Introduction

KDD Cup 2020 Multimodalities Recall 这个赛题主要是淘宝真实业务场景下基于用户 query 的多模态召回问题。参赛选手们需要根据官方提供的 query 和商品图片特征对候选商品进行排序，最终的评价指标是 nDCG@5 。

队伍组成

队伍名称：Vaticinator

队长：张琦 （京东）

队员：王黎翔（东南大学）；李长宇（电子科技大学）；仉鹏（青岛大学）；郑升圆（西安电子科技大学）；王凯（西安交通大学）；许煜东（电子科技大学）；钟磊（深圳大学）；金辰宇（阅文集团）；李敬杰（北京大学）

方案简介

我们的最终方案可以概况为以下的三部分，在后文会逐一进行详细介绍。

1. LXMERT 模型
2. 基于 LightGBM 的 Learning to Rank 模型
3. Ensembling

Part1. LXMERT 模型

模型选取

基于 LXMERT 模型的预训练模型进行 fine-tune。

ROI Select 机制

我们分析了赛题的视觉属性的数据，发现在很多样本中存在大量无关的 box 以及标注有偏差的 box。

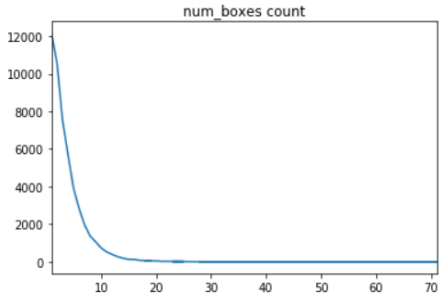

图 1 每个样本中 box 数量统计图

如果不加选择的去进行模型训练的话必然会使模型变得难以很好的收敛且学习到的特征存在偏差。所以本题在 LXMERT 原有模型的基础上加上了 ROI Select 机制。

Target Select 是对所有的 box 类别做统计，我们认为一个 query 语句一定在数据集的 30 多个类别中存在主要类别，所以针对一个样本我们会在 target select 中 mask 非主要类别的 box。Area Select 是筛选出 TOP N 个面积最大的 box，我们通过观察数据发现，当一个样本的 box 数量很多的时候（>20），面积过小的 box 往往是和 query 甚至是和物体类别无关的，所以 Area Select 更关注于明显的 box 以及他的 feature 信息。

对于初始的图像类型数据分别做 Target Select 和 Area Select 后，我们将筛选出的 box 混合到一起，为了避免引入不必要的空间语义信息，这些 box 的顺序会被打乱，并通过设置最大的 box 数（本方案分别测试了 box 数位 5、20），来选出符合条件的 box 以及 box 对应的 feature。

Hard Sample 机制

由于验证集和测试集中样本大多来自同一类型的数据，即当 query 是鞋子时，候选池中的图片也大多为鞋子。导致基于随机选择负样本构建的模型在实际验证过程中分数不理想，因此我们基于 query 相似度选择选取更加符合 valid 和 test 的负样本，在训练集上的操作具体流程如下:

1. query 中最后一个单词（lastword）往往代表当前 query 中的名词（符合大家搜索习惯），因此首先根据 lastword 对 query 分组
2. last_word 分组后计算组内 Top-K 相似度
3. 选取 Top-K'的 query 作为负样本 query（在此我们选取组内相似度后 20%的作为负样本）
4. 实际训练过程中，为防止负样本较少造成的过拟合问题，以一定概率选取 Hard Sample 或随机选取 query 作为负样本

整体流程图

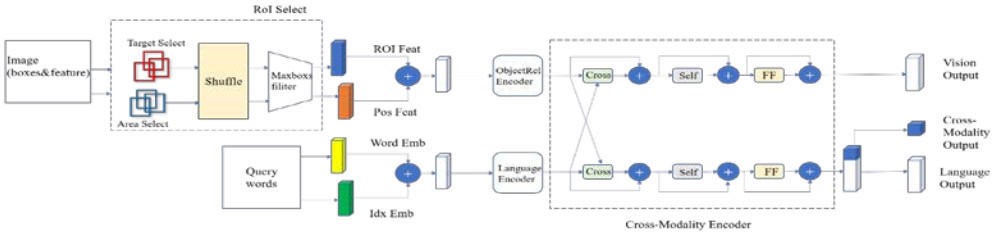

图 2 我们改进的 LXMERT 模型

训练机制

1. LXMERT 预训练模型直接训练
2. 基于 ROI Select 机制的 LXMERT 预训练模型
3. 基于 Hard Sample 机制的 LXMERT 预训练模型

Part2. 基于 LightGBM 的 Learning to Rank 模型

我们采用了基于 LightGBM 的 LambdaMART 方法，通过提取了一系列的特征来对商品进行排序建模，直接进行 nDCG 的优化。

训练数据

官方提供验证集数据中，每个搜索 id 对应多个 target，其中为 5 个正例其余均为反例，根据 LightGBM 模型的特点，验证集数据符合模型要求的数据格式。但由于验证集数据数量过少，所以采取在训练数据中人为构造负样本来模拟真实数据。构造策略为随机选取与搜索 id 不匹配的数据，将搜索 id 于其进行合并将 target 置为 0 。

特征构造

1. 词频特征：运用 CountVectorizer 与 TfidfTransformer 统计验证集与测试集中 query 的词频，作为新的特征，此特征可以从从侧面反映出热门商品与热门搜索词之间的关系
2. 统计特征：围绕搜索 id，产品 id，class 构造统计特征，挖掘产品，搜索词，类别之间的深层关系。例如统计每个 id 对应的 class 数量以及统计商品照片对应的 class 数量，挖掘两者之间最大数量的关系。
3. 概率特征：围绕搜索 id，产品 id，class 构造概率特征，例如计算搜索 id 对应最多类别概率与商品最大概率类别之间的隐含关系等
4. 词向量特征：此赛题的重难点为如何挖掘 query 与商品之间的隐含关系。主办方提供由 CNN 网络得到的商品的高级特征，共 2048 维。query 则为单词短语，主办方并未提供其高级抽象。故我们可以人为训练得到 query 的高级抽象，再将得到的特征与 CNN 得到商品高级抽象进行联合以进行后续下游任务。针对本次赛题，主要采取的策略为：1. 将训练集中大量的 query 短语抽取出来。2. 将从训练集中得到的 query 短语，送入 NLP 模型中得到其高级抽象。3. 利用 NLP 模型得到的特征，自定义 DNN 网络，训练数据为商品的高级抽象，target 为 NLP 得到的 query 高级抽象。4. 训练 DNN 网络，用商品的特征逼近 query 短语特征，得到二者关系。5. 利用训练好的 NLP 模型与 DNN 模型分别对验证集与测试集提取相应的特征，将二者之差作为新的特征。特征提取与构造如下图所示。

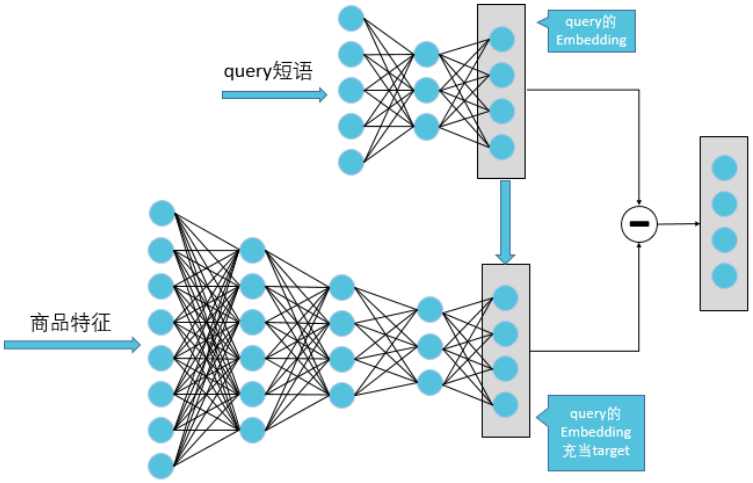

图 3 NLP 特征构造过程

模型训练

基于 bagging 的思想，采用 5 折交叉验证的方式进行模型训练，对得到的 5 个模型做均值融合。

整体流程框图

1. 使用训练集中 query 短语训练 NLP 模型
2. 自定义 DNN 网络使 CNN 得到的商品高级抽象逼近 NLP 模型的输出
3. 使用训练好的 NLP 模型与 DNN 模型作用在验证集和测实际上，将二者差值作为新的特征

4. 使用原始特征构造词频特征，统计特征，概率特征等，并且与上一步得到的特征进行融合
5. 送入 LGB 模型中进行训练

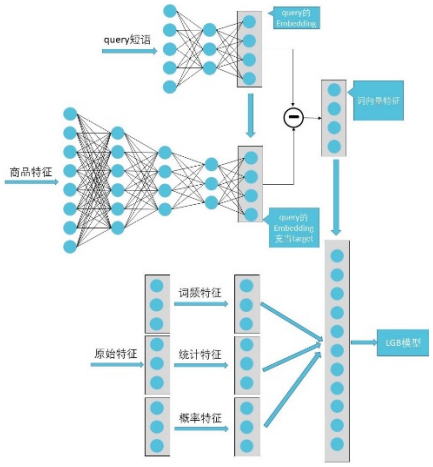

图 4 整体流程框图

Part3. Ensemble

验证集与测试集的相似性、训练集与测试集的差异性

整个训练集包括 3M 对相互匹配的 query 与产品图片的特征，而验证集与测试集则提供了若干 query 信息，每个 query 对应一个 30 左右大小的商品候选池，从结构上训练集与后两者的差异显而易见，另一个差异在于，很多在 train 中大量出现的商品类别，在测试集与验证集并未出现，如 books。

而测试集与验证集在商品类别上具有很强的相似性。不难发现，query 最后一个单词往往就是商品的具体类别，按照最后一个单词进行统计，可以得到训练集与测试集的高频词，如下图所示。

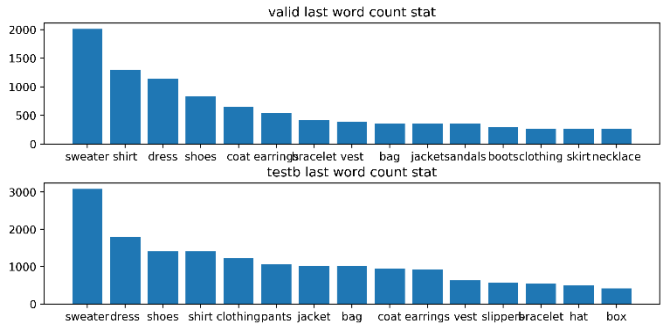

图 5 高频 last word 统计

为了充分利用验证集与测试集的相似性与规模庞大的训练集，本文使用了**迁移学习**与**嫁接**两种技术。

迁移学习在计算机视觉的各个领域被广泛使用，源域与目标域的相似性是这一技术成功应用的关键。本文使用的 lxmert 模型在训练集上微调之后，采用五折 stacking 方法继续在验证集中微调。

嫁接是指当可用于训练的数据中包含与测试集分布不一致的数据集，可以考虑不丢弃这一部分（特别是这部分数据量较大时），而是使用这部分数据先训练一个模型（使用剩余部分做验证，从而使用早停等训练策略），然后对剩余部分（与测试集分布较为一致的部分）以及测试集进行预测，并把输出向量作为特征，在剩余部分继续训练一个模型。这样既可以保证充分利用所有数据，又不影响模型的泛化性能。

集成方案

在嫁接与迁移学习两种思路的引导下，模型的训练过程可以抽象成如下 Pipeline，具体过程如下描述：

Pipeline 步骤
1.对原始特征进行特征工程
2.对训练集进行采样操作
3.在训练集进行微调
4.第一次嫁接生成特征表示
5.在验证机上进行五折的微调
6.第二次嫁接生成特征表示 2

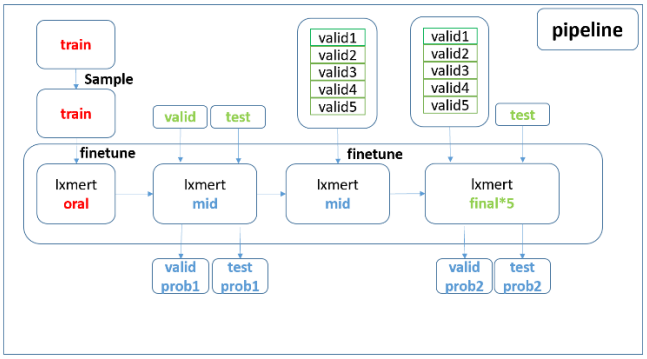

图 6 Pipeline 步骤

在 Pipeline 的基础，通过改变**特征工程的方法**，**采样方式**，**模型结构**，**训练超参**，生成若干组特征表示，最后使用 LightGBM 完成最后集成。

有两点需要说明的是，Pipeline 的模型，也可以是 LightGBM，只需将第一次微调操作改成训练，将第二次微调操作改成继续训练即可。省略在训练集训练这一步，亦可生成有效的特征表示。实际上，在只用图像特征的情况，在验证集上使用 LightGBM 进行训练，即可达到 50+的线上分数，加入 query 等特征后，在验证中 oof(out of fold)的表现可达 0.62+。

最终最高分为 **0.7426**，这其中我们融合了 20 个模型，包括 13 个 LXMERT 模型，4 个 hard_sample_LXMERT 模型，3 个基于特征的 LightGBM 模型

其他尝试

我们也实验了 visul-bert、vlbert、 uniter 等多模态 bert 模型，但由于算力限制、调试困难等等原因，最终没集成到我们的最终方案中。有待以后进行更多的尝试。

联系方式

如有问题请联系：hi@zhangqibot.com

Reference

1. 赛题官网
2. LightGBM
3. Learning to Rank Using Gradient Decent
4. LXMERT: Learning Cross-Modality Encoder Representations from Transformers
5. 结构化数据的迁移学习：嫁接学习
6. 嫁接