加州大学圣巴巴拉分校

项目报告

图片说明

作者：
赛尼基马拉姆

导师：
Prof Yuan Fang Wang

2018 年 8 月 12 日

1 目标

开发可以为给定图像生成适当标题的原型。了解计算机视觉和自然语言处理中的图像字幕问题。演示如何为系统构建不同的组件、预处理和培训程序。

2 简介

检测图像中存在的对象是计算机视觉中的一个基本问题。对象以及它们之间的关系定义了图像。查找图像中存在的对象之间的关系需要检测对象并用自然语言表示它们之间的关系，以获得如图 1 所示的图像说明。图像说明需要用户了解计算机视觉（CV)和自然语言处理（自然语言处理）。

在这个项目中，我们着眼于基于卷积神经网络 (CNN) 和循环神经网络 (RNN) 的生成模型来生成图像说明。该模型被训练以最大化给定训练图像的目标描述句子的可能性。 MS COCO 图像/文本数据集上的实验显示了模型的准确性以及它仅从图像描述中学习的语言的流畅性。

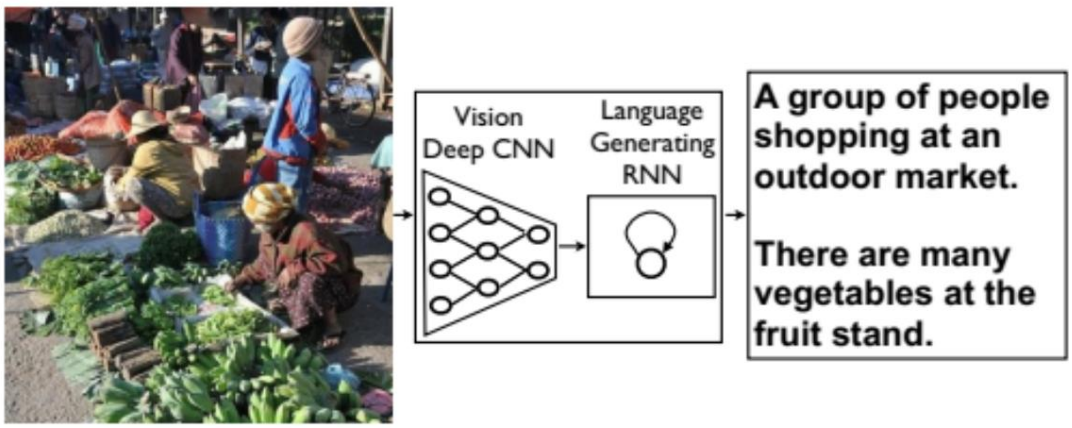

图 1:NIC 模型（视觉 CNN 后跟语言生成 RNN）

该项目基于 CVPR2015 论文“Show
和告诉:神经图像字幕生成器”。

3 型号

当前模型的灵感来自于 NLP 中机器翻译（MT)中通常使用的 Sequence2Sequence 模型。 RNN 用于通过保留时间信息将文本嵌入向量空间。在传统的机器翻译中，输入和输出是不同语言的文本。因此，RNN 被用作 Sequence2Sequence 模型中的编码器/解码器。编码器将输入文本编码到公共向量空间，解码器解码向量空间以在 MT 中生成输出文本。对于图像字幕，输入是图像，输出是文本格式的字幕。

因此，我们需要一个可以将图像编码到公共向量空间的模型。因此，RNN 被 CNN 取代，作为我们改编的 Sequence2Sequence 模型中的编码器，如图 2 所示。
我们的模型是一个基于神经的网络，由一个视觉 CNN 和一个生成语言的 RNN 组成。 CNN 完成的图像编码作为 RNN 模型（LSTM)的第一个输入。

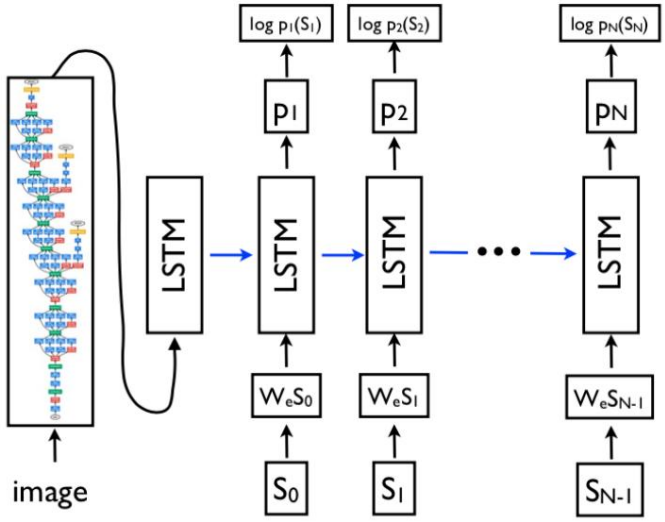

图 2:结合 CNN 图像嵌入器的 LSTM 模型。

.

4 完成的工作

以下部分描述了我们在实现完整模型时所做的工作。我们使用了 tensorflow 和 Python。我们实现的代码可以在这里找到。本节讨论如何构建模型,下一节讨论所涉及的预处理和训练过程。

4.1 编码器

Imagenet 上预训练的 vgg16 CNN 模型用作编码器。 Vgg16 模型由 5 个卷积层、2 个全连接层和 1 个 softmax 层组成,如图 3 所示。

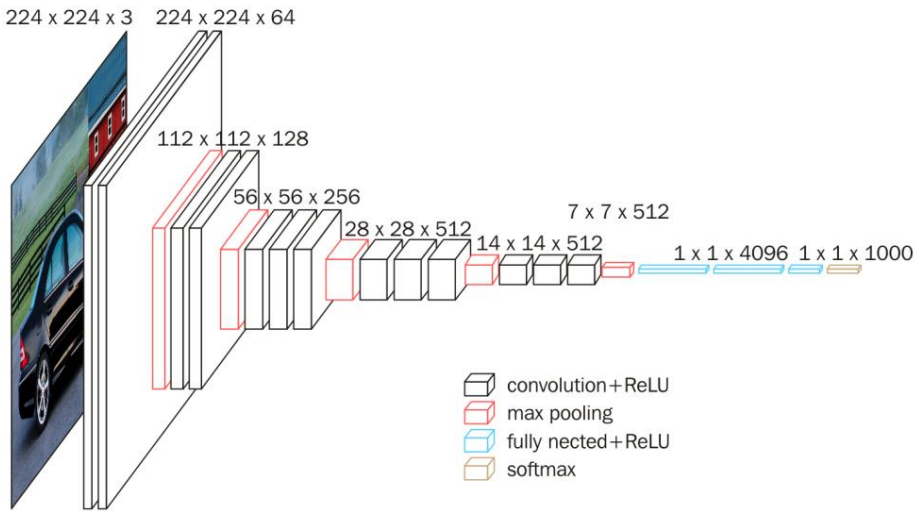

图 3:VGG16 模型

CNN 的参数。模型需要在 tensorflow 中构建,我们需要确保

我们构建的模型和预训练的模型文件中存在的不同参数的名称范围应该完全匹配,以使 tensorflow 中的恢复操作成功。
CNN 的输入是大小为 224*224*3 (RGB) 的图像。第二个全连接层的大小为 4096 的输出被获取,因此我们获得了有关图像中存在的不同对象的更多信息,并将其压缩为 512 的大小并馈送到解码器。

4.2 解码器

如前所述,RNN 用于通过保留时间信息将向量解码为文本。我们使用 LSTM 作为 RNN 模块来缓解梯度下降消失的问题。我们有一个固定长度的 LSTM 单元,因为我们将对每个字幕可以拥有的最大单词数设置一个阈值。需要对文本进行大量预处理,因为任何计算模型都只处理向量。为此,完成了以下预处理步骤。

1. 添加开始和结束标记:由于所有字幕的长度不同,我们需要知道字幕的开始和结束。因此,我们为每个字幕使用一个公共的开始和停止标记。

2. 填充:因为我们在解码器中有恒定数量的 LSTM 单元,每个字幕都必须经过。如果标题中的单词数小于 LSTM 单元数,那么我们用 < unk > 填充其余单词。

3. 词到索引:因为任何计算模型都只处理向量。我们为每个唯一的单词分配一个唯一的索引。单词的索引取决于词汇表中存在的唯一单词的数量。为此,我们首先遍历数据集中存在的所有标题,并列出其中的唯一单词,并为每个单词分配一个索引。在这一步之后,我们可以用每个单词对应的索引的形式来表示一个标题。

4. 嵌入:如果我们用索引来表示每个单词,并用一个热向量来表示每个索引。然后每个向量将是大小词汇,任何计算都将导致稀疏矩阵计算。因此,使用了降维技术,其中每个索引现在表示为 512 维向量。这在 NLP 中通常称为 word2Vec。因此创建了一个嵌入矩阵（We） ,它为每个索引提供了 512 维向量。这就是 CNN 的输出被压缩成 512 大小的向量的原因。

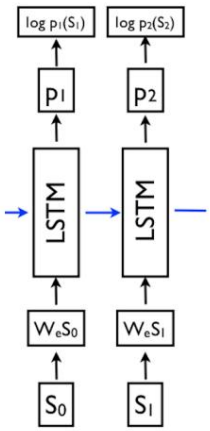

图 4:RNN 模型

在完成上述预处理步骤后,我们将输入提供给 RNN 模型。在图 4 中， S0表示一个词对应的索引。 W0S0对应于索引对应的嵌入向量。当将 LSTM 的先前隐藏状态馈送到当前 LSTM 单元时,这会导致当前输出和当前隐藏状态。 LSTM 的初始隐藏状态为零。我们得到的输出将是一个维度为 512 的向量。这需要转换回单词的索引才能得到预测的输出单词。

在训练期间,我们在每个 LSTM 输入处使用输入词来预测输出,但在评估时,前一个 LSTM 的输出词作为输入馈送到当前 LSTM。

5 实验设置

5.1 数据集

我们使用了包含图像/标题的 MS-COCO 数据集。我们首先对数据集进行预处理,通过解析包含图像 ID 和标题的 cpation JSON 文件来获取图像及其相应的标题。资料说明

数据集	训练	有效	测试
MSCOCO	82783	40504	40775

表 1:MSCOCO 数据集

5.2 培训

为训练完成以下步骤。

1. CNN 组件的初始化权重到预训练模型（VGG16）。
2. 创建词嵌入矩阵。
3. 交叉熵作为损失函数
4. Adam Optimizer,初始学习率为 0.0001,dropout 保持概率为 0.9
5. 所有的RNN输入都是右填充的。最大字幕长度为 20。
6. 词嵌入到 512 维向量中。
7. 使用 32 的 Batch 大小运行 5 个 epoch。
8. BLEU 分数用作评估指标。

CNN 参数在训练期间不会更新。训练时更新的参数是 Word Embedding 矩阵,LSTM 参数。由于每个时期运行 4 小时。由于计算资源的限制,我们只运行了 5 个 epoch 的系统。因此,在两个时期之间不计算验证 bleu 分数以提前停止以避免过度拟合。

6 个结果

6.1 指标

BLEU 分数用作评估指标。由于系统仅运行 5 个 epoch。该指标是根据 5 个 epoch 后生成的模型为验证数据集计算的。以下是 MS COCO 数据集的 BLEU 分数的各种变化。 BLEU 分数逐字比较地面实况和预测文本。 BLUE-1 检查一对一的对应关系，而 BLUE-2 检查真实文本和预测文本中给定词的上下文词。

公制	BLUE-1	BLUE-2	BLUE-3	BLUE-4
值	62.9%	43.6%	29.0%	19.3%

表 2:数据集上的分数

6.2 输出

以下是图像字幕系统生成的一些测试输出。 . .

6.2.1 来自验证 MS-COCO 数据集的图像

可以看出，在左图中，模型能够检测到街道上的公共汽车，包括它的颜色。中间的图像与作为蛋糕的比萨饼有点混淆。在最右边的图像中，该模型能够预测冲浪，而男人和女人之间几乎没有混淆。

6.2.2 来自 flickr 数据集的图像

从左图中可以看出,该模型能够检测到滑板运动,但存在将男孩视为男人的小错误。在中间的图像中,它能够检测到一个婴儿,并且由于这个男人并不清晰可见,所以也许大部分训练集可能有女人抱着一个婴儿,因此模型将它检测为一个女人。在最右边摆动的腿被认为是蝙蝠,而不是女性模型将其检测为男性。

6.2.3 我们拍摄的图像

这些是我们为评估模型而拍摄的图像。左图是研究生宿舍公共休息室的照片。该模型能够检测到椅子排列在一起时的沙发。仔细观察,窗户看起来就像一台正在播放图片的电视。因此,模型将其检测为带沙发和电视的客厅。在中间图像中,模型能够检测到周期和人。在正确的图像模型中,能够检测到准确描述图像的人、沙发和笔记本电脑。

6.3 结论

从这些实验中可以清楚地看出,该模型能够在物体清晰且很少的情况下为图像添加字幕。通过很少的 epoch 运行,该模型能够比基于正式规则的图像字幕方法更好地获得图像的准确描述。