KDD 2020 电商平台挑战赛：多模态检索比赛技术报告

Yuhui Ding
Zhejiang University
Hangzhou, Zhejiang
dingyh@zju.edu.cn

Lei Wu
Zhejiang University
Hangzhou, Zhejiang

Chengxi Li
Zhejiang University
Hangzhou, Zhejiang
chengxili@zju.edu.cn

Jieyu Yang
Zhejiang University
Hangzhou, Zhejiang
yangjieyu@zju.edu.cn

Yue Wang
Zhejiang University
Hangzhou, Zhejiang

摘要

ACM Reference Format:
Yuhui Ding, Lei Wu, Chengxi Li, Jieyu Yang, and Yue Wang. 2020. KDD 2020 电商平台挑战赛：多模态检索比赛技术报告. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

作为挑战赛中国区 B 测试集测试结果排名第十，A 测试集测试结果排名第十一的队伍，我们在这份报告中给出本组提出的对于挑战赛中多模态检索问题的一种技术解决方案。此方案首先使用一个多标签分类任务对图像特征的编码器进行预训练，然后将预训练好的编码器用在两个匹配模型中，最后将两个匹配模型集成。此方案在 B 测试集得到了 0.728 的 nDCG@5 分数，在 A 测试集得到了 0.716 的 nDCG@5 分数，并且单模型在 A 测试集上得到了 0.697 的 nDCG@5 分数。此解决方案中各模型和训练任务的 GPU 需求最多为 4 块 2080Ti，训练时长不长，总体来说是一种很有竞争力的技术解决方案。实现本方案全部的代码[1]和外部材料都已打包上传。

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 引言

作为挑战赛中国区 B 测试集测试结果排名第十，A 测试集测试结果排名第十一的队伍，我们在这份报告中给出本组提出的对于挑战赛中多模态检索问题的一种技术解决方案。此方案首先使用一个多标签分类任务对图像特征的编码器进行预训练，然后将预训练好的编码器用在两个匹配模型中，最后将两个匹配模型集成。此方案在 B 测试集得到了 0.728 的 nDCG@5 分数，在 A 测试集得到了 0.716 的 nDCG@5 分数，并且单模型在 A 测试集上得到了 0.697 的 nDCG@5 分数。此解决方案中各模型和训练任务的 GPU 需求最多为 4 块 2080Ti，训练时长不长，总体来说是一种很有竞争力的技术解决方案。实现本方案全部的代码[1]和外部材料都已打包上传。

2 模型

2.1 图片编码器

本解决方案中图片编码器的设置参考了 [1–3] 等文章中的一些架构设置，接收的输入为若干个边界框和其中内容的 2048 维 FastRCNN 提取的特征，编码器接收边界框特征后做进一步编码，边界框处理时采用线性连接层和 PReLU 激活函数相间连接的方式，初步处理后由一个多层堆叠的多头注意力网络、前馈神经网络和 PReLU 激活函数接收，此处层数可调。边界框特征处理和多头注意力网络中都有 Residual Connection 连接。训练时的 Dropout 设置为 0.1。

输入为图片 $I^{n\times 2048}$ 和边界框 $B^{n\times 5}$，n 代表图片中物体个数，2048 和 5 为维度。边界框 5 个维度分别代

[1] https://code.aliyun.com/zjhndyhnba/KDD-Cup-Multimodalities-Recall.git

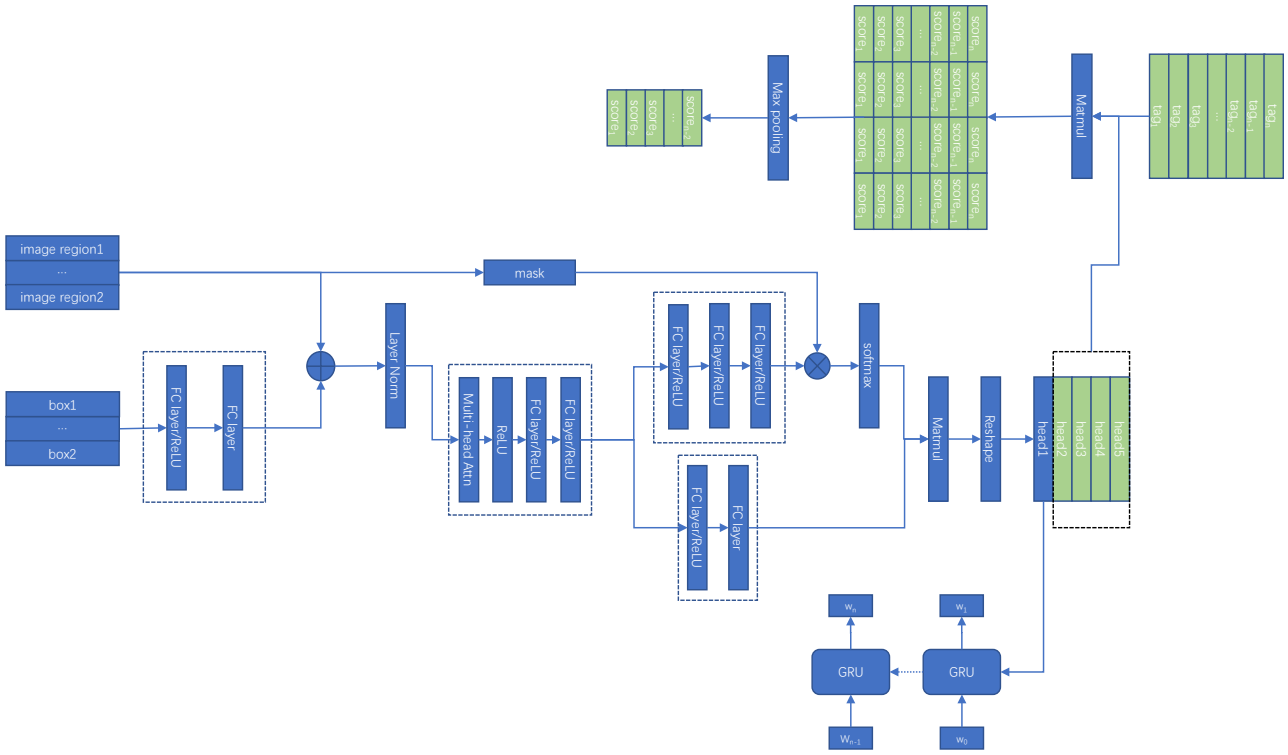

图 1: 预训练模型框架示意图

表左上角和右下角的坐标以及面积。图片和边界框首先进行合并，公式为 $X^{n\times 2048} = I + \mathbf{fc}(B)$，其中 **fc** 为全连接层，将边界框扩展到 2048 维。接着使用多头注意力层 $X^{n\times 2048} = \mathbf{mha}(X)$ 进行变换，接着使用两个全连接层分别产生两个分支 $F^{n\times 1024*5} = \mathbf{fc}(X)$ 和 $W^{n\times 1} = \mathbf{softmax}(\mathbf{fc}(X))$，其中 F 是经过维度压缩后的图片特征，1024*5 代表 5 个 1024 维的头（head），第一个头用于图片标题生成，后四个头用于多标签分类，具体后一节会进行阐述。W 为图片中每个物体的权重，用于对 n 个物体的特征进行加权平均，公式为 $W^T \times F$。

2.2 预训练模型：多标签分类 + 图片标题生成

本模型使用多标签分类和图片标题生成两个任务共同对上述图片编码器进行预训练。

图片标题生成模型使用传统的 Seq2Seq 模型，主体是用 RNN 进行生成。上节提到的图片编码器输出结果的第一个头作为 RNN 起始单元的隐藏层输入，查询文本的单词作为每个 RNN 单元的输入层输入，RNN 单元的输出是下个单词的预测。即给定上文和图片，最大化当前输出单词的概率，公式是 $\mathbf{Prob}(x_i|x_{i-1}x_{i-2}...x_0, h)$，$x$ 代表单词，h 代表图片特征。

多标签分类模型的标签提取由查询文本分词并去除停用词得到，一般而言查询文本的单词都代表了图片中具备的一些属性，这么处理虽然有些粗糙，但实际效果不错。对于分类问题而言，常规做法是通过全连接层输出 k 维的向量，k 为种类数。但实际发现这么做效果不佳，因此我们将所有标签使用 1024 维的向量表达，和图片向量内积产生每个标签的得分。而每个标签使用 bert 词向量进行编码，可进一步加快收敛速度。此外，上一节提到图片编码器生成结果中有 4 个头用于多标签分类，这是因为考虑到标签的词性不同，如形容词、名词等，而一张图包含了多个标签，单

头向量可能无法包含所有标签的信息，因此每张图都和所有标签分别计算四个匹配分，每次选择其中最大的头作为匹配分。

2.3 基于单词（标签）的匹配模型

事实上，多标签分类模型已经可以通过对图片和查询文本中每个词的匹配分进行平均得到最终的匹配分,但是缺点是所有词的权重是相同的。比如一个查询文本“儿童衣服”，显然主体是“衣服”，我们希望“衣服”这个单词的权重要相对高一些。因此，本模型在之前的多标签分类模型基础上，通过神经网络对查询文本钟的每个单词生成一个权重，对每个单词的匹配分进行加权平均作为最终匹配分。在验证集上有 2 点以上的 nDCG@5 分数的提升。

模型使用 pairwise 和 Margin Ranking Loss 进行训练，正负样本使用手工配对和 mini-batch hard mining 产生。

本模型在线下 ndcg@5 分数可达到 0.729，测试集 A 达到 0.697，对单模型来讲效果非常可观。

2.4 基于句子的匹配模型

本模型图片标题匹配是此解决方案集成的模型之一。在比赛提供的训练集和检验集上进行实验时，我们发现基于整句话的嵌入编码和图片编码进行匹配的模型在此多模态检索的某些情形下的表现有可取之处。

模型主体使用一个双向 RNN，图片编码（用于图片文本生成的头）作为 RNN 起始单元的隐藏层输入，查询文本中每个单词输入到 RNN 单元的输入层，使用输出的隐藏层，经过线性变换得到匹配分。

模型优化的损失函数为二分类交叉熵损失(Binary Cross-Entropy Loss)，简单地基于是否匹配来优化网络参数，训练图片标题匹配模型时会冻结图片编码器和单词 embedding（来自 bert）的权重，只更新后续网络结构中的参数。

本模型在线下 nDCG@5 分数可达到 0.71 左右，线上未经测试。后续也尝试了 ViLBert[4]、VLBERT[5] 等模型，可达到和基于单词（标签）匹配模型持平的分数，可惜在 B 测试集上集成后的效果略差，且计算量过于庞大，对服务器的性能有巨大要求，因此没有过多尝试。

3 实验

3.1 采样

由于比赛数据只提供 query-product 正样本对，我们分别对 query 和 product 进行了负采样。采样方法具体为：分别以 query 文本最后一个单词作为类别 c1 和 query 文本训练得到的 embedding 进行 K-Means 聚类得到的类别 c2，以 0.8 的概率在正样本 query 的同一 c1 类别中采样得到 query 负样本，以 0.2 的概率在正样本 query 的同一 c2 中类别中采样得到 query 负样本，并根据 query 负样本得到对应的 product 负样本。负样本用于排序损失函数的学习。

3.2 模型集成

对模型 1 和模型 2 在验证集上的预测分数分别乘以权重 k 和 1-k，并对 k 值从 0 到 1 等分出的 100 个点中进行搜索，得到模型 1 和模型 2 的权重，进行模型集成。

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[2] K. He, X. Zhang, S. Ren, and J. Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 1026–1034.

[3] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[4] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 13–23. http://papers.nips.cc/paper/8297-vilbert-pretraining-task-agnostic-visiolinguistic-representations-for-vision-and-language-tasks.pdf

[5] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. arXiv:1908.08530 [cs.CV]