# 基于全卷积神经网络的关键帧识别

张鹏杰 贾熠寒 马现涛 姜福伟

**摘要：**本文利用一种基于全卷积神经网络的关键帧识别方法，完成视频摘要任务，该方法利用卷积神经网络对视频序列进行特征提取，并通过卷积层实现特征的编码、解码，最终识别出视频的关键帧。实验结果表明，该方法在TVSum数据集上取得了良好的性能，是一种解决关键帧识别任务的有效解决方案。

**关键字：**全卷积神经网络，关键帧，视频摘要

# 第1章 介绍

1.1 应用背景

视频作为一种信息载体正在迅速发展，过去几年间，网络视频的数据量急剧增加，已经成为最重要的视觉数据形式之一。相比于图像，视频的帧与帧之间存在大量信息冗余，而互联网上的视频数量十分庞大，通过人工遍历观看，并提取有效信息是一种不切实际的做法。

因此，利用计算机视觉技术快速浏览大量视频数据，并从中提取感兴趣信息，是一种迫切的现实需求，视频摘要作为一种高效的批量处理技术，可以帮助我们处理数量庞大的视频数据，已成为计算机视觉技术中的一项热门任务。

1.2 视频摘要

关键帧就是几帧普通帧里面具有代表性的一帧，对于整个视频而言，相近帧之间信息冗余度很高，取局部相近帧中最有代表性的一帧作为关键帧，可以降低信息冗余度。

在给定的输入视频中，视频摘要的目标是创建一个更短的视频，以捕捉输入视频的重要信息，本文将视频摘要视为一个关键帧选择问题，用关键帧总结视频中的关键事件。

# 第2章 相关工作

2.1 基于循环模型的关键帧提取

基于循环网络的模型是关键帧提取常用的方法，因为视频是时序数据，不难想到利用LSTM设计双向的递归神经网络来解决该问题，图2-1为利用LSTM搭建的网络结构图，可以看到视频的时序输入与模型结构相符合。

但是，基于循环模型的方法需要在计算完成一个单元后，下一个单元才能得到输入，因此这种结构一次只能计算一帧，严重影响并行性，当视频序列很长时，并不能完美发挥其模型性能。

因此，本文使用全卷积序列模型完成视频摘要任务，参考语义分割模型的设计方法设计视频摘要模型。

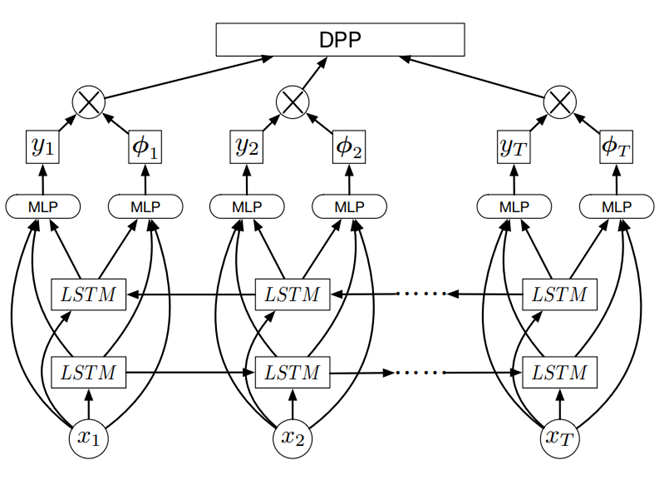


图2-1利用LSTM搭建的网络结构图

2.2 关键帧标记方式

视频摘要有两种标记方式：（1）二进制标签，只有0和1表示这一帧是否为关键帧，其中基于关键帧的标记方式对应一组非连续的帧；基于关键片段则是对应连续的帧组成的时间间隔。（2）重要性分值，对每一帧给定一个分数，分数越高则该帧越关键。本文使用基于关键帧的二进制标签。

本文采用二进制标签标记关键帧，标签为1表征该帧为关键帧。

# 第3章 基于全卷积神经网络的关键帧识别

3.1 模型架构

如图3-1所示，基于全卷积神经网络的模型Sum-FCN主要包括特征提取、下采样编码、上采样解码三部分。

特征提取。给定一个 帧的视频，利用提取每帧的特征向量，其维度记为，并按照2秒一帧的频率将视频均匀下采样至 帧，得到视频的特征描述序列。

下采样编码、上采样解码。将大小为的特征输入全卷积神经网络，经过五个下采样层，三个时序卷积层和两个反卷积层，获得大小为(1,T,C)的输出。其中C为类别数，因采用二进制标签，C为2。下采样的过程使用3\*3的卷积核，对应于语义分割中将长宽较大的图像卷积成更小、通道更多的特征，这里处理的是时间和特征维度。时序卷积层使用1\*1的卷积核，对应于语义分割中对不同通道的特征进行重组、拼接。

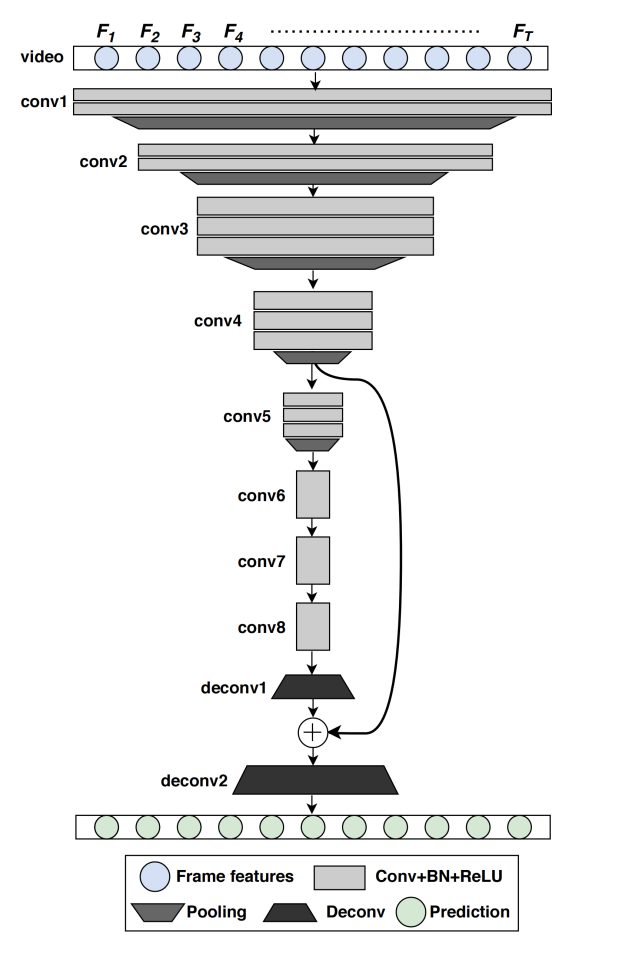


图3-1模型框架Sum-FCN

3.2 损失函数

利用交叉熵作为损失函数：

(1)

其中是第 𝑡 帧预测为 𝑐 的输出结果， 是真实标签。

因为关键帧数量远少于非关键帧，所以需要进行加权平衡，式（2）中是对标签的权重，为标签出现的频率，为各类的频率中值：

(2)

# 第4章 实验与结果

4.1 数据与预处理

现有的视频数据集包括：（1）SumMe：包含25个视频，涵盖运动、假期等场景事件，视频长度1.5到6.5分钟；（2）TVSum：包含10个不同类别的50个YouTuBe视频，视频通常为1到5分钟。

实验中，利用数据集的80%数据进行训练，20%数据进行测试，此外还可以利用YouTuBe数据集、OVP数据集作为增强数据参与训练。

4.2 定量结果

如表4-1所示，本文采用的模型在TVSum上具有最好的性能表现。

|  |  |  |
| --- | --- | --- |
| 数据集 | 方法 | 准确率 |
| TVSum | Zhang et al.(vsLSTM) | 54.2 |
| Zhang et al.(dppLSTM) | 54.7 |
| Mahasseni et al.(supervised) | 56.3 |
| Li et al. | 52.7 |
| SUM-FCN(ours) | **58.9** |

表4-1 模型在TVSum上的定量结果

4.3 定性结果

图4-1展示了SUM-FCN模型在TVSum上的定性结果，可见模型能够从冗长的视频中提取出关键帧作为视频摘要，实现视频压缩的目的。

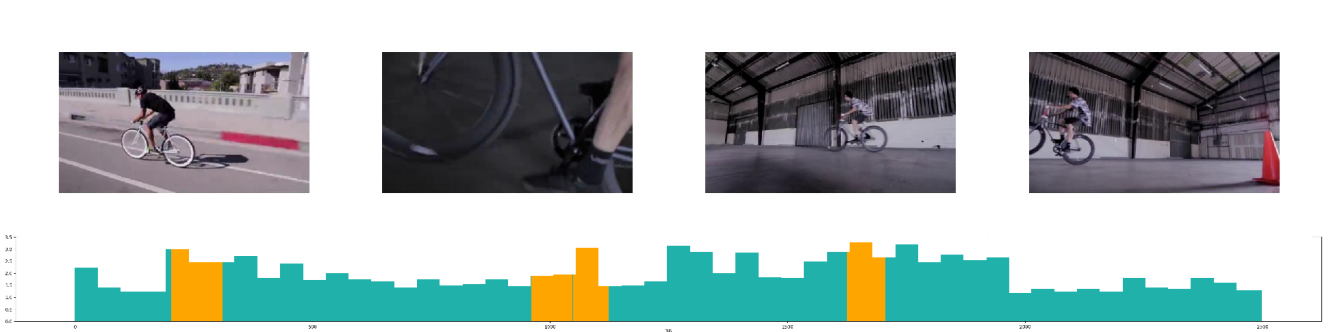


图4-1 模型在TVSum上的定性结果

# 第5章 结论

视频摘要是一种利用人工智能和计算机视觉方法的视频处理技术，从一个长视频中提取出最具代表性和信息量的短视频片段，从而实现对视频内容的快速浏览和理解。

本文利用全卷积神经网络的方法实现了视频的关键帧检测，从而生成了视频摘要，减少了视频的冗余信息，提高了视频的浏览效率。