# 双向图像句映射的深片段嵌入[[1]](#footnote-1)

赵淼译

(北京交通大学电子与信息工程学院，电子科学与技术，15650761201)

## 1 摘要

作者引入了一个可以对图像和句子进行双向转换的可以将视觉和自然语言数据进行嵌入的多模态模型。之前的模型将图像或句子直接映射到一个公共嵌入空间，我们的模型进行工作在一个更精细的层次，将图像片段（目标）和句子片段（类型依赖树）嵌入到一个公共空间。除了先前的工作将目标进行排序外，作者的这种方法可以将加入一种片段对齐目标，可以学习到这些片段之间跨模态的关系。

## 2 介绍

介绍研究的意义：描述图像的内容对于图像字幕的生成或者反向操作都是有用的。基于自然语言查询检索图像的能力可以实现即时图像搜索应用程序。在这项工作中，我们感兴趣的是在一组图像及其相关的自然语言描述上训练一个模型，以便我们可以给定一组图像同时产生一些按贴近程度顺序排列可以描述这幅图像的句子，反之亦然。

这个问题十分有挑战性，因为要求对图像，自然语句以及它们之间的内在关系有详细的理解。即为了理解一幅图像，必须准确识别句子中的所有目标，属性和它们之间的关系，使之逼近复杂的视觉场景。我们的主要贡献在于为深度神经网络制定一个结构化的最大边界的目标，该网络可以学习将视觉和语言数据嵌入共同的多模式空间。与之前的工作不同，我们的模型是将图像（目标）和自然语句（依赖树关系）进行分解嵌入到一个共同的空间中，并明确其中潜在的模态间对应关系的推理。

片段级别的推理使我们可以使用一个新的片段级别损失函数，补充传统的（句子图像排名）损失。广泛的实证评估验证了我们的方法。 特别是，我们报告了在Pascal1K [2]，Flickr8K [3]和Flickr30K [4]数据集上图像语句检索任务的最先进方法的显着改进。 我们计划公开我们的代码。

## 2 相关工作

图像标注和图像检索：一些人专注于描述图像内容，有的将图像映射到一组固定的句子里，或是生成小说标题。与我们的研究是一些更密切相关的方法，它们允许在两个模态进行双向的映射。有将图像和句子进行结合的核规范化方法，这种方法不易扩展，它依赖于一系列图像和句子之间的二次计算核，或是通过目标，动作和场景的方法获得一个公共区间，但是也限制在了目标，动作和方法的范围内进行处理。用条件随机场的方法推理卡通场景与自然语言描述之间的关系。

多模态的表示学习：我们的方法属于多模态学习的范畴。使用深玻尔兹曼机器[15]，对数-双线性模型[16]和主题模型[17,18]，已经开发了几种用于表示图像和句子的联合多模态概率分布的概率模型。与我们的任务更密切的方法是引入一种视觉语义嵌入模型，学习图像和词语之间的排序损失【链接】，类似的还有描述一个依赖树的递归神经网络【链接】，但是这些方法仅从全局方面来推理做关于图像的推理，用一个单独的，固定大小的来自于卷积神经网络的顶层描述整个图像，而我们的模型对于目标如何构成一个复杂的的场景进行简要的推理。

图像和自然语言的神经网络表示：我们的模型将图像像素和一千个单词的表示进行了联系，图像：CNN，语言：LSTM。

## 3 提出的模型

学习和接口的介绍：我们的任务是给定一个图像检索相关的句子，给定句子检索相关图像。我们将在训练中训练我们的模型在一组N个图像和N个对应的描述它们的内容的句子，给定这种对应关系后，我们训练神经网络的权重可以对成对的句子、图像给予较高的分数，反之是较小的分数。

片段嵌入：我们核心的思路是图像是一种复杂的结构，它由多个彼此联系的实体组成，所对应的句子会明确的提到这些实体以及它们之间的关系。在作者的模型中捕获种关系，通过将图像和句子拆成片段和推理它们之间的潜在的对齐。尤其这里我们提出使用图像中检测到的物体作为图像片段和用句子中的依赖树关系作为句子片段。

片段级目标：在过去的相关工作中，神经网络嵌入图像和句子到一个共同的空间中，然后训练参数使真正的图像句子对有一个比其他图像句子对更高的内积。在我们的方法中，我们代替的使用了将句子和图像片段进行嵌入，计算图像和句子的得分作为他们片段的得分的固定函数。因此为了以前工作的排名损失，我们加入了一个更强的片段联合目标。我们将证明这些目标会提供片段互补信息。



图片涉及图像片段生成，句子片段生成以及他们之间获得一个相似性得分矩阵。

### 3.1 句子片段的独立性依赖树

我们想要提取和表示一个句子中视觉可见实体的集合。主要包含视觉实体，他们的属性和关系。这里作者使用依赖树来表示一系列丰富的关系比单个的单词或短语更有用。作者这里放弃了树结构，直接将两个实体和他们之间的关系（R,w1,w2）进行编码映射到嵌入空间中.

|  |  |
| --- | --- |
|  | 1 |

We是一个40000（单词的个数）\*d（转换到的空间的维度）的矩阵，这里取d=200，We通过无监督聚类的方式获得。针对每个关系有各自的权重WR和bR，固定使用ReLU作为f，s维度是交叉验证的。

### 3.2 目标检测作为的图像片段

与句子相似，这里希望描述一系列组成图像的物体。这里介绍了一个假设，大部分句子描述的对象是目标的属性和背景。尤其是这里使用了预训练在ImageNet上RCNN网络。这里利用top-19得分的目标框内的特征向量和整幅图像中的特征向量获得嵌入向量。

这里使用的是分类前的全连接层获得的4096维向量作为特征向量，Wm也许可以使用CNN中的权重进行初始化，但这里作者为了简单起见，放弃了这些向量。

|  |  |
| --- | --- |
|  | 1 |

### 3.3 目标函数

输入是一些训练集，包括N幅图像和对应的句子。在前一部分，作者用参数化的方法将每个句子和图像转化为片段向量。通过内积的方法计算图像句子对的相似度得分，显然多个匹配片段会导致高的得分。

|  |  |
| --- | --- |
|  | 1 |

这里作者设计了两个惩罚项。一是图像句子相似度性与真正的对应关系之间的惩罚项。这是全局排序目标。第二是片段对齐目标，这部分时由所有句子片段在视觉域对齐的表现决定，整个目标函数由两部分的和组成加上一个正则化惩罚项。

这里的θ是神经网络的参数，包括We,WR,Wm,θc.,α,β是超参数。

#### 3.3.1 片段联合目标

片段联合目标从直觉上看，就是如果句子中包含一个片段，那么对应的图片中也应该至少有一个框应该对应这种片段具有高分，而其他框具有一个低分。这种假设可能以多种形式被违反，如三元组不涉及图像中任何可识别的东西，RCNN可能检测不到三元组引用的框，最后其他图像可以包含所描述的视觉概念，但是在相关的句子中被省略。尽管如此，这一目标在许多情况下被满足，假设对应图像与句子之间密集对齐。

|  |  |
| --- | --- |
|  | 1 |

这里是对所有训练集中所有句子和图像片段求和，这里visj表示视觉片段和句子片段之间的联合得分，定义yij作，如果片段在一个对应关系里，否则为1，常数对正例和负例的数目进行正则化。

多实例学习扩展：这里针对前面过强的假设（所有的片段在每个对应的图像句子对之间都有紧密的对应）提出了多实例学习。这里试图推断出潜在的对齐关系。相应的将每个三元组都把关联图像的图像的图像片段放到对应的正例包中，在其他图像中的图像片段放到反例中，这里受到mi-SVM的启发。同时yij不再在是常数。

|  |  |
| --- | --- |
|  | 1 |
|  |  |
| and |  |

这里m返回图像和句子的索引，这里的不等式关系指的是对每个句子片段至少应该有一个正例，如果没有，则在所有正例包中选择一个得分最高的置为正例。

#### 3.3.2 全局排序目标

这里要保证图像句子相似度和标注一致。首先定义图像句子联合得分为他们成对片段的联合得分。

|  |  |
| --- | --- |
|  | 1 |

gk是图像的图像片段的集合，gl是句子l的句子片段的集合。K，都是从1到N.这里以0为截断，大于0被认为是正确的联合。实践中发现加一个平滑项效果更好，取10.三角形是超参数，让真正的图像句子对Skk大于Skl和Slk的分数大于边界，整个目标函数由点积和0阈值组成。

### 3.4 优化

SGD，小批量100，动量0.9整个训练数据作15epochs，最后两个epoch退火0.1，多实例学习和CNNfine\_tune都是有益的，开始的10个epoch片段联合分数和CNN权重被固定。在10个epoch后MIL，CNN开始训练。词嵌入矩阵由于担心过拟合而保持不变。

## 4 实验

数据集：Pascal1K [2], Flickr8K [3] 和and Flickr30K，分别包含100，8000和30000幅图像，每个图像包含5个独立的句子，由机器人标注。

句子数据预处理：这里没有明显的进行过滤，拼写检查或是其他正则化操作。使用CoreNLP获得每个句子的依赖树关系。由于存在许多可能的关系（多达数百个），由于过度拟合关注和实际考虑，我们移除了每个数据集中发生比例少于1％的所有关系类型，。大大减少了关系数目，未在字典中存在的单词移除了。

图像预处理：用RCNN检测所有图像中的目标。这里将ImageNet中的200类检测信息去掉了。只保留4096维的激活值（整幅图像中top19个检测位置）。

双向检索：探索句子和图像之间双向检索的准确率，评价指标有两个，一个输出排序在top-n中的准确率，一个是最接近标注的结果在排序上的中位数。

### 4.1 对照试验

这里和SDT-RNN进行了对比，这种方法是全帧CNN表示。并使用了他们的代码进行了实验，并采取我们的CNN模型来训练他们的方法。

### 4.2 定量的评价

这里作者通过实验获得了以下几条总结。

片段与全局目标互相补充：相对来说全局目标表现更好。

提取目标的表示信息是重要的：相对与全局信息来说，局部的目标检测更为重要，全句的信息不足以表示图像。

依赖树关系：相对单纯的词语关系和双联词结构，依赖关系对句子的组成更重要。

Fine\_tune具有更好的效果。

### 4.3 定性的实验

通过实验显示，我们的方法可以应用到更细致化的图像分类中，把每句话映射到一个关系组中。

限制：这里提到用三元组关系来建模存在很多不足之处，如一个名词短语可能会对其到不同的目标中。还有一些短语无法用三元组建模。RCNN没有考虑到目标的空间信息，很难告诉我们不同的两个人的关系。在语言方面，很多关系不以图像为基础，需要对图像进行句子进行更进一步的分析。

## 5 总结

作者解决的是一个图像和句子的双向检索的问题。网络学习的是多模态的图像和句子的嵌入空间，推理他们内在的对齐关系。这种方法允许我们使用新的片段对齐策略和总体目标。希望进一步扩展到支持计数，物体位置的推理，而不在仅仅局限与片段。

1. 赵淼 电信研1706班 电子科学与技术 17120036 15650761201 [↑](#footnote-ref-1)