摘 要

近些年来，深度学习方法，尤其是深度卷积神经网络在图像应用领域取得了压倒性的优势。如今，由于移动传输技术，信息获取设备和展示设备的高速发展，使得产生的数据量大增，视频和基于视频语义分析也在此基础上快速发展。我设计了一种具有创造性的名为基于残差注意力机制的视频描述生成算法。主要的框架可以分成两个部分：一. 一个深度卷积神经网络(Convolution Neural Network, CNN)的编码器二. 一个基于残差注意力机制的长短时记忆解码器.我首先提取关键的视频帧来作为整个视频的表示，之后利用已经预训练好的卷积神经网络框架（例如VGG网络，GoogleNet网络，ResNet网络）来针对关键帧进行特征提取。在解码过程中，我引入了为了防止梯度消失而构建的长短时记忆单元(Long Short-Term Memory, LSTM)构建了循环神经网络（Recurrent Neural Network, RNN）。我能够根据视频的语义信息生成句子。为了评估设计的方法的效果，我在两个公共基准数据集：微软视频描述库（MSVD）和微软研究视频到文本描述库（MSR-VTT）进行了实验。实验结果表明我的方法在BLEU@N，METEOR以及CIDER的评价标准上优于当前流行的视频描述生成算法的效果。

关键词：注意力机制，注意力机制，视频描述，视频标题生成，深度特征

ABSTRACT

Recently, deep learning approaches, especially deep Convolutional Neural Network (ConvNets), have achieved overwhelming accuracy for image applications such as classification, segmentation and object detection. Today, due to the advances in transmission technologies, capture devices and display techniques, we are witnessing the rapid growth of videos and video content analysis. In this thesis, I propose a novel video caption method named Video Captioning based Residual Attention (VCRA). My framework consists of two components: 1. a deep Convolutional Neural Network (CNN) encoder and 2. a residual attention decoder based Long Short-Term Memory (LSTM) . Specifically I firstly extract key frames to represent the whole video. Then I apply CNN architecture (like VGG, GoogLeNet, ResNet) to extract representative frame level features. In our decode process, I introduce LSTM which is used to avoid gradient vanishing problem in RNN model. I can generate semantic sentences based on semantics of video. To evaluate the proposed model, I perform experiments on two publicly benchmark datasets: MSVD and MSR-VTT. The experiments results demonstrate that our method outperforms the state-of-the-art methods in terms of BLEU@N, METEOR and CIDER.

Keywords: deep learning, attention mechanism, video caption, caption generation, deep feature

目 录

第1章 绪论

1.1 研究工作的背景与意义

随着移动智能终端的普及和网络传输技术的快速发展，人们通过网络产生了各种类型的，海量的多媒体数据。例如，世界著名社交网站Facebook拥有近八千万注册用户一天能产生100PB数据，Flickr网站保存着超过60亿张图片，据Youtube 2017的统计，于2005年4月23日上传了第一视频后，到目前为止有13亿人口在实用Youtube，每一分钟有300小时的视频被上传，Youtube每天的访问量为3千万。如今，海量的多媒体数据随处可见，而样本数量异常大且复杂度极高。数据整体呈现高价值的海量复杂性，这些海量的多媒体数据包括文本数据，音频数据，图像数据，视频数据，甚至非结构化数据等，并推进大媒体数据时代的到来。

在众多的多媒体数据当中，视频所含有的信息量最大，也最复杂。虽然视频数据资源丰富而且蕴含这巨大的潜在价值，但其价值比率低（99%以上无价值）。在这样的数据爆炸的时代，仅仅靠人的力量自然史无法处理如此庞大的数据。这就需要人们设计一系列有效的算法来处理这些庞大的视频数据。海量多媒体内容管理是当前计算机领域的研究热点，而视频的内容分析是多媒体内容管理的关键技术之一。视频数据蕴含着不可估量的社会安全信息，经济价值和应用前景，促使其成为当前计算机科学的研究热点。与此同时，海量，复杂且异构的视频数据也为传统多媒体技术和应用，以及新兴的大媒体技术尤其是视频数据分析带来巨大的冲击和挑战。

针对这些海量的数据，人们最早通过机器学习方法来进行处理，这些方法通过一些基于统计学的算法让机器学习到一些规律并做出判断经验，并去解决一些生活中常见的一些问题。但是由于当时数据量不足和计算机计算能力的限制，基于统计基础的机器学习方法并没有很广泛的应用到实际到生活当中，但是其中的支持向量机（Support Vector Machine，SVM），感知机网络等等算法也都红极一时，为之后的深度学习的发展奠定了良好的基础。后来随着科学技术的发展，计算机的计算能力得到很大的提升，再加上了专门针对科学计算的GPU的加入，人们对海量数据的处理能力有了巨大的飞跃。同时科学研究也达到了一定的程度，2012年，Hinton和他的学生Alex构建了著名的AlexNet[1]，AlexNet是一个7层

的卷积神经网络模型，一举夺得ILSVRC（ImageNet Large Scale Visual Recognition Competition，大型图像识别比赛）的冠军，打败了之前称霸许久的支持向量机。之后深度学习更是一骑绝尘，将之前的传统机器学习方法远远抛在后边。近几年来，深度学习更是发展迅猛，创造了一个又一个的奇迹。同时，科学研究工作者并没有闭门造车，将这些技术应用了人们的实际生活当中，使当前的计算机能够更加智能，更好的为人类生活服务。

近十年来，深度学习在人工智能领域取得了重要的突破，它在图像分割，人脸识别，自然语言处理和物体检测等诸多领域的应用取得了巨大的成功。其在1000类图像（ImageNet）分类上的准确率已经达到了97%，其在人脸识别上的准确率已经超过了人类。然而深度学习在视频分析上还处于起步的阶段。相比于较为简单的图像，视频中含有更深层次的语义信息，动作信息和时序信息，一般的特征提取方法只能学习到较为浅层的语义信息，例如物体的形状，颜色，大小等等，这样使得得到的信息含有极少的语义层面的。深度卷积神经网络的提出，使得计算机能对图像和视频的深层次特征进行更好的挖掘，提出更加贴近语义层的信息。为了提取视频中的时序信息，循环神经网络以及其多种变种被提出。它们将状态在自身网络中循环传递，因此可以接受更广泛的时间序列机构的输入。然而，深度学习尚未达到完美的地步，离最终所谓的完全智能还有很长的一段路要走，还有许多尚待完善的地方，如缺少强有力的理论支撑，缺少强有力的理论支撑，缺少推理和预测能力，存储记忆能力有限，高度复杂性以及非监督深度学习方法的低有效性等。

1.2 本论文的结构安排

本文的章节结构安排如下：

第一章会介绍视频描述生成方法的研究背景以及研究工作的意义，并对本文的行文结构流程做简要的介绍。

第二章详细地介绍一下研究视频描述生成方法所需要的一些基本的理论基础，包括卷积神经网络中的卷积操作，池化操作，激活函数的意义和全连接层，循环神经网络中经典的长短时记忆单元和基于门的循环单元等。之后我们会针对当前国内外的研究现状进行分析，从最初这个任务被提出到目前各大高校，各大公司群雄逐鹿的局面。

第三章主要介绍我们提出的基于残差注意力机制的具体理论推导，我们会将我们模型方法的构思过程完整的展现出来，同时还要展示的一些基础的实现细节，包括前期地文本处理方法，词向量地生成方法，关键帧地提取和特征的提取过程，模型的层次化结构，句子的生成等等整个过程，画出整个模型的框架图。

第四章首先介绍评价标准的数据集微软研究视频描述库（MSVD）[2] 和微软研究视频到文本库（MSR-VTT）[3]，之后我们会介绍当前较为流行的评价标准BLEU@N，METEOR和CIDER，并针对生成的视频描述进行量化的分析和评价，展示基于残差思想的模型结果并和当前流行的方法进行比较。之后会展示针对某些视频生成的句子，我们会针对这些句子展开分析，分析一下我们模型的优点和缺点。

第五章对整个模型方法进行总结，并进一步指出了现有视频描述生成算法研究中目前存在的问题及相应的解决思路。

第2章 理论基础及研究现状

深度学习是神经网络的一个重要的部分，在解决图像，语音，文本等各种问题已经取得了很大的成就。深度学习在具体实现上有很多变化，核心是特征学习，目的是通过分层网络获取不同层次和水平的语义特征信息，避免了以往需要人工提取特征的重要难题，达到自动学习数据特征表达的效果。深度学习中有许多框架，包含许多重要的算法：1）自动编码器（AutoEncoder）；2）多层感知器神经网络（Multi-Layer Perception，MLP）；3）卷积神经网络（Convolutional Neural Network，CNN）；4）循环神经网络（Recurrent Neural Network，RNN）。对于不同的问题（图像，语音，文本），需要选用不同网络模型才能达到更好效果。2012年，Hinton获得ImageNet冠军之后，利用CNN在解决图片或视频分类问题上掀起一阵风潮，并取得了卓越的成绩。在之后的ImageNet挑战赛中，VGG-Net，GoogLeNet和ResNet网络在近几年内各领风骚，达到了前所未有的高度。循环神经网络的出现，改变了语音和文本这些时序信息的传统处理方法，并且取得了很好的效果。在我们的视频标题生成过程中，会用到卷积神经网络和循环神经网络，所以接下来我们会着重介绍卷积神经网络（CNN）和循环神经网络（RNN）。

2.1 理论基础

接下来我们会针对我们要构建模型的基础知识进行简单的介绍，主要包括以下几个方面：构成卷积神经网络的卷积（Convolution）操作，池化（Pooling）操作，激活函数（Activation Function）和全连接（Fully Connection）层。构成循环神经网络的长短时记忆单元（LSTM）和基于门的循环单元（GRU）。

2.1.1 卷积神经网络（CNN）

由于卷积神经网络在处理图像过程中具有稀疏连接，权值共享和平移不变等特性，近几年在图像分类，物体检测，物体检测，图像检索等领域取得很大的发展。一般卷积神经网络包括以下几个操作：1）卷积操作；2）池化操作；3）激活函数（非线性映射）；4）全连接层。所以接下来我们针对这些操作进行详细描述。

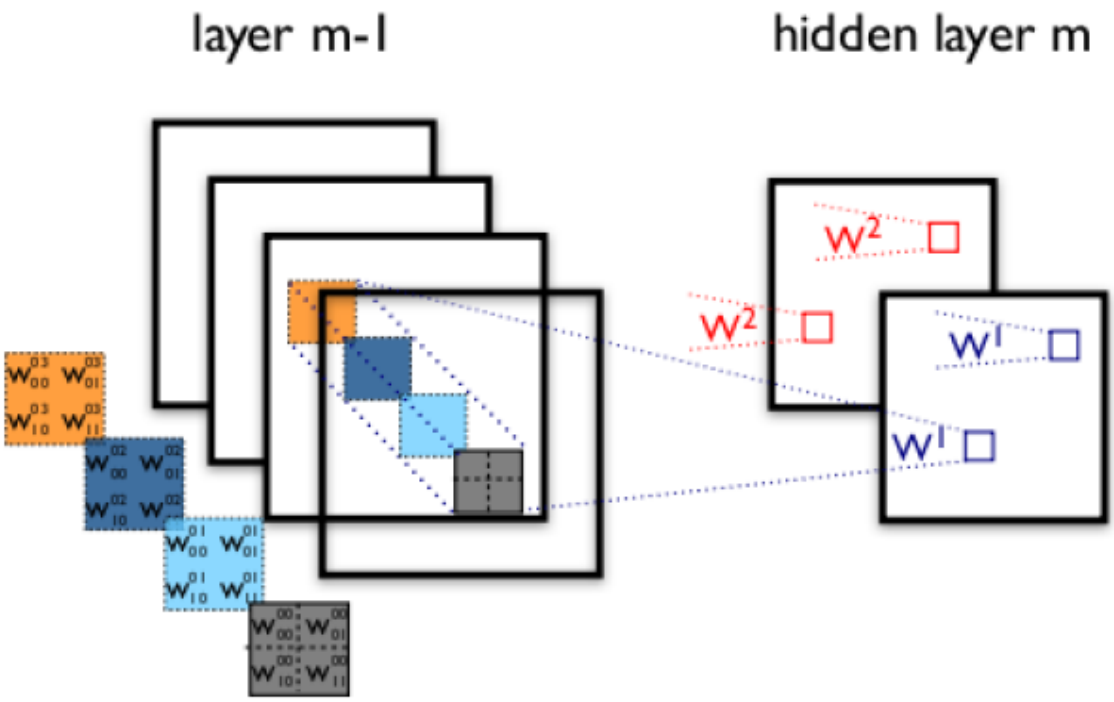


图 2-1 卷积示意图

卷积操作：卷积操作是卷积神经网络中十分重要的一部分，在对图像进行卷积操作中一般包含隐藏层和权值参数，最初在信号处理方面应用广泛，后因为其良好的效果在深度神经网络中起到了至关重要的作用。卷积操作是研究者根据人类视觉系统选取的一种特征提取手段。由于图片有非常多的像素点，如果针对所有的像素都做全连接的话，网络会因参数过多而无法进行工作。卷积的作用就是通过卷积核（又称为感受野）来提取局部图像的局部信息，并通过逐层的网络向上传播，这同样也符合人类视觉系统的工作流程，使得网络中的参数大大减少，训练的复杂度也大大减小。如图所示，卷积操作具有以下特性：1）局部感知，在图2-1可以看出，第m隐藏层种的两个特征图 h^0 和 h^1 的输出，是对m-1隐藏层中特征图用大小为 2×2 大小卷积映射过去的（图中标有蓝紫色虚线的映射）；2）权值共享，在图2-1可以看出，第m-1隐藏层中的特征图中的每个局部卷积操作共享卷积核参数（第m层标有蓝紫色正方形框的两个共享卷积核参数）。因为卷积的局部感知和权值共享的特性，这样能够大大的减少神经网络需要训练的权值参数的个数。

池化参数：池化层又被称为降采样层，能够使得特征映射之后的像素再此减少。池化在卷积神经网络中具有以下特点：1）降低特征维度。如果我们用卷积之后的特征做分类，会发现该特征维度相当大，这样做会使模型更加复杂，出现

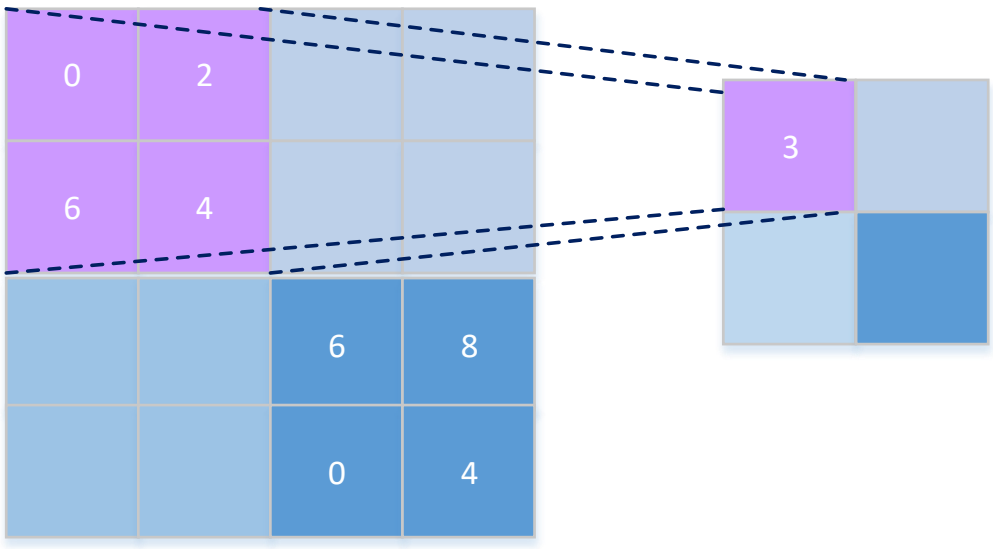


图 2-2 平均值池化示意图

过拟合现象。所以才引进池化操作，降低特征维度大小。2）平移不变性。图像具有局部和静态特性，对图像进行局部特征统计，并不会影响图像的特征表达。并且池化操作可以分为平均值池化和最大值池化。如图2-2，利用大小为 2×2 的核，步伐（stride）为2，对左特征图进行平均值池化，图中的不同颜色区域在池化过程中的一个区域，如虚线所表示的，平均值等于3。

激活函数（非线性映射）：神经网络是神经网络的基本组成单位，每一个神经单元都意味着一个运算单元，其实也就是通过一个函数完成映射的过程。如果网络中没有激活函数，或者激活函数是线性的，神经网络的作用则完全表现不出来。常用的映射函数或者激活函数有Sigmoid，tanh，RELU[4]（Rectified Linear Unit，修正线性单元）等等，他们在不同的场景中发挥着不同的作用。它们的函数图像如图所示2-3，Sigmoid和tanh是传统的神经网络中最常用的两种激活函数，从数学上来看，非线性的Sigmoid对中央区的信号增益较大对两侧区的信号增益小，但是这两种激活函数容易引起梯度消失的问题。后来根据神经学的研究，又发现了一种新的ReLU的新型激活函数，不仅大大的减小了计算量，同时也避免了梯度消失的问题。其实卷积神经网络的概念早在上世纪九十年代Lecun就提出过，但是因为当时他选择的激活函数并不是如今的ReLU，而是Sigmoid，因此其效果非常有限。

全连接层：如图2-4所示，全连接层的高一层的单元由低一层的神经单元的值

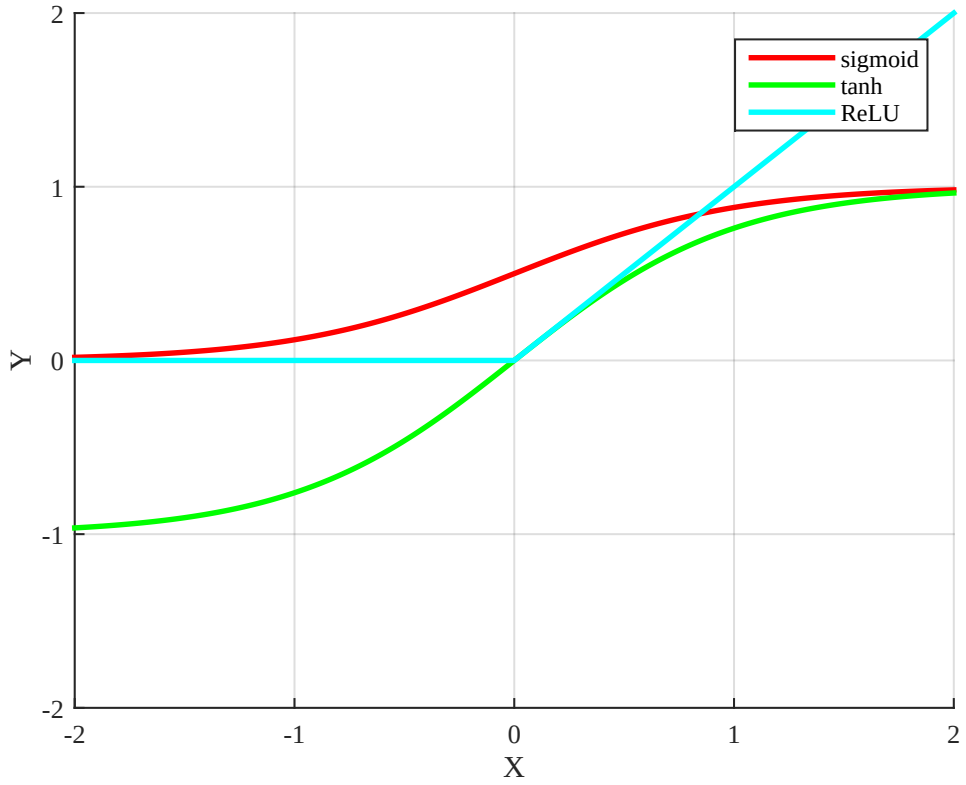


图 2-3 激活函数Sigmoid, tanh和ReLU示意图

加权得到，并在最终经过一个激活函数，通过公式来表示就是：

$$z_t = \phi(x * W + b) \tag{2-1}$$

其中W和b是需要训练得到的全连接层参数，ϕ是激活函数。全连接层的主要目的其一是为了改变维度，将输入的特征变换为所想要的维度，另外的目的使能够增加网络的非线性能力。由多层的全连接层连接起来能够构成多层的感知器网络。与这种模型相似的就是Softmax回归模型，它是一个非常经典的多分类，我们视频描述生成过程中预测上万个单词同样会用到Softmax回归。

在Softmax回归中，我们解决的是多分类问题。对于给定的测试输入x，我们想用假设函数针对每一个类别j估算出概率值$p(y = j|x)$。也就是说，我们想估计x的每一种分类结果出现的概率。因此，我们的假设函数将要输出一个k维的向量，该向量元素的和为1，来表示这k个估计的概率值。具体地说，我们的假设函

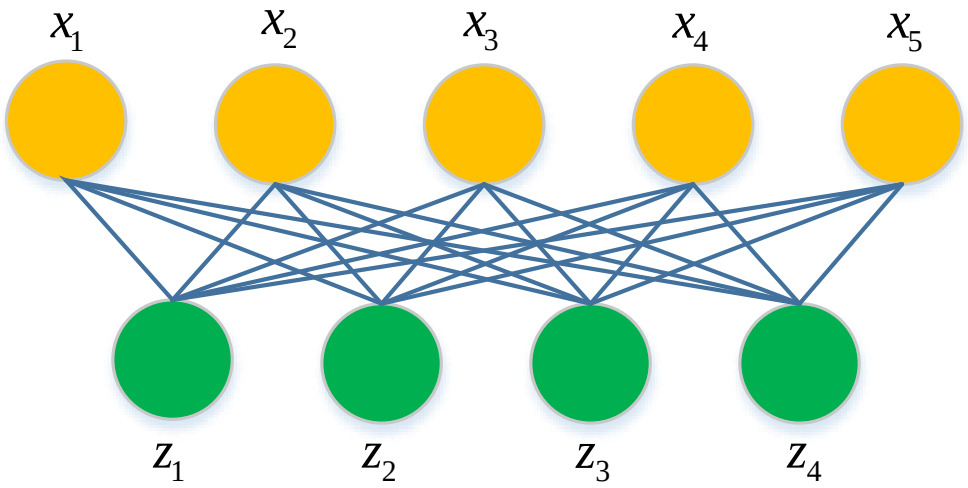


图 2-4 全连接层示意图

数$h_\theta(x)$形式如下：

$$h_\theta(x^i)=\begin{bmatrix} p(y^{(i)}=1|x^{(i)};\theta) \\ p(y^{(i)}=2|x^{(i)};\theta) \\ \cdots \\ p(y^{(i)}=k|x^{(i)};\theta) \end{bmatrix}=\frac{1}{\sum_{j=1}^{k}e^{\theta_j^T x^{(i)}}}\begin{bmatrix} e^{\theta_1^T x^{(i)}} \\ e^{\theta_2^T x^{(i)}} \\ \cdots \\ e^{\theta_k^T x^{(i)}} \end{bmatrix} \tag{2-2}$$

其中$\theta_1,\theta_2,\cdots,\theta_k$是参数模型；$\frac{1}{\sum_{j=1}^{k}e^{\theta_j^T x^{(i)}}}$归一化所求概率，使其所有概率和为1。

CNN实际是一个层级递增的结构，像人理解文章一样，先逐字逐句，再段落大意，再全文理解，CNN也是从像素，边缘，局部形状一直到整体形状的感知。我们所选用的GoogLeNet是多个卷积层，池化层以及全连接层组合而成的，共有22层。此网络的构造受到了Network in Network[5]的思想的启发并用此来降低网络中的参数量，同时也大大降低了计算量同时也能够增加网络的深度和宽度。我们用GoogLeNet Inception-v3[6]版本的网络，并在ImageNet中预训练出模型的参数，为我们后续的提取视频特征做好准备。

2.1.2 循环神经网络（RNN）

相比于卷积神经网络，循环神经网络是一个特殊的结构，如图2-5。循环神经网络(RNN)早在很多年前被提出，RNN的特别之处是用一层内部隐藏层处理一个变长的输入序列和一个变长的输出序列。循环神经网络常常被用来处理序列数据，例如自然语言，视频信息等等。在传统的神经网络模型中，同一层的隐藏层中层与层之间是无连接的，所以这种普通的神经网络对于很多具有时序信息的问题无

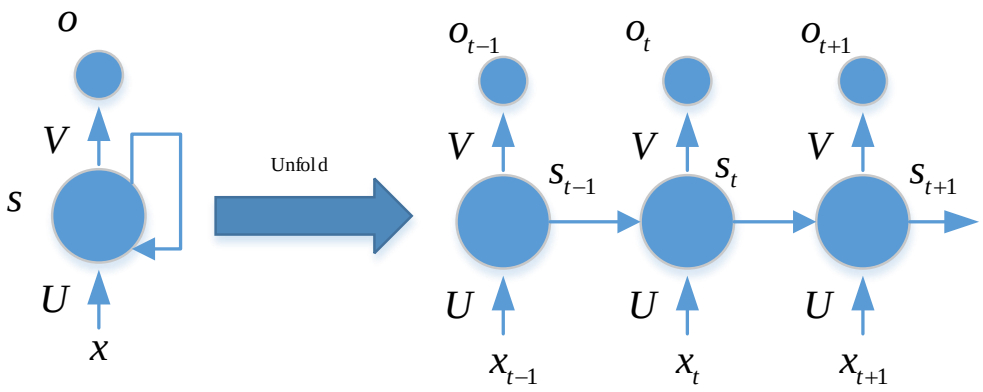


图 2-5 循环神经网络（RNN）示意图

能无力。例如，你要预测句子的下一个单词是什么，一般需要用到前面的单词作为已知量，因为一个句子中前后单词并不是独立的。循环神经网络能够准确的获取当前的输入与之前的输出的时序关系。具体的表现形式为网络会对前面的信息进行记忆并应用于当前输出的计算中，即隐藏层之间的节点不再无连接而是有连接的，并且隐藏层的输入不仅包括输入层的输出还包括上一时刻隐藏层的输出。理论上，循环神经网络能够对任何长度的序列数据进行处理。但是在实践中，为了降低复杂性往往假设当前的状态只与前面的几个状态相关，其实循环神经网络和普通的神经网络并没有太大的不同，循环神经网络可以被认为是相同网络的多重复制结构，每个网络把消息传给其继承者。

但是普通的循环神经网络存在长期依赖缺失，并且原始循环神经网络中的梯度消失问题长久没有得到解决。在此基础之上，有许多RNN的变体应运而生，应用较为流行的两种是基于门的循环单元（Gated Recurrent Unit, GRU）[7]和长短时间记忆网络（LSTM）[8]。LSTM是一种循环神经网络特殊的类型，可以学习长期依赖信息。LSTM同样是循环结构，但是重复的模块拥有一个不同的结构。不同于普通的循环神经网络单元，这里是有“门”机制，以一种非常特殊的方式进行交互。如图所示2-6，LSTM的关键就是细胞状态，水平线在图上方贯穿运行。细胞状态类似于传送带。直接在整个链上运行，只有一些少量的线性交互。信息在上面流传保持不变会很容易。LSTM有通过精心设计的称作为“门”的结构来去除或者增加信息到细胞状态的能力。门是一种让信息选择式通过的方法。他们包含一个Sigmoid神经网络层（图中带有σ符号的图标）和一个元素相乘（图中带有$\times$符号的图标）乘法操作。Sigmoid层输出0到1之间的数值，描述每个部分有多少量可

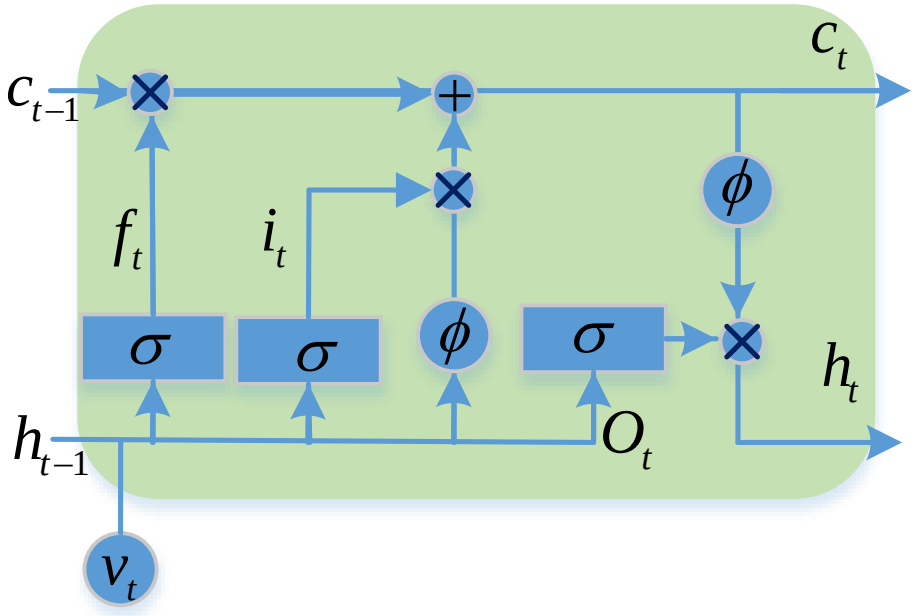


图 2-6 LSTM示意图

以通过。0代表“不许任何量通过”，1就指“允许任意量通过”。LSTM拥有三个门，来保护和控制细胞状态：1）遗忘门（Forget Gate，如图中f_t所示）：遗忘门决定此LSTM单元要遗忘何种信息，它以h_{t-1}和x_t为输入，在细胞状态m_t输出一个介于0和1之间的数。其中0表示完全遗忘，1表示完全保留。2）输入门（Input Gate，如图i_t所示）：输入门决定单元中要存储哪些信息。它以h_{t-1}和x_t为输入，之后由一个Sigmoid层来决定输入进入单元的信息。其次一个tanh层常见一个新的候选变量g_t，它可以加在状态之中。在下一步我们会结合两者的状态来更新当前的状态。3）输出门（Input Gate，如图中o_t所示）：输出门决定输出哪些内容。输出由细胞的状态决定，但是是一个过滤后的版本。首先通过Sigmoid层来判断哪些单元要输出哪些信息。然后用tanh处理单元状态，最后将其与Sigmoid的输出值做

点乘，得到我们输出的值。所以LSTM单元的门机制的公式为：

$$f_t = \sigma(W_{vf}v_t + U_{hf}h_{t-1} + b_f) \tag{2-3}$$

$$i_t = \sigma(W_{vi}v_t + U_{hi}h_{t-1} + b_i) \tag{2-4}$$

$$o_t = \sigma(W_{vo}v_t + U_{ho}h_{t-1} + b_o) \tag{2-5}$$

$$m_t = \phi(W_{vm}v_t + U_{hm}h_{t-1} + b_m) \tag{2-6}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ m_t \tag{2-7}$$

$$h_t = o_t \circ \phi(c_t) \tag{2-8}$$

其中W,U,b都是要学习的参数，σ表示的是Sigmoid函数，ϕ表示的是tanh函数。 对于基于门的循环单元（Gated Recurrent Unit，GRU），计算流如图所示2-7，相比于长短时记忆单元改动较大，它将忘记门和输入门合成了一个单一的更新门。同样还混合了细胞状态和隐藏状态，和其他一些改动。最终的模型比标准的LSTM模型要简单，也是非常流行的变体。具体的计算公式如公式2-12，相对于LSTM来说参数的数目明显减少许多，因此在训练的过程中，其收敛的速度要快于LSTM。

$$z_t = \sigma(W_z * [h_{t-1}, x_t]) \tag{2-9}$$

$$r_t = \sigma(W_r * [h_{t-1}, x_t]) \tag{2-10}$$

$$\tilde{h}_t = \tanh(W * [r_t * h_{t-1}, x_t]) \tag{2-11}$$

$$h_t = (1 - z_t) \cdot h_{t-1} + z_t * \tilde{h}_t \tag{2-12}$$

2.2 视频描述生成算法的国内外研究历史与现状

视频描述生成算法是视频内容理解研究中典型代表课题之一，涉及到多个学科领域，融合了计算机图形图像处理，模式识别理论，自然语言处理等多个学科。具体到描述生成算法，也涉及到统计学习，机器学习，深度学习，自然语言处理[9]等领域的各个方面的知识，是一个典型的交叉学科。在本论文中，我们将从深度学习和视频描述生成两个方面分析相关技术之间的关系以及国内外的发展状况。

深度学习让机器理解视频（Video Analysis）：近几年来，随着深度学习的快速发展，深度神经网络已经深入到计算机视觉的各个领域，如图片分类[1, 10]，物体追踪[11, 12]，目标检测[13, 14]，视频描述生成和视觉问答[15]等，其在ImageNet

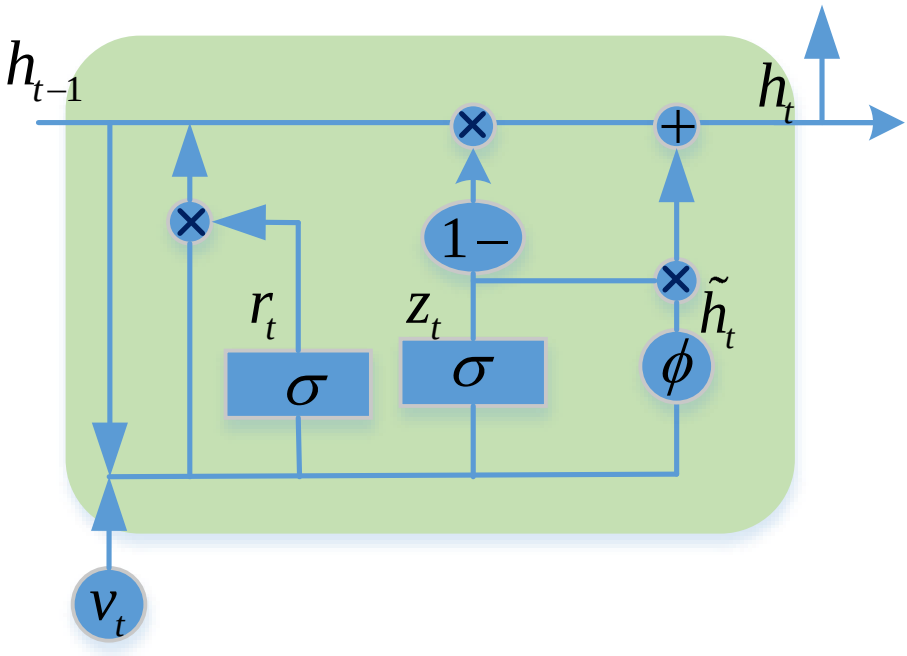


图 2-7 GRU示意图

图像分类上的准确率已经达到了97%，其在人脸识别上的准确率已经超过了人类。这些方法大多选取现有的比较流行的卷积神经网络框架，如AlexNet[1]，GoogLeNet[16]，VGGNet[17]，ResNet[18]等，作为基本的特征的提取方法。除了上面的成就之外，实践证明卷积神经网络的发展也为视频内容的理解做出了突出的贡献，比如视频物体检测[13]和视频追踪[19]。

视频物体检测，到目前为止最成功的还属美国国家标准与技术研究院(National Institute of Standards and Technology)，早在2010年他们就开放了关于视频物体检查的的各种竞赛，每年报名参赛的队伍来自全球各地的知名高校和研究员，例如麻省理工大学、斯坦福大学、牛津大学、新加坡国立大学、香港中文大学、北京大学等。这项竞赛要求参赛队伍在指定的视频数据中检索出相似的物体。在数据方面，牛津大学提供了两个公开的数据集，专门为图片搜索研究者提供，一个是OX_FORD5K，另外一个是PARIS6K，并且为每个数据集提供了55个查询目标。目前这两个数据集的检索精度已经提高到了85%左右。精度的提高主要源自于深度学习的快速发展。在过去的几年里，大量的研究者还在用浅层的特征来描述图

片特征，比如SIFT，HOG等等。2011年后，由于卷积神经网络的迅速推广，研究者们用深度特征来代替原有的浅层特征，性能得到了巨大提升。目前研究者们研究的重点是把实例搜索应用到视频上，比如TRECVID竞赛中的INS任务。如果实例搜索算法在视频上有所突破，这项技术将被大量应用在监控中，并将在智能交通中也起着举足轻重的作用。此外，视频物体跟踪分为单物体跟踪与多物体跟踪，针对以上两个分类分别以OTB、MOT两个challenge为主。到目前而言现，单物体跟踪算法以SANet (CNN+RNN)为首，其表现优异，在OTB数据集评测标准下可达到Precision plots: 0.928， Success plots: 0.692的优秀表现。以MOT2016报告而言，综合表现考虑，多物体跟踪算法以LMP为首，各项指标均靠前，在鲁棒性能上相对于单物体跟踪较弱，主要是因为其应用场景更为复杂，对于算法的要求更为苛刻。

作为多媒体内容分析的重要子领域，视频目标跟踪[19]是一个复杂且困难的研究课题，因为现实环境中有太多因素会对跟踪目标进行干扰。经过数十年的努力，虽然对一些简单场景已经能够很好处理，但面对更多更复杂的环境时跟踪效果仍不理想。深度学习方法的出现，为构建更加鲁棒的目标外观模型提供了可能，但为了设计出高精度，高鲁棒性和实时性的视频分析算法，仍然需要开展大量研究工作，目前的研究重点和发展趋势主要集中于以下几点：1) 深度学习与在线学习的融合。视频内容的深层次理解本质上是一个在线学习问题，最显著的特点是在线数据集是在不断扩充的。深度学习应用中所采用的先逐层训练而后全局微调的训练方式在纯粹的在线环境是否真正适用，如何避免陷入局部极小值，都是值得深入研究的问题。2) 构建适合视频复杂内容理解的深度网络。需要在目标表征能力和实时性之间有所权衡，既要保持深度学习特征学习的优势，同时也要兼顾跟踪的高实时性要求。例如，卷积神经网络中的降采样等损失空间信息的操作都是应用于视频分析任务的障碍，因此要进行必要改进，才能使深度网络真正适用于视频分析问题。3)视频内容理解数据平台的创建。目前建立模型的训练与测试数据平台并举行定期的比赛，已经成为图像与视频研究的流行趋势。因此如何根据视频内容进行理解是研究的特点，建立起大规模、具有代表性、测试方法严谨、适合深度网络训练、测试的内容理解视频数据平台，仍然是一个值得研究的课题。4) 递归神经网络的应用。尽管应用于一般性目标及开放环境的视频内容理解问题困难较大，但作为对于时间序列建模的重要深度模型，递归神经网络仍然可以在视频分析中有所作为。

从内容理解到视频描述生成（VideoCaption）:视频描述生成的方法在深度学

习发展的大背景下产生，并在岁后得到了快速的发展。2014年，百度[20]提出了视频描述生成这一全新的任务，之后便受到了整个计算机视觉领域的关注，如微软、谷歌、Facebook、Adobe 等商业公司，以及国内的知名高校如清华大学、浙江大学、南京大学、哈工大、中科院等，以及美国的斯坦福大学、卡耐基梅隆大学、加州理工大学等。知名高校和工业界巨头的加入，使得视频描述生成这一研究课题得到了迅速的发展，每年的最好成绩也不断的被刷新。

对于人类来说，看懂视频似乎是再简单不过的事情了。从出生拥有视觉开始，人眼所看到的世界就是连贯动态的影像。视野中每一个动态的形象都被我们轻易的识别和捕捉。但这对于计算机来说就没那么容易了。对于计算机来说，画面内容的识别、动作的捕捉，都要经过复杂的计算才能得出。当计算机从视频中识别出一些关键词后，由于语义和句子结构的复杂性，还要涉及词汇的词性、时态、单复数等表达，要让计算机将单个的词汇组成通顺准确的句子也是难上加难。那么让计算机看懂视频都要经过哪一步呢?首先，识别视频里的内容。目前的图像识别研究大多基于CNN(Convolutional Neural Networks，卷积神经网络)，首先，计算机识别出物体的种类，例如人、动物或其他物品；第二阶段，计算机获取物品在图像中的精确位置，这两个阶段分别回答了“是什么”和“在哪里”的问题。但在视频识别过程中，则需利用RNN (Recurrent Neural Networks，循环神经网络)将静态的图片加上时间的维度使其连贯，从而实现对视频内容中的静态物体和动作的识别。当计算机回答出“是什么”，“在哪里”和“做什么”的内容之后，就需要把这些分裂的词汇组成一个合乎人类表达规范的句子2-8。而在将计算机识别出来的内容组成句子的环节中，相关性(relevance)和连续性(coherence)是两个关键。相关性表示的是句子结构中的元素与视频内容的相关性，例如保证视频中所出现的客观物体的准确性。而连续性则是保证计算机最后“说”出来的句子要合乎语法，保证句子的连贯性。目前最好的成绩是由普渡大学，Facebook研究院和百度深度学习研究院共同创建的一个方法:分段循环神经网络(h-RNN[21])的方法，这个方法在MSVD数据集上的实验效果中BLEU@4 达到了49.9%，Meteor达到了32.6%，已经超越了之前微软亚洲研究院(MSRA)和众多高校的最好成绩。

目前关于视频描述生成的研究大多集中于结合关注机制进行并提出了很多基于关注机制的理论和算法，但是目前基本上所有的关注机制都基于已有的视频或图像属性生成的算法，其可以事先生成一系列的关键属性或关键词[22]，但是这些关键属性和关键词与整个视频描述生成是独立进行的，因此其优化工式很难最有效的描述视频的内容信息。到目前为止，视频描述生成还处于初级发展阶段，

图 2-8 视频描述生成示意图

针对该任务的应用目前寥寥无几，但其具有广泛的应用空间，例如，针对几段监控视频，可以通过视频描述生成方法快速的得到一段对视频中出现物体的文本信息描述。当然，视频描述生成的研究阶段还是属于初级阶段，只能对视频的一致信息进行总结和描述，但是可以想象当技术发展的成熟到一定程度，就能够根据一部电影直接生成电影讲述的具体信息并且能够自动生成影评，这样的技术也必定能够极大的推动人工智能的发展，为机器智能化作出重要的贡献。

2.3 本文的主要贡献与创新

本论文以注意力方法和长段时间记忆深度学习模型为研究重点，主要创新点与贡献有如下几点：1）我们提出了一个新的模型，叫做基于残差思想的注意力机制（Residul Attention，Res-Att）的视频描述生成算法，该方法结合时间注意力机制，对视频中的文本信息和视频帧特征都进行了动态加权处理。引入了注意力机制，使得在生成描述的过程中动态变化其权重。2）我们在处理注意力全中的时候引入了残差网络的思想,提出了一个新的调节机制，时间注意力机制用来决定什么时候来看那些视觉信息，而调节机制能够决定什么时候利用视觉信息，什么时候依赖句子模型。残差的思想能够降低注意力误差所导致的叠加的误差。 3）实验结果表明，我们的方法相比于之前的其他方法在评价标准BLEU@N，METEOR和CIDER上来说达到了最好的效果。

2.4 本章小结

本章主要从两个方面来介绍了基于深度学习的视频描述算法的理论基础和当前国内外针对这个任务的研究现状，其中理论基础主要包括卷积神经网络，循环神经网络，广泛应用且非常重要的激活函数和全连接层这四个部分。之后会讲述视频描述生成这个任务最初被提出，当前各大高校，各大公司都投入重金到这个研究领域予以重视，我们将重点介绍当前的发展状况，最后我们针对我们提出的方法的贡献和创新性做了简要的介绍。

第3章 基于深度学习的视频描述生成方法

在之前的研究中，计算机视觉[23]（Compoter Vision）和自然语言处理[24]（Natural Language Processing）是两个不同的领域，两者之间也并不存在联系，但是随着深度学习的发展，计算机视觉和自然语言处理的模型逐渐趋于统一，所有模态的信息之间的边界逐渐模糊，这就促使计算机视觉和自然语言之间的语义鸿沟逐渐被打破。在此背景之下，图片描述生成，视频描述生成和视觉问答系统应运而生，并取得了瞩目的成就。相对于图片描述生成方法，视频描述生成显然难度更大，因为视频中包含更多的物体和动作信息。针对像视频这样的时序信息，大部分研究工作都通过长短时记忆网络（Long Short-Term Memmory,LSTM）来对视频进行描述。由LSTM单元组成的循环神经网络（RNN）能够让通过网络参数学习到什么时候输入哪些信息，什么时候忘记之前输入的哪些信息，什么时候更新哪些信息，同时还能避免深度学习中经常出现的梯度消失的问题，并且在这些任务中取得了不错的效果。因此受到之前工作的启发，我们沿用他们使用的LSTM来处理视频的时序信息。 最近注意力机制在神经网络领域异常火热，在机器翻译，视觉描述生成和视觉问答系统中取得了不凡的成就。注意力网络通过设置一个变长的存储空间，从而避免直接将元数据压缩成一个固定长度的向量，所以我们需要提供一个随机访问元数据机制。其实注意力机制就是利用神经网络中的隐含层来计算分类分布，改分布体现了对原数据的软选择。所以我们利用LSTM结合注意力机制去挖掘视频在时间结构上的显著信息。最近的工作直接将卷积模型和LSTM模型相连接，比如姚力等[25] 直接对视频中的视觉信息用与训练好的卷积神经网络提取特征，结合注意力机制，输入到递归网络中解码成句子来描述该视频。但是视频描述生成不仅需要对视频序列的动态时间注意力机制到自然语言过程的建模，而且需要考虑视频内容和句子语义信息之间的关系。

在本篇论文中，我们提出了一种新颖的框架，叫做基于残差思想注意力机制的视频描述生成算法，如图3-1所示。首先，为了提取有意义的语义特征，我们采用GoogLeNet[16]的一个升级版Inception-v3神经网络框架对视频的每一帧提取CNN特征。然后我们通过一层的LSTM来编码词嵌入的语义信息，然后将LSTM的隐藏层信息与提取的视频帧信息，进行第一次的融合，然后通过一个软注意力

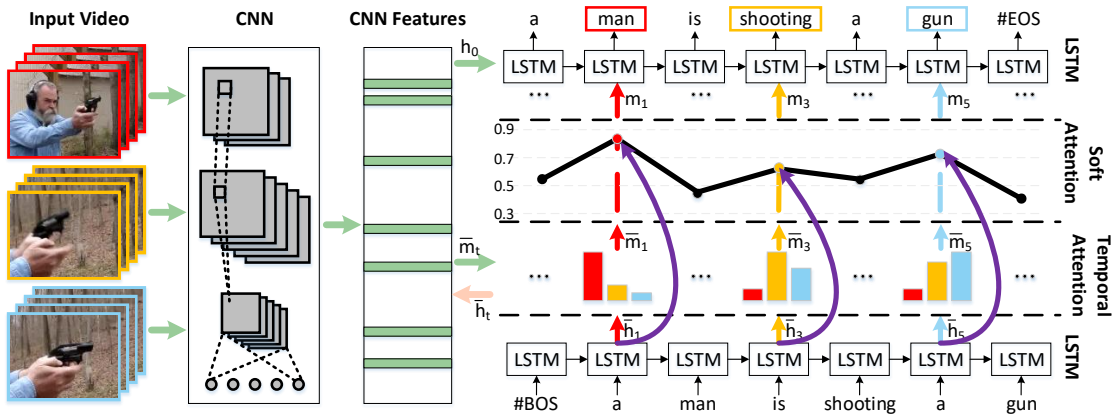


图 3-1 提出的基于残差注意力机制的视频描述生成方法的框架图

机制将其与词嵌入信息在此的进行融合，最终生成与视频语义相关的句子。注意力机制的引入使得在句子生成的过程中，每一个视频帧的权重是不一样的，这样会使得生成的句子有更高的概率出现视频中出现的物体，动作等信息，能更全面的概括视频的信息。两层的注意力机制能够让视频中出现的主体信息在其他的信息面前非常突出，这点相对于单层的注意力机制生成最终的描述效果更好。残差思想的引入，能够使得网络中学习到的误差不会依次的向后传播，通过残差的思想能够使得误差降低。这两个机制的结合保证了我们实验的结果，接下来我们会从下面几个方面详细地介绍我们的方法。

3.1 基于残差原理的注意力机制的视频描述生成方法

3.1.1 特征提取

为了获取视频中的语义信息，我们需要首先提取视频当中的语义信息，我们将通过现有的模型来提取视频的语义信息。自从2012年Hinton提出了卷积神经网络之后，深度网络一直处于快速发展的阶段，2014年GoogLeNet和VGG刷新了之前的深度网络，相比于之前的AlexNet，GoogLeNet和VGG的网络深度都远大于之前的网络，GoogLeNet更是达到了22层。2015年微软亚洲研究院的何凯明，更是创新性的提出了152层的残差网络，在此打破之前ImageNet的记录。如此惊人的效果证明了卷积神经网络的高效性，经过迁移学习的方法验证，这些深度网络在数据量丰富的ImageNet上预训练之后具有很好的普遍性，能够提取出具有很高抽象层次的语义信息。我们选用了GoogLeNet的一个修改版本，主要所做的修改就是将之前的7×7的卷积核分解成了两个一维的卷积（1×7，7×1），这样的好

处，既可以加速计算（多余的计算能力可以用来加深网络），又可以将一个卷积称分成两个卷积操作，使得网络的深度进一步的增加，增加了网络的非线性能力。GoogLeNet由两种基本模块构成，如图所示3-2，由于整个GoogLeNet的模型过大无法展示，但是整个22层的网络都是由这两种基本模块构成的。

3.1.2 注意力机制

近些年来，通过LSTM构成的循环神经网络对时序信息成功的进行了建模，并成功应用在了机器翻译，语音识别和计算机视觉视频描述生成等任务中，同时也解决了之前出现的梯度消失的问题，并且能够学习到更长的时间依赖。因为视频和自然语言都是时序信息，因此LSTM是我们处理数据的一个基本子单元。在LSTM的基础之上添加上注意力机制在这两天逐渐成为一种趋势，因为视频在不同的时间具有大量的冗余信息，每一帧所带有的语义信息的数量也是不相同的，注意力机制就是针对每一帧给予不同的权重信息，在生成不同的单词的时候每一帧的权重也不相同。基于注意力机制[25]的LSTM的定义如下：

$$i_t = \sigma(W_i y_t + U_i h_{t-1} + A_i c_t + b_i) \tag{3-1}$$

$$f_t = \sigma(W_f y_t + U_f h_{t-1} + A_f c_t + b_f) \tag{3-2}$$

$$o_t = \sigma(W_o y_t + U_o h_{t-1} + A_o c_t + b_o) \tag{3-3}$$

$$g_t = \phi(W_g y_t + U_g h_{t-1} + A_g c_t + b_m) \tag{3-4}$$

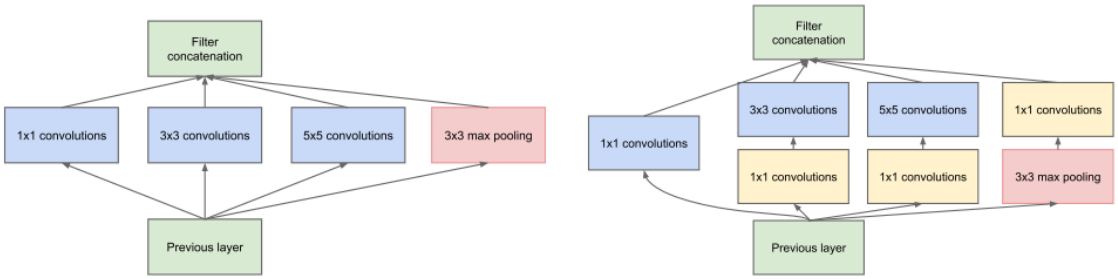
$$m_t = f_t \circ m_{t-1} + i_t \circ g_t \tag{3-5}$$

$$h_t = o_t \circ \phi(m_t) \tag{3-6}$$

其中，W,U,A和b是LSTM中要学习的参数，y_t代表每一个时间t输入LSTM中的输入值，σ表示Sigmoid激活函数，ϕ代表tanh激活函数，$\circ$代表元素级别乘法操作，c_t表示上下文向量，为视频描述生成提供了大量的视觉语义信息。

上下文向量信息是一个重要的因素，因为现实生活中提取到的视频很有可能其视频长度不一样，也有可能其播放的帧率也是不相同的。要解决的时频时常的多样性的问题，最简单的策略就是对所有的帧特征向量做平均，然后将其输入到每一个LSTM单元中，计算过程如公式3-7所示：

$$c_t = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} v_i \tag{3-7}$$



(a) Inception module, naïve version (b) Inception module with dimension reductions

图 3-2 GoogLeNet的基本模块框架

然而这个策略虽然将视频帧特征压缩成一个向量，但是却忽略了视频特征的时序性，从而导致一定的信息损失。注意力机制允许模型关注视频中的关键帧和关键元素，因为视频的性质，这些关键元素可能是多连续帧，我们在每个时刻动态的对视频帧特征进行动态加权，如公式3-8所示

$$c_t = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \alpha_t^i v_i \tag{3-8}$$

其中α表示针对每一个视频帧的特征，需要在每一个时间t都要进行计算，并且$\sum_{i=1}^{n} \alpha_i^t = 1$，我们称之为时间t的注意力权重。

在给定所有已经生成词的情况下，像$z_1, \cdots, z_{t-1}$,注意力权重α_i^t反映了在第t时间在第i帧特征的权重。因此我们设计一个函数，将LSTM解码之前的隐藏层状态，该隐藏状态总结了前面多有已经产生的词，同时输入视频帧特征V和h_{t-1}做运算，得到没有进行归一化的相关得分：

$$\varepsilon_t = W^T \tanh(W_a h_{h-1} + U_a V + b_a) \tag{3-9}$$

其中W^T，W_a，U_a和b_a是同LSTM编码解码过程中参数一起学习。一旦得到相关的ε_t，之后我们通过softmax将其进行归一化得到注意力权重：

$$\alpha_t = softmax(\varepsilon_t) \tag{3-10}$$

我们将计算相关得分和注意力权重过程成为注意力机制。该注意力机制在解码过程中通过提高相关帧注意力权重达到视频中的部分帧信息。然而我们没有明确的强制选择关注哪些部分信息，而是让注意力机制通过LSTM神经网络学习挖掘视频中的时间结构。此注意力机制最早应用于自然语言处理领域用作机器翻译工作，并取得了令人满意的成绩。姚力[25]创新性的将这种机制引入计算机视觉领域，通过获取视频的语义信息得到视频的描述同样也取得了非常好的效果。

3.1.3 基于残差原理的注意力机制的视频描述生成方法

注意力机制虽然能够一定程度上能够在生成句子的时候对视频帧特征进行加权计算，并且也有利用多层的注意力机制的实例实现了较好的效果，但是利用多层的注意力机制的话，在低层出现的误差会依次传递到高层，导致误差再一次扩大，最终也会使得最终的效果并不使人满意，这一点已经在传统的网络上面得到了验证，何凯明[18]也是基于这样的思想从而创造出了残差网络，解决了传播过程中出现的误差叠加的问题，从而能创造出更加深层的网络，取得更好的效果。受到残差思想的启发，如图3-3所示，残差网络在向上传递的过程当中，同时增加一个短接线路，这样能够保证在传递过程中传递的误差会大大的减小。同样，我们在使用双层注意力机制的时候同样也引入残差思想，设计了一种基于残差思想的注意力机制，在将语义信息向更上一层传播的同时增加一个短接机制，这样就能够消除语义信息传播过程中误差叠加的现象，降低语义信息在传播过程中的误差，这样能够让最终生成的句子也更加贴合真实的句子。

我们的模型方法同样是基于基本的编码-解码框架，包括两个神经网络：1）编码网络，2）解码网络。编码网络ϕ_E将给定的输入视频x编码成一个连续空间表达集合$V=\{v_1,\cdots,v_N\}=\phi_E(x)$，其中$\phi_E$通常是一个CNN神经网络。$N$是视频特征向量个数，$v_i\in R^M$是第$i$帧的一个$M$维向量，在这里我们选择LSTM作为解码网络$\phi_D$将视频特征解码成一个语句描述$z=z_1,\cdots,z_T$，其中$T$表示句子长度。并且LSTM基于之前的隐藏状态$h_{t-1}$，当前输入$y_t$和视频编码得到的特征$V$来更新它的隐藏层状态$h_t$，并预测当前生成单词：

$$\begin{pmatrix} h_t \\ z_t \end{pmatrix} = \phi_D(y_t, h_{t-1}, V) \tag{3-11}$$

其中LSTM递归的更新它的内部隐藏状态，直到遇到结束符号才停止。 我们设计的模型如图3-4所示，$z_0,z_1,z_2,...$表示的是将描述中的单词转换为的词嵌入。底层的LSTM能够将单词生成的词向量进行编码，并且通过RNN层能够保证语句的词法和语法信息。之后我们将循环神经网络中的隐藏层状态和视频的视觉信息进行融合，并生成融合语义的视觉信息，此时的视觉信息已经包含一定程度的时序信息。但是要想生成句子语法信息更加重要，并且为了减少传播的误差，我们之后通过一个软注意力机制与第一个LSTM的隐藏层状态融合，预测出生成的下一个单词。框架的具体介绍大致可以分成以下一个部分：

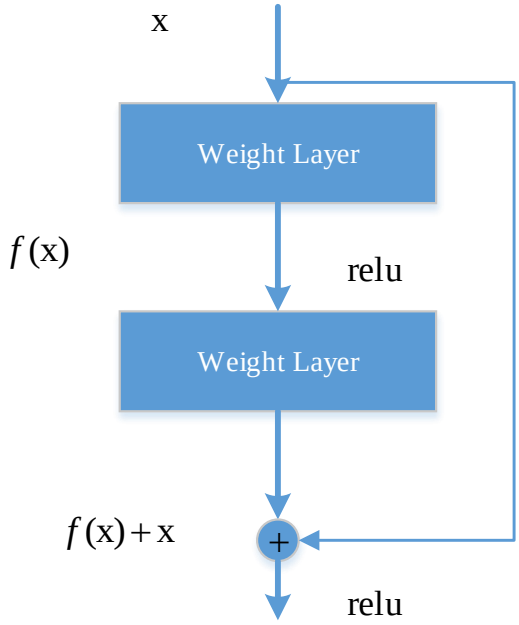


图 3-3 残差网络中通过增加短接（shortcut）以减小传播过程中的误差

底层LSTM：根据当前词y_t，之前的隐藏状态h_{t-1}和记忆细胞状态m_{t-1}来循环更新它的内部隐藏状态，定义如下：

$$\begin{aligned} h_0, m_0 &= [W_{ih}; W_{ic}] Mean(\{v_i\}) \\ h_t, m_t &= LSTM(y_t, h_{t-1}, m_{t-1}) \end{aligned} \tag{3-12}$$

其中$y_t = E[y_t]$表示单词的特征向量y_t，$Mean(\cdot)$表示对给定矩阵做均值池化，v_i表示视频中第i帧的特征向量，W^{ih}和W^{ic}是要学习的参数。

注意力融合层：根据底层的LSTM的输出h_t，以及输入的视频帧视觉信息V，我们可以得到融合后的特征，用公式表示如下：

$$\begin{aligned} \alpha_t &= Attention(V, h_t) \\ r_t &= \sum_{i=1}^{N} v_i \alpha_t^i \end{aligned} \tag{3-13}$$

通过视觉特征V和将词向量经过一层LSTM的隐藏层状态h_i进行结合，得到针对视频帧信息的权重，然后通过求加权和得到视觉信息和语义信息的融合r_t。$Attention$函数如3.1.2部分所描述的那样获得注意力权值。

顶层LSTM：有些模型通过双层的注意力机制以期望达到更好的结果，但是并没有关注到，底层的误差会在向上传播的过程中扩大，借鉴与残差的思想，我

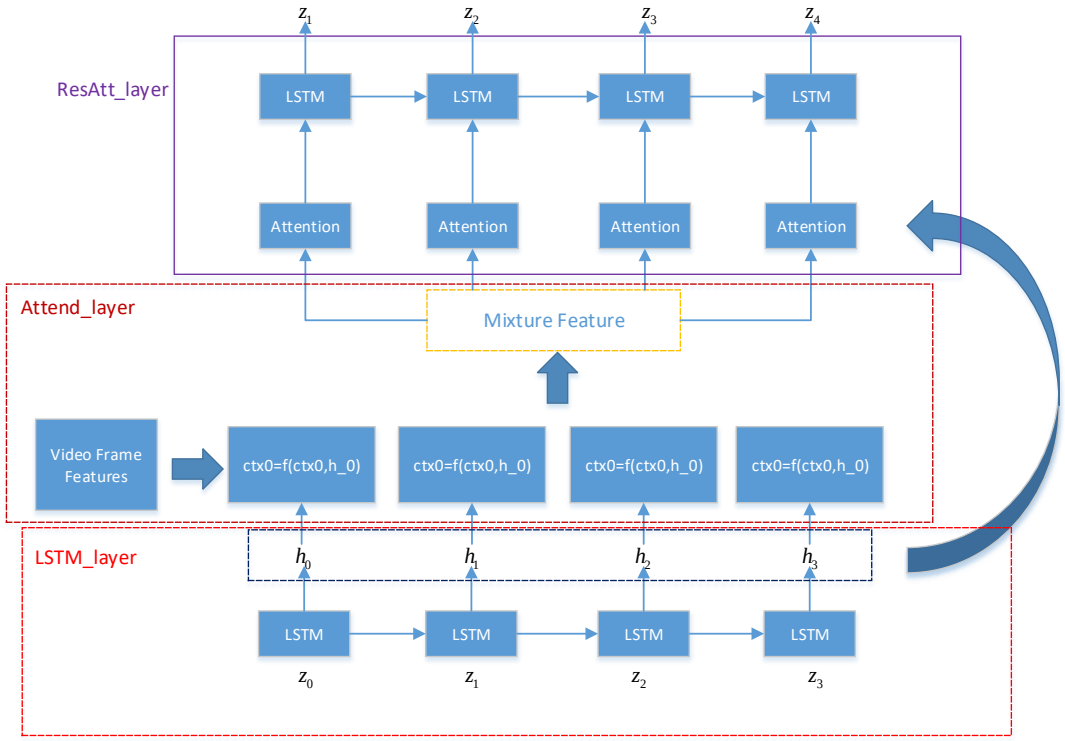


图 3-4 基于残差注意力机制的视频描述生成方法的具体实现框架

们在构建第二层注意力机制层的时候，加入了残差思想中的短接机制，来降低向上传播的误差。具体的计算公式如下所示：

$$\begin{aligned} \beta_t &= Attention(r, h_t) \\ c_t &= \sum_{i=1}^{N} \beta_t^i r_i \end{aligned} \tag{3-14}$$

将注意力融合层与第一层LSTM产生的隐藏层单元再次融合，通过注意力机制求得针对于融合之后的特征进行加权求和，将加权求得的特征放入到解码LSTM中以方便得到最终产生的句子，公式如下：

$$\begin{aligned} h'_0, m'_0 &= [W'_{ih}; W'_{ic}] Mean(\{r_i\}) \\ h'_t, m'_t &= LSTM(h'_{t-1}, m'_{t-1}) \end{aligned} \tag{3-15}$$

通过顶层的LSTM我们会得到对应的隐藏层单元h'_t，这个就是我们产生单词并构成句子的关键。

句子生成：为了得到下一个单词z_t，我们要用底层的LSTM的隐藏状态h_t得到数据词库的概率分布：

$$p_t = softmax(U_p \phi(W_p h_t + b_p) + d) \tag{3-16}$$

其中U_p，W_p，b_p和d是需要学习的参数；$softmax$函数可以将p_t丽姐为已经产生的词$y_{<t}$和视频中的视觉信息V，产生对下一个词的概率分布：

$$P(z_t|z_{<t},V,\Theta) \tag{3-17}$$

其中Θ表示模型要学习的参数。

为了学习我们模型中的参数Θ，我们用负对数似然估计方法来定义我们的损失函数：

$$\min_{\Theta} -\sum_{t=1}^{T} \log \mathrm{P}(z_t|z_{<t},V,\Theta) \tag{3-18}$$

其中T表示的是生成描述的长度。

得到模型的损失函数之后，我们就能够通过求取损失函数的导数来对函数进行优化，让损失函数的值尽可能的减小，从而使得生成的句子与真实的视频描述尽可能的相同。优化函数的方法非常多，有传统的随机梯度下降（Stochastic Gradient Descent，SGD），自适应时刻估计（Adaptive Moment estimation，Adam），Adadelta等等。随机梯度下降非常的简单，但是选取非常恰当的学习速率是非常困难的，并且SGD很容易陷入局部最优的情况。Adam对内存的需求比较小，还能够为不同的参数计算不同的自适应学习速率，适用于大型的数据集和高维空间。Adadelta相对于其他的优化方法来说，训练初期和中期的速度较快，能够很快的优化损失函数。但是到了训练的后期，很容易导致在局部最小值的附近抖动。在本文的实验部分，我们选取了Adadelta作为我们的优化方法。

3.2 本章小结

本章主要介绍了我们构建自己的模型所需要的深度学习基础，主要包括卷积神经网络中的卷机操作和池化操作，循环神经网络中的LSTM基本单元。之后我们还介绍了经典的注意力机制，这也是我们修改的基础。在讲述完这些基础之后，我们通过四个方面来介绍我们所设计的模型：底层LSTM，注意力融合层，顶层LSTM，句子生成。最后通过公式推导计算出了我们最终的损失函数，通过损失函数的优化就能够不断的优化我们的模型，使生成的句子更加贴合视频语义。

第4章 实验实现及其结果分析

4.1 数据集介绍

我们利用两个大规模的公共数据集来验证我们的模型，并在这两个数据集与当前表现最好的几个算法进行对比，以证明我们方法的优越性。

4.1.1 微软视频描述库

微软视频描述库的英文全称是：MicroSoft Video Descriptions(MSVD)[2]，这个视频描述库是微软提供的数据集，包括1970个视频片段，拥有大约80000个视频描述，大约包含16000个不重复的单词，每个视频大约有20个自然语言描述，很适合用于视频描述生成的这个任务。我们将整个数据集分为训练集，验证集和测试集，分别有1200，100，670个视频。因为这个数据集发布的时间较早，因此许多方法都将此数据集作为基准数据集来进行测试，我们能够在这个数据集上对比多个方法。

4.1.2 微软研究视频到文本

微软研究视频到文本的简称为MSR-VTT[3]。在2016年，微软发布了目前最大的可以用作视频描述生成的数据集。这个数据集包含10000个网络视频片段，每个视频片段由20个自然语言句子进行标注。而且这些视频中包含的都是常见的种类，具有不同的视觉内容。总共包含约200000个视频-句子对。由于这个数据集出现的时间较晚，许多方法并未公布其在这个数据集上的效果，因此我们在这个数据集上只和两种方法进行对比。

4.2 模型实现细节

系统环境硬件配置：我们执行代码所搭建的硬件环境服务器的具体配置如下：

CPU：Intel(R) Xeon(R) CPU E5-2650 v3 2.30GHz，

内存：256G，

显卡：Nvidia TITAN X Pascal 12G 显存，GPU是显卡的大脑，GPU的全称是Graphic Processing Unit，中文翻译为图形处理器，它决定了该显卡的档次和大部分性能。CPU的特点是什么都能干，但都不够专，运算能力不高，而显卡就很专，对于运算图形方面的浮点运算比CPU强10倍以上。

系统环境软件配置：

CUDA 8.0：CUDA是NVIDIA推出的通用并行计算架构，该架构使GPU能够解决复杂的计算问题，利用显卡强大的浮点运算能力来完成以往需要CPU才能完成的任务。它包含了CUDA指令集架构以及GPU内部的并行计算引擎，由CUDA优化过的程序能够在支持CUDA的处理器上以超高性能运行。

CUDNN 5.0：其全称为CUDA Deep Neural Network labrary，是NVIDIA专门针对深度神经网络设计的一套GPU计算加速库，被广泛用于各种深度学习框架，例如CAFFE，Tensorflow，Theano，Torch，CNTK等。这个工具主要是用来对GPU的科学计算进行加速，安装之后能够使得GPU的使用率更高。

OpenCV 3.1：OpenCV是基本的操作视频或者图像信息的开源库，有一系列C函数和C++类构成，实现了很多计算机视觉方面很多通用算法。我们在实验当中主要是通过OpenCV这个工具来处理视频的信息，取出视频当中的每一帧来进行单独的处理。

CAFFE[26]：CAFFE的全称是Convolutional Architecture for Fast Feature Embedding，它是一个清晰，高效的深度学习框架，它是开源的，支持命令行，Python和MATLAB的接口，能够在CPU和GPU上运行。我们本文主要是通过已经在ImageNet上预训练好的CAFFE模型，利用这个预训练好的模型来提取视频的特征。我们所利用的模型是GoogleNet　Inception v-3，提取到的每一帧对应一个2048维的特征向量。

Theano：是一个开源的深度学习的库，相对于之前的深度学习库，大大的简化了程序。同时它还能自动的对函数进行求导，降低了初学者学习深度学习的门

槛。我们利用Theano实现我们模型的框架模型，具体的代码我们会在后续过程上传到Github上。

预处理：我们首先将所有语句描述转换成小写字符，然后通过NLTK工具中的$wordpunct_tokenizer$方法对语句分词并去掉无用的标点符号。对于MSVD数据集，最终产生了15903个单词作为训练单词库，针对每一个单词，先将其转换成One-hot 的向量，然后根据对应向量将单词转换为词嵌入。同[25]一致,我们首先提取视频中的前360帧，然后等间距取28帧，并将这28帧送入Inception-v3网络提取pool3层特征。最后每个视频得到28*2048维的特征。

训练细节：在训练过程中，为了解决句子的变长问题，我们在每个句子最前面增加一个<BOS>符号作为句子的开始符，在每个句子的最后增加一个<EOS>符号作为句子的结束符，并选择30作为句子的最大长度。如果句子的长度小于30，用<EOS>来代替句子中的空位置。在测试过程中，我们输入<BOS>开始符到我们的模型框架中，开始视频描述生成过程。对每一个词的生成过程。另外，所有LSTM单元大小都设置为512，根据以往的经验，词嵌入特征向量维度也设置成512。我们通过训练集中的所有成对的视频-句子，以64作为最小批次来优化我们的目标损失函数。我们采用具有自适应学习速率的adadelta方法作为优化方法。为了防止过拟合的现象出现，我们利用概率为0.5的dropout层来防止我们的模型过拟合，并设置最大的梯度值为10以防止梯度爆炸。直到训练500轮或者连续20次验证数据及上效果不再提升，就结束我们的训练过程。

语句生成：给定一个视频，有多种方法可以生成句子描述。1）Sampling：每次只选取概率最大的单词，然后将该单词输入到网络产生下一个单词，直到产生特定的结束符<EOS>或者达到句子的最大长度而停止运行。2）BeamSearch：在每一个时间t，迭代地考虑前k个最好地句子作为时间t+1地输入。最终产生k个句子，然后根据Y=$argmax_{Y'}Pr(Y'|V)$选取最好的一个句子作为我们视频描述的语句。实验过程中我们采用的是BeamSearch的策略，且k=5。

评价方法：我们采用BLEU@N[27]，METEOR[28]和CIDER[29]。其中BLEU-@N是一种流行的机器翻译评价指标，用于分析预测语句和实际的语句中n元组共同出现的程度，由IBM于2001年提出。主要测量N-gram（我们生成的标题语句和标签数据之间的相同之处）的比例，N常常取1，2，3，4。METEOR标准于2004年由Lavir发现在评价指标中召回率的意义后提出。METEOR测量基于单精度的加权平均数和单字召回率，其目的是解决BLEU@N标准中固有的缺陷，同时基于词网络考虑词之间的相似性。

4.3 实验结果分析

我们在MSVD和MSR-VTT两个数据集上验证了我们的模型，并且与当前表现优秀的一些视频描述生成算法进行了对比，结果如表4-1和4-2所示，接下来我们将对其中的对比算法做简单的介绍，并对实验的结果进行分析。

4.3.1 对比方法介绍

为了突出我们算法相对于其他算法的优越性，我们将会用我们的方法跟其他做视频描述生成算法的效果进行对比，以衬托我们算法在各种评价指标上都好于其他的算法。主要对比的算法包括以下几个：

多层长短时记忆单元的感知网络（MP-LSTM[30]）:多层长短时记忆单元的感知网络（MP-LSTM）是美国德克萨斯州大学于2015年提出的一种生成视频描述生成算法，其用深度卷积神经网络和两层LSTM神经网络连接，直接将视频转换成自然语言句子。为了得到视频的特征表达，该方法对所有的视频帧特征做均值池化，而忽略了视频中的时间动态信息。虽然此方法的效果并不算是非常优秀，但是其开辟了将卷积神经网络和循环神经网络结合的先河，为之后的视频描述生成提供了很好的一种思路。

软注意力模型（SA[25]）：软注意力模型这篇文章是著名的蒙特利尔大学在2014年ICCV会议上发布的论文，首次的将注意力机制由自然语言处理领域引入到计算机视觉领域。为了挖掘视频的局部结构，SA提出两种特征：1）从GoogLeNet中提取视频帧级别的特征；2）从3D-ConvNet中提取视频片段级别的特征。然后用一些描述子，包括HOG，HOF和MBH，然后将这些描述子拼接成一个向量作为视频特征。进一步提出时间注意力机制去学习特定时间区域的权重，产生更具有表达性的语句。又因为作者在发布论文的第一时间将代码公之于众，因此许多视频描述生成方法的后来者都是在姚力的代码基础之上进行的修改，他在本领域做出的贡献非常卓越。

序列到序列-视频到文本（S2VT[31]）：序列到序列模型最早也是机器翻译领域的思想，由于视频信息作为一个序列信息，生成的描述语句同样也是一个序列信息，后被研究者将其思想应用至视频描述生成领域。S2VT是一个端到端序列到序列的视频标题生成模型。首先输入视频帧和光流描述子到深度卷机网络中，然后该方法利用两层LSTM生成视频语句。后来的许多序列到序列的任务大都沿用现有的模型或者在现有的序列到序列的基础之上进行修改。

视觉语义嵌入的长期记忆（LSTM-E[32]）：微软亚洲研究院的视觉语义嵌入的长期记忆方法（LSTM-E）给定之前学习到的单词和视频视觉特征，利用LSTM预测下一个单词。并且LSTM-E提出了一个视觉-语义嵌入方法，加强整个句子和整个视频内容之间的关系，其在原来的基础之上增加了一个相关性损失，能够尽最大的可能保证生成句子和真实句子之间的差别最小。

基于段落式的循环神经网络（p-RNN[33]）：基于段落式的循环神经网络（p-RNN）实现一个级联RNN框架，可以实现用一段话（包括多个句子）描述一个视频片段。该框架包含两个生成器：1）句子生成器产生一个简短的句子来描述视频中特定时间片段和特定区域；2）段落生成器用句子嵌入作为输入，用另外一个循环神经网络输出段落状态，该状态初始化句子生成器。其中句子生成器和段落生成器都采用循环神经网络。

分层循环神经网络编码器（HRNE[21]）：为了挖掘视频中的时间动态信息，该方法提出了一个级联循环神经编码网络，然后结合注意力机制做视频标题生成任务。浙江大学提出了该创新性的方法，作者发现了双层长短时记忆网络无法捕获长时间的依赖，因此将更高层次的更抽象的信息提取出来。

4.3.2 在微软研究视频描述库数据集上的结果分析

我们的基于残差的注意力机制的视频描述生成算法和其他对比算法在MSVD数据集上的效果如表4-1所示，因为我们使用的是单特征，因此我们也只于其他方法的单特征效果进行对比。同目前最好的p-RNN进行比较，我们的方法在BLEU@4评价标准能够提高4.5%，在METEOR评价标准下也能提高1.6%。相比于单个简单的

表 4-1 实验结果以及和当前几个最好的模型在MSVD数据集上的结果比较，V表示的是VGG特征，G表示的是GoogLeNet特征

模型	BLEU-1	BLUE-2	BLEU-3	BLEU-4	METEOR	CIDER
S2VT(V)	-	-	-	-	29.2%	-
HRNE(G)	78.4%	66.1%	55.1%	43.6%	32.1%	-
LSTM-E(V)	74.9%	60.9%	50.6%	40.2%	29.5%	-
SA	-	-	-	40.3%	29.0%	-
p-RNN(V)	77.3%	64.5%	54.6%	44.3%	31.1%	62.1%
Res-Att(G)	**80.6%**	**69.4%**	**59.6%**	**48.8%**	**32.7%**	**72.2%**

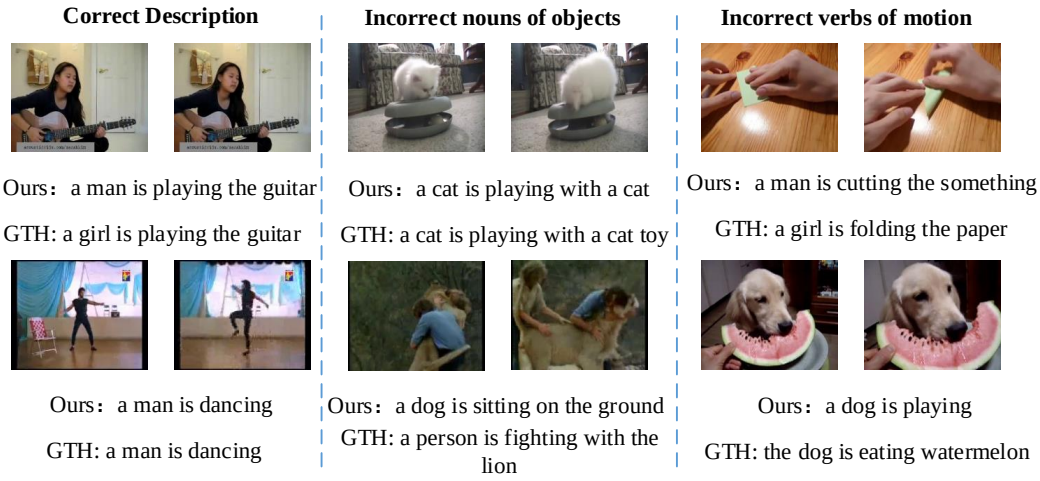


图 4-1 视频描述生成效果展示图：Ours表示我们本模型所生成的句子，GTH表示的是视频对应的真实的句子

注意力机制方法相比，更是在BLEU-4和METEOR上提高了8.5%和3.7%，突出了我们基于残差注意力机制方向的正确性和我们所设计算法的优越性。

我们模型的具体效果展示如图4-1所示，算法生成的描述基本都符合视频的语义描述，但是并不完美仍存在一些瑕疵，例如都某些名词或者动词理解的有偏差等等。在经过训练之后，基于残差注意力机制的模型生成的句子效果如图所示4-1，经过一定的训练之后，模型已经能够大致的识别视频中出现的主语和动词，并能够生成较为准确的句子描述。然而我们的模型并不完美，依然存在着生成的名词或者动词不符合视频语义信息的现象，例如第二列第一个出现的”a cat is playing with a cat”，对于最后一个生成的单词明显出现了偏差。第二列的第二个例子显然出现的问题更加的严重，其一是将视频中的”lion”理解成了”dog”，同时将其中的动作等语义信息完全理解错了。当然还有一些将视频中的动作理解错误的现象，对于第三列第一个视频，模型对”folding”和”cutting”两个单词理解，但是我们不得不承认两个动作确实是有相似之处，但是如果考虑到后边的宾语同样也可能会对前面的动词产生影响，预计生成句子过程中会避免一些动作信息的错误理解。第三列的第二个视频生成的描述同样与真实的句子差别较大。虽然我们的模型并没有达到十全十美，但是根据生成的这些句子和单词给我们接下来的工作指明了方向。

4.3.3 在微软研究视频到文本库数据集上的结果分析

如表4-2所示，由于这个数据集是1026年提出来的，好多模型并未在这个数

表 4-2 实验结果以及和当前几个最好的模型在$MSR-VTT$数据集上的结果比较，V表示的是VGG特征，C表示的是C3D特征

模型	BLEU-4	METEOR
MP-LSTM(V)	34.8%	24.8%
MP-LSTM(C)	35.4%	24.8%
MP-LSTM(V+C)	35.8%	25.3%
SA(V)	35.6%	25.4%
SA(C)	36.1%	25.7%
Res-Att(G)	**38.8%**	**25.8%**

据集上做过验证，因此我们只能对比MP-LSTM和SA两种方法，在较大的数据集上我们所设计的基于残差思想的注意力机制的方法同样能够在各种评价标准上优于其他的方法，同时也拥有更高的鲁棒性，因为生成的句子的质量主要可以由BLEU@4和METEOR两个指标来体现，因此在大数据集上，我们只通过这两个评价指标来衡量。针对只是用单特征的MP-LSTM，我们的方法分别高于它4.0%和1.0%，同样我们也使用了能够表现动态特征的C3D特征以及两种特征相结合来对比，我们的方法依然优于他们的模型。同时我们的方法还和软注意力机制进行了对比，我们的模型同样能够在BLEU@4上高于其2-3个百分点，在METEOR上也略微好于软注意力机制模型。

4.4 本章小结

本章我们主要从实验的方法来介绍模型，首先介绍了验证模型的两个数据集：MSVD和MSR-VTT。其次我们介绍了我们在通过代码实现模型的过程中的实现细节。之后通过在两个数据集上的结果显示我们所设计的模型的效果要好于当前所公布的其他模型，这也说明我们设计方法方向的正确性和模型的优越性。

第5章 全文总结与展望

5.1 全文总结

在人工智能快速发展的大环境下，视频描述生成算法也成为一个火热的研究话题。考虑到传统注意力机制的误差在传播过程中不断扩大的不足，同时加上受到残差思想的启发，我们设计了一个基于残差思想的注意力机制的视频描述生成模型，此方法吸取了注意力机制和残差思想的长处，并进行了创新。同时该模型也达到了非常好的效果，超过了近两年的其他模型的效果，这也说明我们这种基于残差思想的注意力机制方法的正确性和自身模型的优越性。

5.2 后续工作展望

本文围绕视频描述生成算法这个课题进行了探索，并围绕该问题提出了自己的模型和想法，并在两个标准数据集上验证了我们方法的优越性。然而这个研究课题是综合了计算机视觉和自然语言处理两个方向。因此还有许多能够优化的地方，在本文研究工作的过程中，发现了许多仍然可以改进的地方： 1）.本文主要注重的是对每一帧特征的进行加权，但是忽略了在对视频进行一句话的描述过程中主要描述的是视频中出现的主体，而并不是视频中出现的每一个物体，这个信息也将会在生成句子的过程中发挥重要的作用。2）注意力机制在视频描述生成算法中发挥着重要的作用，但是在应用的过程中也存在着一定得问题，例如有些时候，视频中的某个词会因为视频中的主体而在生成的语句中出现多次：”A cat is playing with a cat.”，如果在未来注意力机制的使用当中注重视频的时空信息的话，这样的问题就能够避免。

参考文献

[1] A. Krizhevsky, I. Sutskever, G. E. Hinton. Imagenet classification with deep convolutional neural networks[M]. 2012, 1097–1105

[2] D. L. Chen, W. B. Dolan. Collecting highly parallel data for paraphrase evaluation[M]. 2011, 190–200

[3] J. Xu, T. Mei, T. Yao, et al. Msr-vtt: A large video description dataset for bridging video and language[M]. 2016, 5288–5296

[4] X. Glorot, A. Bordes, Y. Bengio. Deep Sparse Rectifier Neural Networks.[M]. 2011, 275

[5] M. Lin, Q. Chen, S. Yan. Network in network[J]. arXiv preprint arXiv:1312.4400, 2013

[6] C. Szegedy, V. Vanhoucke, S. Ioffe, et al. Rethinking the inception architecture for computer vision[M]. 2016, 2818–2826

[7] K. Cho, B. Van Merriënboer, C. Gulcehre, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014

[8] S. Hochreiter, J. Schmidhuber. Long short-term memory[J]. Neural computation, 1997, 9(8):1735–1780

[9] R. Pascanu, C. Gulcehre, K. Cho, et al. How to construct deep recurrent neural networks[J]. arXiv preprint arXiv:1312.6026, 2013

[10] J. Yang, K. Yu, Y. Gong, et al. Linear spatial pyramid matching using sparse coding for image classification[M]. 2009, 1794–1801

[11] G. Malkomes, C. Schaff, R. Garnett. Bayesian optimization for automated model selection[M]. 2016, 2900–2908

[12] S. S. Blackman. Multiple-target tracking with radar applications[J]. Dedham, MA, Artech House, Inc., 1986, 463 p., 1986

[13] R. Girshick. Fast r-cnn[M]. 2015, 1440–1448

[14] S. Ren, K. He, R. Girshick, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[M]. 2015, 91–99

[15] K. Kafle, C. Kanan. Answer-type prediction for visual question answering[M]. 2016, 4976–4984

[16] C. Szegedy, W. Liu, Y. Jia, et al. Going deeper with convolutions[M]. 2015, 1–9

[17] K. Simonyan, A. Zisserman. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014

[18] K. He, X. Zhang, S. Ren, et al. Deep residual learning for image recognition[M]. 2016, 770–778

[19] Y. Xiang, A. Alahi, S. Savarese. Learning to track: Online multi-object tracking by decision making[M]. 2015, 4705–4713

[20] J. Mao, W. Xu, Y. Yang, et al. Deep captioning with multimodal recurrent neural networks (m-rnn)[J]. arXiv preprint arXiv:1412.6632, 2014

[21] P. Pan, Z. Xu, Y. Yang, et al. Hierarchical recurrent neural encoder for video representation with application to captioning[M]. 2016, 1029–1038

[22] Q. You, H. Jin, Z. Wang, et al. Image captioning with semantic attention[M]. 2016, 4651–4659

[23] X. Li, L. Gao, X. Xu, et al. Kernel based latent semantic sparse hashing for large-scale retrieval from heterogeneous data sources[J]. Neurocomputing, 2017

[24] D. Britz, A. Goldie, T. Luong, et al. Massive Exploration of Neural Machine Translation Architectures[J]. arXiv preprint arXiv:1703.03906, 2017

[25] L. Yao, A. Torabi, K. Cho, et al. Describing videos by exploiting temporal structure[M]. 2015, 4507–4515

[26] Y. Jia, E. Shelhamer, J. Donahue, et al. Caffe: Convolutional architecture for fast feature embedding[M]. 2014, 675–678

[27] K. Papineni, S. Roukos, T. Ward, et al. BLEU: a method for automatic evaluation of machine translation[M]. 2002, 311–318

[28] S. Banerjee, A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[M]. 2005, 65–72

[29] R. Vedantam, C. Lawrence Zitnick, D. Parikh. Cider: Consensus-based image description evaluation[M]. 2015, 4566–4575

[30] S. Venugopalan, H. Xu, J. Donahue, et al. Translating videos to natural language using deep recurrent neural networks[J]. arXiv preprint arXiv:1412.4729, 2014

[31] S. Venugopalan, M. Rohrbach, J. Donahue, et al. Sequence to sequence-video to text[M]. 2015, 4534–4542

[32] Y. Pan, T. Mei, T. Yao, et al. Jointly modeling embedding and translation to bridge video and language[M]. 2016, 4594–4602

[33] H. Yu, J. Wang, Z. Huang, et al. Video paragraph captioning using hierarchical recurrent neural networks[M]. 2016, 4584–4593

[34] G. Li, S. Ma, Y. Han. Summarization-based video caption via deep neural networks[M]. 2015, 1191–1194

致 谢

在我完成自己的本科毕业设计，有许多人对我给予了许多帮助，在此我对这些人表示真心的感谢。首先需要感谢的是申恒涛教授和高联丽副教授，他们让我拥有了在大媒体计算中心学习和进行科学研究的机会。尤其要感谢的是高联丽老师，她在我的学业和生活上给予我非常大的帮助，尤其是在我前期构思论文的框架中，给了我许多有建设性的意见和建议。在写论文的阶段，老师更是不厌其烦的给我讲解论文的写作技巧，修改我论文中不恰当的地方。高联丽老师严谨的科研态度，大气的为人风范对我的影响非常大，值得我在以后的学习生活中学习。在此，对高联丽老师表示崇高的敬意。

此外，还需要感谢的有大媒体计算中心的宋井宽教授，邵杰教授，沈复民副教授，杨阳教授，在我学习中给予的帮助，他们勤勉严谨的治学态度同样也对我产生了深远的影响。同时还要感谢的有我们组的同学们，他们是郭招，王轩瀚，曹良富，何涛，郭昱宇，陈岱渊，感谢他们给予我学习上的动力和生活上的关怀。

Joint Modeling Embedding and Translation to Bridge Video and Language

Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, Yong Rui
Microsoft Research, Beijing

1.1 Abstract

Automatically describing video content with natural language is a fundamental challenge of multimedia. Recurrent Neural Networks (RNN), which models sequence dynamics, has attracted increasing attention on visual interpretation. However, most existing approaches generate a word locally with given previous words and the visual content, while the relationship between sentence semantics and visual content is not holistically exploited. As a result, the generated sentences may be contextually correct but the semantics (e.g., subjects, verbs or objects) are not true. This paper presents a novel unified framework, named Long Short-Term Memory with visual-semantic Embedding (LSTM-E), which can simultaneously explore the learning of LSTM and visual-semantic embedding. The former aims to locally maximize the probability of generating the next word given previous words and visual content, while the latter is to create a visual-semantic embedding space for enforcing the relationship between the semantics of the entire sentence and visual content. Our proposed LSTM-E consists of three components: a 2-D and/or 3-D deep con- volutional neural networks for learning powerful video representation, a deep RNN for generating sentences, and a joint embedding model for exploring the relationships between visual content and sentence semantics. The experiments on YouTube2Text dataset show that our proposed LSTM-E achieves to-date the best reported performance in generating natural sentences: 45.3% and 31.0% in terms of BLEU@4 and METEOR, respectively. We also demonstrate that LSTM-E is superior in predicting Subject-Verb-Object (SVO) triplets to several state-of-the-art techniques. 1.

1.2 Introduction

Video has become ubiquitous on the Internet, broadcasting channels, as well as personal devices. This has encouraged the development of advanced techniques to analyze the semantic video content for a wide variety of applications. Recognition of videos has been a fundamental challenge of multimedia for decades. Previous research has predominantly focused on recognizing videos with a predefined yet very limited set of individual words. Thanks to the recent development of Recurrent Neural Networks (RNN), researchers have strived to automatically describe video content with a complete and natural sentence, which can be regarded as the ultimate goal of video understanding. Figure A-1 shows the examples of video description generation. Given an input video, the generated sentences are to describe video content, ideally encapsulating its most informative dynamics. There is a wide variety of video applications based on the description, ranging from editing, indexing, search, to sharing. However, the problem itself has been taken as a ground challenge for decades in the research communities, as the description generation model should be powerful enough not only to recognize key objects from visual content, but also discover their spatio-temporal relationships and the dynamics expressed in a natural language as well.

Despite the difficulty of the problem, there have been a few attempts to address video description generation, and image caption generation, which are mainly inspired by recent advances in machine translation using Recurrent Neural Networks (RNN). The standard RNN is a nonlinear dynamical system that maps sequences to sequences. Although the gradients of the RNN are easy to compute, RNN models are difficult to in logic but the subject “man” is not relevant to the video content. To address the above issues, we leverage the semantics of the entire sentence and visual content to learn a visual-semantic embedding model, which holistically explores the relationships in between. Specifically, we present a novel Long Short-Term Memory with visual-semantic Embedding (LSTM-E) framework to bridge video content and natural language, as shown in Figure 2. Given a video, a 2-D and/or 3-D Convolution Neural Networks (CNN) is utilized to extract visual features of selected video frames/clips, while the video representation is produced by mean pooling over these visual features. Then, a LSTM for generating video sentence and a visual-semantic embedding model are jointly learnt based on the video represen-

Input Video:

Output Sentence:

- **LSTM:** a man is riding a horse.
- **LSTM-E:** a woman is riding a horse.
- **Humans:** a woman gallops on a horse. / a woman is riding a horse along a road. / the girl rode her brown horse.

图 A-1 Examples of video description generation. Input: a short video. Output: a natural language sentence describing the main content of the input video.

tation and sentence semantics. The spirit of LSTM-E is to generate video sentence from the viewpoint of mutual reinforcement between coherence and relevance. Coherence expresses the contextual relationships among the generated words with video content which is optimized in LSTM, while relevance conveys the relationship between the semantics of the entire sentence and video content which is measured in the visual semantic embedding.

In summary, this paper makes the following contributions: 1） We present an end-to-end deep model for automatic video description generation, which incorporates both spatial and temporal structures underlying video. 2） We propose a novel Long Shot-Term Memory with visual-semantic Embedding (LSTM-E) framework, which considers both the contextual relationship among the words in sentence, and the relationship be- tween the semantics of the entire sentence and video content, for generating natural language of a given video. 3） The proposed model is evaluated on the popular Youtube2Text corpus and outperforms the-state-of- the-art in terms of both Subject-Verb-Object (SVO) triplet prediction and sentence generation.

连接视频与语言的联合建模嵌入与翻译

潘英伟，梅涛，姚婷，李厚强，芮勇
微软亚洲研究院

1.1 绪论

用自然语言自动描述视频内容是多媒体的根本挑战。循证神经网络（RNN），其专门处理动态的模型序列，已经引起视觉描述越来越多的关注。然而，大多数现有的方法是在给定的前一个单词和视觉内容的前提下生成一个单词，而句子语义和视觉内容之间的关系并没有被全面地利用。结果，生成的句子可以是上下文正确的，但是语义（例如，主题，动词或对象）不是真的。本文提出了一个新颖的统一框架，命名为具有视觉语义嵌入的长期记忆（LSTM-E），可以同时探索长短时记忆单元（LSTM）的学习和视觉语义嵌入。前者旨在在当地最大化生成给定前一单词和视觉内容的下一个单词的概率，而后者则是创建一个视觉语义嵌入空间，用于强化整个句子的语义与视觉内容之间的关系。我们提出的LSTM-E由三个组成部分组成：用于学习强大的视频表示的2维和/或3维深度约束神经网络，用于生成句子的深RNN和用于探索视觉之间的关系的联合嵌入模型内容和句子语义。在YouTube-2Text数据集上的实验表明，我们提出的LSTM-E分别达到了生成自然句子的最佳报告性能：分别为BLEU@4和METEOR的45.3％和31.0％。我们还表明，LSTM-E在预测主题-动词-对象（SVO）三元组方面优于几种最先进的技术。

1.2 介绍

视频在网络世界，电视频道以及个人设备当中逐渐变得更加普遍，这也促进了通过先进的设备来对广大应用领域的视频语义内容进行分析。近几十年来，我们也已经认识到了视频的识别已经成为一个多媒体领域的一个基本问题。之前的研究主要集中在通过预先限定的一些词来识别视频。由于近些年来循环神经网络的发展，研究者已经努力研究出了一种能够通过一个完整的句子来自动的描述视

频内容的技术，它的最终目标能够理解视频当中的语义信息。如图1当中所展示的就是视频描述生成的例子。给定一个输入视频，生成的句子就能够描述视频当中的具体内容，理想情况下能够动态的将视频内容压缩成一个语句。当前有大量的基于视频描述的应用基于包括编辑，索引，查找，共享等技术。然而，视频描述生成问题本身也已经在研究领域成为一个基础的研究课题，视频描述生成模型应该要足够强大，不仅能够识别出视觉上的关键内容，还要能够发现时空信息的关系并且能够通过一个自然的语句来动态的表述视频内容信息。

无论遇到了怎样的困难，受启发于近些年来运用循环神经网络来实现机器翻译的良好效果，仍然有许多视频描述生成和图像标题生成的方法的大胆尝试。标准的循环神经网络是一个非线性的，能够将一个时序信息映射到到另一个时序信息的动态系统。虽然循环神经网络的梯度计算十分简单，但是要想让循环神经网络模型懂得逻辑依然非常困难，尽管主语的“人”和视频的内容并不相关。为了解决上述问题，我们利用整个句子和视觉内容的语义来学习一个视觉语义嵌入模型，从整体上探讨其间的关系。具体来说，我们提出一个新颖的长短期内存与视觉语义嵌入（LSTM-E）框架来桥接视频内容和自然语言，如图所示。给定一个视频，一个二维或三维卷积神经网络（CNN）用于提取所选视频帧/剪辑的视觉特征，而视频表示是通过平均集中在这些视觉特征上产生的。然后，基于视频表示和句子语义联合学习用于生成视频语句的长短时记忆单元和视觉语义嵌入模型。LSTM-E的精神是从连贯性和相关性之间相互加强的角度来产生视频句子。一致性表示生成的词与LSTM中优化的视频内容之间的语境关系，而相关性传达了整个语句的语义与视觉语义嵌入中测量的视频内容之间的关系。

总而言之，本文作出以下贡献： 1）我们提出了一个自动视频描述生成的端到端深度模型，其中包含视频的空间和时间结构。 2）我们提出了一种新颖的具有视觉语义嵌入框架的长短时记忆模型，它考虑了句子中的单词之间的语境关系，以及整个句子的语义与视频内容之间的关系，用于产生给定视频的自然语言。3）所提出的模型是在受欢迎的Youtube2Text语料库上进行评估，在主语-动词-对象（SVO）三元组预测和句子生成方面效果都是最优异的。