中国人工智能学会
Chinese Association for Artificial Intelligence

2022 杭州国际人才交流与项目合作大会
2022 HANGZHOU INTERNATIONAL HUMAN RESOURCES EXCHANGE AND COOPERATION CONFERENCE

余杭人才 YUHANG TALENTS

京东

2022 | HANGZHOU GLOBAL AI INNOVATION CONTEST

杭州全球人工智能技术创新大赛

赛道二：商品标题实体识别

虎牙181469的团队

目录

整体设计

赛题描述

背景：京东商品标题包含了商品的大量关键信息，商品标题实体识别是 NLP 应用中的一项核心基础任务，能为多种下游场景所复用，从标题文本中准确抽取出商品相关实体能够提升检索、推荐等业务场景下的用户体验和平台效率。

数据：本赛题数据来源于京东商品标题，共包含百万量级无标注样本、4万训练样本和1万测试样本。

任务：利用模型抽取出商品标题文本中的实体。与传统的实体抽取不同，京东商品标题文本的实体密度高、实体粒度细，赛题具有特色性。

整体设计

评价指标

本赛题采用实体级别的 micro F1 值作为排名依据。

$$P = \frac{|S \cap G|}{|S|} \quad R = \frac{|S \cap G|}{|G|}$$

$$F_1 = \frac{2PR}{P+R}$$

复赛阶段，将根据模型在单卡 GPU（NVIDIA T4，或者同等算力的 GPU 卡）上的推理耗时，对 micro F1 值进行惩罚

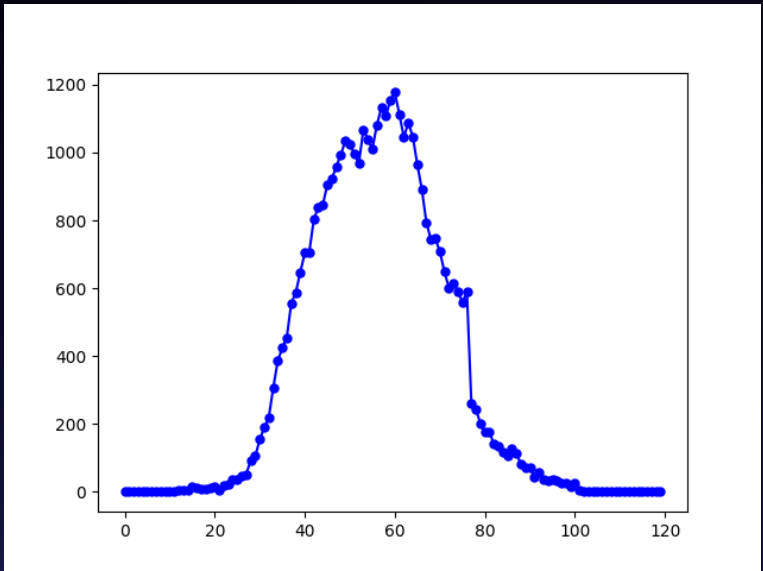
$$F_1 = \begin{cases} F_1 & if\ tinference \leq 360 \\ F_1\left(1 - \frac{t_{inference} - 360}{2000}\right) & if\ tinference > 360 \end{cases}$$

整体设计

数据分析

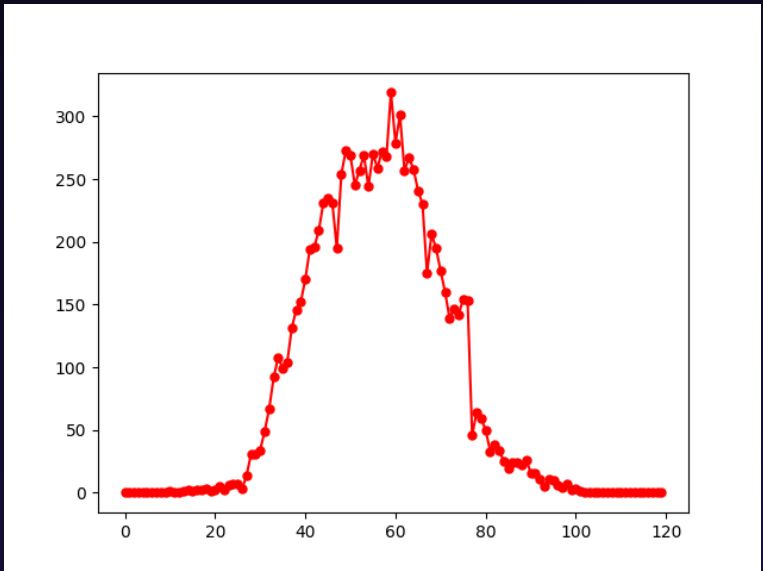
训练集长度分布

经过分词器分词得到的文本最大长度101，文本长度集中在60上下，整体数据分布不长。



训练集长度分布

与训练集的分布近似。

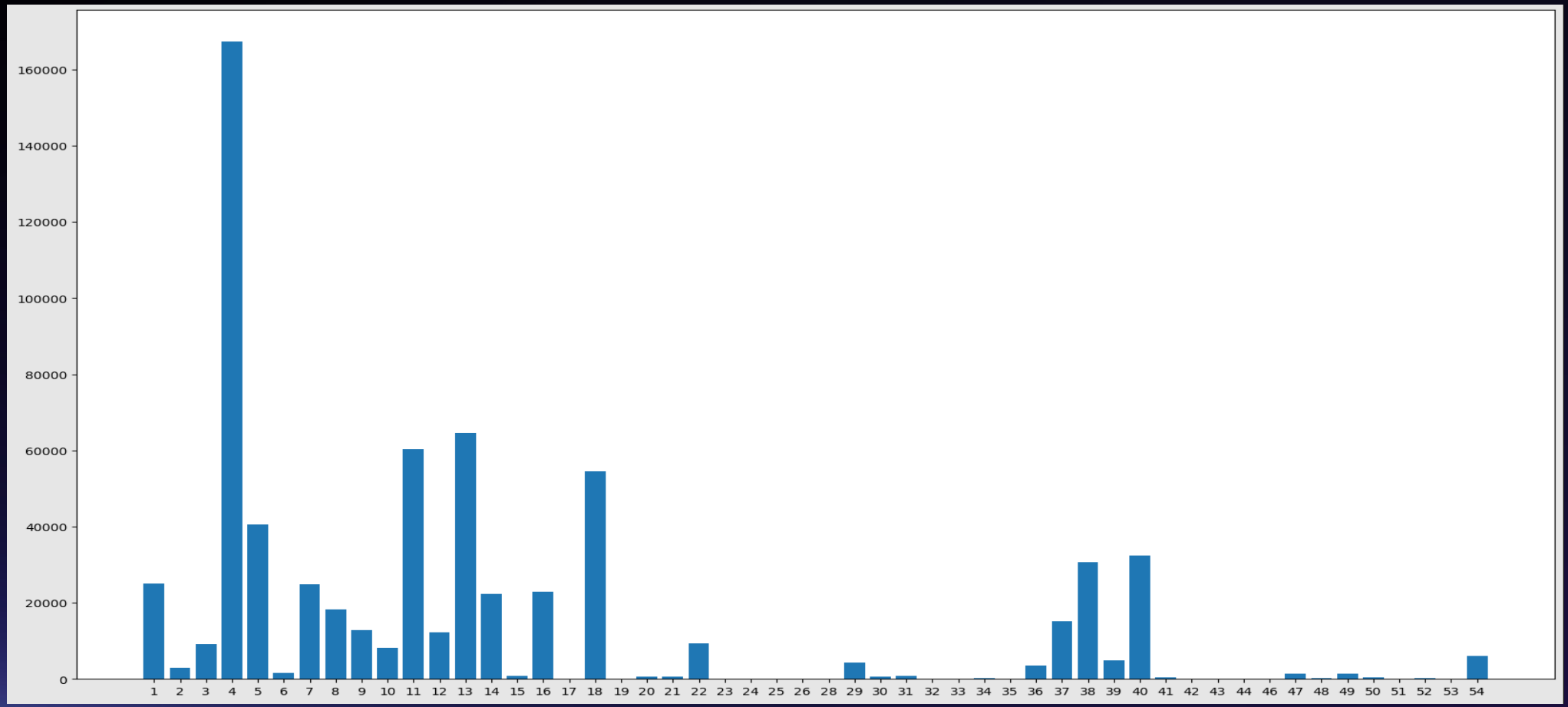


整体设计

数据分析

标签分布

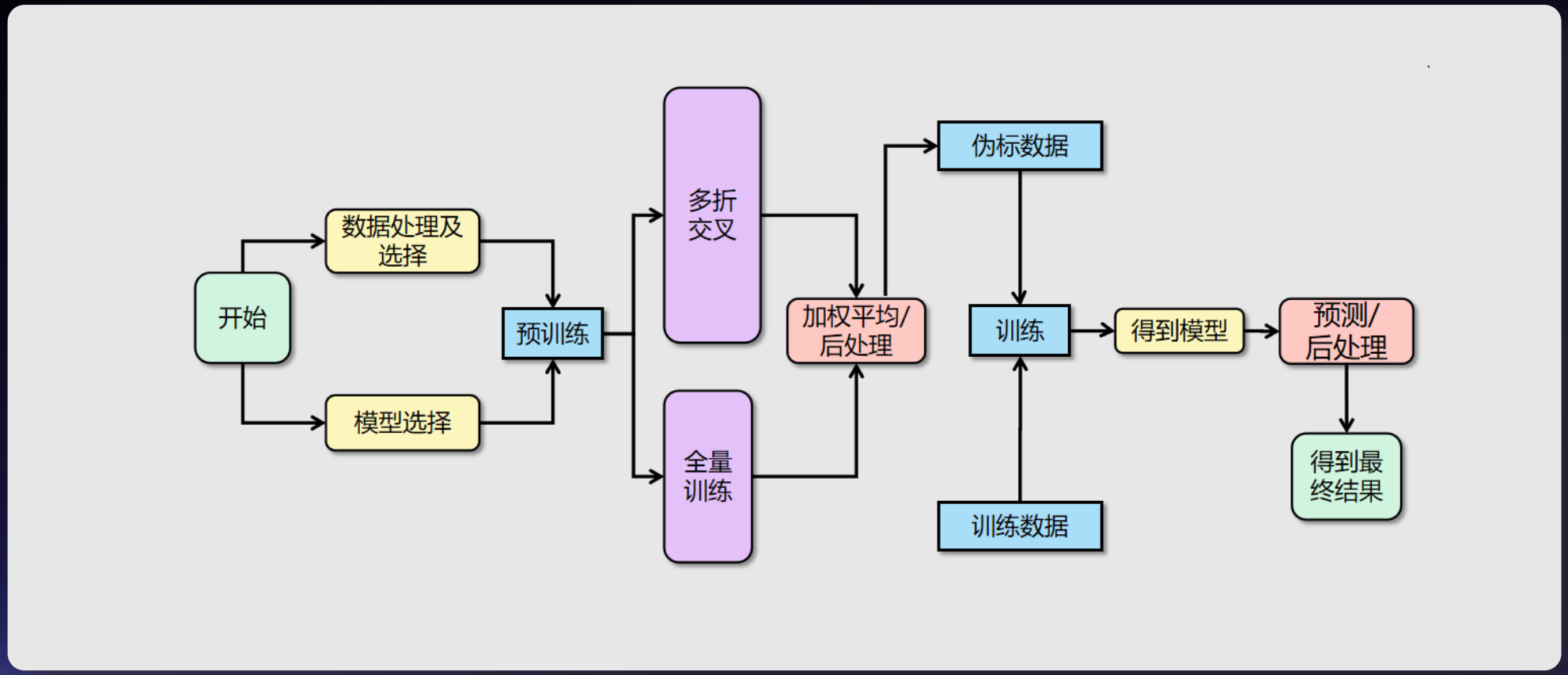
本赛题标签数量较多，共有52个，实体不仅仅与实体词有关，而且与当前标题所售卖商品有关，标签最多 4 有 167271 个，最少的 26 只有 1 个，标签分布不均衡。



整体设计

总体流程

本团队方案主要包括数据处理、预训练、伪标、训练、预测、后处理等步骤，流程图如下：



整体设计

模型选择

结构选择

本赛题的标题文本实体密度高，粒度细，由于赛题禁止模型融合，所以尽早选取合适的模型结构至关重要，我们在赛初尝试了多种结构的模型，如BERT-CRF，BERT-SPAN，GlobalPointer 等。其中 GlobalPointer 在解决实体密度高的任务中更具优势，线上效果也更出色，因此，我们选用 GP 作为此赛题的基本结构。

权重选择

预训练权重的选择上，尝试了 roberta，macbert，UER，nezha 等权重，其中NeZha的效果最好，我们对比了 nezha-large 和 nezha-base，效果接近，考虑到效率等反面，我们最终选用 nezha-cn-base 作为预训练权重。

整体设计

预训练

训练数据

100w无标注数据，4w训练数据，初赛A榜和B榜2万测试数据

MASK策略

在本赛题中尝试了常规 MASK，Whole Word MASK 以及N-gram MASK，经对比，发现采用常规 MASK 效果最好，采用动态MASK策略，训练时随机掩盖掉 15% 的token，每次迭代都会生成新的MASK文本，增强模型鲁棒性。

训练策略

混合精度训练：由于数据量过大，使用混合精度训练，加快模型训练过程。

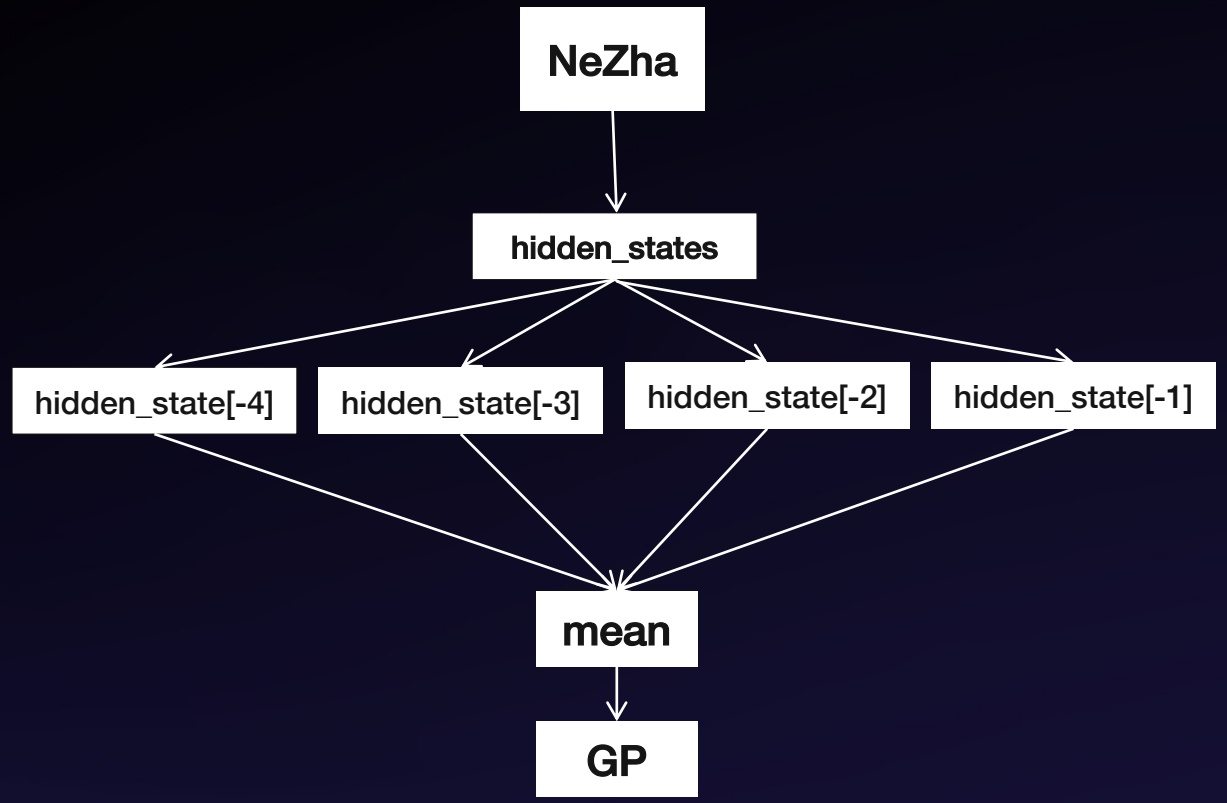
学习率warmup与衰减： 训练初期使用小学习率，使模型慢慢趋于稳定，相对稳定后使学习率上升到预设值， 之后使用余弦衰减使得学习率慢慢下降，训练过程平滑。

学习率设置：在数据量较大以及加载已有的模型权重作为初始化的情况下，使用较小的学习率进行预训练，起到微调的作用，使得模型能够慢慢适应电商标题领域的数据。

整体设计

微调

后接结构



经过实验，取 nezha 的最后四层隐藏层取平均，效果优于只选某一层 hidden state。

整体设计

微调

训练数据

4w 训练数据

训练策略

对抗训练： 对抗训练是一种引入噪声的训练方式，提升模型的鲁棒性。团队尝试了FGM、PGD等对抗方法，在本赛题中，我们将两种方法进行结合，采用FGM和PGD交叉训练的方式，效果优于PGD，且训练速度比单独使用PGD快。

SWA： 随机加权平均， 一种通过梯度下降改善深度学习泛化能力的方法,而且不会要求额外的计算。我们对训练保存的第3到第5个epoch的权重进行平均，相比取最优epoch提升明显。

整体设计

微调

训练策略

学习率warmup与衰减： 训练初期使用小学习率，使模型慢慢趋于稳定，相对稳定后使学习率上升到预设值，之后使用余弦衰减使得学习率慢慢下降，训练过程平滑。

权重衰减：限制网络权值的大小，防止过拟合现象。

Spatial Dropout：在SpatialDropout中，整个嵌入通道都将被丢弃，而对embeding进行dropout将丢弃整个单词的所有通道，有时丢失一个或多个单词会完全改变含义。

整体设计

微调

伪标签

为充分利用无标签数据，我们尝试对其打伪标并将这些数据加入微调训练中。

伪标策略：在原始数据上训练，使用多折交叉验证训练出 10 个模型，用全量数据训练一个模型，之后对其进行 logits 加权平均，解码+后处理，尽可能的得到高质量伪标数据。

伪标筛选：按照置信度进行筛选，最终得到30w伪标。

参数设置

-epoch：6
-batch size：16
-learning：NeZha层4e-5，非NeZha层5e-4
-dropout rate：0.3，训练时以一定概率丢弃某些神经元，防止过拟合
-weight decay：0.01
-warmup step：(total step)/epoch

整体设计

模型推理

解码策略

对于GP输出的（ent_type_size, max_seq_len, max_seq_len）的三维矩阵，解码得到所有的实体，因为本赛题不存在嵌套实体，故针对解码出来的嵌套实体的 case，采用高分过滤原则，首先将所有嵌套实体归纳到一个集合，对每个实体的分数进行排序，只取最高分的实体作为最终的输出。

例：[‘绘图文具盒’，‘文具’ ， ‘文具盒’]，scores : [0.468，0.562，0.768]，选文具盒作为最终结果。

后处理

后处理是比赛中常用且有效的一个点，本赛题实体标签分布广且训练集标注质量不佳，难以针对 badcase 设计效果明显的规则进行约束。通过对数据的观察，我们从标点符号和单字实体下手，对斜杠、空格，横杠、括号等标点符号进行不同的处理。主要采用的有提升的后处理方式有以下几种：

1). 符号方面以斜杠为例，斜杠是分隔符的一种，根据经验，其大概率代表着将两种实体分隔开，因此我们对测试集中斜杠分类为 / I-X 的改为 / O；
2). 针对出现在训练集中的单字实体，进行保存，对模型预测的结果进行限制，若预测的单字实体没有出现在训练集，则进行过滤；

目录

创新

- **对抗训练**：与常规对抗不同的是，我们不在单独的使用一种攻击策略，而是将 FGM 和 PGD 混合使用，交叉训练，训练时长低于单独使用 PGD，效果优于 PGD。
- **伪标签**：使用全量加多折进行融合预测，提高伪标质量，筛选30w伪标签加入训练，充分利用无标签数据，而不只是对其进行预训练。
- **解码策略**：设计针对此赛题的解码方法，解决类似嵌套实体的情况。
- **后处理**：从标点符号和单字实体出发，设计规则的后处理对实体进行约束。

落地

- 我们的模型以end-to-end的方式进行实体识别任务，可以无差别的识别出嵌套实体和非嵌套实体，通用性强，不局限于本赛题数据，可以广泛落地于自然语言处理领域中涉及实体识别的各类任务中，如知识图谱，推荐系统，问答系统等。
- 推理速度较快，在复赛1w条数据上用时270~280秒，远低于赛题限制，没有模型融合，没有使用large预训练，模型参数量小，训练所需资源少。
- 可扩展性强，我们的模型结构简单，并没有后接复杂的结构，可以在整体架构不变的情况下，针对不同的任务设计相应的结构进行适配，迁移方便。
- 深度学习发展迅速，效果好，但也不是万能的，其很难学习到数据的全部特性，我们用规则对其加以约束，系统鲁棒性得以提升。

目录

方案总结

阶段总结

（1）模型选取阶段。尝试多种不同的解码结构进行实验对比，选取最优的解码结构；
（2）预训练设置阶段。采用掌握的多种预训练任务进行尝试，选取最优的策略进行预训练；
（3）微调阶段。根据赛题的限制，采取伪标签的策略，将多模型的信息集成到单个模型上进行学习；
（4）预测阶段。针对赛题数据特点，设计多种不同的后处理方案。

优缺点

优点：方案较充分的利用了无标签数据，训练所需资源较少，推理速度快，鲁棒性高。

缺点：模型在少数据量情况下表现不好，少样本/零样本学习能力较弱，整体结构简单，处理复杂任务的能力欠缺。

展望

科学研究：可以更具针对性的设计相关预训练任务、引入NLP新范式Prompt、设计更合理的结构等。

应用角度：更多的训练数据、便捷的部署和迁移等，让实体识别任务可以精确高效的服务于检索、推荐等业务场景下，增加用户体验和平台效率。

中国人工智能学会
Chinese Association for Artificial Intelligence

2022 杭州国际人才交流与项目合作大会
2022 HANGZHOU INTERNATIONAL HUMAN RESOURCES EXCHANGE AND COOPERATION CONFERENCE

余杭人才
YUHANG TALENTS

京东

感谢观看