相关性特征在图片搜索中的实践

沈炎军 智能推荐系统 2019-06-30

图片来自网络

文章作者：沈炎军 搜狗

内容来源：搜狗图片搜索-八层会议室知乎专栏

一、引言

图片搜索是搜索引擎中比较重要的一块内容，相比于网页搜索，图片搜索需要结合文本、图像等多维度的特征，来对检索结果进行排序、去重等工作。在学术界，经常会提到多模态（Multi-modal）、跨模态（cross-modal）检索，图片搜索就是其中的典型场景。

本文主要分为两个部分，第一部分是图片排序中，相关性特征计算用到的一些模型，主要是用来计算文本检索词（query）和图片（doc）的相关性；第二部分是 attention 机制在图片排序特征融合上的实践。这两部分工作都是为图片搜索服务的，第一部分可以理解为一些常规方式，第二部分是结合 attention 的一种思路。

此外，本文不涉及图像检索（以图搜图、识图）等内容，说明下以免混淆。

二、图片搜索常用方法介绍

做过搜索引擎都知道，搜索最基础的两部分：召回+排序，召回功能由索引完成，排序就是对候选 doc 计算相关性分值，然后根据分值完成最终的检索结果排序。本文着重要介绍的就是排序部分的工作。

排序部分工作不是由一个模型完成的，用一个模型计算 query 和 doc 的相关性分值就直接排序，这样太简单粗暴了，也很难在工业上达到较好的效果。首先需要说明的是，图片搜索中，一个 doc 不单单包含图像信息，还包括一系列的文本域（ title、content、surround.. ）、站点（ site ）等信息，所以最终的排序是根据这些信息的综合排序。

因此大多数设计模式是，通过基础模型学习不同的特征维度来表示各个域的相关性，如 query 和 doc 文本相关性、和图像相关性、站点质量、图片质量等特征，然后使用模型将这些特征综合计算得到排序分值。这里我们关注的重点是相关性特征的表示。

2.1 相关性特征整体介绍

传统的相关性特征中，最为经典的就是 BM25 ，根据 tf 和 idf 来计算 query 和 doc 文本的匹配度，主要考虑在 term 空间上的匹配。此外还有 matchrank、PLS（ Partial Least Square ）、RMLS（ Regularized Mapping to Latent Space ）等。

随着深度学习的兴起，近几年的主要都考虑使用 DL 来计算 query 和 doc 相似度，有些地方也将其称为 deep matching 。在 DL 方法的应用上，根据模型框架的侧重不同，也可以分为两类方向，一类侧重于表示学习（ Representation learning，图1 ），一类侧重于匹配方法学习（ Matching function learning，图2 ），我们可以根据下面图1、2对两种方法做进一步的理解。

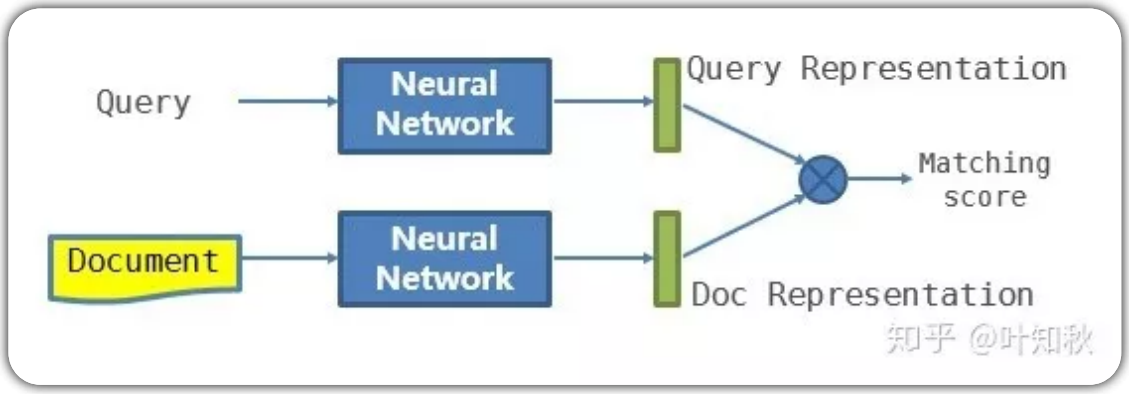

图1 Representation Learning

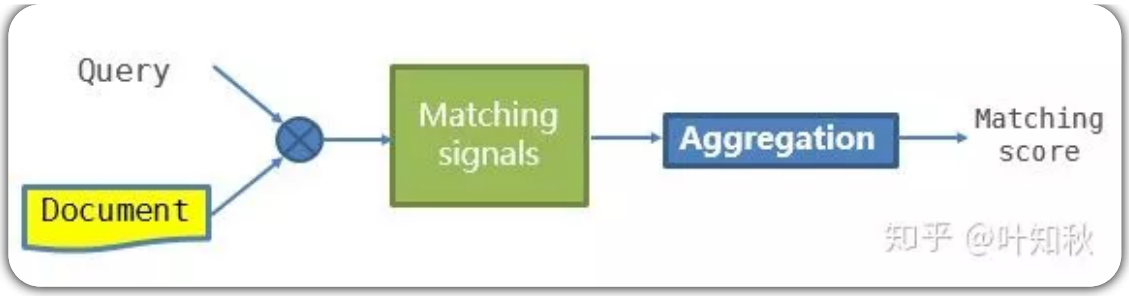

图2 Matching Function Learning

根据图1、2我们可以看到，在 RL 上，模型将 query 和 doc 分别根据神经网络各自得到一个潜在空间的表示，一般是一个 float 型的向量，然后使用 cosine 、欧氏距离等计算相关性。而 MFL 类模型是先计算 query 和 doc 的交互信息，在交互信息基础上通过模型计算相关性得分。

在图3中，列出了近几年两个方向的一些比较有代表性的论文。在 RL 方向上，列出了有 Transformer 和 BERT ，这里有同学可能会有疑问，因为这两个模型在初始论文中，并不是用来计算文本相似度的。在笔者看来，RL 模型的核心就是 query 和 doc 各自向量表示学习，Transformer 的 encoder 过程、BERT 的 token \ position embedding 最终都是要解决句子的向量表示问题，包括 EMLo ，虽然侧重点不一样，但都可以用在两个文本序列的相似度学习上。

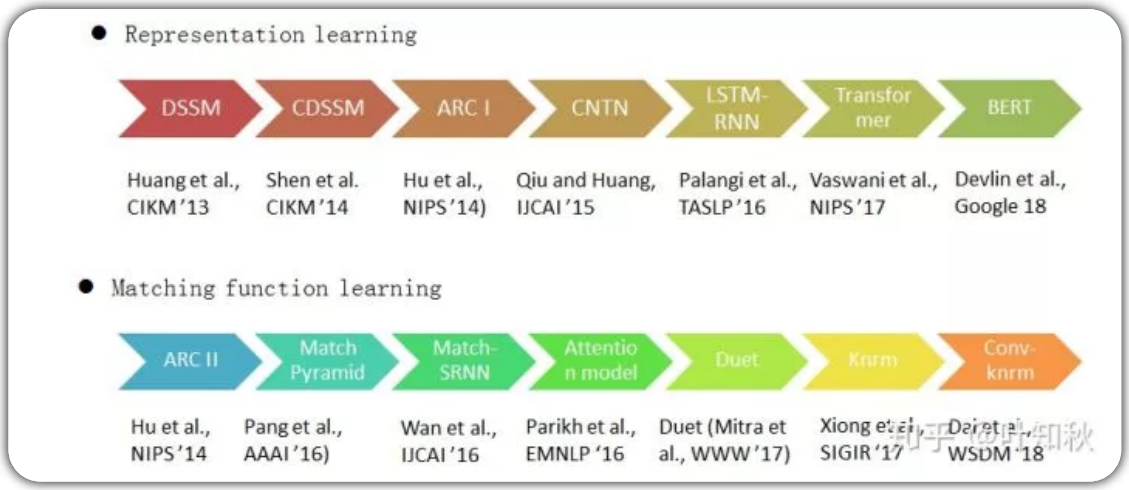

图3 在RL和MFL方向部分论文

接下来我们各自挑选一个比较典型的模型做下简单的介绍，便于我们理解两类模型的思路。

2.2 Representation learning

这里介绍下 CDSSM 模型，这是个很经典的模型，模型结构如图4所示。

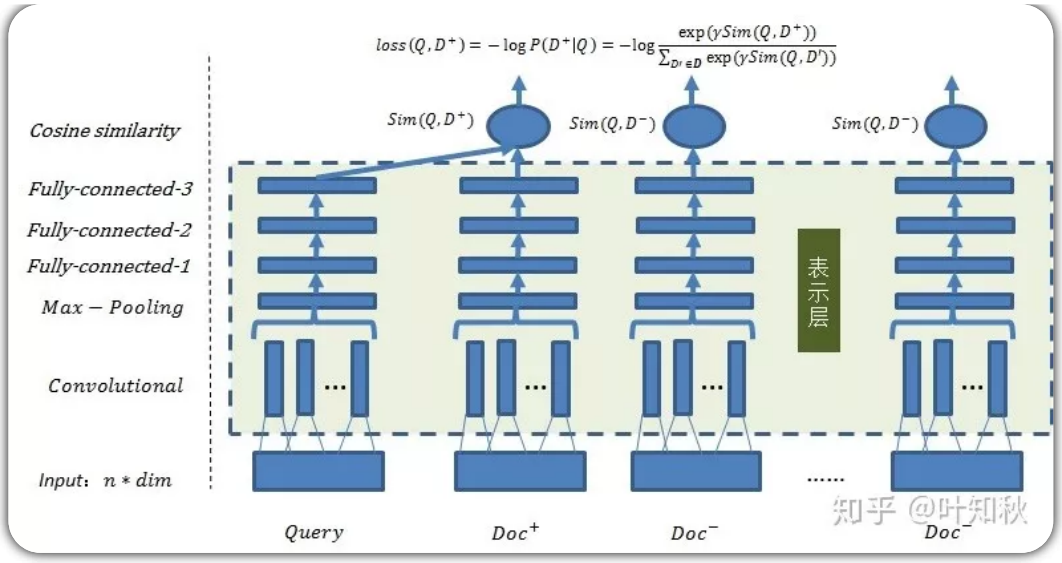

图4 CDSSM 模型

在 CDSSM 中，一个训练样本，由一个 query 、一个正样本 doc 和 n 个负样本 doc 组成，表示层的参数是共享的，网络结构是卷积 + max 池化 + 3个全连接层实现，sim 值计算只用 cosine 距离，loss 函数如图中所示。

可以看到，CDSSM 模型是一个典型的“双塔”结构，即 query 端和 doc 端的计算是独立的。DSSM 是 CDSSM 的简化版，没有加卷积层；而 ARC I、CNTN 也是用了卷积+全连接的基础结构，在目标函数细节上有优化。LSTM-RNN 从名字就知道，是用 lstm 替换了卷积层。

在论文中，CDSSM 模型的参数是共享的，但实际试验中，其实你可以选择 query 端和 doc 端不共享，因为 query 和 doc 文本的分布并不相同，在实际检索场景中，query 相对较短，doc 较长；更进一步地，在原始论文中，RL 模型都是用于学习 query 和 doc 文本的相似度，即文本序列的相似度。

其实我们将 doc 端的模型结构替换，比如替换成 vgg 、resnet 等（可以使用开源模型而不必 fine tune），在此基础上简单增加几个全连接层+激活层等，doc 端的输入变成图像特征，这样就可以计算 query 和 doc 端图像特征的相似性，同样可以得到不错的效果。图5是我们通过这种方式得到的 query 和图像相关性的 demo 结果。

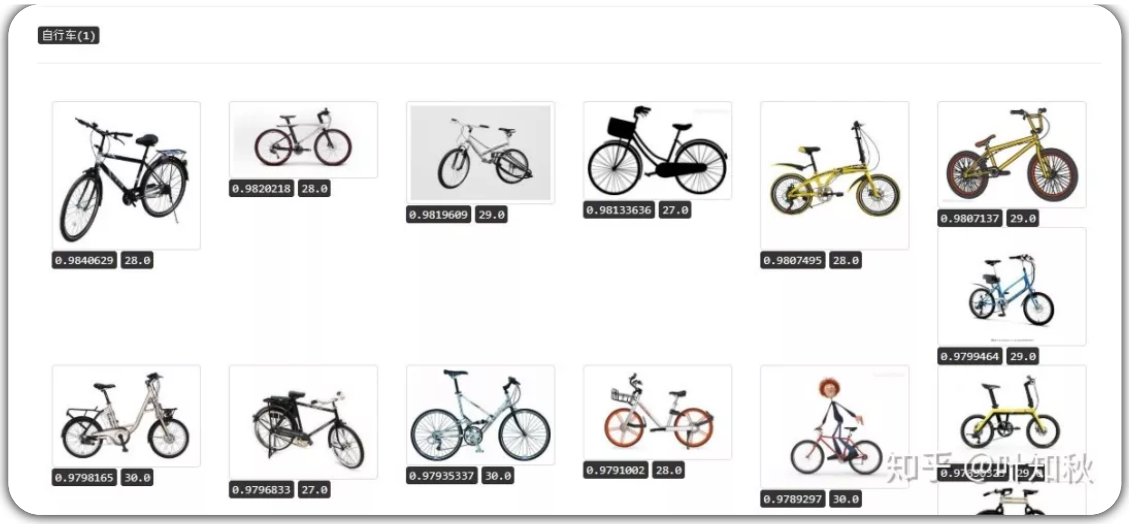

图5 query 和 doc 图像特征相关性示例：自行车

因此可见，RL 模型只是一个大框架，在这个框架的基础上，解决的是，如何更好的将 query 和 doc 特征映射到同一个潜在空间，不同的训练数据、特征表示，可以得到不同的相关性特征。

2.3 Matching function learning

MFL 的模型在 query 和 doc 的交互信息表示上，方式不尽相同，但整体思路还是一致的。这部分我们介绍下 k-nrm 模型。模型的基本模型如图6所示。

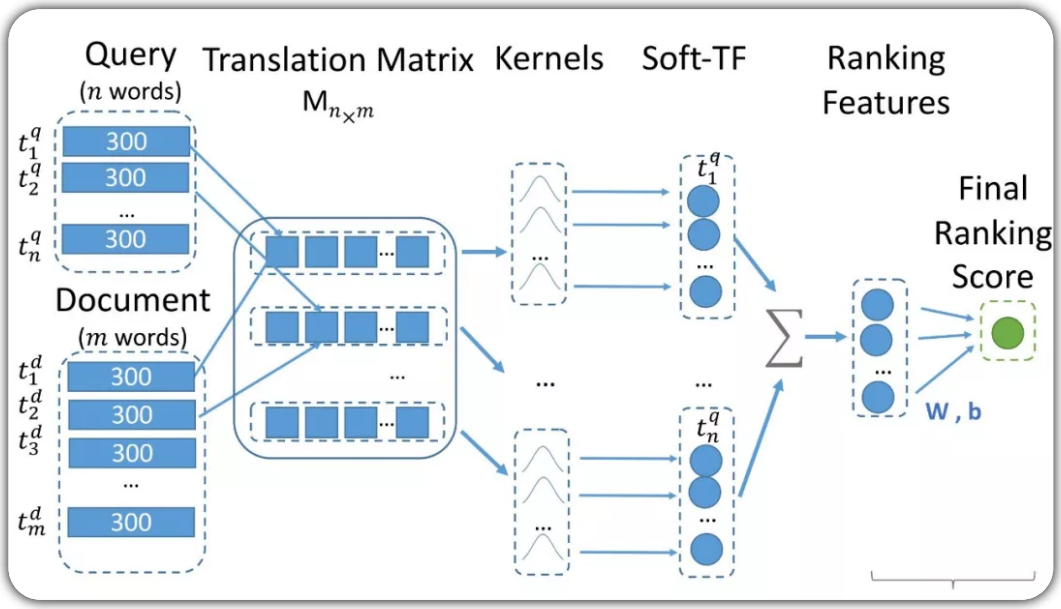

图6 K-NRM 模型结构

K-NRM 是一个基于 kernel 的神经网络排序模型，根据 query 和 doc 的向量特征（vector）生成一个平移矩阵（translation matrix），使用 kernel 提取出泛化的匹配特征（soft match feature），然后使用 LTR 模型来计算 score 。该模型是 end-to-end ， loss 函数使用的是 pairwise loss 。

此外，作者根据用户点击行为对 word2vec 做了 tune（click2vec），效果很明显。这里多说两句，笔者比对过 word2vec 和 click2vec，click2vec 表现出来一个很好的特质是更倾向于子类别下的相似词，举个例子就很好理解，在 w2v 中“北京”最相近的词一般是“上海”、“南京”这些城市名，但是在 click2vec 中，最相近的词会有“海淀”、“昌平”这类词；再比如“华为”在 w2v 中相近词是“摩托罗拉”、“爱立信”这些，但在是 click2vec 中，相近词是“mate”、“p10plus”等。

具体哪种词向量更好，针对不同的任务需要具体尝试，况且现在都习惯对词向量做 fine tune 。但就检索任务来说，click2vec 似乎更符合直观理解，毕竟用户检索“北京”出个海淀的图片没问题，出个南京的图片就说不过去了。这块也是挺有意思的一个任务，以后有机会再单独总结。

模型部分其实是比较简单的，需要注意两点，①是平移矩阵到最终匹配特征的过程；②是高斯 kernel 的使用，如图7所示。高斯 kernel 的使用有点类似卷积，可以理解为 filter size 变化的卷积核。论文中还使用 max、mean 等方式来做 pooling 。

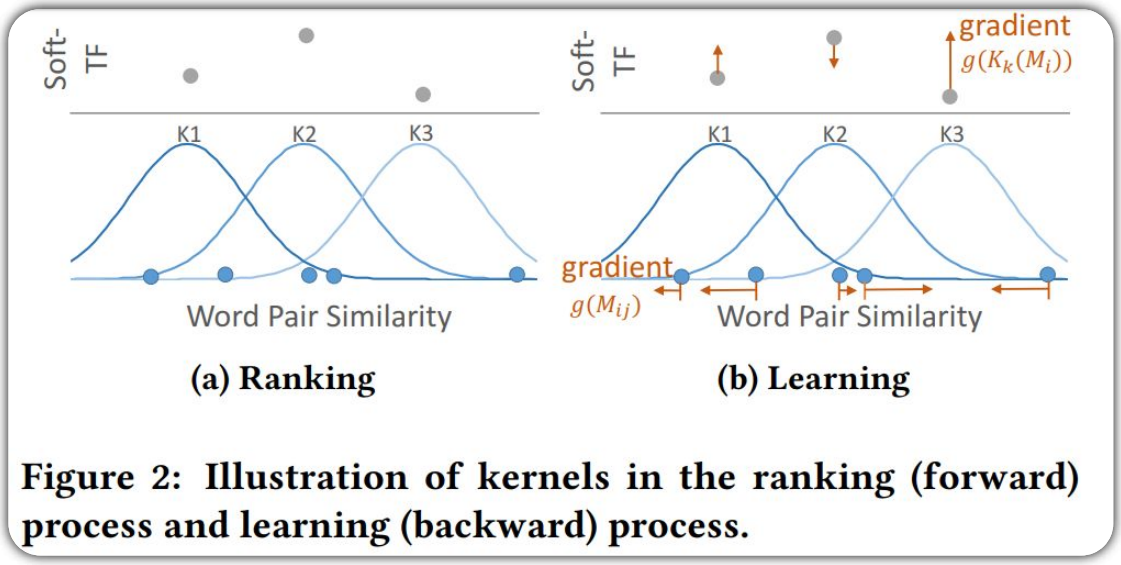

图7 高斯 kernel

看完高斯 kernel 我们梳理下整体流程，图6表示得挺清晰，这里说明下数据流的 shape 变化。query 是 n*300，doc 是 m*300，translation matrix 是 n*m，经过 kernel pooling 后变成 n*k（每行的 1*m 经过一个 kernel 变成一个数值），然后累加得到 k 维的 Soft-TF features 。具体公式论文中很详细，可参看论文：

http://www.cs.cmu.edu/~zhuyund/papers/end-end-neural.pdf

这时候我们再看图3中下半部分的论文模型，从 ARC II 到 Conv-KNRM，都是要通过 query 和 doc 的交互来计算最终的相似度，不同之处在于交互的方式有变化，但整体思路是一样的。

2.4 小结

对比 Representation learning 和 Matching function learning 两类方法，从直观上理解，一般认为后者的思路更好，因为 query 和 doc 的交互发生的初始阶段，这样可以尽量多的保留二者之间的相互作用信息。在论文的一些指标表现上确实如此，但是在工业的使用方法上大都倾向于使用前者。

原因是 RL 类的模型更“独立”，即 doc 端的特征在 inference 阶段是不依赖于检索 query 的，对于搜索引擎而言，这种“独立性”大大提高了计算效率。当然，MFL 不是不能用，可以再粗排序的基础上，只对top部分重排的时候使用，这样就能避免响应时间过高。

针对上面罗列的算法，大部分都有开源的代码，这里要感谢 NTMC 的成员们，在 GitHub 上的开源项目 MatchZoo：

https://github.com/NTMC-Community/MatchZoo

对相关模型进行了实验效果比对，如图8所示。这个项目没有的模型，在 GitHub 上大都也有开源代码。

Models	NDCG@3	NDCG@5	MAP
DSSM	0.3412	0.4179	0.3840
CDSSM	0.5489	0.6084	0.5593
ARC-I	0.5680	0.6317	0.5870
ARC-II	0.5647	0.6176	0.5845
MV-LSTM	0.5818	0.6452	0.5988
DRMM	0.6107	0.6621	0.6195
aNMM	0.6160	0.6696	0.6297
DUET	0.6065	0.6722	0.6301
MatchPyramid	0.6317	0.6913	0.6434
DRMM_TKS	0.6458	0.6956	0.6588

图8 MatchZoo实验结果

三、Attention 机制在图片搜索中应用

第二部分介绍的两类方法，大致就是计算 query 和 doc 相关性的两类思路，这些方法都是计算 query 和 doc 单一相关性，即 query 和 doc 文本、query 和 doc 图像等。得到这些基础的相关性特征后，然后再使用 ltr 模型（如 lr\svmrnak）来计算最终的排序分值。

这里有个问题，就是 ltr 模型是一个和检索 query 无关的模型。

拿 lr 模型来说，比如有10个基础相关性特征，经过训练之后，lr 模型就有10个固定的权重。稍加思考就知道，对于不同 query 权重应该是变化的，比如“5月伤感图片”、“老虎简笔画图片”这两个 query，前者应该更倾向于语义特征，因为很难定义什么样的图像叫伤感图像，但后者应该更倾向于图像特征，至少该是个简笔画图片。

后来看到有研究使用 Attention 机制来解决这个问题的，感觉是个很好的思路。大体想法是，分别计算 query 和 <doc文本，doc图像> 整体的相关性，然后根据 query 和 doc 的本身特征，学到两种相关性的权重。

AMC 算法 - Attention guided Multi-modal Correlation，地址：

https://arxiv.org/pdf/1704.00763.pdf

是我看到比较有代表性的，以此为例做介绍。

3.1 AMC 模型介绍

针对不同的模态相关性（这里就 query 和图片为例），AMC 模型根据内部（intra）和交互（inter）网络来学习两个内容：① doc 表示的重点；② 如何平衡各个模态的重要程度，其中 intra 和 query 无关，主要是为了找到 doc 中文本和图像中最有信息的部分；inter 是通过 query 意图等信息，来平衡不同模态相关性的重要性。

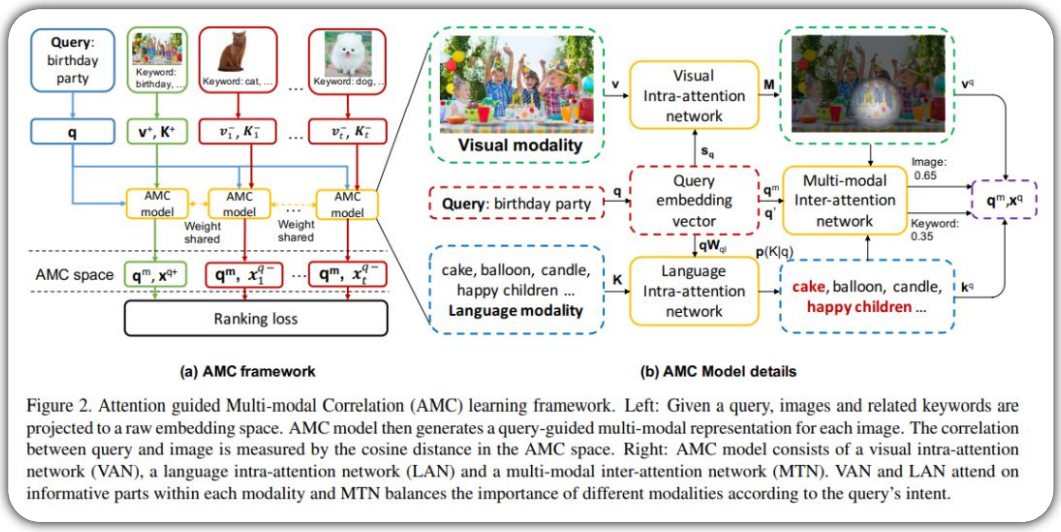

Figure 2. Attention guided Multi-modal Correlation (AMC) learning framework. Left: Given a query, images and related keywords are projected to a raw embedding space. AMC model then generates a query-guided multi-modal representation for each image. The correlation between query and image is measured by the cosine distance in the AMC space. Right: AMC model consists of a visual intra-attention network (VAN), a language intra-attention network (LAN) and a multi-modal inter-attention network (MTN). VAN and LAN attend on informative parts within each modality and MTN balances the importance of different modalities according to the query's intent.

图9 AMC 模型框架图

根据图9所示，我们可以看到 AMC 模型，其实包含了三个部分 VAN、LAN 和 MTN，VAN 是 doc 图像特征的子模型，LAN 是 doc 文本特征的子模型，而 MTN 是通过 query 向量表示和 v、k 作用计算权重，计算出 doc 的 attention 向量 X。训练数据使用点击，构造正、负样本。大致公式如图10所示，具体见论文。

MTN	VAN	LAN
$\mathrm{q}^m = f(\mathrm{W}_{qm}\mathrm{q} + \mathrm{b}_{qm})$ (2) $[c_v, c_k] = \langle \mathrm{q}', [\mathrm{v}^q, \mathrm{k}^q] \rangle, \ \mathrm{q}' = f(\mathrm{W}'_{qm}\mathrm{q} + \mathrm{b}'_{qm})$ (3) $\mathrm{x}^q = p_v \mathrm{v}^q + p_k \mathrm{k}^q, \ [p_v, p_k] = \sigma([c_v, c_k])$ (4) 解释： q'表示query的意图；<.,.>表示cosine距离；f表示非线性的激活函数；	$\mathrm{M} = \sigma(\mathrm{s}_q * \mathrm{v}'), \ \mathrm{s}_q = f(\mathrm{W}_{qs}\mathrm{q} + \mathrm{b}_{qs})$ (5) $\mathrm{v}^q = \mathrm{AvgPool}(\mathrm{M} \odot \mathrm{v}')$ (6) 解释： v'表示图像映射特征；*表示卷积操作；q是query特征；⊙表示对应元素点乘；	$s(q, K, \mathrm{W}_{ql}, \mathrm{W}_{kl}, \mathrm{W}_l) = (\mathrm{q}\mathrm{W}_{ql})\mathrm{W}_l(\mathrm{K}\mathrm{W}_{kl})^\top$ (7) $\mathrm{k}^q = p(K\|q)^\top \mathrm{K}\mathrm{W}_{kl}, \ p(K\|q) = \sigma(s(q, K))$ (8) 解释： 公式（7）计算query（q）和关键词（k）的相关性，3个W分别映射到d维空间，然后根据中间W计算similarity；

图10 AMC 三个子模型的公式推导

作者还给出了示例，根据图11，我们可以很清晰的看到 AMC 模型的结果。针对不同的 query 和 doc ，图像特征（visual）和文本特征（language）的权重是不同的。

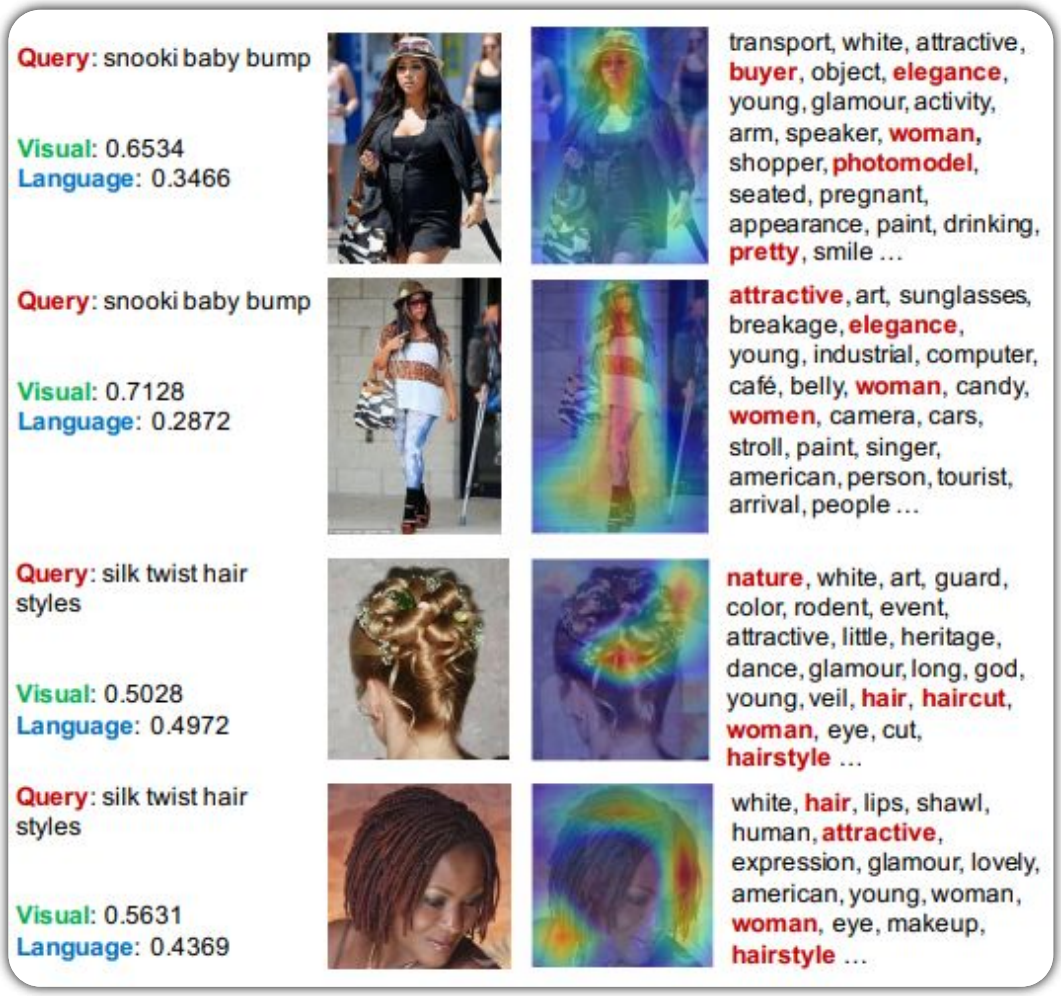

图11 AMC 模型应用实例

3.2 AMC 一点想法

这篇论文证明了使用 query 给图像、文本特征加入 attention 的可行性，最重要的是证明了 query 对于文图、文文特征，动态调整重要程度对结果提升明显，但是，工业上直接照抄恐怕难以复制，因为 query 和 doc 的交互较多，对搜索这种对响应时间要求高的场景，资源要求略高，当然借鉴这种思想；

另外，该模型中，动态调整权重的方案是根据 query 和 doc 文本/图像特征的 sim 作为权重，在物理上不太好解释。当然，模型训练的 fine tune 后计算出的不一定是 sim 值，但直觉上，query 对文本/图像权重调整应该只和 query 本身相关即可。我们单纯用 query 做过一些实验，效果并不好，当然也可能是相关性特征本身不平衡的问题，这块后续会继续验证。

四、总结

本文主要介绍了图片搜索（以文搜图）场景下的常用方法，有些在工业上已有很成熟的应用，有些尚待验证，尤其是近两年的各种文本领域的深度模型，如何在文本表示更好的基础上，应用到图文相关性问题上，还需要很多实验和尝试。

针对文中的各种问题，欢迎大家指正、交流。

作者介绍：

沈炎军，现就职于搜狗搜索 - AI 研究部，副研究员，工作内容为图片搜索中排序算法、相关性特征挖掘，研究方向包括信息检索、NLP 等。

——END——

「 更多干货，更多收获 」

推荐技术随谈（附交流视频和下载链接）

一文带你看懂智能推荐系统原理

推荐系统的十二大评价指标总结

如何将知识图谱特征学习应用到推荐系统？

今日头条推荐系统原理

深度学习与推荐系统完结篇（知识、论文、源码、数据集与行业应用）

智能推荐之：什么是A/B测试（定义、步骤、应用场景及作用）

个性化推荐研究人点之用户画像

京东购物在微信等场景下的智能推荐算法应用与实践

feed流设计：那些谋杀你时间的APP

【推荐算法】基于关联规则的推荐算法及业务实践