融合两级相似度的跨媒体图像文本检索

李志欣[1],凌 锋[1],张灿龙[1],马慧芳[2]
(1. 广西师范大学广西多源信息挖掘与安全重点实验室,广西桂林 541004;
2. 西北师范大学计算机科学与工程学院,甘肃兰州 730070)

摘 要: 为了更好地揭示图像和文本之间潜在的语义关联,提出了一种融合两级相似度的跨媒体检索方法,构建两个子网分别处理全局特征和局部特征,以获取图像和文本之间更好的语义匹配. 图像分为整幅图像和一些图像区域两种表示,文本也分为整个语句和一些单词两种表示. 设计一个两级对齐方法分别匹配图像和文本的全局和局部表示,并融合两种相似度学习跨媒体的完整表示. 在 MSCOCO 和 Flickr30K 数据集上的实验结果表明,本文方法能够使图像和文本的语义匹配更准确,优于许多当前先进的跨媒体检索方法.

关键词: 卷积神经网络; 自注意力网络; 两级相似度; 跨媒体检索
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2021)02-0268-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20191037

Cross-Media Image-Text Retrieval with Two Level Similarity

LI Zhi-xin[1], LING Feng[1], ZHANG Can-long[1], MA Hui-fang[2]
(1. *Guangxi Key Laboratory of Multi-Source Information Mining and Security, Guangxi Normal University, Guilin, Guangxi* 541004, *China*;
2. *College of Computer Science and Engineering, Northwest Normal University, Lanzhou, Gansu* 730070, *China*)

Abstract: To better reveal the latent semantic correlation between image and text, this paper proposes a cross media retrieval method by fusing two level similarity, which constructs two subnets to deal with global features and local features respectively so as to obtain better semantic matching between image and text. The image representation is divided into whole image and some image regions, and the text representation is also divided into whole sentence and some words. A two level alignment method is designed to match the global and local representation of image and text, and the two similarities are fused to learn the complete cross-media representation. The experimental results on MSCOCO and Flickr30K datasets show that the proposed method can make the semantic matching of image and text more accurate, and is superior to many state-of-the-art cross-media retrieval methods.

Key words: convolutional neural network; self-attention network; two level similarity; cross-media retrieval

1 引言

跨媒体既表现为各种文本、图像、音频、视频等复杂媒体对象的混合并存,又表现为各类媒体对象形成的复杂关联关系和组织结构. 跨媒体检索的主要挑战在于不同媒体类型具有异构的形式,相同的内容信息可能涵盖在不同种类的多媒体数据. 只有对这些多媒体数据进行融合分析等操作,才能够全面、正确地理解这些跨模态综合体之间所蕴涵的内容信息.

由于不同模态的特征通常具有不一致的分布和表示,因此需要缩减模态间的语义鸿沟. 基本的学习方法是建立一个公共子空间,然后将所有的数据投影到该空间中进行学习. 然而,对于某种模态中的实例,可能存在不同模态的多个近似语义实例相匹配的情况,所以仅通过公共子空间匹配文本和图像还不够. 为此,本文提出了一种基于两级网络的跨媒体图像文本检索方法 GRLR(Global Representation and Local Representation),能够融合全局和局部相似度进行更精确的检索,其创新之处主要体现在三个方面:

(1) 提出一个跨媒体两级模型,结合各种模块完成

收稿日期:2019-09-10;修回日期:2020-09-20;责任编辑:李勇锋
基金项目:国家自然科学基金(No. 61663004, No. 61966004, No. 61866004, No. 61762078);广西自然科学基金(No. 2019GXNSFDA245018, No. 2018GXNSFDA281009, No. 2017GXNSFAA198365)

检索任务;

(2)设计了全局和局部两种损失函数,以获取图像和文本的全局和局部特征;

(3) 基于图像和文本的特征表示,设计了两级相似度并进行综合累加,在一定程度上实现信息互补,从而在跨媒体检索中得到更好的效果.

2　相关工作

目前针对图像-文本匹配做了大量的研究. Kiros 等[1]利用卷积神经网络(Convolutional Neural Networks, CNN)和循环神经网络(Recurrent Neural Network,RNN)分别对图像和文本进行编码,在实验结果上取得了显著提升. Feng 等[2]通过编码器将图像和文本特征分别映射到一个公共子空间,并使用 L2 范数度量图像和文本之间的相似度. Wang 等[3,4]和 Vendrov 等[5]引入文本之间的三元组损失并加入了结构保留. Nam 等[6]为图像和文本两种模态都加入了注意力,并在公共子空间使用三元组损失度量相似度. Faghri 等[7]利用了三元组损失的硬负效应,实现了端到端的训练. Huang 等[8]通过图像和文本之间的排名损失和文本生成的损失共同学习得到一个较好的图像表示. Lee 等[9]分别在图像和文本两端利用注意力机制,获得更好的潜在语义对齐.

GRLR 方法结合了全局和局部两种相似度,更全面地描述了图像和文本的语义,可以更好地探索图像和文本之间的潜在语义对齐. 一方面,将不同模态的数据从各自独立表示的空间映射到一个公共子空间,可以计算不同模态实例的相似度,从而减小不同模态数据相同语义特征之间的距离;另一方面,受到基于排名的方法[7~9]启发,使用了基于三元组的损失函数,可以加大不同模态涵盖不同语义特征之间的距离,缩小不同模态涵盖相同语义特征之间的距离,最终得到图像和文本更好的匹配.

3　跨媒体两级模型

本文构建的跨媒体两级模型结构如图 1 所示,包含用于提取全局表示和局部表示的两个子网. 本模型不仅利用自注意力网络获取图像的全局宏观表示,还利用注意力机制获取文本的局部表示,并提出了两级相似度的融合方法用于相互提升,使得跨媒体关联学习实现信息互补.

3.1　全局表示处理

3.1.1　图像的全局表示

将每幅输入图像 i_m 都调整为 256×256 的大小,送到 CNN 以利用其高维信息. 采用的网络与 ResNet-152[10]具有相同的配置,在大规模数据集 ImageNet 上预先训练. 将最后一个图像特征进行均值池化,即获得图像全局特征 $\boldsymbol{x}$. 随后,将图像全局特征通过自注意力网络[11]进行训练,如图 2 所示.

首先,根据前面得到的图像全局特征 $\boldsymbol{x}$,可计算 $f(\boldsymbol{x})=\boldsymbol{W}_f\boldsymbol{x}$ 和 $y(\boldsymbol{x})=\boldsymbol{W}_y\boldsymbol{x}$,分别表示图像特征乘上不同权重矩阵 $\boldsymbol{W}_f$ 和 $\boldsymbol{W}_y$ 后得到的两个特征空间. 利用 softmax 函数计算相关性 $\alpha_{j,i}$,表示模型 j 区域的图像内容与 i 区域的相关程度,即:

$$\alpha_{j,i}=\frac{\exp(b_{ij})}{\sum_{i=1}^{n}\exp(b_{ij})} \tag{1}$$

其中:$b_{ij}=f(\boldsymbol{x}_i)^{\mathrm{T}}y(\boldsymbol{x}_j)$.

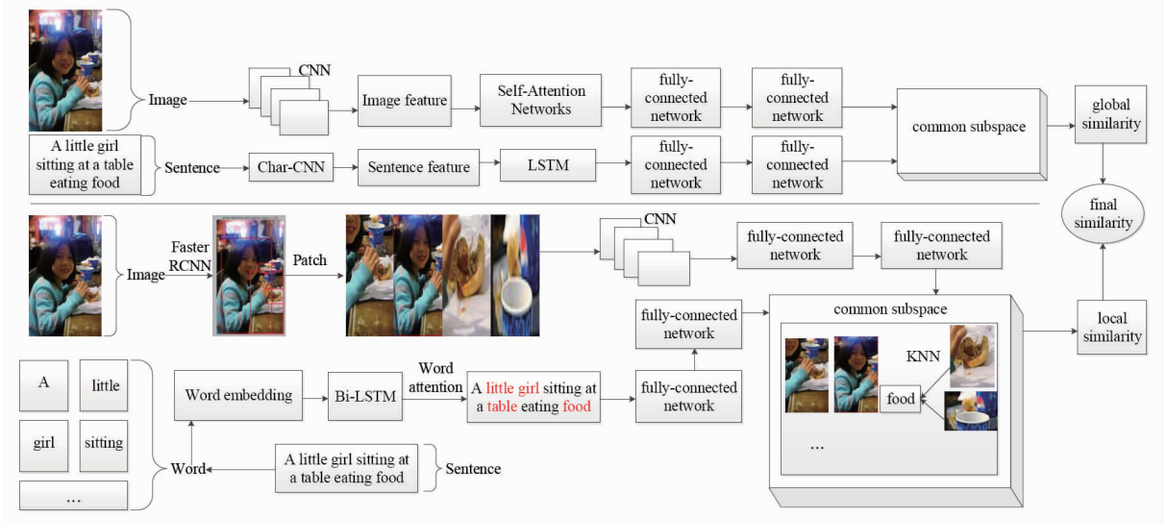


图1　跨媒体两级模型结构图

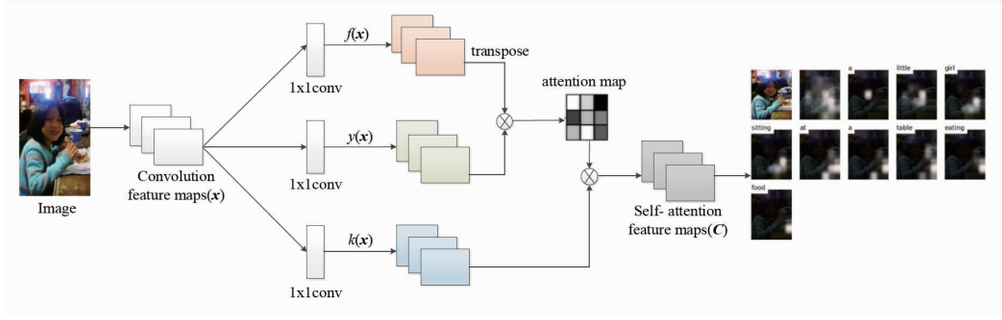


图2 自注意力网络的结构图

其次,计算自注意力网络的输出 $\boldsymbol{c}_j$,即:

$$\boldsymbol{c}_j = \sum_{i=1}^{N} \alpha_{j,i} k(\boldsymbol{x}_i) \tag{2}$$

其中:$k(\boldsymbol{x}_i) = \boldsymbol{W}_k \boldsymbol{x}_i$,$\boldsymbol{W}_k$ 是权重参数矩阵. $\boldsymbol{C} = (\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_j, \cdots, \boldsymbol{c}_n)$ 可以把全部信息整合到一起.

最后,结合图像原始特征和注意力层的特征,得到输出 $\boldsymbol{g}_i = \lambda \boldsymbol{c}_i + \boldsymbol{x}_i$ 作为图像的全局表示,这里 λ 的数值设置为 0.1.

3.1.2 文本的全局表示

每个输入文本 t_k 可组成一个字符序列,可构建一个字符卷积网络(称为 Char-CNN)[12] 来处理文本. 从最后一个激活层生成一个表示序列,并将它们送到 RNN 中从字符层面进行文本分类,这样就可提取出高层抽象语义特征,而不需使用预先训练好的词向量和语法句法结构等信息. 每个输入文本 t_k 经过 Char-CNN 输出的文本序列是 P,作为长短期记忆网络(Long Short Term Memory,LSTM)[13] 的输入,隐藏单元的输出为 $\boldsymbol{H}_i = \{\boldsymbol{h}_1^i, \cdots, \boldsymbol{h}_m^i\}$,则可学习得到文本的全局表示 $\boldsymbol{g}_t = \frac{1}{m}\sum_{k=1}^{m} \boldsymbol{h}_k^i$.

Char-CNN 将每个语句都视为字符序列,其长度设置为 300,长度大于 300 的语句被截断,而长度小于 300 的语句则被零填充. Char-CNN 中有三个卷积层,参数组合为(256,4)、(512,4)和(2048,4),括号中参数表示内核的数量和宽度.

3.2 局部表示处理

3.2.1 图像的局部表示

将每幅图像 i_m 通过 Faster R-CNN[14] 处理后可获得若干边界框,即得到所有候选图像区域,并依据评分选取前 5 个进行计算. 将选中的区域送入 ResNet-152 网络进行训练,在最后一个图像特征中进行均值池化以获得这些图像区域的特征[15]. 这些特征代表一个图像内的 n 个不同的区域,即得到图像的局部表示 $\{\boldsymbol{l}_i^1, \cdots, \boldsymbol{l}_i^n\}$,其中 i 代表图像序号.

3.2.2 文本的局部表示

为了学习文本的局部表示,如果只是利用 LSTM 对语句进行建模,就会无法编码从后到前的信息. 在进行更细粒度的分类时,需要注意情感词、程度词、否定词之间的交互,所以需要编码从后到前的信息. 这里使用双向 LSTM(即 Bi-LSTM)[16] 来捕捉文本双向的语义依赖.

对于某一个语句 $Y = \{y_1, y_2, \cdots, y_i, \cdots, y_m\}$ 中的第 i 个单词,显示词汇表中对单词的检索,并通过词嵌入矩阵 $\boldsymbol{W}_E$ 表示为:

$$\boldsymbol{W}_E \cdot y_i = \boldsymbol{W}_E \omega_i,\ i \in [1, m] \tag{3}$$

这里将单词嵌入到 300 维向量中,并使用双向 LSTM 通过总结某一个语句中两个方向的单词信息. 双向 LSTM 包含两个组件:前向 LSTM 从 ω_1 到 ω_n 方向读取语句 Y;后向 LSTM 从 ω_n 到 ω_1 方向读取语句 Y. 即:

$$\overrightarrow{\boldsymbol{h}_i} = \overrightarrow{LSTM}(y_i),\ i \in [1, m] \tag{4}$$

$$\overleftarrow{\boldsymbol{h}_i} = \overleftarrow{LSTM}(y_i),\ i \in [1, m] \tag{5}$$

最后一个词的特征 $\boldsymbol{e}_m$ 是通过平均前向隐藏状态 $\overrightarrow{\boldsymbol{h}_i}$ 和后向隐藏状态 $\overleftarrow{\boldsymbol{h}_i}$ 来定义,总结了以 ω_i 为中心的语句的信息,即:

$$\boldsymbol{e}_m = \frac{(\overrightarrow{\boldsymbol{h}_i} + \overleftarrow{\boldsymbol{h}_i})}{2},\ i \in [1, m] \tag{6}$$

这里词嵌入提取单词的输出维数是 2048. 将词嵌入的输出作为 Bi-LSTM 的输入,可以获得隐藏单元的输出,表示为 $\boldsymbol{E} = \{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_i, \cdots, \boldsymbol{e}_m\}$. 这是某一个语句中 m 个不同的文字片段,可作为解释语句上下文的最终特征.

此外,文本局部表示的目标是使模型专注于一些重要的图像区域,因此可使用注意力机制[17] 来捕获有用的文本片段. 每个集合中的元素代表在输入信息中某个空间位置上的输入信息,输出 e_m^t 就是当前空间位置 t 下,某个上下文表示对应的注意力. 那么,第 m 个文字片段的得分为:

$$e_m^t = f_{ATT}(z_{t-1}, \boldsymbol{e}_m, \{\alpha_j^{t-1}\}_{j=1}^{m}) \tag{7}$$

其中 z_{t-1} 是第 $t-1$ 个空间位置下 LSTM 的隐状态的输入. 通过 softmax 进行归一化后,每一个输入的上下文表示对应的权重和为 1,文本片段生成的注意权重(即得分)为:

$$\alpha_m^t = \frac{\exp(e_m^t)}{\sum_{j=1}^{m}\exp(e_j^t)} \tag{8}$$

具有较大注意力的文字片段更可能包含一些描述相应视觉对象的关键词. 因此,经过 Bi-LSTM 和注意力机制处理后,可以获得某一个语句中的局部表示 $\frac{1}{m}\sum_{k=1}^{m}\alpha_k^t \boldsymbol{e}_k$.

假设有 n 个文本,那么可以从 Bi-LSTM 的隐藏单元获得输出,表示为 $\boldsymbol{E}' = \{\boldsymbol{e}_1^n, \cdots \boldsymbol{e}_i^n, \cdots, \boldsymbol{e}_m^n\}$,代表 n 个语句中,每个语句有 m 个不同的文字片段. 经过双向 LSTM 和注意力机制处理后,可以获得 n 个语句中的文本局部表示 $\boldsymbol{l}_t = \frac{1}{m}\sum_{k=1}^{m}\sum_{j=1}^{n}\alpha_k^t \boldsymbol{e}_k^j$,这可作为最终的文本局部表示.

在全局和局部表示处理网络的最后都添加了两个全连接网络,将图像和文本的特征向量维度变换为 1024. 这两个网络作为跨媒体语义对齐部件将异构特征映射到公共子空间中.

3.3　跨媒体两级对齐

全局和局部表示基于三元组损失函数[18]进行计算,其核心是锚示例、正示例、负示例共享模型. 通过这个模型,可以将锚示例与正示例尽可能的聚集在一起,并远离负示例. 三元组损失函数表示为 $Loss_{\text{triplet}} = \max(d(a,p) - d(a,n) + \text{margin}, 0)$,其中 a 是锚示例,p 是正示例,n 是负示例.

基于三元组损失,设计的目标函数定义如下:

$$\begin{aligned} L_{\text{global}} &= \frac{1}{N}\sum_{n=1}^{N} L_1(i_+^n, t_+^n, t_-^n) + L_2(t_+^n, i_+^n, i_-^n), \\ L_1(i_+^n, t_+^n, t_-^n) &= \max(0, \alpha - d(\boldsymbol{g}_{i+}^n, \boldsymbol{g}_{t+}^n) + d(\boldsymbol{g}_{i+}^n, \boldsymbol{g}_{t-}^n)), \\ L_2(t_+^n, i_+^n, i_-^n) &= \max(0, \alpha - d(\boldsymbol{g}_{i+}^n, \boldsymbol{g}_{t+}^n) + d(\boldsymbol{g}_{i-}^n, \boldsymbol{g}_{t+}^n)) \end{aligned} \tag{9}$$

其中 L_1 和 L_2 表示模型训练时匹配的全局图像-文本对的相似度,与不匹配的图像-文本对的相似度之间的差异要尽可能大. $d(\cdot)$ 表示图像-文本对之间的点积,即它们的相似度. $(\boldsymbol{g}_{i+}^n, \boldsymbol{g}_{t+}^n)$ 表示匹配的图像-文本对,而 $(\boldsymbol{g}_{i+}^n, \boldsymbol{g}_{t-}^n)$ 和 $(\boldsymbol{g}_{i-}^n, \boldsymbol{g}_{t+}^n)$ 表示不匹配的图像-文本对. n 表示图像-文本对的数量,α 表示边际参数,N 是从训练集中采样的三元组的数量.

局部对齐的目标是在一对图像和文本中找到文本局部表示 $\boldsymbol{l}_t$ 与多个图像局部表示 $\{\boldsymbol{l}_i^1, \cdots, \boldsymbol{l}_i^n\}$ 之间的最佳匹配. 也就是针对每个文本局部表示从多个图像局部表示中选择 K 个最近邻,这里 K 值设定为 3 时能够更好地匹配图像和文本的局部表示. 目标函数如下:

$$L_{\text{local}} = \max\left(0, \alpha - \frac{1}{K}\sum_{k=1}^{K} d(\boldsymbol{l}_{t+}, \boldsymbol{l}_{i+}^k) + \frac{1}{K}\sum_{k=1}^{K} d(\boldsymbol{l}_{t+}, \boldsymbol{l}_{i-}^k)\right) \tag{10}$$

最后,设计了图像 i_m 和文本 t_k 之间的跨媒体综合相似度,它结合了全局和局部两种相似度,在 1024 维公共子空间中计算:

$$\begin{aligned} sim1 &= d(\boldsymbol{g}_i, \boldsymbol{g}_t), \\ sim2 &= \frac{1}{K}\sum_{k=1}^{K} d(\boldsymbol{l}_i^k, \boldsymbol{l}_t), \\ sim(i_m, t_k) &= \theta \cdot sim1 + (1-\theta) \cdot sim2 \end{aligned} \tag{11}$$

这里输入是图像-文本对转换的全局和局部特征,输出是整体相似度. θ 是 GRLR 定义的学习参数,大小在 0.3 和 0.7 之间.

4　实验结果分析

GRLR 在 MS-COCO 和 Flickr30K 数据集上进行了实验验证,其性能普遍高于当前国际先进方法. 与代表性方法 SCO[8]相比,在 Flickr30K 数据集上(基于 Recall@1),GRLR 基于文本查询图像的性能提高了 2.3%,基于图像查询文本的性能提高了 12.7%. 在 MS-COCO 数据集上(基于 Recall@5),GRLR 基于文本查询图像的性能提高了 0.7%,基于图像查询文本的性能提高了 1.2%.

4.1　数据集和评估指标

Flickr30K 数据集包含 31784 幅图像,每幅图像用 5 个语句注释. 按照 Karpathy 等人[19]的分割方法,测试集和验证集各 1000 幅图像,其余为训练集.

MS-COCO 数据集包含 123287 幅图像,每幅图像有 5 个注释语句. 同样是测试集和验证集各 1000 幅图像,其余用于训练.

实验使用了三个性能评估指标:Recall @ K(K = 1,5,10)的评分表示检索结果取前 K 个实例时所获得的召回率,其值越高模型性能越好. Med R 是第一个检索到的真实语句或图像的中位数,其值越低模型性能越好. sum 用来评估跨模态检索的整体性能,定义为模型中所有评价指标为 R@1 和 R@10 的评分之和,即:

$$sum = \underbrace{R@1 + R@10}_{\text{Text}\rightarrow\text{Image}} + \underbrace{R@1 + R@10}_{\text{Image}\rightarrow\text{Text}} \tag{12}$$

4.2　检索结果定量分析

在 Flickr30K 和 MSCOCO 数据集上的召回率结果分别如表 1 和表 2 所示. 可以看出,GRLR 不仅在处理类似 Flickr30K 的小数据集性能良好,在处理类似 MS-COCO 的大型数据集时性能仍然保持优越性,充分说明了该方法的可扩展性和鲁棒性.

在 Flickr30K 和 MSCOCO 数据集上使用 1000 幅测试图像后,将得到的 Med R 和 sum 值与其他方法进行了比较,结果分别如表 3 和表 4 所示. 表中的数据指标很好地反映了 GRLR 的综合能力.

表 1 在 Flickr30K 数据集上的召回率结果(%)

方法	Text→Image			Image→Text		
	R@1	R@5	R@10	R@1	R@5	R@10
DSPE[3]	29.7	60.1	72.1	40.3	68.9	79.9
VSE ++[7]	39.6	70.1	79.5	52.9	80.5	87.2
Embedding Net[4]	29.2	59.6	71.7	40.7	69.7	79.2
SCO[8]	41.1	70.5	80.1	55.5	82.0	89.3
DAN[6]	39.4	69.2	79.1	55.0	81.8	89.0
DCCA[20]	26.8	52.9	66.9	27.9	56.9	68.2
SM-LSTM[21]	30.2	60.4	72.3	42.5	71.9	81.5
GRLR	43.4	73.5	82.5	68.2	89.1	94.5

表 2 在 MS COCO 数据集上的召回率结果(%)

方法	Text→Image			Image→Text		
	R@1	R@5	R@10	R@1	R@5	R@10
VSE ++[7]	52.0	–	92.0	64.6	–	95.7
SCO[8]	56.7	87.5	94.8	69.9	92.9	97.5
order-embeddings[5]	39.6	75.3	86.7	48.5	80.9	90.3
DCCA[20]	6.6	20.9	32.2	6.9	21.1	31.8
SM-LSTM[21]	40.7	75.8	87.4	53.2	83.1	91.5
GXN[22]	56.6	–	94.5	68.5	–	97.9
HM-LSTM[23]	36.1	–	86.7	43.9	–	87.8
GRLR	58.6	88.2	94.9	68.9	94.1	98.0

表 3 在 Flickr30K 数据集上得到的 Med R 和 sum 值

方法	Text→Image	Image→Text	sum
	Med R	Med R	
VSE ++[7]	**2.0**	**1.0**	259.2
Embedding Net[4]	–	–	220.8
UVS[1]	6.0	4.0	181.6
DAN[6]	**2.0**	**1.0**	262.5
SM-LSTM[21]	3.0	2.0	226.5
GRLR	**2.0**	**1.0**	**288.6**

表 4 在 MS COCO 数据集上得到的 Med R 和 sum 值

方法	Text→Image	Image→Text	sum
	Med R	Med R	
DSPE[3]	2.0	1.0	265.8
VSE ++[7]	1.0	1.0	304.6
order-embeddings[5]	2.0	2.0	260.4
SM-LSTM[21]	2.0	1.0	272.8
GXN[22]	1.0	1.0	317.5
HM-LSTM[23]	3.0	2.0	254.5
mCNN[24]	3.0	2.0	242.3
GRLR	1.0	1.0	320.4

4.3 训练可视化结果

图 3 给出在 MS-COCO 数据集上的本文训练阶段,模型评分 sum 随训练数据量不同的变化趋势图,其中 x 轴表示训练数据量,y 轴表示 sum 的值. 选择 sum 评分的原因在于 sum 通常用来评估跨模态检索的整体性能,sum 的值越大,表明本文方法的性能越好. 可以看到,当训练数据量达到 80K 及以上时,sum 的值趋向于稳定状态. 在数据量大约为 109K 时,sum 达到峰值约为 320.

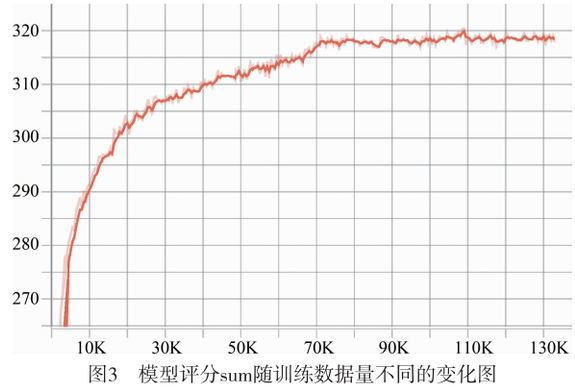
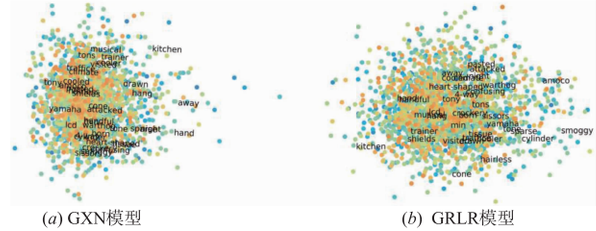


图3 模型评分sum随训练数据量不同的变化图

4.4 词嵌入可视化结果

GRLR 学习了嵌入矩阵 $\boldsymbol{W}_E \cdot x_i = \boldsymbol{W}_E \boldsymbol{\omega}_i, i \in [1, m]$,通过将一些选定的单词向量投影到二维空间来可视化单词嵌入. 图 4 给出 GXN 模型[22]和 GRLR 模型的词嵌入可视化结果,图中不同颜色的点代表不同单词的嵌入. 可以看到,GXN 模型中很多词不能得到很好的区分,而 GRLR 模型可以学习更多的单词嵌入,而且词嵌入的分布更分散. 这说明 GRLR 模型能更好的识别和区分不同的语义信息.



(*a*) GXN模型 (*b*) GRLR模型

图4 两种模型的词嵌入可视化结果对比

5 结束语

本文提出了跨媒体两级模型 GRLR,探讨了图像和文本这两种不同模态之间的匹配方法. 该方法分别使用全局和局部表示来捕获跨媒体相关性学习中的不同信息,提出两个层次的跨媒体对齐,并设计两种相似度的有效结合,可以促进两种模态彼此学习更准确的相关性. 最后,在两个基准数据集上设计了跨媒体检索实

验,并验证了本文方法的有效性和可扩展性.

参考文献

[1] Kiros R, Salakhutdinov R, Zemel R S. Unifying visual-semantic embeddings with multimodal neural language models[OL]. http://arxiv.org/abs/1411.2539,2014-11-10.

[2] Feng F, Wang X, Li R. Cross-modal retrieval with correspondence autoencoder[A]. Proceedings of the 22nd ACM International Conference on Multimedia[C]. New York, USA:ACM,2014.7-16.

[3] Wang L, Li Y, Lazebnik S. Learning deep structure-preserving image-text embeddings[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition [C]. Los Alamitos, USA: IEEE Computer Society, 2016. 5005-5013.

[4] Wang L, Li Y, Huang J, et al. Learning two-branch neural networks for image-text matching tasks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(2):394-407.

[5] Vendrov I, Kiros R, Fidler S, et al. Order-embeddings of images and language[OL]. http://arxiv.org/abs/1511.06361,2015-11-19.

[6] Nam H, Ha J W, Kim J. Dual attention networks for multimodal reasoning and matching[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Los Alamitos, USA: IEEE Computer Society, 2017.299-307.

[7] Faghri F, Fleet D J, Kiros J R, et al. VSE++:Improving visual-semantic embeddings with hard negatives[A]. Proceedings of the 28th British Machine Vision Conference [C]. Durham, UK:BMVA,2017.121-132.

[8] Huang Y, Wu Q, Song C, et al. Learning semantic concepts and order for image and sentence matching[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Los Alamitos, USA:IEEE Computer Society,2018.6163-6171.

[9] Lee K H, Chen X, Hua G, et al. Stacked cross attention for image-text matching[A]. Proceedings of the European Conference on Computer Vision[C]. Cham, Swit:Spriger, 2018.201-216.

[10] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Los Alamitos, USA: IEEE Computer Society, 2016. 770-778.

[11] Zhang H, Goodfellow I, Metaxas D N, et al. Self-attention generative adversarial networks[A]. Proceedings of International Conference on Machine Learning [C]. New York, USA:ACM,2019.7354-7363.

[12] Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification[A]. Advances in Neural Information Processing Systems[C]. Cambridge, UK: MIT Press,2015.649-657.

[13] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural computation,1997,9(8):1735-1780.

[14] Ren S, He K, Girshick R, et al. Faster R-CNN:Towards real-time object detection with region proposal networks [A]. Advances in Neural Information Processing Systems [C]. Cambridge, UK:MIT Press,2015.91-99.

[15] Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Los Alamitos, USA:IEEE Computer Society,2018.6077-6086.

[16] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[OL]. http://arxiv.org/abs/1508.01991,2015-08-09.

[17] Mnih V, Heess N, Graves A. Recurrent models of visual attention[A]. Advances in Neural Information Processing Systems[C]. Cambridge, UK: MIT Press, 2014. 2204-2212.

[18] Hoffer E, Ailon N. Deep metric learning using triplet network[A]. Proceedings of International Workshop on Similarity-Based Pattern Recognition [C]. Cham, Swit: Springer,2015.84-92.

[19] Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Los Alamitos, USA:IEEE Computer Society, 2015.3128-3137.

[20] Yan F, Mikolajczyk K. Deep correlation for matching images and text[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Los Alamitos, USA:IEEE Computer Society,2015.3441-3450.

[21] Huang Y, Wang W, Wang L. Instance-aware image and sentence matching with selective multimodal LSTM[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Los Alamitos, USA: IEEE Computer Society,2017.2310-2318.

[22] Gu J, Cai J, Joty S, et al. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Los Alamitos, USA:IEEE Computer Society,2018.7181-7189.

[23] Niu Z, Zhou M, Wang L, et al. Hierarchical multimodal LSTM for dense visual-semantic embedding[A]. Proceedings of the IEEE International Conference on Computer Vision [C]. Piscataway, USA: IEEE, 2017. 1881

–1889.

[24] Ma L, Lu Z, Shang L, et al. Multimodal convolutional neural networks for matching image and sentence[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Piscataway, USA: IEEE, 2015. 2623–2631.

作者简介

李志欣(通信作者) 男,1971 年 10 月出生,广西桂林人. 现为广西师范大学计算机科学与信息工程学院教授、博士生导师. 研究领域为图像理解、机器学习与跨媒体计算.

E-mail:lizx@gxnu.edu.cn

凌 锋 男,1993 年 11 月出生,广西崇左人. 广西师范大学计算机科学与信息工程学院硕士研究生. 研究方向为机器学习与跨媒体检索.

E-mail:lingfeng93@126.com

张灿龙 男,1975 年 10 月出生,湖南娄底人. 现为广西师范大学计算机科学与信息工程学院教授. 研究领域为目标跟踪与模式识别.

马慧芳 女,1981 年 7 月出生,甘肃兰州人. 现为西北师范大学计算机科学与工程学院教授. 研究领域为数据挖掘与机器学习.