# 使用深度递归神经网络将视频翻译成自然语言

## Translating Videos to Natural Language Using Deep Recurrent Neural Networks

### ——Subhashini Venugopalan、Huijuan Xu、Jeff Donahue 、Marcus Rohrbach、Raymond Mooney 、Kate Saenko（Computer Science）

**一、科学问题**

**1.1 本文所涉及科学问题**

（填写所读论文涉及的科学问题，所谓科学问题一般为本论文的大方向）

使用深度递归神经网络使视频生成文本

**1.2 同行专家如何解决**

（个人感觉主要为国内外当前研究现状，一般在论文的引言中都有介绍。）

Donahue等人的研究。他们将模型的一个版本应用于视频生成文字领域，但是没有提出端到端的单一网络，而是运用中间角色表示。此外，他们仅在一小组预定义对象和参与者的烹饪视频的狭窄领域中显示了结果。

**1.3 本文所解决的问题**

（解决什么问题，效果如何，一般摘要及总结中都会有）

本文解决了计算机视觉符号的基础问题。能够创建出词汇量较大的开放域视频的句子描述。

**1.4 本文解决方案效果**

（使用本文的方法可以达到怎样的效果，一般在总结中会提到）

该模型利用神经网络对整个过程从像素点到句子进行描述，能够对整个网络进行训练和调整。在广泛的实验评估中，证明了此方法比相关方法产生更好的句子。我们也证明了利用图像描述数据可以提高性能。

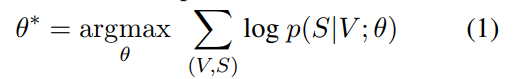
**二、研究内容**

**2.1 理论与方法介绍**

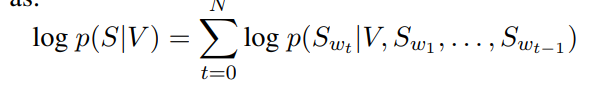
（论文主要研究内容的提出，主要技术路线、理论与方法介绍）

工作原理是首先在一个图像上应用一个特征变换来生成一个固定维度的向量表示。然后，使用序列模型，特别是递归神经网络(RNN)，将向量“解码”成一个句子(即单词序列)。

在给定视频V和模型参数为（theta）的情况下，最大化句子S的log可能性，从而确定给定视频最可能的描述



假设S的生成模型按顺序生成每个单词，则句子的log概率由单词的log概率之和给出，可以表示：



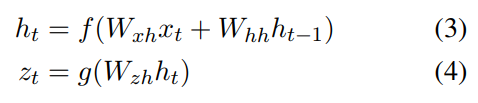
序列模型适用于，然后我们选择一个递归神经网络。由（theta）参数化的递归神经网络将输入xt和前面看到的表示为隐藏层或记忆的单词，ht-1映射到输出zt和使用非线性函数f更新状态ht:



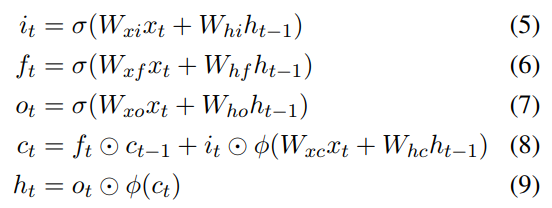
当（h0=0）。使用非常有效的长短时记忆(LSTM)网络作为序列模型，因为它在语音识别、机器翻译和生成图像的句子描述等这一相关的任务上表现出了优越的性能。

使用两层LSTMs(一个LSTM堆叠在另一个LSTM之上)：

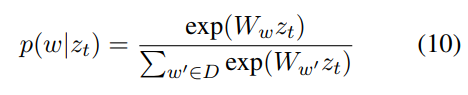
标准的RNNs学会映射输入序列(x1,…,xt)到一个隐藏状态序列(h1,…,ht)，以及从隐藏状态到输出的一个序列(z1,…,zt)，基于以下的递归性:



LSTM模型的核心是一个记忆细胞c，它在每个时间步长上，编码到该步长之前观察到的输入的知识。该细胞由门调节，门均为s型函数，范围是[0,1]，可多次应用。这些门决定了LSTM是将值从门中保留(如果该层计算为l)还是丢弃(如果计算为0)。考虑到它的当前输入(xt)、遗忘门(f)允许LSTM忘记它以前的细胞(ct-1)以及输出门(o)决定将多少记忆传输到隐藏层(ht)，所有这些都是使LSTM能够学习复杂的长期依赖关系。LSTM的递归定义为:



LSTM的第一层xt是视觉特征向量与前一层编码的单词。对于LSTM xt的第二层是第一层的zt。因此，推理也必须按照h1=fw(x1,0)，h2=fw(x2, h1)的顺序进行，直到模型在最后一步T发出语句结束(EOS)令牌。在我们的模型中，第二层LSTM单元的输出(ht=zt)用于获得发出的单词。应用Softmax函数，得到了词汇D中单词w的概率分布

。|

模型：

HVC。是Thomason描述的最高视觉置信度模型。该模型使用强大的视觉检测器来预测45个主语、218个动词和241个宾语的可信度。

FGM。Thomason还提出了一种因子图模型(FGM)，该模型利用因子图将从文本语料库中挖掘的知识与从HVC模型中提取的视觉置信度相结合，并根据每一种形式的概率推理来确定最可能的主语、动词、宾语和场景元组。然后，他们使用一个简单的模板从元组生成一个句子。在本文中，我们将模型的输出量与由HVC和FGM模型预测的主语、动词、宾语进行了比较，由主谓宾三元组生成句子。

我们的LSTM模型。我们提出了四个主要的模型。LSTM—YT是我们在YouTube视频数据集上训练的基础双层LSTM模型。LSTM—YTflickr是在Flickr30k数据集上训练的模型，并在YouTube数据集上进行微调，如本节所述3.3。LSTM-YTcoo首先在COCO2014数据集上进行训练，然后在视频数据集上进行微调。我们的最后一个模型LSTM-YTcocoflickr基于Flickr和COCO模型的组合数据进行训练，并在YouTube上进行调优。为了比较图像数据集和YouTube数据集之间内容的重叠，我们使用仅针对Flickr图像(LSTMflick)和COCO图像(LSTMcoco)训练的模型，并评估它们在测试视频中的性能。

**2.2 验证分析与实验效果**

（论文中的实验分析和实验效果）

序列的LSTM直接输出句子，所以我们从生成的句子的依赖关系解析中提取主谓宾。在表1和表2中，我们给出了主谓宾单词的准确性，并将我们的模型与任何有效的真值三元组和在每个视频的人类描述中发现的最频繁的三元组的性能进行了比较。

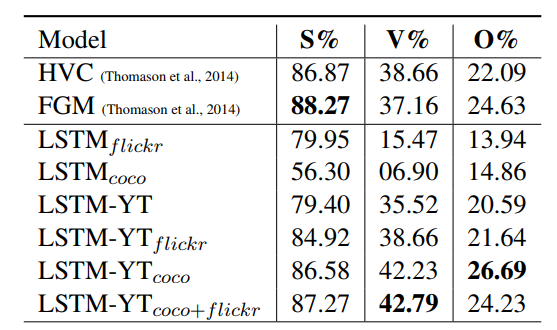


表1：主谓宾准确性：主谓宾的二进制精度与基本真实描述中的任何有效主谓宾三元组相比。我们使用依存句法分析分析其模型从句子输出中提取主谓宾的值。如果模型给出的主谓宾与人类描述一致，那么就是正确的。

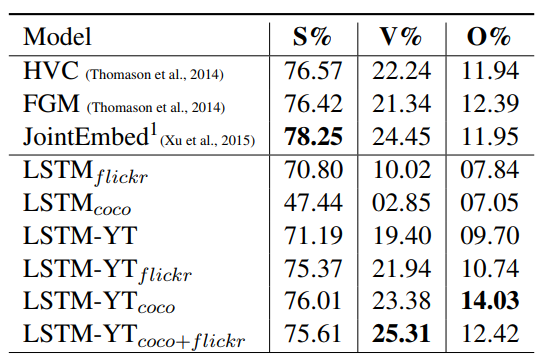


表2：主谓宾准确性：SVO的二进制准确率与最常见的SVO三元组的基础真值进行比较。我们通过使用依存句法分析，提取一个输出的句子中的SVO值。只有在输出被人类描述中最常被提及的SVO时，该模型才是正确的。

句子生成。为了评估生成的句子，我们使用BLEU和METEOR对所有基础的真实句子进行评分。

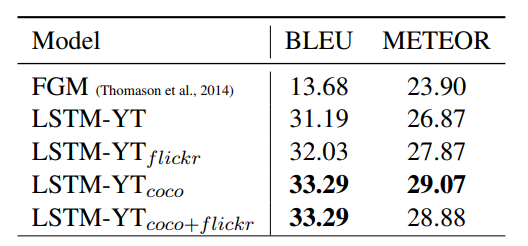


表3：BLEU评分为4分，与自动评估指标的METEOR评分比较生成的质量。所有值都报告为百分比(%)。

人类评估。

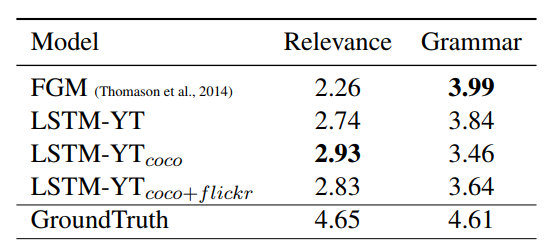


表4：人类评价平均得分。根据句子与给定视频的相关性，句子的评分在1到5之间。句子的语法正确性评分在1到5分之间。数值越大越好。

单独的帧。为了评价均值池的有效性，我们对视频中的单个帧进行了训练和测试。

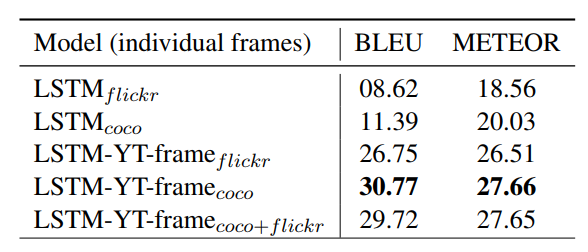


表5:BLEU评分与METEOR评分比较Flickr30k和COCO上训练的模型生成句子的质量，并在视频的随机帧上进行测试。LSTM—YT—帧模型在Youtube视频数据集中的单个帧上进行了微调。所有值都报告为百分比(%)。

**三、论文存在问题及后续研究重点**

**3.1 论文存在问题**

在更好地利用视频中的时间信息方面存在不足。

**3.2 后续研究重点**

着重于视频的时间信息方面，改进更加优良的算法生成文本。

**四、该问题相关研究成果**

**4.1 相关论文一**

**（1）题目**：Generating image descriptions using dependency relational patterns（利用依赖关系模式生成图像描述）

**（2）作者介绍**：Ahmet Aker and Robert Gaizauskas. 2010.

**（3）摘要**：本文提出了一种通过汇总包含多个web文档的地理标记图像与图像位置相关的信息方法来自动地理解图像。因依赖关系模型而存在偏差，包含通常提供的功能的句子的模式模型和不同的场景类型，如教堂、桥等。我们的结果表明依赖模式模型会导致显著较高ROUGE分数，这两种n-gram语言模型的得分都比以前的维基百科总结研究报告要高。还使用依赖关系模式生成的总结与没有依赖模式生成的总结相比，可以生成更具可读性的总结。

**4.2 相关论文**

**（1）题目**：Collecting Highly Parallel Data for Paraphrase Evaluation(为释义评估收集高度并行的数据)

**（2）作者介绍**：David L. Chen and William B. Dolan. 2011.

**（3）摘要**：由于缺乏标准的数据集和评估指标，释义领域无法取得过去15年机器翻译领域所享有的那种快速进展。我们通过提出一种新的数据收集框架来解决这两个问题，该框架以相对较低的成本和大规模生成高度并行的文本数据。该数据高度并行的特性允许我们使用简单的n-gram比较来度量候选释义的语义充分性和词汇差异。除了简单和高效的计算,实验结果表明,这些指标与人类的判断紧密相关。

**4.3 相关论文三**

**（1）题目**：A deep convolutional activation feature for generic visual recognition（一种用于通用视觉识别的深度卷积激活功能）

**（2）作者介绍**：Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2013.

**（3）摘要**：近二十年来，意大利理性主义建筑理论通过数字建模进行了研究。在许多案例研究中，模型的产生和随之而来的三维视图的表示常常被认为是架构研究的结果。实际上，作为一个关键工具的数字模型必须被视为一个“起点”。

# 弱监督密度的视频描述

## Weakly Supervised Dense Video Captioning

### Zhiqiang Shen, Jianguo Li , Zhou Su, Minjun Li，Yurong Chen , Yu-Gang Jiang, Xiangyang Xue（2017 IEEE Conference on Computer Vision and Pattern Recognition）

**一、科学问题**

**1.1 本文所涉及科学问题**

自动地描述一个包含多种信息和不同描述句子的视频片段。该方法不需要对视频区域序列对应的句子进行显式标注，只基于弱视频的句子标注。

**1.2 同行专家如何解决**

视频的多句子描述近年来在各种作品中得到应用。这些作品大多侧重于生成一个长描述，首先用动作定位或不同级别的细节对视频进行时间分割，然后为这些片段生成多个描述，并将其与自然语言处理技术连接起来。然而，这些方法只考虑了时间分割，忽略了帧级区域的集中和区域级对象的运动序列。Yu等人的既考虑了时间因素，也考虑了空间因素，但仍然忽略了句子与视觉位置的关联或对齐。

**1.3 本文所解决的问题**

自动地描述一个包含多种信息和不同描述句子的视频片段。

**1.4 本文解决方案效果**

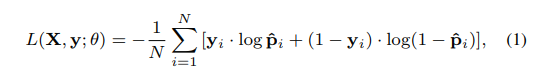
我们的方法可以产生多种信息和多样化的标题。在MSR-VTT挑战中，我们最好的符号描述输出以显著的优势超过了最先进的方法。

**二、研究内容**

**2.1 理论与方法介绍**

本文试图利用时间和空间信息，进一步探索句子和区域序列之间的对应关系，以便更准确地建模。采用多标签学习(MLL)方法学习基于CNN的视觉输入与多个概念标签之间的映射关系。采用双向配方的S2VT模型来提高编码器的质量，在实验中性能优于普通S2VT模型。

采用多实例多标签学习(MIMLL)来训练我们的词汇模型，该模型可以看作是一个词汇检测和深层词汇分类的结合。它以{Xi, Yi}为输入对，其中每个XI是一个带有一组单词yI={yI1,…,yik}。在MIMLL中，每个实例通常都有一个或多个单词标签。例如，我们可以用“woman”。描述女性的词典中的“people”、“human”或其他同义词(参见图3中的一个测试组)。现在，我们为一个实例包定义损失函数。由于每个包都有多个单词标签，我们采用交叉熵损失来测量多标签误差:



其中（theta）为模型参数，N为袋数，yi为袋子Xi的标签向量，pi是对应的概率向量。当袋子中的所有实例都为负数时，我们将袋子弱标记为负数，因此使用噪音或公式来组合袋子中单个实例为负数的概率:



其中是第i个袋子中w为正的概率。我们定义了一个s型函数来模拟单个单词概率：



其中ww为权重矩阵，bw为偏置向量，为logistic函数。在我们的词法—FCN模型中，我们使用最后一个池层(pool5 for ResNet-50)作为实例Xij的表示。

lexical - fcn模型建立了框架区域和词汇标签之间的映射关系。lexical- fcn的第一步是从视频理解训练集构建词汇表，提取整个训练集中每个单词的词性。lexical- fcn的第二步是训练具有上述MIMLL损失的CNN模型。

为了获得密集的描述，我们需要将句子与ROI序列(感兴趣的区域)联系起来。我们直接从这16个非常粗的网格区域开始考虑信息区域序列生成问题

将不同帧之间的区域进行匹配并按顺序连接，以产生区域序列。

我们将区域序列生成任务表示为一个子集选择问题，其中我们从一个空集合开始，然后依次添加一个信息量最大的集合。并将每帧的相干区域划分为子集，同时保证了不同区域序列之间的多样性。设Sv为视频v的所有可能区域序列的集合，A为区域序列子集，即A包含于Sv。我们的目标是选择一个区域序列A\*，它对应一个目标R:



其中xv为视频v的所有区域特征表示。我们将R(Xv,A)定义为线性组合目标

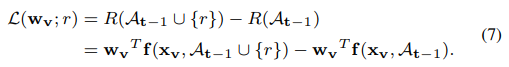


其中f = [finf, fdiv, fcoh]T，描述了区域序列的三个方面，即信息丰富，多样化和连续性。当Sv随视频长度呈指数增长时，Eq-4的优化问题很快变得棘手。我们限制目标f为单调子模函数，而wv为非负的。这使我们能够以一种有效的方式找到一个接近最优的解。

简要介绍了子模极大值的概念，并给出了权值wv的学习方法。如果集合函数满足收益递减特性，则称为子模函数。这意味着，给定一个函数f和任意集合A包含于B包含于Sv \r, f为子模如果满足:



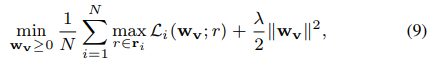
子模函数的线性组合仍然是非负权的子模。详情请参考[28,22]。子模函数具有许多与凸函数或凹函数相似的性质，这是我们想要的优化。以前的工作[28,22,9]表明，使用贪心算法使子模函数最大化可以很好地逼近最优解。在本文中，我们应用了一种常用的CELF方法。我们定义了一个边际增益函数为：



CELF算法从空序列Ao = 空集开始，将步骤t的区域rt添加到区域序列中，使边际收益最大化:



其中St为帧-t中的区域集。给定N对已知对应关系{(r,s)}，用如下公式优化wv:



其中最大项是广义链接损失，这意味着真值或oracle选择的区域r的得分应该比其他区域高一些。

WTA计划四步给一个准确的句子。首先,我们从年代中提取词汇标签基于词汇V,并形成一个词汇子集Vs。第二，我们获得的概率词w属于Vs，第i个区域序列由Piw=maxjPijw决定，其中pijw是第j帧中单词w的概率，它实际上来自于每个区域的Lexical- fcn输出。第三。我们用阈值（theta）来阈值piw，即如果piw <（theta）(在我们的研究中（theta）= 0.1)，则重新定义piw =0。最后，计算匹配分数，通过：



并通过i\* = argmaxifi获得最佳区域序列。这一目标表明，我们应该生成在句子中有高分单词的区域序列。

3.2.3子模块功能

根据子模函数的性质，我们描述了如何定义这三个部分如下。

区域序列的信息量定义为各区域信息量之和:



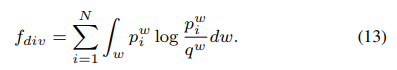
如果视频级别的句子注释在训练用例中或由oracle知道，我们将用Eq-10替换该定义，Eq-10限制句子词汇量Vs。

一致性是为了保证区域序列的时间一致性，因为区域范围的显著变化可能会混淆语言模型。与视觉跟踪中的一些工作类似，我们尝试选择时间变化最小的区域，并将一致性部分定义为：



式中Xrt为区域rt在第t步Xrs处的特征，是前面(t-1)步骤中的区域特征之一，表示两个正规化特征向量之间的点生成操作。在实际应用中，我们还将区域rt的搜索空间限制在前一步的9个邻域位置以内。

多样性度量候选区域序列与所有现有区域序列之间的差异程度。假设{piw}Ni=1为已有N个区域序列的概率分布，qw为候选区域序列的可能性分布，不同的完备性定义为Kullback-Leibler散度：



我们首先选择信息最丰富的区域序列，并将其提供给语言模型(LM)用于句子输出。然后，我们迭代地选择一个区域序列，最大限度地利用多样性来生成多个句子输出。

我们使用双向编码器扩展了原始的S2VT，不同的视频种类有非常不同的视觉模式和句子风格。分类清晰的语言模型定义为：



其中c为类别标签，v为视频特征表示，s为预测语句。P（s|c,v）为c类和视频v的概率条件，P(c|v)为视频v属于一个种类c的先验置信度，可由一般的视频分类模型得到。分类清晰的语言模型可以看作是最大后验估计。

**2.2 验证分析与实验效果**

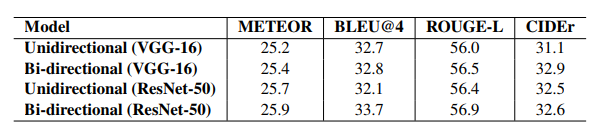


表1:MSR-VTT验证集上双/单向编码器单句描述精度。

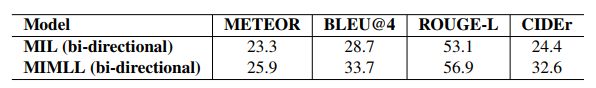
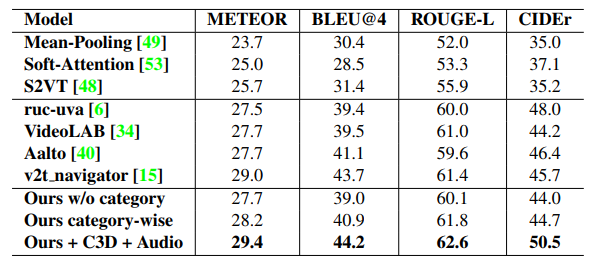


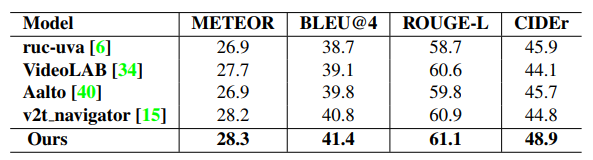
表2: MSR-VTT验证集上MIL和MIMLL单句描述精度。

对于单个描述的评估，这个基准测试的作者提出了基于机器翻译的度量，如METEOR , BLEU@1-4 , ROUGE-L和CIDEr。对于密集的视频描述结果，我们提出了自己的评估协议来验证结果。所有的训练和测试都是在一个拥有12GB内存的Nvidia TitanX GPU上完成的。在测试阶段，我们的模型是有效的。它可以在泰坦x GPU上以840ms左右的速度处理一个30帧的视频剪辑，其中CNN 特征提取用570ms，区域序列生成用90ms，语言模型用180ms。

ResNet-50作为网络结构，采用了双向编码器模型和MIMLL模型及分类语言模型。

还从两个方面对结果进行了定量评估:1)自动结果与oracle结果之间的性能差距，2)密集描述的多样性。





于我们的自动方法，需要对输出的句子进行排序，以获得前5或前10个句子。与[40]类似，我们以一种监督的方式训练了一个评价者网络，因为子模块最大化并不能保证句子的质量降序。表5列出了使用三种策略对MSR-VTT的验证集进行比较的结果:(1)句子区域序列关联的oracle和句子重新排序的oracle(简称OSR + ORE);(2)密集视频描述+ oracle重新排名(密集+简短的ORE);(3)全自动密集视频描述法(DenseVidCap)。

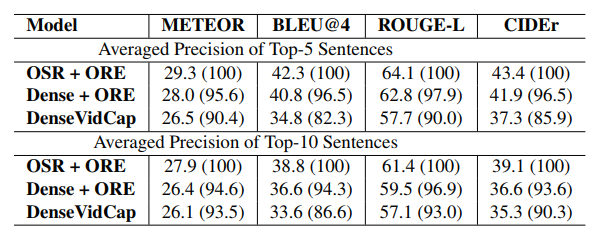
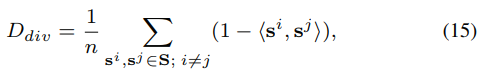


表5:在MSR-VTT验证集上生成的前5/10个句子的平均精度。OSR表示oracle用于区域序列关联，ORE表示oracle用于句子重新排序。括号中的值表示与完全oracle结果(OSR+ORE)的相对百分比(%)。

结果表明，密集视频描述+ oracle对前5个句子的全oracle结果(OSR+ORE)在所有指标上的相对准确率均可达到>95%，对前10个句子的全oracle结果的相对准确率为>93%。完全自动化的方法(我们的DenseVidCap)在前5名和前10名的设置上都能始终达到oracle结果82%以上的相对准确度。这是非常令人鼓舞的，因为性能差距不是很大，特别是考虑到我们的模型是用弱注释数据训练的。造成这种差距的一个重要原因是，与oracle的重新排名相比，评价者网络不够强大，这是进一步提高性能的一个方向。

生成描述的多样性对于密集的视频描述至关重要。我们从标题的相对相似度来评价多样性。一种常见的解决方案是确定描述对之间的相似性，或者一个描述与一组其他描述之间的相似性。在这里，我们从明显的语义关联性来考虑相似性。我们使用潜在语义分析(LSA)，它首先生成句子的词包(BoW)表示，然后将其映射到LSA空间来表示一个句子。该方法在测量文档距离[20]方面取得了较好的效果。在此基础上，计算了句子中两个LSA向量的余弦相似度。最后，计算分集为:



其中S为基数为n的句子集，(S i， S j)表示S i与Sj之间的余弦相似度。

通过与oracle结果的比较和本小节的多样性评价，我们证明了我们的方法确实可以产生良好的稠密覆盖。

**三、论文存在问题及后续研究重点**

**3.1 论文存在问题**

没有很好地利用上下文之间的密集描述，为输入的视频剪辑产生一个一致的故事

**3.2 后续研究重点**

着重于描述构建一个完整的故事。

**四、该问题相关研究成果**

**4.1 相关论文一**

**（1）题目**： Describing novel object categories without paired training data（深层复合描述:描述没有配对训练数据的新对象类别）

**（2）作者介绍**：L. Anne Hendricks, S. Venugopalan, and et al

**（3）摘要**：近年来，深度神经网络模型在图像配图任务上取得了良好的效果，但它们在很大程度上依赖于配图和句子配图的语料库来描述上下文对象。在这项工作中，我们提出了深层复合Captioner(DCC)来处理生成在成对的图像-句子数据集中不存在的新对象的描述的任务。我们的方法通过利用大型对象识别数据集和外部文本语料库，以及在语义相似的概念之间传输知识来实现这一点。目前的深层标题模型只能描述成对的图像句子语料库中包含的对象，尽管它们是用大型对象识别数据集(即ImageNet)预先训练的。相反，我们的模型可以组成描述新对象及其与其他对象交互的句子。我们证明我们的模型描述小说概念的能力通过实证评估其enformance MSCOCO ImageNet并展示定性结果的图像对象成对。

**4.2 相关论文**

**（1）题目**：From captions to visual concepts and back (从描述到视觉概念，然后再回来)

**（2）作者介绍**：H. Fang, S. Gupta, and et al

**（3）摘要**：提出了一种自动生成图像描述的新方法:视觉检测器、语言模型和直接从图像标题集学习的多模态相似模型。我们使用多个实例学习来训练视觉检测器来识别标题中常见的单词，包括许多不同的词性，如名词、动词和形容词。单词检测器输出作为最大熵语言模型的条件输入。语言模型从一组超过400,000个图像描述中学习，以捕获单词使用的统计信息。我们通过使用句级特征和深度多模态相似模型对标题候选词进行重新排序，从而获得全局语义。我们的系统在官方的Microsoft COCO基准测试中是最先进的，BLEU-4的得分为29.1%。当人类判断比较系统描述的测试集,系统描述有同等或更好的质量。

**4.3 相关论文三**

**（1）题目**：Sequence to sequence-video to text（序列到序列-视频到文本）

**（2）作者介绍**：S. Venugopalan, M. Rohrbach, and et al

**（3）摘要**：现实世界的视频往往具有复杂的动态特性;生成开放域视频描述的方法应该对时间结构敏感，允许可变长度的输入(帧序列)和输出(单词序列)。为了解决这一问题，我们提出了一种新颖的端到端序列到序列模型来生成视频描述。为此，我们利用了递归神经网络，特别是LSTMs，它在图像标题生成方面表现出了最先进的性能。我们的LSTM模型以视频句子对为训练对象，学习将视频帧序列与单词序列相关联，从而生成视频片段中事件的描述。我们的模型自然能够学习框架序列的时间结构以及生成句子的序列模型，即语言模型。我们评估几种变异模型,利用不同的视觉特性在一组标准的YouTube视频和两个电影描述数据集(M-VAD和MPII-MD)。