Multimodalities Recall via Ensembling Transformer-based Models

Kai Zuo
Meituan-Dianping Group
zuokai02@meituan.com

Chao Ma
Meituan-Dianping Group
machao32@meituan.com

Dongshuai Li
Meituan-Dianping Group
lidongshuai@meituan.com

Zuo Cao
Meituan-Dianping Group
zuo.cao@dianping.com

Xing Xu
University of Electronic Science and Technology of China
xing.xu@uestc.edu.cn

ABSTRACT

本文旨在针对 KDD Cup 2020 Challenges for Modern E-Commerce Platform: Multimodalities Recall 赛事，构建跨模态检索系统，根据手机淘宝平台的用户查询，对候选图像集合进行相关性排序，并通过 nDCG@5 指标对排序结果进行评估。本方案由 MTDP_CVA 团队提出，共包含了3个部分的内容，分别是：数据分析，模型构建与训练，以及结果融合和后处理。最终，该方案在 KDD Cup 2020 的testB 数据集上取得了第二名的成绩。

KEYWORDS

神经网络，多模态检索

ACM Reference Format:
Kai Zuo, Chao Ma, Dongshuai Li, Zuo Cao, and Xing Xu. 2020. Multimodalities Recall via Ensembling Transformer-based Models. In *San Diego '20: ACM SIGKDD Conference on Knowledge Discovery and Data Mining, August 23–27, 2020, San Diego, CA.* ACM, New York, NY, USA, 3 pages.

1 数据分析

本章对比赛数据进行分析，初步确定后续模型训练过程中使用数据的方式。

首先，对数据中图像包含的box的最大数量和query文本的最大长度进行了统计，结果见表1和表2。从表中可以看出，query的最大长度为21，图像中box的最大数量为91。在模型中，直接设置box的最大数量为91会导致模型规模过于庞大，不利于训练和微调。因此，本方案进一步对图像中box数量的分布进行了可视化分析，结果如图1所示。从图中可以发现，绝大多数图像所包含的box数量都小于20。因此在模型训练的过程中，我们首先尝试设定box的最大数量为20，后续在不影响结果的情况下进一步调整为了10。

其次，观察数据中的query字段，可以发现绝大多数query都是"多个形容词+名词"的形式，因此query的最后一个词[1]一定程度上代表了该query的类别。因此，在构建训练数据的过程中，应当将数据的上述性质考虑在内。

[1]这个词在本方案中被称为主体词.

Table 1: train数据中图像所包含box数量的统计

count	mean	std	min	25%	50%	75%	max
3,000,000	3.7709	3.3161	1	2	3	5	91

Table 2: train数据中query长度的统计

count	mean	std	min	25%	50%	75%	max
3,000,000	3.8642	3.2809	1	3	4	5	21

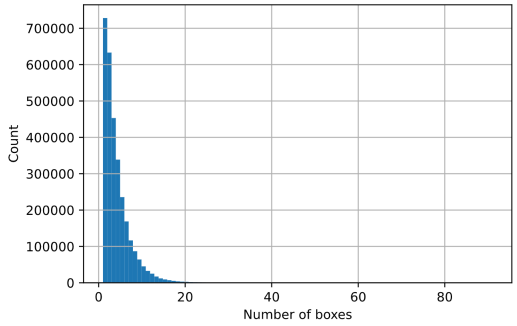

Figure 1: train数据中图像所包含box数量的分布

2 模型构建与训练

鉴于近年来，基于Transformer模型的算法在跨模态检索以及VQA等相关领域取得了突出成果，因此，本方案从上述领域最新的研究成果中，选择了两种模型来处理本次比赛的图文匹配任务，分别是LXMERT [4] 和ImageBert [3]，其中ImageBert采用了两种不同的模型结构。模型的细节如下：

1）LXMERT: 采用了与LXMERT论文中相同的特征网络，并在图像特征部分加入了图像所包含box的类别文本特征。特征网络的预训练权重使用了LXMERT的作者所提供权重文

San Diego '20, August 23–27, 2020, San Diego, CA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00

件[2]。在特征网络之上，使用了两层全连接来得到二分类的输出，其中第一个全连接层之后使用GeLU [2] 进行激活，然后通过LayerNorm [1] 进行归一化处理。在第二个全连接层之后使用Cross Entropy Loss计算损失，从而训练网络。

2) ImageBert A: 与原始ImageBert结构的差异如下：a）根据赛题中的要求，初步判定position的信息对于实际的相关性帮助不大，故去掉了ImageBert中Position Embedding部分的结构；b）不对图像特征和query的部分做掩码，仅训练相关性匹配任务，不进行MLM等其他任务的训练；c）将Segment Embedding统一编码为0，不对图像特征和query文本单独进行编码。d）在[CLS]位输出query与product的匹配关系，通过Cross Entropy Loss计算损失。依据上述策略，选用BERT-Base模型对变量初始化，在此基础上进行FineTune。

3）ImageBert B：与原始ImageBert结构的差异如下：a）对query文本，根据其位置信息得到Position Embedding，对图像特征，认为所有box的Position Embedding相同；b）不对图像特征和query的部分做掩码，仅训练相关性匹配任务，不进行MLM等其他任务的训练；c）文本的Segment Embedding编码为0，图像特征的Segment Embedding编码为1；d）在[CLS]位输出query与product的匹配关系，通过Cross Entropy Loss计算损失。依据上述策略，同样选用BERT-Base模型对变量初始化，在此基础上进行FineTune。

在完成模型构建后，接下来生成数据进行模型训练。本方案首先采用如下的策略生成训练样本：对于每一个batch的数据，按照1:1的比例选择正负样本，其中，正样本为train中的原始数据，负样本为通过对正样本中query进行替换产生，替换的query是按照一定策略从train数据中抽取得到的。为了提高模型学习效果，本方案在构建负样本的过程中进行了难例挖掘，以提高模型对于相近的正负样本的区分度。本方案首先对valid数据中box所属类别进行了统计分析，结果如下：

1）负样本的图像中至少有一个box的类别与正样本一致的数据占所有数据的97.74%;

2）负样本的图像中至少有一个box的类别与正样本一致，且该类别不是others类别的数据占所有数据的48.99%。

从上述结果可以发现，在构建负样本难例时，应当考虑与正样本的相关关系。通过使负样本图像所包含box的类别与正样本之间存在重合，同时以一定概率要求该重合类别不是others类别，从而构建一部分与正样本较为相似的负样本，提高模型对相似正负样本的区分度。

其次，本方案对valid 数据集query字段的主体词进行了统计，可以发现valid数据中负样本query的主体词大多与正样本query的主体词一致。因此，应当在构建负样本过程中，使得部分负样本的主体词与正样本一致，提高模型对这类相似正负样本的区分度。

根据上述数据分析结果，负样本所使用query的生成策略见表3。使用上述策略生成训练数据，训练后的模型在valid数据集上的效果如表4所示。

在完成初步的模型训练后，接下来使用不同的损失函数对模型进行进一步的微调，如AMSoftmax Loss [5]、Multi-Similarity Loss [6]等。AMSoftmax Loss通过权值归一化和特征归一化，在缩小类内距的同时增大类间距，从而提高了模型效果。Multi-Similarity Loss将深度度量学习转化为样本对的加权问题，采用采样和加权交替迭代的策略实现了自相似性，负相对相似性和正相对相似性三种，能够促使模型学习得到更好的特征。具体而言，所采用的策略如下：

[2]https://github.com/airsplay/lxmert

Table 3: 负样本**query**抽取策略

抽取比例	抽取策略
50%	抽取一条主体词相同的query
20%	抽取一条query，其对应图像box的label与正样本存在重合
20%	抽取一条query，其对应图像box的label与正样本存在重合，且重合的label不是others类别
10%	随机抽取一条query

Table 4: 初步训练后模型在**valid**数据集上的效果

Model	nDCG@5
LXMERT	0.7049
ImageBert A	0.6953
ImageBert B	0.6930

1）对于LXMERT，在特征网络后加入Multi-Similarity Loss，与Cross Entropy Loss 组成多任务学习网络，进行模型微调。

2）对于ImageBert A，在最后一层全连接层中加入了Multi-Similarity Loss，与Cross Entropy Loss一并对模型进行训练，来提高模型的整体效果。

3）对于ImageBert B，使用AMSoftmax Loss代替Cross Entropy Loss。

经过微调，各模型在valid数据集上的效果如表5所示。为了进一步提高模型效果，本方案根据train数据中query字段与testB中数据的query字段的相似程度，对train数据集进行了过采样。其采样规则具体如下：

1）对query在testB中出现过的样本，或与testB中的query为包含关系的样本，根据其出现的次数，按照反比例进行过采样；

2）对query未在testB中出现过的样本，根据其重复的词的数量，抽取了重复词的数目为前10的样本，每条样本过采样50次。

本方案首先使用过采样得到的训练样本对LXMERT模型进行进一步微调。其次，训练数据中存在不同query的单词相同但是顺序不同的情况，为了增强模型的鲁棒性，本方案以30%的概率对除了倒数第一个单词之外的其他单词进行随机打乱，以70%的概率对除了最后两个单词之外的其他单词进行随机打乱，将该数据生成策略与过采样策略相结合，生成训练数据对ImageBert A模型进行了进一步微调。对于ImageBert B模型，本方案从train数据中过滤了query主体词未在testB数据中出现、或出现过主体词但所有形容词都未在testB中出现过的数据，使用剩余数据对模型进行微调。训练后各模型在valid数据集上的效果如表6所示。

为了充分利用所有标记数据的信息，本方案进一步使用了valid数据集对模型进行FineTune。为了避免过拟合现象，最终提交结果只对ImageBert B模型进行了上述操作。

Table 5: 增加或更换损失函数对模型微调后在**valid**数据集上的效果

Model	nDCG@5
LXMERT	0.7094
ImageBert A	0.7098
ImageBert B	0.6980

Table 6: 重新采样数据对模型微调后在**valid**数据集上的效果

Model	nDCG@5
LXMERT	0.7159
ImageBert A	0.7015
ImageBert B	0.7150

在query-product样本对的相关性的预测阶段，本方案对testB数据的query 字段进行统计，发现其中"sen department"这一形容词在testB数据中大量出现，但在train数据中从未出现，只出现过"forest style"一词。为了避免这组同义词对模型预测带来的影响，选择将testB数据中query字段所包含的"sen department"替换为"forest style"，该策略下的结果记为"ImageBert C"。

3 结果融合和后处理

经过上述的模型构建和训练，本方案共得到了4个样本对相关性得分的文件。接下来对模型结果进行ensemble，并对最终得分进行后处理，得到query所对应的product候选集的相关性排序。

在ensemble阶段，本方案选择对不同模型所得相关性分数进行加权求和，来作为每一个query-product样本对的最终相关性得分，各模型[3]的权值为3:3:2:2。

在得到所有query-product样本对的相关性得分之后，接下来将根据样本对之间的相关性得分对query所对应的多个product进行排序。通过对valid数据的分析，可以发现，有部分product出现在了多个query-product样本对之中，本方案对这部分样本进行了进一步处理。首先，考虑到同一个product通常只对应一种query描述，因此认为同一个product只与相关性分数最高的query相关。使用上述策略对ImageBert A模型在valid数据集上所得结果进行后处理，模型的nDCG@5分数从0.7098提升到了0.7486。

受此启发，进一步分析数据发现，同一product对应的多条query往往差异较小，其语义也是比较接近的，这导致了训练后的模型对这类样本的区分度还不够理想，没有区分度的相关性分数会一定程度上引起模型效果的下降。因此，本方案对这类样本对进行了筛选，筛选策略如下：对包含同一个product的多个query-product样本对，如果这些样本对的相关性分数中，top 1的相关性分数超过top 2的相关性分数一定阈值，则保留top 1所对应的query-product样本对，删除其他样本对；反之则删除所有包含该product的样本对。使用上述策略对ImageBert A模型在valid数据集上所得结果进行后处理，当选定阈值为0.92时，模型的nDCG@5 分数从0.7098提升到了0.8352。因此，本方案对所有模型ensemble 后的相关性得分采用了同样的筛选策略，生成了最终的相关性排序。

最终，本方案在testB上的得分为0.843，位列第二名。

[3]按照LXMERT、ImageBert A、ImageBert B、ImageBert C的顺序。

REFERENCES

[1] Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).
[2] Hendrycks, D., and Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016).
[3] Qi, D., Su, L., Song, J., Cui, E., Bharti, T., and Sacheti, A. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966* (2020).
[4] Tan, H., and Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* (2019).
[5] Wang, F., Liu, W., Liu, H., and Cheng, J. Additive margin softmax for face verification. *arXiv preprint arXiv:1801.05599* (2018).
[6] Wang, X., Han, X., Huang, W., Dong, D., and Scott, M. R. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 5022–5030.