保留多个条件概率最优的输出，从而扩大搜索空间，提高生成序列的质量。Beam Search 在每一个时间步都会考虑多种可能的输出路径（即候选词），并为每条路径计算一个得分。然后，它将这些得分最高的若干条路径保留下来，作为下一步的输入。这个过程会持续到达到预设的条件或满足停止准则为止，详细超参数见第 4 节实验部分参数设置。

在本赛题中，我们构建的 MoE 单模型便可解决两个类型的任务，线上成绩可达 96 分，较其他队伍有明显领先，但因该赛题并未限制模型集成，为保险起见，团队在后续依然选择通过集成来提高线上的分数。针对该赛题，我们提出了针对 OCR 的二级异构集成策略，在正式介绍方案前，先介绍几个常用集成策略。

3.7.1 基础集成策略介绍

(1) Logits/probs 平均。Logits/probs 平均是最常见的一种集成策略。其主要做法为将多个模型给类别预测的概率进行（加权）平均，将平均后最大概率对应的类别作为最终分类类别。其效果好，适用性强，但该策略不能有效利用多个模型协作能力，模型过多时会导致分类性能的下降、资源需求率高、利用率低。

(2) 投票。投票主要用于分类。其将多个模型预测的结果进行投票，票数高者为最终结果。该策略使用简单，但适配任务类型少，且单模型效果本身一般，故最终投票得出的结果上限不高。

(3) 筛选。筛选可以理解为重排序，是一种常用于文本生成的集成方法。通过将每个模型与其他模型产出的文本分别计算 BLEU/CER 等评测指标作为分数，选出最鲁棒的一条文本作为最终结果。该策略使用简单，与投票一样，单模型本身效果一般的时候上限较低。

3.7.2 异构二级集成

经验地说，不同结构的模型具有互补性，因此异构集成比同构集成可以带来更好的效果（实验结果表明亦是如此）。针对本赛题的异构二级集成综合了 probs 平均和筛选两种策略，充分发挥了多模型的协作能力。针对任务二，我们使用一个 MoE 模型和 3 个普通模型，将其分为三组，每组包含 1 个 MoE 和 2 个普通模型，组内模型不完全一致，确保差异性。第一级集成为组内模型逐 token 加权平均，产出三个结果，之后进行第二级基于 CER 的文本筛选，筛选出鲁棒性更高的结果。具体流程见图 3-11。

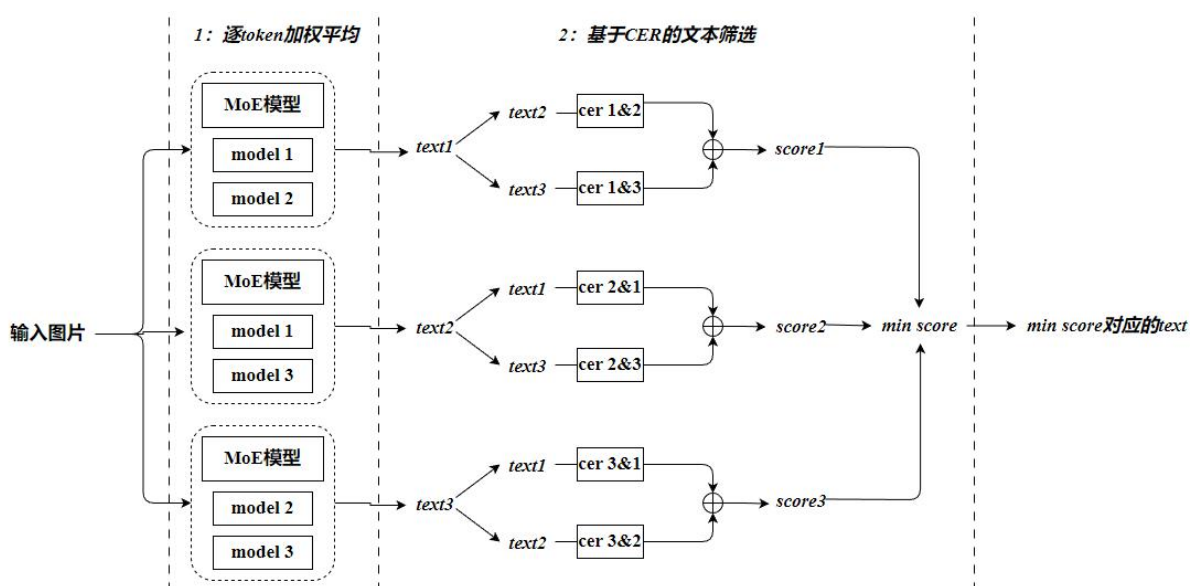
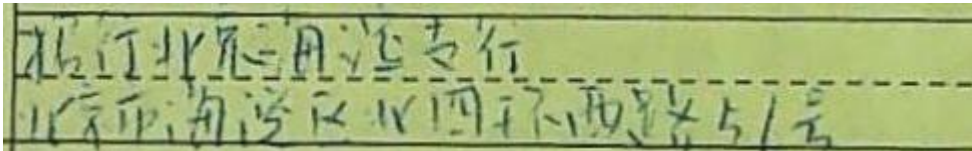

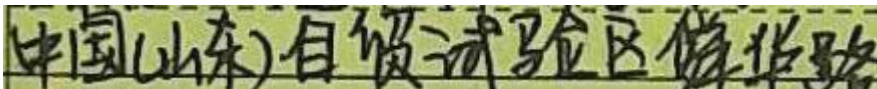
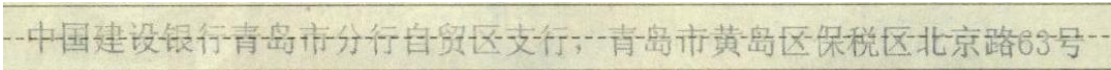


图 3-11 异构二级集成流程图

3.8 后处理

后处理是比赛常用的提分方法，通过对 bad case 的观察来总结出修正 bad case 的规则，针对本赛题，本团队在初赛 A 榜期间尝试了多种类型的规则修正，在分数较高的情况下仍有有 0.1（未单独测试 B 榜）分的提升。示例如下：

表 3-2 bad case 后处理示例

连续 重复	
修正前：北京北京市西湖区支行 北京市海淀区北四环西路 56 号	
修正后：北京市西湖区支行 北京市海淀区北四环西路 56 号	
多余 前缀	
修正前：付加信息及用途:从监督账户转至投资人账户	
修正后：从监督账户转至投资人账户	
括号 缺失	
修正前：中国山东)自贸试验区货华路	
修正后：中国(山东)自贸试验区货华路	
字形 错误	
修正前：中国福建省厦门大炬	
修正后：中国福建省厦门火炬	
虚线 背景	
修正前：""中国建设银行青岛市分行自贸区支行，青岛市黄岛区保税区北京路 63 号"	
修正后：中国建设银行青岛市分行自贸区支行，青岛市黄岛区保税区北京路 63 号	

3.9 其他尝试

(1) EMO Loss。文本生成任务的训练可以理解为逐 token 的分类任务，默认损失也是交叉熵，它有着简单高效的特点，但在某些场景下也暴露出一些问题，如偏离评价指标、过度自信等，相应的改进工作也有很多。Ren 等人^[10]基于最优传输思想提出了新的改进损失函数 EMO(Earth Mover Distance Optimization)，实验结果显示其能大幅提高 LLM 的微调效果，在小模型上的继续预训练实验上，EMO loss 相比交叉熵（MLE）的提升最多的有 10 个点。其通过 Embedding 算相似度，来为“近义词”分配了更合理的损失，从而使得模型的学习更加合理。

本团队在竞赛初期尝试了该 loss，效果与常规交叉熵 loss 接近，参考苏神对该 loss 的讲解^[11]，猜测是由于前期的预训练还不够充分，加上对 EMO 与 MLE 的加权探索不够细致，收敛速度和最终效果未能达到平衡。该 loss 本身潜力巨大，相信对其进行更深入的探索，能够明显提升实验效果，受限于 b 榜提交次数，本团队后续没有对其深耕。

(2) 检索增强。Wang 等人^[12]提出了一种非常简单的 NLP 性能增强法则 REINA（REtrieving from the traINing datA）：即通过 BM25 算法在训练集中检索和输入 x 最相似的数据，并把检索到的最相似的 K 个数据直接拼接接到原输入 x 后再重新训练模型，便能有效地提升 NLP 模型的性能。

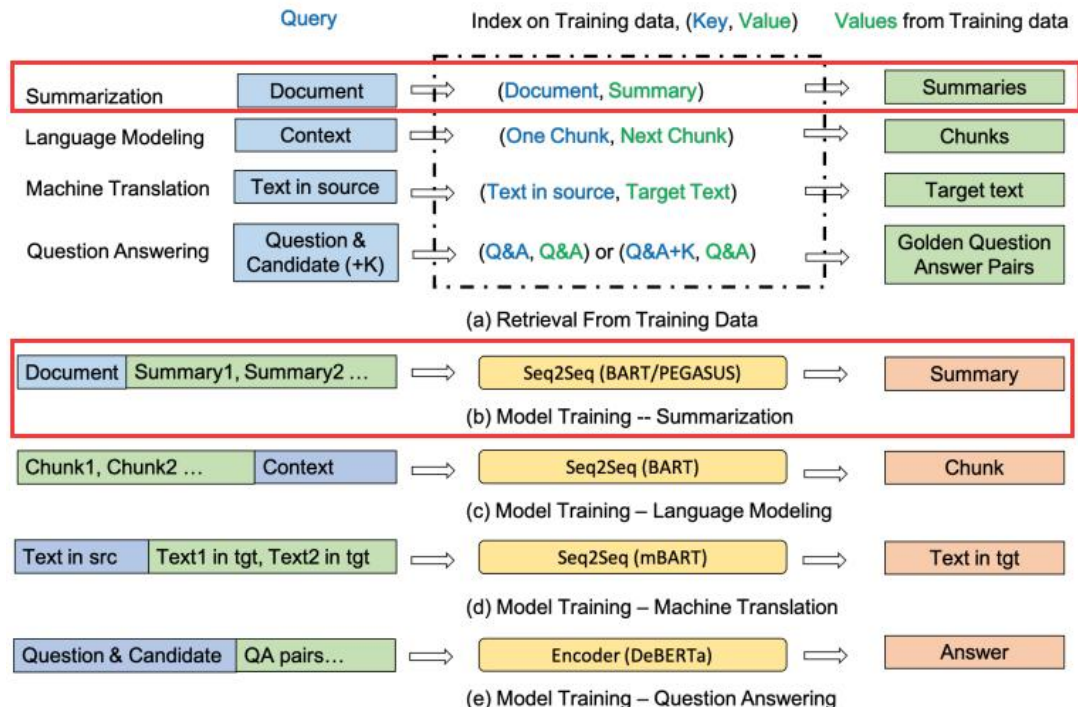


图 3-12 REINA 检索增强

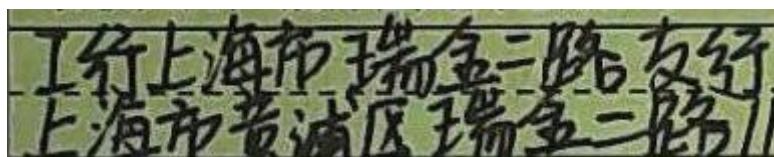
本团队曾在全球人工智能技术创新大赛上充分地利用了这一技术，最终获得了榜单第二名的成绩。在此赛题中，我们尝试对 encoder 提取的图片的特征进行相似度检索，将从训练集中检索到的相似特征拼接到最后，在用一个 MLP 将其恢复到原本维度，最

后送入 decoder 进行解码。实验效果略差于不进行检索增强，猜测是由于数据太少，很多样本检索到的特征并不相似，从而引入噪音，且图像粒度和文本粒度不同，由于时间原因，团队未能深入探索这一技术在 OCR 上的使用。

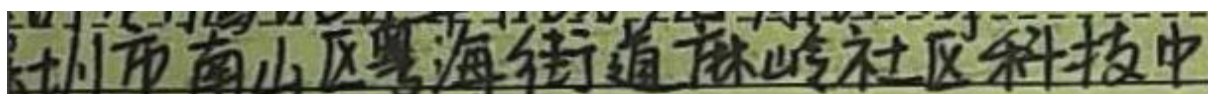
（3）基于检索的银行/地址文本纠错

团队搜集了大量地址数据和银行（包含各地支行）数据，尝试通过检索相似的数据对生成的地址和银行文本纠错。通过线下的观察，可以发现很多错误语句的确可以被修正，但该赛题 OCR 数据中存在很多图片其包含的数据或银行文本本身不完整，其 label 也是不完整的（从 OCR 任务本身考虑，标注没有任何问题），具体示例如表所示。

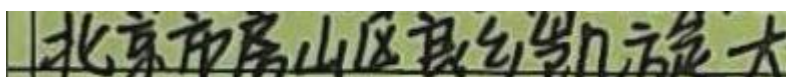
表 3-3 图片本身信息不完整示例



Label: 工行上海市瑞金二路支行 上海市黄浦区瑞金二路 11



Label: 圳市南山区粤海街道麻岭社区科技中



Label: 北京市房山区良乡凯旋大

这些样本被纠正后反而会造成分数下降，故该方案未能应用在本赛题中。

4. 实验

4.1 预训练实验

4.1.1 实验设置

我们在大量不同场景的 OCR 数据集上对 TrOCR-Base 进行预训练,提升其在通用场景下的识别能力,这样在本赛题任务上以少量的数据进行微调即可得到非常好的成绩。预训练设备为 3*p100 (16G), 预训练参数见表 4-1。

表 4-1 预训练参数设置

参数	参数值	含义
Epoch	5	预训练轮数
Batch size	12	批量大小
Max target length	80	最大文本长度
Eval steps (验证)	5000	每 5000step 评估一次模型
Num beams (验证)	2	Beam search 搜索方式
no_repeat_ngram_size (验证)	3	用于控制重复词的生成
length_penalty (验证)	2.0	长度惩罚

4.1.2 实验结果及分析

表 4-2 为预训练 4w step 后验证集上的效果,通过上表可以发现随机初始化 Encoder 效果极差,而使用英文的 Encoder 权重进行初始化后,OCR 识别效果得到极大的提升,即预训练好的 Encoder 进行初始化可以加速模型的收敛,提高图像的特征能力,也对后续文本生成有了较大改善。

表 4-2 Encoder 随机初始化 vs 开源英文权重迁移

	Cer	Acc
随机初始化 Encoder	1.0285	0.0016
英文 Encoder 权重进行初始化	0.1035	0.6846

4.2 微调实验

4.2.1 实验设置

我们在一张 3090（24G）上进行微调，微调时参数如表 4-3 所示：

表 4-3 微调参数设置

参数	参数值	含义
Pretrain step	910000	使用的预训练权重对应步数
Epoch	5	微调轮数
Batch size	8	批量大小
Max target length	70	最大文本长度
Eval step	500	模型评估间隔数
Save step	500	模型保存间隔数
lmd	0.1	管理器的 loss 权重

我们使用 Beam Search 为解码时搜索策略，推理参数如表 4-4 所示：

表 4-4 推理参数设置

参数	参数值	含义
Batch size	8	推理批量大小
Num beam	8	候选解数量
Max length	80	最大文本长度
Min length	3	最小文本长度
No repeat ngram size	5	模型评估间隔数
Length penalty	2.4	长度惩罚
Repetition penalty	0.8	重复度惩罚

4. 2. 2 实验结果及分析

由于 B 榜测评为线上实施测评，选手无法获取 B 榜数据，因此不能针对 B 榜做消融实验，此处我们给出 A 榜上分路线图以及 B 榜测评对比实验。

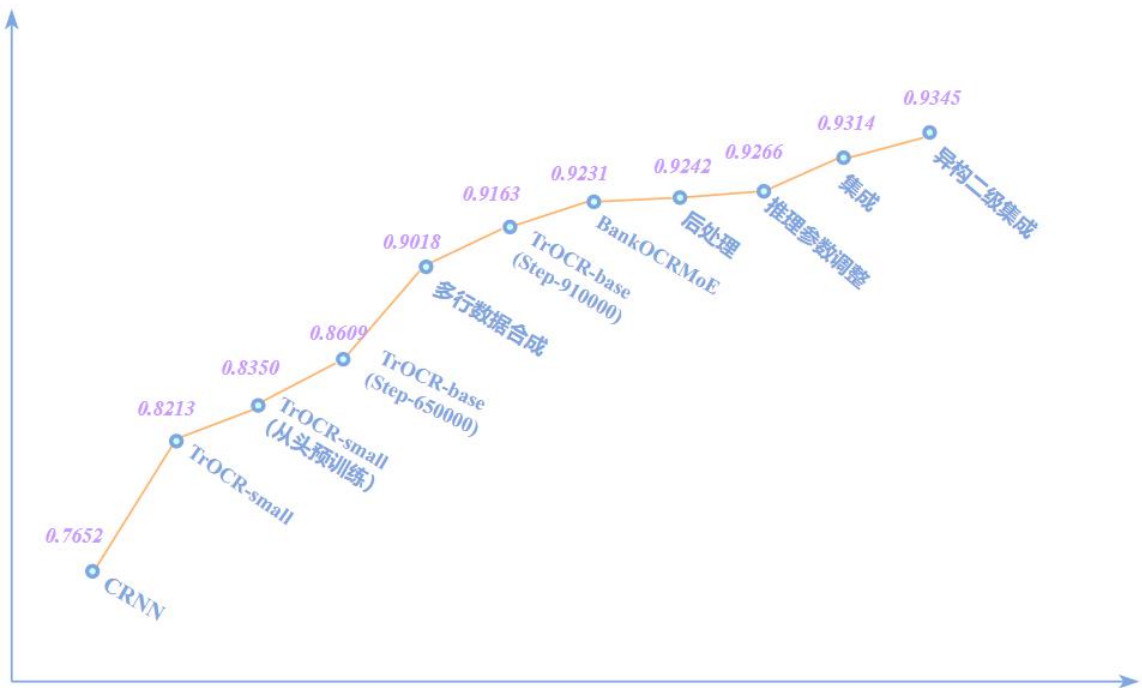


图 4-1 A 榜上分路线图

表 4-5 B 榜对比实验

Methods	Score
单模 BankOCRMoE（非全量数据）	0.95738
单模 BankOCRMoE（全量数据+数据增强）	0.96041
任务 2 异构二级集成	0.96235
任务 1 prob 集成+任务 2 异构二级集成	0.96342

推理结果。从 B 榜线上分数看，我们的方案单模即可达到 0.96，处于领先地位，且仅使用一个模型便可解决两个任务，此外，我们的二级集成策略在高分情况下仍带来较大提升。图 4-1 显示了我们的方案在 A 榜的大致上分情况，方案鲁棒性极高，切榜时能够稳住分数，且迁移能力强，面对不同的数据量均能有好的表现。

推理速度。我们的方案单模线上推理时长为 693s, num_beam=4 的情况下预计为 350s, 分数几乎与 num_beam=8 时相同，方案推理速度较快，可以根据实际应用场景动态调整 num_beam，灵活性高。

5. 总结

在本文中，我们介绍了我们在第四届厦门国际银行数创金融杯建模大赛使用的方案，包括基础模型结构选型及调整、预训练、模型结构设计、数据增强、推理策略等一系列提升分数的方法。总的来说，我们的贡献/创新点如下：

(1) 我们设计了中文 TrOCR-base, 使用英文 TrOCR-base 对 Encoder 部分进行参数初始化, 收集了约 800w 外部数据对其进行预训练, 效果相对开源的中文 TrOCR-small 有明显提升, 我们计划将该权重开源, 供更多 OCR 学者参考和使用。

(2) 我们提出了 BankOCRMoE, 这是一种在银行多任务 OCR 场景下的新方法, 其利用专家的混合来关注到不同任务, 应用具有注意力引导模块的管理器来指导专家的训练, 并为每个专家分配合理的注意力分数。其可以很方便的迁移到其他数据集和任务, 只需让不同专家关注不同部分即可。实验结果表明 BankOCRMoE 在本赛题相对传统模型拥有显著优势, 达到了最先进的效果。

(3) 我们观察到了传统数据增强存在 easy sample 对 hard sample 的空间挤压现象, 为预训练和微调设计了不同的数据增强方法。

(4) 我们设计了针对 OCR 的异构二级集成策略, 其效果优于传统集成, 使得方案整体鲁棒性进一步提升, 切榜时更加稳定。

我们希望我们在该比赛的方案能够为其他选手带来更多思路及灵感。在此衷心地感谢赛题方提供的宝贵数据以及所有工作人员的辛苦付出。

6. 参考文献

- [1] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
- [2] Fu X, Ch'ng E, Aickelin U, et al. CRNN: a joint neural network for redundancy detection[C]//2017 IEEE international conference on smart computing (SMARTCOMP). IEEE, 2017: 1-8.
- [3] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
- [4] Li M, Lv T, Chen J, et al. Trocr: Transformer-based optical character recognition with pre-trained models[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(11): 13094-13102.
- [5] Li C, Liu W, Guo R, et al. PP-OCRV3: More attempts for the improvement of ultra lightweight OCR system[J]. arXiv preprint arXiv:2206.03001, 2022.
- [6] Shao Y, Geng Z, Liu Y, et al. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation[J]. arXiv preprint arXiv:2109.05729, 2021.
- [7] Lewis M, Liu Y, Goyal N, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[J]. arXiv preprint arXiv:1910.13461, 2019.

- [8] Jacobs R A, Jordan M I, Nowlan S J, et al. Adaptive mixtures of local experts[J]. Neural computation, 1991, 3(1): 79-87.
- [9] Zhou Y, Liu X, Zhou K, et al. Table-based fact verification with self-adaptive mixture of experts[J]. arXiv preprint arXiv:2204.08753, 2022.
- [10] Ren S, Wu Z, Zhu K Q. EMO: Earth Mover Distance Optimization for Auto-Regressive Language Modeling[J]. arXiv preprint arXiv:2310.04691, 2023.
- [11] 苏剑林. (Oct. 13, 2023). 《EMO: 基于最优传输思想设计的分类损失函数 》 [Blog post]. Retrieved from <https://kexue.fm/archives/9797>
- [12] Wang S, Xu Y, Fang Y, et al. Training data is more valuable than you think: A simple and effective method by retrieving from training data[J]. arXiv preprint arXiv:2203.08773, 2022.