计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(12):1−8 [doi: 10.15888/j.cnki.csa.006130]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于语义蕴含关系的图片语句匹配模型①

柯　川, 李文波, 汪美玲, 李　孜

(中国科学院 软件研究所, 北京 100190)

摘　要: 本文提出一种基于蕴含关系的图片语句匹配模型 IRMatch, 旨在解决图片语句两种不同模态语义之间的非对等匹配问题. 在利用卷积神经网络分别对图片和语句进行语义映射的基础上, IRMatch 模型通过引入最大软间隔的学习策略挖掘图片与语句之间的蕴含关系, 以强化相关图片语句对在公共语义空间中位置的邻近性, 改善图片语句匹配得分的合理性. 基于 IRMatch 模型, 本文实现一种图文双向检索方法, 并在 Flickr8k、Flickr30k 以及 Microsoft COCO 数据集上与基于已有图片语句匹配模型的图文双向检索方法进行了比较. 实验结果表明, 基于 IRMatch 模型的检索方法在上述三个数据集上的 R@1, R@5, R@10 以及 Med r 均优于基于已有模型的检索方法.

关键词: 图文非对等匹配; 蕴含关系; 最大间隔学习; 图文双向检索; 卷积神经网络

引用格式: 柯川,李文波,汪美玲,李孜.基于语义蕴含关系的图片语句匹配模型.计算机系统应用,2017,26(12):1–8. http://www.c-s-a.org.cn/1003-3254/6130.html

Image Sentence Matching Model Based on Semantic Implication Relation

KE Chuan, LI Wen-Bo, WANG Mei-Ling, LI Zi

(Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: In this paper, we propose a model called IRMatch for matching images and sentences based on implication relation to solve the nonequivalent semantics matching problem between images and sentences. The IRMatch model first maps images and sentences to a common semantic space respectively by using convolutional neural networks, and then mines implication relations between images and sentences with a learning algorithm by introducing maximum soft margin strategies, which strengthens the proximity of locations of related images and sentences in the common semantic space and improves the reasonability of matching scores between images and sentences. Based on the IRMatch model, we realize approaches of bidirectional image and sentence retrieval, and compare them with approaches using existing models for matching images and sentences on datasets Flickr8k, Flickr30k and Microsoft COCO. Experimental results show that our retrieval approaches perform better in terms of R@1, R@5, R@10 and Med r on the three datasets.

Key words: nonequivalent match between images and sentences; implication relation; maximum margin learning; bidirectional image and sentence retrieval; convolutional neural network

1　引言

图片和自然语言语句 (以下简称语句) 的关联在图片字幕生成、图片检索等图片相关应用中扮演着不可或缺的角色[1-4]. 图片和语句关联的关键是在图片与语句之间建立合理的匹配, 其实质为一个多模态匹配问题, 具体来说语义相关的图片-语句对的匹配得分应该高于语义不相关的图片-语句对的匹配得分.

目前已有的图片-语句匹配方法主要有两大类, 一类是将图片和语句映射到一个公共的语义空间, 然后进行两者之间的匹配; 另一类是采用诸如典型相关分析 (Canonical correlation analysis, CCA)[5,6]、深度学习[1]等方式来建立图片和语句之间的关联. 在已有的这些

① 基金项目: 国家“863”项目 (2013AA01A603)
收稿时间: 2017-03-22; 修改时间: 2017-04-13; 采用时间: 2017-04-24

© 中国科学院软件研究所 http://www.c-s-a.org.cn

方法中, 图片和描述它的语句通常被看作是语义上对等的. 然而, 我们发现图片与描述它的语句在语义上并非简单的对等关系. 图 1 显示了 Microsoft COCO[7]、Flickr30K[8]与 Flickr8K[5]数据集中描述同一图片的 5 条语句之间语义相似程度[9]的统计情况. 从图 1 可以看出, 上述三个数据集中 5 条语句语义彼此之间都相似的图片数占数据集中图片总数的比例分别为 8.0%, 6.6% 以及 15.3%(请见图中横坐标为 10 的数据), 这表明描述同一幅图片的不同语句之间往往是弱相似或者不相似的. 这是因为描述同一幅图片的不同语句可能是出于不同的描述视角, 例如在表 1 中, 右侧的语句“a girl sits on a bar stool”与“dark nightclub with chairs”都描述了左侧的图片, 但是二者的语义相似度很低. 这说明, 在语义上图片与描述它的语句之间并非对等的关系, 而是蕴含关系[3]. 如果按照对等关系进行图片与语句的匹配, 那么势必会将弱相似或不相似的语句看作相似的, 显然是不合适的.

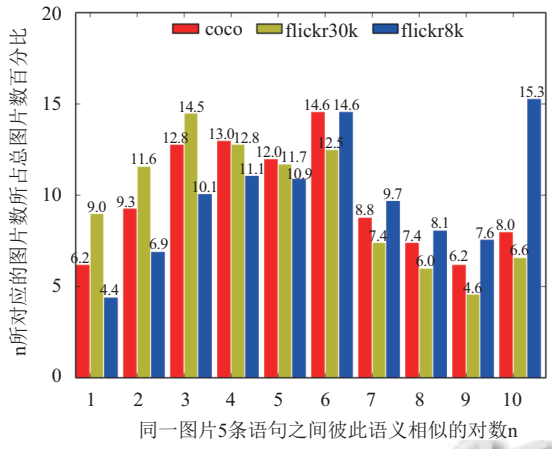

图 1 Microsoft COCO、Flickr30K 以及 Flickr8K 数据集中语句相似度的统计

表 1 图片与 5 条描述语句示例

图片	语句
	there are several people in a dark bar type room including one girl on a stool
	a girl sitting in a dark bar
	dark nightclub with chairs
	a girl sits on a bar stool
	a woman sits at a dark bar

本文基于图片与语句在语义上的这种蕴含关系提出一种新的图片语句匹配模型, 称为 IRMatch 模型. IRMatch 模型利用卷积神经网络 (Convolutional neural network, CNN) 分别实现图片与语句的语义映射, 在此基础上, 在图片语句对得分学习中通过引入最大软间隔的策略挖掘图片与语句之间的语义蕴含关系, 以强化相关图片语句对在公共语义空间中位置的邻近性, 改善图片语句匹配得分的合理性. 基于 IRMatch 模型, 本文实现了图文双向检索方法, 并在 Flickr8K[5]、Flickr30K[8]以及 Microsoft COCO[7]数据集上与基于已有图片语句匹配模型的图文双向检索方法进行比较. 实验结果表明, 基于 IRMatch 模型的检索方法在上述三个数据集上的 R@1、R@5、R@10 与 Med r 均优于基于已有模型的检索方法.

本文第 2 节对相关工作进行介绍, 第 3 节描述所提出的 IRMatch 模型, 第 4 节给出实验结果, 第 5 节对全文加以总结.

2 相关工作

当前的图片语句匹配方法[1,3,5,6,10-15]主要有两大类: 一类方法是将图片和语句映射到同一语义空间, 然后在该空间中进行两者之间的语义匹配. Socher 等[14]提出使用语义依赖树递归神经网络 (SDT-RNN) 来将语句映射到图片所在语义空间, 然后图片与语句之间的关联可以通过该空间上的距离来度量; Klein 等[12]使用 Fisher vector(FV) 作为语句的表示; Kiros 等[11]提出了 Skip-thought vectors(STV) 来对语句进行编码以与图片进行匹配; Wang, Jian 等[15]利用 WCNN 提取语句特征, 利用 CNN 提取图片深度特征, 将两者映射到同一公共空间, 并使用 one vs more 的学习策略进行学习; Karpathy 等[10]的工作在一个更加精细的水平, 他们将图片的片段 (对象) 与语句的片段 (类型依赖关系树) 嵌入到一个公共空间中从而对两者的关联性进行度量; Plummer 等[13]使用实体来实现区域到短语 (RTP) 的对应关系, 从而用于图片-语句建模.

另一类方法利用诸如 CCA, 深度学习等方法来挖掘图片和语句之间的语义关联. Hodosh 等[5]提出核典型相关分析 (Kernel canonical correlation analysis, KCCA) 用于发现图片和语句之间共享的特征空间; Yan 等[6]将全连接层堆叠在一起来表示语句, 同时使用深度典型相关分析 (DCCA) 来匹配图片和语句; Vendrov, Ivan 等[3]采用 Gated recurrent unit(GRU) 来提取语句的特征, 并将图片和语句的关系看作是一种偏序关系, 并

© 中国科学院软件研究所 http://www.c-s-a.org.cn

在此关系的基础上度量图片和语句的关联性. Ma, Lin 等[1]使用 m-CNNs 将图片与语句在 word、phrase 以及 sentence 级别进行匹配, 从而实现图片与语句在局部以及全局的混合匹配.

上述两类已有方法通常将图片和描述它的语句看作是语义上对等的, 而本文所提出的 IRMatch 模型挖掘图片与语句之间的语义蕴含关系, 通过 CNN 将图片与语句映射到公共语义空间, 之后基于最大软间隔的策略进行图片语句的关联学习.

3 IRMatch 模型

3.1 模型概述

图片语句匹配的目标是语义相关的图片语句对的匹配得分高于语义不相关的图片语句对的匹配得分[1]. 解决思路通常有两种: 一种是首先对图片和语句进行表示学习, 之后再利用典型相关分析等方法进行图片和语句的语义关联学习[5,6], 另一种是将图片与语句映射到一个公共的语义空间, 之后再学习图片语句对的匹配得分[10-15]. 其中第二种思路的优势在于图片语句的表示学习和关联学习是同时进行的而不是分离的, 使得图片语句匹配过程的整体性更强. 因而本文所提 IRMatch 模型采用第二种思路进行图片语句匹配, 步骤如下:

(1) 设 I 为图片集, S 为语句集, 建立映射 $p: I \to R^k$, $q: S \to R^k$, 以将 I 中图片与 S 中语句映射到公共语义空间 R^k 中, 其中 k 是公共空间 R^k 的维度.

(2) 令得分函数 $f: R^k \times R^k \to R$ 量度图片与语句语义映射的匹配度, 即若图片与语句越匹配则得分函数的值越大. 进而, 基于 p, q, f 定义损失函数 L, 并通过求解以 L 为目标函数的最小化问题学习图片语句对的匹配得分. 本文将匹配得分函数 f 视作超参数.

更具体地, IRMatch 模型利用卷积神经网络 CNN 分别实现图片与语句的语义映射, 在此基础上, 在图片语句对得分学习中通过引入最大软间隔的策略挖掘图片语句之间的语义蕴含关系, 以强化相关图片语句对在公共语义空间中位置的邻近性, 改善图片语句匹配得分的合理性.

下面分别针对基于 CNN 的图片、语句的语义映射与基于最大软间隔的图片语句对匹配得分学习进行详细的介绍.

3.2 基于 CNN 的图片语义映射述

近年来 CNN 已经展现了其超强的图片特征学习能力[16-19], 因而本文也采用 CNN 进行图片语义映射. 如图 2 所示, CNN 可由卷积层、池化层以及全连接层等组成, 其中卷积层提取图像的特征, 池化层针对原始特征信号进行抽象, 以减少训练参数, 而全连接层主要负责分类与回归.

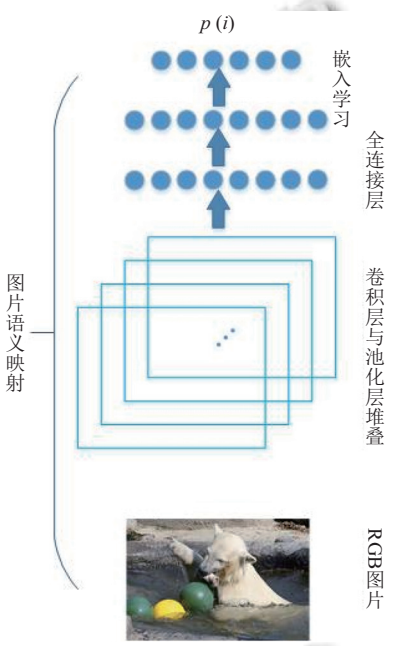

图 2 图片语义映射架构

借鉴文献[3]中的思想, IRMatch 模型中用于图片语义映射所采用的 CNN 是具有 19 层的 VGG 网络[18], 其包含 19 个卷积层、4 个池化层以及 3 个全连接层. 此 CNN 以 RGB 图片作为输入, 使用其第二个全连接层的输出作为图片表示[3], 其中图片深度特征的维度为 4096, 之后学习一个矩阵 W_i, 将所提取的 4096 维的图片特征向量映射到 R^k 中. 将此实现图片语义映射的 CNN 程序记作 $iCNN$, 其以图片为输入, 输出为 4096 维向量, 则对任意的 $i \in I$ 有:

$$p(i) = W_i \cdot iCNN(i) + b_i \quad (1)$$

其中 W_i 是 $k \times 4096$ 矩阵, b_i 表示偏置, $p(i)$ 表示 R^k 中图片映射点.

3.3 基于 CNN 的语句语义映射

为了在图片语句对匹配计算中使语句的表示与图片的表示具有一致的形式, IRMatch 模型采用 CNN 进行语句语义映射.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

借鉴文献[20]中的思想, 用于语句语义映射所采用的 CNN 具有一个卷积层与一个最大池化层, 如图 3 所示.

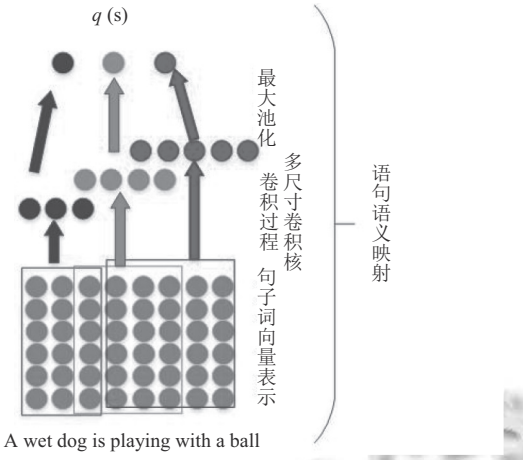

图 3 语句语义映射架构

输入语句中单词的表示方式与文献[20]一致, 即用对应的词向量表示. 输入语句由词嵌入矩阵 (图中蓝色的部分) 表示, 词嵌入矩阵中单词的顺序与该单词在语句中的顺序一致.

卷积过程采用了不同尺寸的卷积核, 如图 3 所示, 紫色的卷积核的宽度是 3, 黄色的卷积核的宽度是 4, 红色卷积核的宽度是 5, 它们的长度与词向量的长度是一致的, 例如图中紫色卷积核卷积的输入由图中紫色的圆点表示. 卷积核可以看作是不同长度短语的特征提取器, 使得整个卷积过程可以提取语句局部的语义特征. 卷积过程的步长均为 1. 而所有卷积核的卷积输出均是一个向量, 最大池化层对卷积输出的结果进行池化. 池化的宽度分别是每个卷积核卷积输出向量的长度, 这样就把每个卷积核卷积的输出池化成为一个点, 最终整个 CNN 所提取特征的维度就是所有卷积核的个数. 将此实现语句语义映射的 CNN 程序记作 $sCNN$, 其以语句为输入, 输出为 R^k 中向量, 则对任意的 $s \in S$ 有:

$$q(s) = sCNN(s) \tag{2}$$

$sCNN$ 在提取语句特征方面具有如下优势: 一是所提取语句特征的维度仅取决于卷积核的个数, 而与语句的长度无关; 二是卷积和池化操作考虑了语句的序列与结构的信息, 因而很容易处理词汇量很大的数据集, 而输出的维度不依赖于词汇量的大小.

3.4 基于最大软间隔的图片语句对匹配得分学习

函数 f 以图片和语句在公共空间 R^k 中的映射为输入计算图片语句对的匹配得分. 具体的, IRMatch 模型将 f 视作超参数并选用余弦相似度函数作为 f 来计算匹配得分, 即:

$$cos(p(i),q(s)) = \frac{p(i)q(s)^T}{\|p(i)\|\|q(s)\|} \tag{3}$$

在此基础上, 定义如下排序损失函数 L:

$$L = \sum_{(i,s)} \left(\begin{array}{l} \sum\limits_{s^-} max\{0, \mu - \cos(p(i),q(s)) + \cos(p(i),q(s^-))\} \\ + \\ \sum\limits_{i^-} max\{0, \mu - \cos(p(i),q(s)) + \cos(p(i^-),q(s))\} \end{array} \right) \tag{4}$$

其中 s 与 s^- 分别表示与图片 i 语义相关及不相关的语句; i 与 i^- 分别表示与语句 s 语义相关及不相关的图片, μ 表示相关的图片语句对的匹配得分比不相关的图片语句对的匹配得分所高出的间隔值. 进而, 通过求解以 L 为目标函数的最小化问题, 学习图片语句的匹配的分.

为了增大相关图片语句对的匹配得分, 并且缩小不相关图片语句的匹配得分, IRMatch 模型采用最大间隔策略进行学习. 最大间隔学习的典型形式为最大硬间隔学习[10-15], 即对任意的 $i \in I, s \in S$ 有:

$$cos(p(i),q(s)) - cos(p(i),q(s^-)) \geq m \tag{5}$$

$$cos(p(i),q(s)) - cos(p(i^-),q(s)) \geq m \tag{6}$$

其中公式 (5) 表示图片 i 与其相关语句 s 的匹配得分比与其不相关的语句 s^- 的得分至少大于 m, 公式 (6) 表示语句 s 与其相关图片 i 的匹配得分比与其不相关的图片 i^- 的得分至少大于 m. m 为间隔值, 即相关图文得分比不关图文得分大的最小值.

最大硬间隔策略将图片与语句当作对等关系来处理. 然而, 当图片与语句之间是语义蕴含关系时, 以图片对应多个描述语句为例, 可能会出现描述同一幅图片的语句之间存在语义不相似的情况, 而如果继续采用公式 (5), (6) 的硬间隔的学习策略, 就会导致在公共空间 R^k 中这些彼此之间语义不相似的语句很难同时出现在相应图片的附近. 如图 4 第二列所示, 三角形表示图片, 圆表示语句, 最大硬间隔的学习策略会导致公共空间中某些语句映射点无法出现在图片映射点的附近 (图中矩形框中). 为了解决这个问题, IRMatch 模型中引入松弛变量, 使用软间隔学习策略, 这样可以容忍一定偏差, 将图片在公共空间的映射点尽量靠近所有语句的映射点 (如图 4 第三列所示) 即对任意的 $i \in I, s \in S$ 有:

© 中国科学院软件研究所 http://www.c-s-a.org.cn

$$cos(p(i),q(s))-cos(p(i),q(s^-))\geq m-\varepsilon_{s^-};\varepsilon_{s^-}\geq 0 \quad (7)$$

$$cos(p(i),q(s))-cos(p(i^-),q(s))\geq m-\varepsilon_{i^-};\varepsilon_{i^-}\geq 0 \quad (8)$$

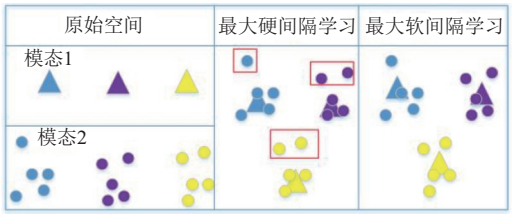

图 4 最大间隔学习示意图

而基于最大软间隔学习的损失函数定义如下:

$$L=\sum_{(i,s)}\left(\begin{array}{l}\sum\limits_{s^-}max\{0,m-\varepsilon_{s^-}-cos(i,s)+cos(i,s^-)\}\\+\\\sum\limits_{i^-}max\{0,m-\varepsilon_{i^-}-cos(i,s)+cos(i^-,s)\}\end{array}\right) \quad (9)$$

其中$\varepsilon_{i^-}\geq 0,\varepsilon_{s^-}\geq 0$, (i,s)表示相关的图片语句, s^-与 i 不相关, i^-与 s 不相关. 超参数 m 的值由交叉验证选择. 本文使用 Adam[21]优化算法来进行模型训练, 同时 ReLU[22]作为整个卷积神经网络的激活函数.

4 实验

本文在 Flickr8k[5]、Flickr30k[8]以及 Microsoft COCO[7]数据集上进行了图文双向检索任务的实验, 以将所提图片语句匹配模型与文献[1-4,6,10-14,23-28]所提出的模型进行了比较.

4.1 实验设置

4.1.1 数据集

本文选择如下公开图片语句基准数据集进行图文双向检索任务的实验.

(1) Flickr8K[5]: 此数据集由采自 Flickr 的 8000 张图片组成, 每张图片对应 5 句描述图片内容的语句. 此数据集提供了标准的训练集、校验集以及测试集划分.

(2) Flickr30K[8]: 此数据集由采自 Flickr 的 31783 张图片组成, 每张图片对应 5 句描述图片内容的语句. 其中大部分图片的内容与人类活动有关. 本文采用和[28]中相同的划分方法划分训练集、校验集以及测试集.

(3) Microsoft COCO[7]: 此数据集包含 82783 张训练图片以及 40504 张校验图片. 每张图片对应 5 句描述图片内容的语句. 本文采用和文献[26]中相同的划分方法来划分训练集, 校验集以及测试集.

4.1.2 评价指标

本文采用 Med r 与 R@K 评价图文双向检索的结果[10]. Med r 表示与查询最相关的结果在结果列表中的平均排名, 其值越小越好. R@K(K=1, 5, 10) 表示在前 K 个结果中出现正确结果的百分比, 其值越大越好.

4.1.3 参数设置

在训练过程中, 本文采用公式 (9) 定义的损失函数. 训练 batch-size 设为 250, 即每一次从数据集中采样 250 对不同的相关图片-语句对, 对于每一张图片本文获得 249 与之不相关的语句, 同理对于每一个语句本文也可以获得 249 个与之不相关的图片. 使用 Adam 优化算法训练 25-40 个 epochs, 并且设置初始学习率为 0.001, 采用提前停止策略防止训练过拟合. 公共空间的维度 k 设置为 1200, 词向量的维度设置为 300, 间隔 m 的值设置为 0.5. 这些超参数, 包括学习率以及 batch-size, 都是通过校验集进行选择的.

4.2 实验结果分析

我们分别实现了 IRMatch 模型采用最大硬间隔策略 (记为 IRMatchH) 与最大软间隔策略 (记为 IRMatchS) 时的图文双向检索方法, 之后在数据集 Flickr8k、Flickr30k 以及 Microsoft COCO 上, 计算所实现的图文双向检索方法的 Med r 与 R@K(K=1, 5, 10), 并与文献[1-4,6,10-14,23-28]所提出的方法在上述三个数据集上的结果进行对比, 分别如表 2, 表 3, 表 4 所示.

4.2.1 IRMatchH 与已有方法比较

总体来看, 基于 IRMatchH 的检索方法的结果优于基于已有图片语句匹配方法的检索方法的结果. 尤其是在 Flickr30k 数据集上, 所有指标均优于已有方法. 这说明 CNN 能够有效的提取语句的语义信息. 本文采用 *sCNN* 来对语句进行特征建模, 使用了宽度为 1 到 6 的卷积核. 应用了不同宽度的多个卷积核, 相当于可以提取蕴含 1 到 6 个词的短语蕴含的语义信息. 除了具有提取不同长度短语的能力, 该模型还能考虑到语句语序信息以及结构信息. 池化层中的最大池化操作能够对上述语义信息进行筛选. 将语句和图片映射到同一空间中后, 使用余弦相似度在公共空间 R^k 中直接计算语句和图片之间的相似度, 从而完成两者之间的关联.

Flickr8K 数据集上, FV[12]在语句检索 (以图片检索语句) 任务中 R@1 指标略高于 IRMatchH 的结果, 除此之外, 两者在各个指标上均取得了最好的结果. 这说明当训练数据不是很充分的时候, 本文的模型依然能

© 中国科学院软件研究所 http://www.c-s-a.org.cn

够很好地对语句和图片进行建模, 并且完成两者之间的匹配. 而当数据充分时, 在 Flickr30k 以及 Microsoft COCO 数据集上, IRMatchH 远好于 FV[12]的结果. IRMatchH 在这三个数据集上的试验结果有效的证实了 CNN 在提取语句语义信息方面的优越性.

4.2.2 IRMatchS 与 IRMatchH 比较

从表 2, 表 3, 表 4 可以看出, IRMatchS 的结果好于 IRMatchH 的结果, 尤其是在 Microsoft COCO 数据集. 这组对比试验表明最大软间隔的学习方式能够有效的解决图片语句非对称匹配问题. 本文采用公式 (9) 作为训练模型中的损失函数, 引入松弛变量将硬间隔转变成软间隔. 由于 Flickr8K、Flickr30K 以及 Microsoft COCO 这三个数据集中图片和语句之间语义之间并非是对等关系, 而是一种蕴含关系, 并且描述同一图片的语句语义之间存在不相似或者弱相似的情况. 因此, 若采用硬间隔 (IRMatchH) 的方式将彼此之间不相似的语句投影在公共空间 R^k 中的点, 无法都临近对应图片的映射点, 必然导致某些语句和图片没有匹配在一起. 而采用软间隔的方式, 容忍一定的偏差, 可以将图片和不同语义的语句关联在一起, 因此可以提高匹配的性能以及泛化能力. 对比 IRMatchS 以及 IRMatchH 在上述三个数据集上的实验结果, 可以佐证这种软间隔 (IRMatchS) 的学习策略能够很好的解决图片和语句之间的语义非对等问题.

表 2 Flickr8k 数据集上图文双向检索比较结果

	语句检索				图片检索			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Random Ranking	0.1	0.6	1.1	631	0.1	0.5	1.0	500
DeViSe[25]	4.8	16.5	27.3	28.0	5.9	20.1	29.6	29
SDT-RNN[14]	6.0	22.7	34.0	23.0	6.6	21.6	31.7	25
MNLM[27]	13.5	36.2	45.7	13	10.4	31.0	43.7	14
MNLM-vgg[27]	18.0	40.9	55.0	8	12.5	37.0	51.5	10
m-RNN[28]	14.5	37.2	48.5	11	11.5	31.0	42.4	15
Deep Fragment[10]	12.6	32.9	44.0	14	9.7	29.6	42.5	15
RVP(T)[23]	11.6	33.8	47.3	11.5	11.4	31.8	45.8	12.5
RVP(T+I)[23]	11.7	34.8	48.6	11.2	11.4	32.0	46.2	11
DVSA(DepTree)[26]	14.8	37.9	50.0	9.4	11.6	31.4	43.8	13.2
DVSA(BRNN)[26]	16.5	40.6	54.2	7.6	11.8	32.1	44.7	12.4
DCCA[6]	17.9	40.3	51.9	9	12.7	31.2	44.1	13
NIC[4]	20.0	*	61.0	6	19.0	*	64.0	5
FV(Mean Vec)[12]	22.6	48.8	6.0	6	19.1	45.0	60.4	7
FV(GMM)[12]	28.4	57.7	70.1	4	20.6	48.5	64.1	6
FV(LMM)[12]	27.7	56.6	69.0	4	19.8	47.6	62.7	6
FV(HGLMM)[12]	28.5	58.4	71.7	4	20.6	49.4	64	6
FV(GMM+HGLMM)[12]	31.0	59.3	73.7	4	21.3	50.0	64.8	5
m-CNN(ENS)[1]	24.8	53.7	67.1	5	20.3	47.6	61.7	5
IRMatchH	28.7	60.8	72.8	3	24	54.8	69	4
IRMatchS	30.0	60.1	74.1	4	24.2	55.0	69.2	4

5 结束语

本文提出一种新的基于语义蕴含关系的图片语句匹配模型 IRMatch, 能够很好的解决图片和语句语义之间的非对等匹配问题. 该模型使用两种不同的卷积神经网络 *iCNN* 与 *sCNN* 来对图片以及语句进行语义映射, 从而将两者投影到同一公共空间 R^k 中, 有利于两种不同模态数据的直接比较, 而且模型采用最大软间隔的学习策略来学习图片语句之间的匹配得分, 强化了相关图片语句对在公共语义空间中位置的邻近性, 改善了图片语句匹配得分的合理性. 本文分别在 Flickr8K, Flickr30K 以及 Microsoft COCO 数据集上进行了实验, 实验表明基于所提 IRMatch 模型的图文双向检索方法的结果优于基于已有模型的检索方法的结果.

未来我们将重点针对语句对应多个图片的语句图片蕴含关系的模型进行研究.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

表 3 Flickr30k 数据集上图文双向检索比较结果

	语句检索				图片检索			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Random Ranking	0.1	0.6	1.1	631	0.1	0.5	1.0	500
DeViSe[25]	4.5	18.1	29.2	26	6.7	21.9	32.7	25
SDT-RNN[14]	9.6	29.8	41.1	16	8.9	29.8	41.1	16
MNLM[27]	14.8	39.2	50.9	10	11.8	34.0	46.3	13
MNLM-vgg[27]	23.0	50.7	62.9	5	16.8	42.0	56.5	8
m-RNN[28]	18.4	40.2	50.9	10	12.6	31.2	41.5	16
Deep Fragment[10]	14.2	37.7	51.3	10	10.2	30.8	44.2	14
RVP(T)[23]	11.9	25.0	47.7	12	12.8	32.9	44.5	13
RVP(T+I)[23]	12.1	27.8	47.8	11	12.7	33.1	44.9	12.5
DVSA(DepTree)[26]	20.0	46.6	59.4	5.4	15.0	36.5	48.2	10.4
DVSA(BRNN)[26]	22.2	48.2	61.4	4.8	15.2	37.7	50.5	9.2
DCCA[6]	16.7	39.3	52.9	8	12.6	31.0	43.0	15
NIC[4]	17.0	*	56.0	7	17.0	*	57.0	7
LRCN[24]	*	*	*	*	17.5	40.3	50.8	9
RTP(joint training)[13]	31.0	58.6	67.9	*	22.0	50.7	62.0	*
RTP(SAE)[13]	36.7	61.9	73.6	*	25.4	55.2	68.6	*
RTP(wighted distance)[13]	37.4	63.1	74.3	*	26.0	56.0	69.3	*
FV(Mean Vec)[12]	24.8	52.5	64.3	5	20.5	46.3	59.3	6.8
FV(GMM)[12]	33.0	60.7	71.9	3	23.9	51.6	64.9	5
FV(LMM)[12]	32.5	59.9	71.5	3.2	23.6	51.2	64.4	5
FV(HGLMM)[12]	34.4	61.0	72.3	3	24.4	52.1	65.6	5
FV(GMM+HGLMM)[12]	35.0	62.0	73.8	3	25.0	52.7	60	5
m-CNN(ENS)[1]	33.6	64.1	74.9	3	26.2	56.3	69.6	4
IRMatchH	40.0	66.6	77.5	2	29.9	61.1	73.0	3
IRMatchS	38.7	67.4	79.0	3	29.7	61.0	72.7	3

表 4 Microsoft COCO 数据集上图文双向检索比较结果

	语句检索				图片检索			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Random Ranking	0.1	0.6	1.1	631	0.1	0.5	1.0	500
m-RNN-vgg[2]	41.0	73.0	83.5	2	29.0	42.2	77.0	3
DVSA[26]	38.4	69.9	80.5	1	27.4	60.2	74.8	3
STV(uni-skip)[11]	30.6	64.5	79.8	3	22.7	56.4	71.7	4
STV(bi-skip)[11]	32.7	67.3	79.6	3	24.2	57.1	73.2	4
STV(cmbi-skip)[11]	33.8	67.7	82.1	3	25.9	60.0	73.6	4
FV(Mean Vec)[12]	33.2	61.8	75.1	3	24.2	56.4	72.4	4
FV(GMM)[12]	39.0	67.0	80.3	3	24.2	59.2	76.0	4
FV(LMM)[12]	38.6	67.8	79.8	3	24.9	58.8	76.5	4
FV(HGLMM)[12]	37.7	66.6	79.1	3	24.9	58.8	76.5	4
FV(GMM+HGLMM)[12]	39.4	67.9	80.9	2	25.1	59.8	76.6	4
m-CNN(ENS)[1]	42.8	73.1	84.1	2	32.6	68.6	82.8	3
ORDER-EMBEDDINGS[3]	46.7	*	88.9	2	37.9	*	85.9	2
IRMatchH	46.8	79.3	89.0	2	37.76	73.5	86.1	2
IRMatchS	48.3	79.7	89.4	2	38.3	74.2	86.4	2

参考文献

1 Ma L, Lu ZD, Shang LF, *et al*. Multimodal convolutional neural networks for matching image and sentence. Proc. of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile. 2015. 2623–2631.

2 Mao JH, Xu W, Yang Y, *et al*. Deep captioning with multimodal recurrent neural networks (m-RNN). Proc. of the

© 中国科学院软件研究所 http://www.c-s-a.org.cn

International Conference on Learning Representations. San Diego, USA. 2015.
3 Vendrov I, Kiros R, Fidler S, *et al*. Order-embeddings of images and language. Proc. of the International Conference on Learning Representations. San Juan, Puerto Rico. 2016.
4 Vinyals O, Toshev A, Bengio S, *et al*. Show and tell: A neural image caption generator. Proc. of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA. 2015. 3156–3164.
5 Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research, 2013, 47(1): 853–899.
6 Yan F, Mikolajczyk K. Deep correlation for matching images and text. Proc. of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA. 2015. 3441–3450.
7 Lin TY, Maire M, Belongie S, *et al*. Microsoft COCO: Common objects in context. Proc. of the 13th European Conference on Computer Vision. Zurich, Switzerland. 2014. 740–755.
8 Young P, Lai A, Hodosh M, *et al*. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Trans. of the Association for Computational Linguistics, 2014, 2(4): 67–78.
9 Mueller J, Thyagarajan A. Siamese recurrent architectures for learning sentence similarity. Proc. of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16). Phoenix, Arizona, USA. 2016. 2786–2792.
10 Karpathy A, Joulin A, Li FF. Deep fragment embeddings for bidirectional image sentence mapping. Proc. of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada. 2014. 1889–1897.
11 Kiros R, Zhu YK, Salakhutdinov R, *et al*. Skip-thought vectors. Proc. of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada. 2015. 3294–3302.
12 Klein B, Lev G, Sadeh G, *et al*. Associating neural word embeddings with deep image representations using Fisher Vectors. Proc. of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA. 2015. 4437–4446.
13 Plummer BA, Wang LW, Cervantes CM, *et al*. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. Proc. of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile. 2015. 2641–2649.
14 Socher R, Karpathy A, Le QV, *et al*. Grounded compositional semantics for finding and describing images with sentences. Trans. of the Association for Computational Linguistics, 2014, 2(4): 207–218.
15 Wang J, He YH, Kang CC, *et al*. Image-text cross-modal retrieval via modality-specific feature learning. Proc. of the 5th ACM on International Conference on Multimedia Retrieval. New York, NY, USA. 2015. 347–354.
16 He KM, Zhang XY, Ren SQ, *et al*. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. Proc. of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile. 2015. 1026–1034.
17 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proc. of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada, USA. 2012. 1097–1105.
18 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Proc. of the International Conference on Learning Representations. San Diego, USA. 2015.
19 Szegedy C, Liu W, Jia YQ, *et al*. Going deeper with convolutions. Proc. of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA. 2015. 1–9.
20 Kim Y. Convolutional neural networks for sentence classification. Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar. 2014. 1746–1751.
21 Kingma DP, Ba J. Adam: A method for stochastic optimization. Proc. of the International Conference on Learning Representations. Banff, Canada. 2015. 1–13.
22 Dahl GE, Sainath TN, Hinton GE. Improving deep neural networks for LVCSR using rectified linear units and dropout. Proc. of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vancouver, BC, Canada. 2013. 8609–8613.
23 Chen XL, Zitnick CL. Mind's eye: A recurrent visual representation for image caption generation. Proc. of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA. 2015. 2422–2431.
24 Donahue J, Hendricks LA, Guadarrama S, *et al*. Long-term recurrent convolutional networks for visual recognition and description. Proc. of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA. 2015. 2625–2634.
25 Frome A, Corrado GS, Shlens J, *et al*. DeViSE: A deep visual-semantic embedding model. Proc. of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada, USA. 2013. 2121–2129.
26 Karpathy A, Li FF. Deep visual-semantic alignments for generating image descriptions. Proc. of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA. 2015. 3128–3137.
27 Kiros R, Salakhutdinov R, Zemel RS. Unifying visual-semantic embeddings with multimodal neural language models. arXiv: abs/1411.2539, 2014.
28 Mao JH, Xu W, Yang Y, *et al*. Explain images with multimodal recurrent neural networks. arXiv: abs/1410.1090, 2014.

© 中国科学院软件研究所 http://www.c-s-a.org.cn