

DeepSeek : 基于内容的图像搜索和检索

Tanya Piplani

加州大学伯克利分校信息学院 ta
nyapiplani @ berkeley. edu

Abstract

大多数互联网内容由视频和图像等数字媒体构成。随着像素成为互联网上大多数交易的基础货币，拥有一个相对简便的方式来浏览这些信息海洋变得越来越重要。每分钟上传到YouTube的视频时长达到400小时，而Instagram、Facebook等平台上也浏览着数以千万计的图片。受近期深度学习领域进展的启发以及其在图像字幕等多种问题上的成功应用，[Karpathy 和 Fei-Fei, 2015](#)) and ([徐等人。 , 2015](#))、机器翻译 ([Bahdanau 等人。 , 2014](#)) , word2vec , 跳过思想 ([Kiros 等人。 , 2015](#)) , 等等，我们在后续部分描述了两种方法，介绍了基于自然语言处理的深度学习模型DeepSeek。该模型允许用户输入他们想要搜索的图像类型描述，系统则会检索所有与查询在语义和上下文上相关的图像。

1 Introduction

图像搜索是一个非常具有挑战性的问题，目前仍然是活跃的研究领域。所有主要玩家如亚马逊、谷歌、苹果等都提供了相应的解决方案。然而，这些方案各自存在局限性。例如，亚马逊的图像搜索使用计算机视觉来检索相似图片，虽然在大多数情况下较为准确，但最大的问题是用户需要输入基于图片的查询，这在很多情况下可能难以获取。苹果在其设备中通过使用短语如“带有.....的照片”等选项来搜索图片，但这仍然不能完全解决问题。

“食品”，“生日”等。由于受到可以准确处理的令牌数量限制，表达能力严重受限。此外，这种搜索当然也受限于设备上的图像数量。一些其他解决方案，如Google的图像搜索，使用元数据，这可能会相当误导。

为了克服所有这些问题，我们提出了一种端到端的图像搜索和检索方法，通过基于文本的查询使用自然语言处理技术。在下一节中，我们将描述所使用的数据和方法。

2 相关工作

在基于内容的图像检索领域已经做了很多工作。([周等人。 , 2017](#)) 特别是，两项开创性的工作为大规模多媒体数据库基于内容的视觉检索的重大进展铺平了道路。第一项工作是引入不变局部视觉特征SIFT (Scale-Invariant Feature Transform) 。 [Lowe , 2004](#)) 。第二件工作是引入视觉袋 (BoW) 模型 ([Sivic 和 Zisserman , 2003](#)) 。通过利用信息检索技术，BoW模型基于图像中局部特征的量化构建紧凑的图像表示，并且可以轻松适应经典的倒排文件索引结构以实现可扩展的图像检索。图像表示源自这样一个事实：基于内容的视觉检索中的核心问题是图像比较。为了便于比较，将图像转换为某种特征空间。其动机在于实现隐式的对齐，以便消除背景影响以及潜在的变换或变化的影响，同时保持内在的视觉内容可区分性。传统上，视觉特征是根据启发式方法设计的，并可以分为局部特征和全局特征。

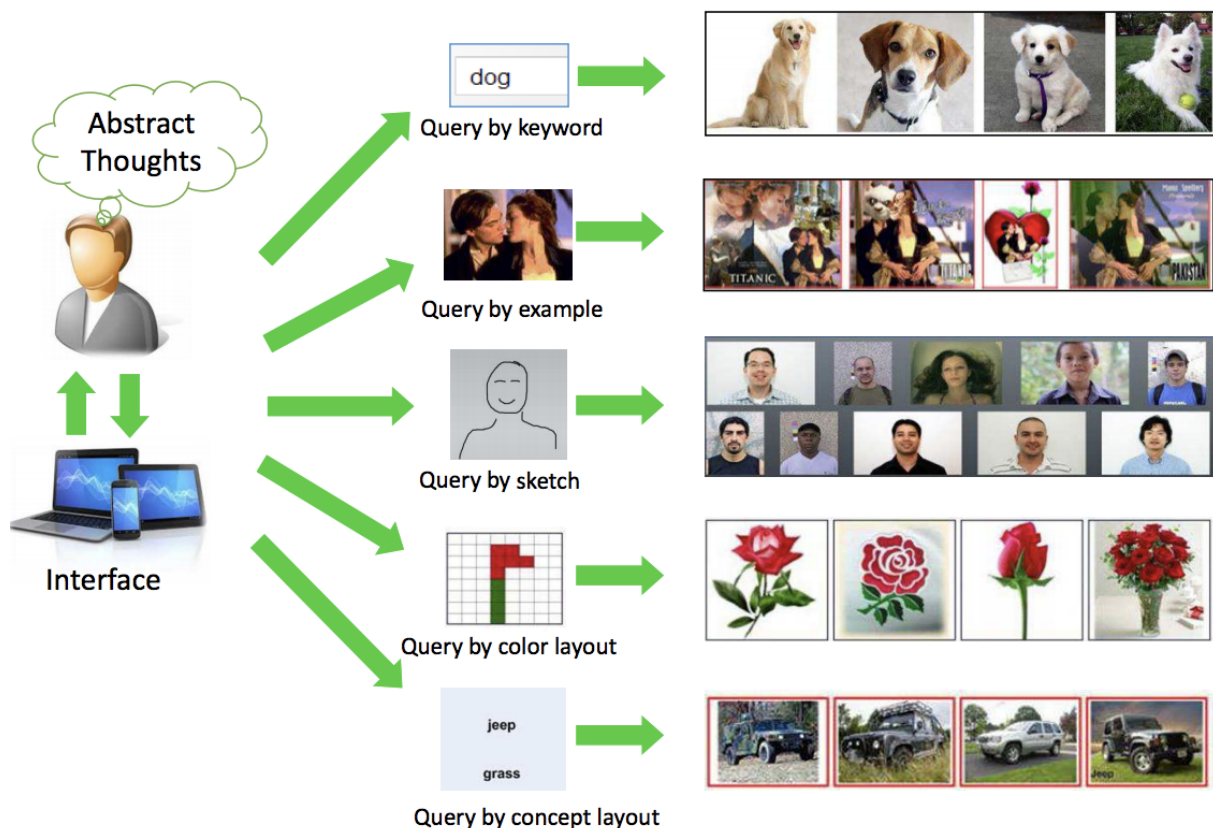


图 1：具有相应检索结果的不同查询方案的示意图。

除了手工制作的特征之外，近年来基于学习的特征也得到了发展。这些系统的主要问题是查询的制定是以输入图像的形式进行的，该图像被转换为特征表示；基于此，对数据集中的图像进行排名。这种方法从根本上说是不完善的，因为它要求用户最初有一个类似的图像。还有一些方法通过输入文本在网页上搜索图像，例如Google。但在这里，更多的是关注文本检索（text retrieval）。谢等人^[1]，2014 并且返回与检索文本相关的图像。这本质上是一种弱学习方法，问题是以间接方式解决的。

因此，以上所有系统在努力提供基于文本的语义图像检索的有效解决方案方面均存在不足。为此，我们提出了一种方法，在同一空间中语义嵌入文本和图像，以便根据文本查询更高效、更准确地检索图像。

3 我们的方法

在本节中，我们将描述我们想要尝试的两种方法，以解决上面详述的问题。

3.1 基于标题的检索

在图像生成描述领域，已经开展了大量的工作。图像-描述生成问题涉及生成描述图像中场景的单一短语。最先进的卷积神经网络架构用于提取基于图像的特征，并将输入转换为一个富含语义的向量。该向量可以用于训练不同的任务，如分类或检测，也可以作为其他任务管道的一部分。为了我们的目的，我们将使用它来初始化一个语言模型。该语言模型是一种基于长短期记忆（LSTM）的架构，旨在构建对语言的理解。

$$P(S|h, x, S)$$

这个词在哪里

$$t-1 \quad t-1 \quad t$$

$$th$$

，是 LSTM 在时间的隐藏状态

$$t-1$$

$$t-1 \quad x$$

并且是 LSTM 单元格在时间的输入

$$t$$

$$t$$

在每个时间步，softmax 用于输出词汇表中所有单词的概率。

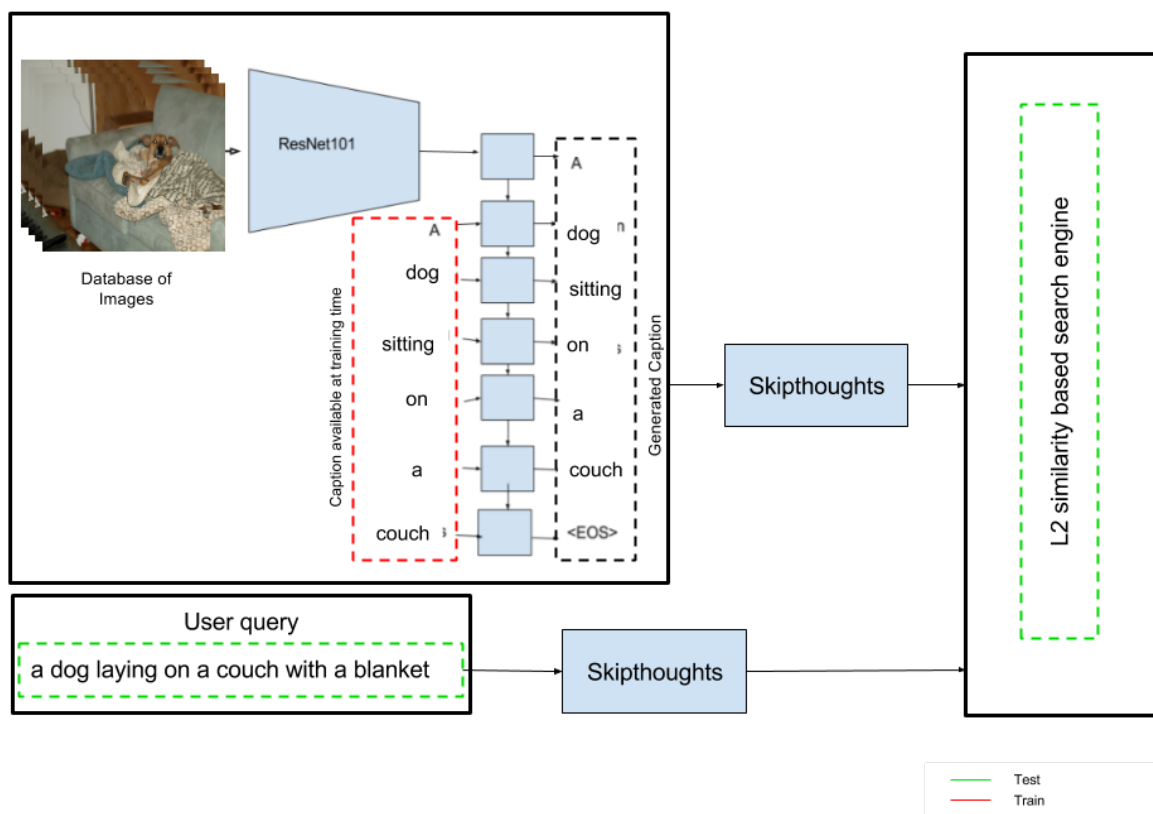


图 2：基于字幕的图像检索管道的端到端检索系统。

对于我们的实验，我们使用 ResNet - 101 (他等。 , 2015) 作为特征提取的骨干网络。我们使用从MS-COCO目标检测预训练得到的权重初始化该网络。 Lin et al. , 2014) 任务，并将其微调以适应MS-COCO数据集上的Caption生成任务。生成Caption后，我们使用skip-thought (跳过句子) 方法。 Kir os 等人。 , 2015 将字幕转换为向量嵌入的模型。一旦将字幕转换为向量，我们也会对用户提供的查询进行相同的处理。然后通过最小化查询向量与图像关联的字幕之间的L2距离来进行图像检索。

$V(x)$ 这个向量 i

$$V(x) = CNN(x) \quad (1)$$

$i = 1, \dots, N$

where $\{x_i\}$ 是中的图像之一

2. 图像图像检索的标题也被投影到一个语义特征表示空间中，使用skipthoughts模型。

3.2 嵌入空间检索

此模型由以下组件组成：

1. 使用预初始化的先进卷积神经网络 (ResNet-101) 从图像帧中提取语义信息，构建表示图像内容的特征。我们称之为

$$U(c^k) = \Gamma(c) \quad (2)$$

c^k 是

哪里是 - 相关的标题

然后对这两个特征应用投影以创建一个嵌入空间，并且同时可以通过模型这些向量的L2距离来进行学习。

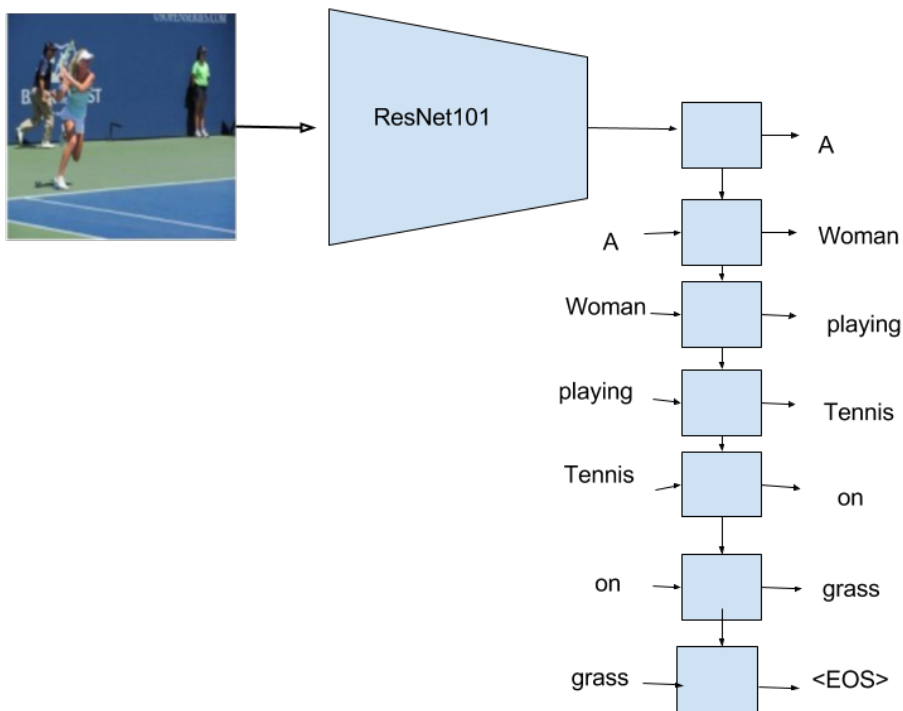


图 3：标题生成

4 实验

4.1 标题生成

对于我们的第一种方法，我们首先训练一个 Caption 生成模型。这里的思路是将图像转换为其语义丰富的等价文本表示。生成的 Caption 随后被转换成向量（使用 skipthoughts 方法，稍后会解释）。因此，每张图像间接地被转换为其语义丰富的特征表示。

4.1.1 数据集

我们在 MS COCO 数据集上训练了一个标题生成模型 (Lin et al., 2014)。这个数据集包含复杂的日常生活场景图像，这些场景中包含自然环境下的常见物体。物体通过实例分割进行标注，以辅助精确的目标定位。该数据集包含 2.5 万个标记实例，共计 32.8 万张图像，涵盖了 91 种 4 岁儿童能够轻易识别的物体类型。每张图像由人类注释并附有 5 个描述。2014 年的版本包含 8 万张图像。

$$E(V(x)) = W V(x) + b \quad (3)$$

and

$$E(W(c^k)) = W U(c^k) + b \quad (4)$$

在哪里都 $E_v(V(x_i))$ and $E_u(W(c^k)) \in \mathbb{R}^d$.

4. 目标函数为 de fi Ned，如下所示：-

$$L(E(U(c^k)), E(V(x))) \quad (5)$$

$$\frac{1}{2} \|E(U(c^k)) - E(V(x))\|_2^2 \quad (6)$$

于是我们得到了一个空间，在这个空间中，图像与其相关的说明文字彼此靠近。然后可以在这个空间中投影用户的查询，并根据它们的 L2 距离检索图像。

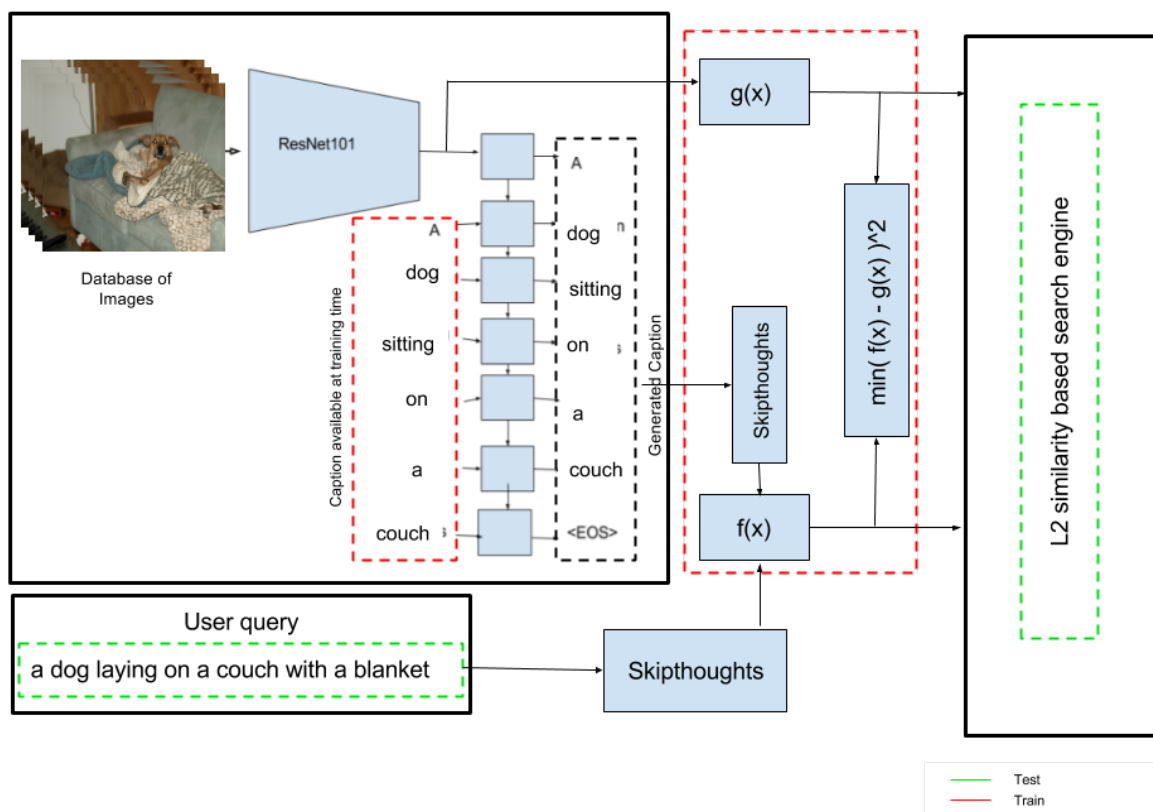


图 4：嵌入空间图像检索的端到端检索系统

用于训练的数据量为40k，用于验证的数据量为20k，用于测试的数据量为20k。我们遵循相同的设置。对于图像检索，我们使用与caption生成相同的测试集，即MS COCO数据集中20k张图像。对于基于caption的检索，除了训练caption生成模型本身外，没有专门的训练阶段。对于基于嵌入的检索模型，使用MS COCO数据集中80k张图像作为训练集。一旦提取了向量，我们定义它们之间的L2损失。如方程所示。

运行 70k 次迭代，CIDEr 评分为 0.7。

For the embedding space learning problem, the setup is similar to above. The loss defined in equation 5

使用Adam优化器和批量大小为128个向量来最小化损失。整个训练集包含80,000张图像，与MS COCO数据集用于生成字幕的训练集数量相同。学习率为1e-3，并在整个训练过程中指数衰减。训练允许运行直到收敛。我们还发现将梯度削减到标准是有用的，并且可以更早期地稳定训练。

4.1.2 设置和培训

对于Caption生成模型，我们使用批量大小为128和尺寸为224x224的图像。训练在Nvidia Titan-X GPU上进行。我们使用具有动量的Adam优化器，并采用默认设置。

$\beta = 0.99$ $\beta = 0.9999$

(and). The learning rate

设置为1e-3并以指数方式衰减。CNN的学习率设置为小一个数量级。LSTM的梯度在训练过程中被裁剪至10。训练允许进行到收敛。截至本报告撰写之时，我们能够

4.2 Skipthought 向量

skipthought向量是一种无监督学习通用分布式句子编码器的方法。通过书籍中连续的文本内容，首先训练一个编码-解码模型，尝试重建经过编码段落周围的句子。具有语义和句法属性相似性的句子因此被映射到类似的向量表示。接下来介绍一种简单的词汇扩展方法来编码单词。

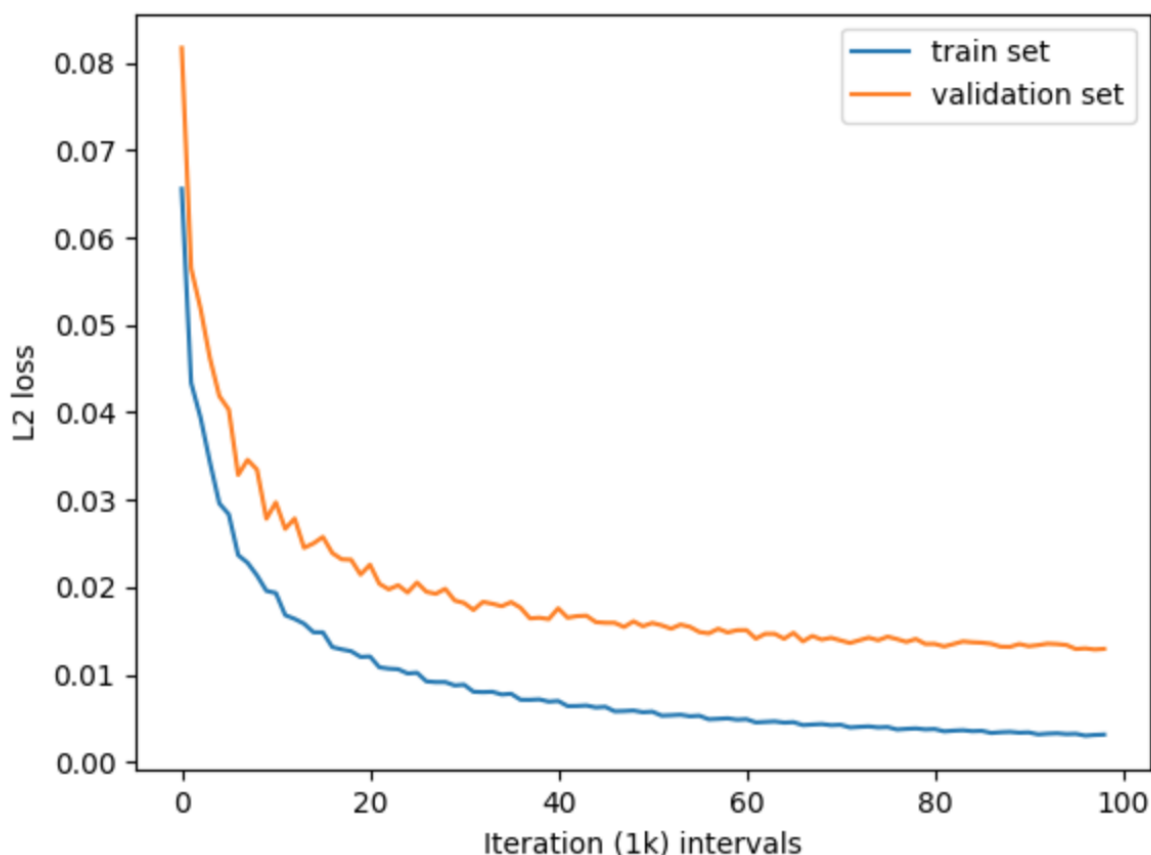


图 5：嵌入学习任务的训练和验证集的 L2 损失曲线

那些没有被视为培训的一部分，使我们能够将词汇量扩展到一百万词。最终结果是一个即用型编码器，可以生成高度通用的句子表示，这些表示既稳健又在实际应用中表现良好。对于skipthought向量生成，我们使用

完全相同的设置与原始论文一致。我们也没有自行训练Skip-Thought Vectors模型，目前使用作者官方GitHub页面提供的模型。有关该模型训练的更多信息，请参阅他们的论文。Kiros 等人。 , 2015)

4.3 用于检索的嵌入空间

为了同时学习图像和文本的联合嵌入空间是一项具有挑战性的任务，但这一问题已在多项工作中被尝试解决。我们从 (弗罗姆等人。 , 2013)，并通过两个并行网络学习联合空间嵌入，这些网络采用CNN (包含在caption生成模型中的一部分) 的输出作为输入。这些向量已经非常具有语义丰富性，但随后会被进一步嵌入-

$E(V(x))$
ded 到一个向量 $\in \mathbb{R}^d$
在维度中 -
 v_i

局部嵌入空间。在训练时，skip thought向量采用由caption生成模型生成的captions。

$E(W(c))$
向量 $k) \in \mathbb{R}^d$ 对于生成的 cap -

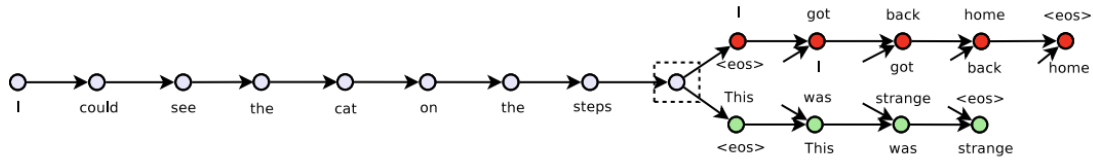
u_i^c
tion k_i
我们训练集中的图像也是

5. 评价

d
5.1 定义
嵌入到维度空间中。

为了评估管道的不同部分，采用了不同的定量指标。caption生成使用MS COCO服务器进行评估，该服务器采用包括BLEU、Meteor、Rouge-L和CIDEr等指标。我们将模型与MS COCO caption生成排行榜上的现有最先进的系统进行了比较。

为了评估我们的图像检索系统，我们提出了以下最相关的指标。我们计算在@ k @ k 三个不同的层次。精度 (p) 是指 -



(s_i, s_{i+1}, s_{i+2})

图 6：skip-thoughts 模型。给定一个元组 连续的句子，第 i 个

s_i, s_{i+1}, s_{i+2}

s_i, s_{i+1}

一本书的句子，句子被编码，并试图重建前一个句子和下一个

s_i, s_{i+1}

s_i

句子。在此示例中，模型重建了句子“我在家里”和“我可以看到猫”

s_{i+1} 我们的模型输出 0.928 0.320 0.693 1.092

台阶。这很奇怪。未附加的箭头连接到编码器输出。颜色指示哪些组件共享参数。{ eos } 是句子的结尾标记。

表 1：本文所述字幕生成模型的定量评估与 MS COCO 排行榜上的最新技术。

fi Ned 如下：-

$$\frac{tp @ k}{tp @ k + fp @ k} =$$

$tp @ k$

其中真正的积极事件 () 是 de fi Ned

k

When out of the images that the system retrieve, the correct image is retrieved. And the false position -

$fp @ k$

活动事件 当系统超出图像范围时，如果系统允许检索出 none 个正确的图像，则定义为正确的图像标题配对属于 MS COCO 中提供的图像标题。

C

数据集。因此给定与关联的标题

j

x

测试集中的 20k 图像中的图像

j

MS COCO，我们同时执行基于字幕的重新检索和嵌入空间检索。

C

LOW. We take the caption associated with the

j

x

图像 1 并嵌入标题。

从上面的表 2 中我们可以看到，基于嵌入的检索性能略低于基于字幕的检索性能。

MS COCO 数据集的检索性能。所有图像组成检索数据集也通过我们的管道处理以生成相应的嵌入。尽管基于嵌入的检索中有更多的优势，这是因为模型的目标函数是使用嵌入来检索图像。

j

C

由我们的系统检索查询，那么

j

$tp @ k$

被视为真正的积极事件 for the

$p @ k$

模型	$p@1$	$p@3$	$p@5$	时间 (秒)
ESR	68.3	85.7	91.2	2.1
CBR	72.9	84.9	90.5	1.7

表 2：两者的定量评估 - 本论文中描述的图像检索方法 - per.

但是检索到的第一张图片可能不是最近的，因此正确的图片可能会出现在前五张图片中，但可能不是第一张检索到的。请注意，上述检索操作是

没有 GPU 优化，这将极大地改善这两种方法的时机。

5.2 Qualitative

在图 7 中，我们将比较两个系统（即嵌入空间检索和生成式描述检索）在同一查询下的输出，以 qualitative 的方式评估哪个系统表现更好。

5.3 结论与未来工作

7

如图所示，两种方法都能获得良好的定性结果。有趣的是，这两种模型不仅捕捉到了正确的对象，还捕捉到了场景中各种对象之间的正确语义关系。

在未来，我们希望探索基于三元组损失的学习嵌入方法以应用于 Embedding Space Retrieval。这将使我们的模型不仅能够了解关于 caption 的相关信息。



(a) 嵌入空间检索的输出



(b) 基于字幕的检索输出

图7：查询“一只狗在房间内观看电视”的两种方法的输出。有趣的是，这两种模型不仅捕捉到了对象的具体语义，还捕捉到了它们之间的关系。在上述结果中，我们看到不仅有动物和电视出现在场景中，而且在这两张图片中，这些动物实际上都在观看电视。

相似的图像与不相似的配对类似，这将大大规范模型，并允许学习更多语义丰富的特征。另一个使模型更加用户友好的方向是整合知识图谱的概念，这将使我们能够关于实体进行推理。这可以让我们回答类似于“Hector正在看电视”而不是“一只狗正在看电视”的查询。我们还希望对检索系统进行GPU优化，以使其在两种技术方面显著更快。

参考文献

Dzmitry Bahdanau, Kyunghyun Cho 和 Yoshua Bengio。2014 年。通过联合学习对齐和翻译实现神经机器翻译。CoRR abs/1409.0473. <http://arxiv.org/abs/1409.0473>.
 Andréa Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. 设计：一种深度视觉语义嵌入模型。In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahra-man, and K. Q. Weinberger, editors, 神经信息处理系统的进展 26, Curran Associates, Inc., 第 2121 - 2129 页。

<http://papers.nips.cc/paper/5204-deep-visual-semantic-embedding-model.pdf>

何开明、张祥玉、任少卿、孙坚。2015 年。基于深度残差学习的图像识别。CoRR abs/1512.03385.

<http://arxiv.org/abs/1512.03385>.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual - semantic alignments for generating image descriptions. In IEEE 计算机视觉与模式识别会议论文集 第 3128 - 3137 页。

Ryan Kiros, Yukun Zhu, and Ruslan Salakhutdinov. 2015. 跳过思想向量。CoRR abs/1506.06726.

<http://arxiv.org/abs/1506.06726>.

林宗义、迈克尔·梅尔和谢尔盖·贝隆吉。2014 年。Microsoft COCO：上下文中的常见对象。CoRR abs/1405.0312.

<http://arxiv.org/abs/1405.0312>.

David G. Lowe. 2004. 来自尺度不变关键点的独特图像特征。Int. J. Comput. Vision 60(2):91-110. <https://doi.org/10.1023/B:VISI.0000029664.99615.94>.

J. Sivic and A. Zisserman. 2003. 视频 google：一种用于视频中对象匹配的文本检索方法。In 第九届 IEEE 计算机视觉国际会议论文集, pages 1470 - 1477 vol.2.

<https://doi.org/10.1109/ICCV.2003.1238663>.

H. Xie , Y. Zhang , J. Tan , L. Guo 和 J. Li
。 2014 。

用于图像检索的上下文查询扩展 .

IEEE 多媒体学报 16(4):1104– 1114. <https://doi.org/10.1109/TMM.2014.2305909>

Kelvin Xu, Jimmy Ba, and Ryan Kiros. 2015. 显示 , 参加和讲述 : 带有视觉注意力的神经图像字幕生成 . *CoRR* abs/1502.03044.

<http://arxiv.org/abs/1502.03044> .

周文刚、李厚强、齐天。 2017 年。

基于内容的图像检索的最新进展 : 文献调查 . *CoRR* abs/1706.06064.

<http://arxiv.org/abs/1706.06064> .