Software Engineering and Applications 软件工程与应用, 2023, 12(3), 433-443
Published Online June 2023 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2023.123043

Hans汉斯

基于空洞卷积的多维度注意力视频摘要

秦佳林

上海理工大学，光电信息与计算机工程学院，上海

收稿日期：2023年3月18日；录用日期：2023年5月31日；发布日期：2023年6月12日

摘　要

智能手机、摄像机的普及导致视频数据每天呈指数级上升，准确可靠的视频摘要技术对视频概括、视频浏览和视频检索等具有重大意义。目前主流的视频摘要方法主要基于LSTM (Long Short-Term Memory)与卷积神经网络。但这些方法有固有的局限：一是LSTM每个时间步只能处理一帧，训练速度慢，不利于并行化，二是卷积网络中重复的下采样与最大池化操作导致大量细节信息丢失。基于上述问题本文提出一种基于空洞卷积的多维度注意力模型DCMAN (Dilated Convolutional Multi-Dimension Attention Network)。该模型首先利用级联空洞卷积网络提取视频的短期时间信息和视觉特征，不使用最大池化，同时设置合适的膨胀系数保证网络感受野不受影响。其次，空间与通道注意力的结合捕获视频的长期依赖，给每个视频帧赋予对应的重要性分数。最后，更多的跳连接结构融合更多尺度的信息，不一样的初始化注意力查询向量带来更完整的视频信息。实验在四个公共的视频摘要数据集上进行，实验结果表明本文提出的DCMAN模型明显优于与其它最新的视频摘要方法。

关键词

视频摘要，空洞卷积，深度学习，自注意力机制，计算机视觉

Video Summarization Using a Dilated Convolutional Multi-Dimension Attention Network

Jialin Qin

School of Optical-Electrical Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Mar. 18th, 2023; accepted: May 31st, 2023; published: Jun. 12th, 2023

Abstract

The popularity of smartphones and video cameras have led to an exponential increase in video

文章引用: 秦佳林. 基于空洞卷积的多维度注意力视频摘要[J]. 软件工程与应用, 2023, 12(3): 433-443.
DOI: 10.12677/sea.2023.123043

data every day. Accurate and reliable video summarization techniques are of great significance for video summarization, video browsing, and video retrieval. The current mainstream video summarization methods are based on LSTM (Long Short-Term Memory) with convolutional neural networks. However, these methods have inherent limitations: first, LSTM can only process one frame per time step, which is slow in training and not conducive to parallelization, and second, the repetitive down-sampling and maximum pooling operations in convolutional networks lead to the loss of a large amount of detailed information. Based on the above problems, this paper proposes a multidimensional attention model based on cavity convolution DCMAN (Dilated Convolutional Multi-Dimension Attention Network). The model firstly extracts short-term temporal information and visual features of the video using cascaded dilated convolutional network without using maximum pooling, while setting appropriate expansion coefficients to ensure that the network perceptual field is not affected. Second, the combination of spatial and channel attention captures the long-term dependence of the video, assigning a corresponding importance score to each video frame. Finally, more hop-connected structures fuse more scales of information, and different initialized attention query vectors bring more complete video information. Experiments are conducted on four public video summarization datasets, and the experimental results show that the DCMAN model proposed in this paper significantly outperforms with other state-of-the-art video summarization methods.

Keywords

Video Summarization, Dilated Convolution, Deep Learning, Self-Attention, Computer Vision

1. 引言

近十几年来，随着互联网的日益普及，拍摄设备如手机、相机的普及以及成本的下降，视频正逐渐成为信息交流的媒介以及大众获取网络信息的重要形式之一，视频的数量每秒正以指数级爆炸形式的速度增长，因此开发一些智能技术来高效地检索，分析大量的视频就变得非常迫切。视频摘要[1]就是最有前景的解决上述问题的技术之一。视频摘要目的是将一个视频压缩成一个简短的摘要。与原视频相比，摘要保持主要的语义信息的同时，大大缩短了原视频的长度。

视频摘要首次被 Pfeiffer [2]提出后，由于其巨大的实际应用意义，引发了众多学者的关注。早期的视频摘要技术大多基于手工标准方法[3] [4]，随着近年来深度学习的发展，许多基于深度学习的视频摘要方法被相继提出，并且取得不错的效果。这些方法主要可以分为无监督与有监督，由于没有人工注释作为参考，无监督的方法不能很好的代表原视频的内容，因此，有监督的方法往往比无监督方法效果更好。本文将采取有监督的方法。最有代表性的有监督视频摘要是基于 LSTM [5] (Long Short-Term Memory)。但 LSTM 有它固有的局限性，每一帧都需要等待前一帧处理完，不能实现并行化处理以充分利用 GPU 硬件，并且这种网络结构受视频长度影响，无法捕获长距离的时间依赖性，还存在梯度消失等问题。卷积神经网络(Convolutional Neural Networks, CNN) [6]是另外一种主要的视频摘要方法，但这类网络中存在大量的下采样和最大池化操作，这将导致大量视频细节信息的丢失，且无法通过上采样恢复。自注意力[7]近年来也被用于处理视频摘要问题，得益于其高效的计算效率，取得了不错的效果，但目前基于自注意力的视频摘要方法都忽略了通道间的注意力。受上述工作启发，本文提出一种基于空洞卷积的多维度注

意力网络(Dilated Convolutional Multi-Dimension Attention Network, DCMAN)。

该网络摒弃传统的 LSTM 结构，使用级联空洞卷积网络替代普通的卷积网络，空洞卷积网络不包含下采样与最大池化，通过设置合作的膨胀系数扩大网络感受野，设置 padding 大小保证原特征图大小不变，设置更多的跳连接结构融合上下不同尺度的信息，此外，卷积网络融合了包含空间与通道的多维度自注意力网络捕获视频的长期依赖性。本文的主要贡献如下：

1) 提出一种级联空洞卷积结构，该网络不包含下采样与最大池化操作，最大限度保留了视频的细节信息，通过设置 padding 保证原特图大小不变，并且设置合适的膨胀系数扩大网络的感受野。该网络还包含更多的跳连接结构，融合了更多不同尺度的上下文信息，提高摘要的丰富性。

2) 在卷积网络后融入为高质量视频摘要分配重要性分数的注意力感知模块，该模块包含空间与通道注意力，更大程度提高了视频摘要的质量。

3) 在四个公共数据集上进行大量实验，实验结果表明本文的方法优于目前最新的方法。

2. 国内外研究现状

人工智能，机器学习，深度学习是当今学术界的热门词汇，深度学习已经被证实是普适性最先进的技术之一，目前优秀的视频摘要方法大多都基于深度学习。Zhang 等人[8]使用 LSTM 建模视频帧之间不确定时间依赖性，并通过行列式点过程增加视频帧的多样性表示。Zhao 等人[9]提出双层 LSTM 结构。第一层用来提取和编码输入特征序列，第二层通过第一层编码后的信息选择视频的关键片段。在此基础上，Zhao 等人[10]新增了一个训练好的组件辨别镜头级别视频的时间结构，并通过这些知识生成关键镜头形式的视频摘要。Ji 等人[11]引入自注意力机制，自适应地调整当前状态对上下文状态的注意力权重，学习对视频摘要更重要的视频帧，这项工作在文献[12]得到提升，在网络中插入了一个语义保留网络。Lal [13]等人提出带有卷积 LSTM 的编码解码结构，通过镜头检测机制增强摘要的视觉丰富性。

Rochan 等人[14]摒弃 LSTM 结构，在语义分割与视频摘要之间建立了联系，建立了全卷积序列网络-FCSN 处理视频摘要问题，实现了网络的并行化处理，但是该网络忽略了潜在的时间依赖性，一样无法捕获长期依赖，并且网络中重复的下采样与最大池化操作导致很多细节信息丢失，且无法通过上采样恢复。Fajtl 等人[15]提出一种序列到序列的纯注意力网络，对于整个视频序列，通过简单矩阵乘法就可以获得每个视频帧的重要性，大大降低了计算复杂度，取得了不错的效果，但确忽略了通道间的注意力。Gupta 等人[16]将卷积与注意力结合在一起，Liang 等人[17]将卷积，注意力，LSTM 以生成对抗的方式融合在一起。但他们卷积网络中一样包含了重复的下采样与最大池化操作以及忽略了通道间的注意力。

3. 整体设计

3.1. 模型架构

本文将基于有监督的视频摘要视作为序列到序列的预测问题，并提出 DCMAN 模型，与传统的编码解码网络不同，DCMAN 无需固定长度的中间隐藏层状态，这样对于较长的视频序列不会导致较高的信息丢失。模型架构如图 1 所示，我们认为视频是帧的集合，每一帧类似于一张图像，因此会包含一定量的冗余信息，因此，本文不是直接处理原始视频，而是先执行预采样，目的是在不丢失任何信息的情况下降低模型的计算成本。如图 1 所示，模型的输入即预采样后的特征定义为 $F=[F_0,F_1,\cdots,F_N]\in\mathbb{R}^{F*d}$。其中 F 表示视频的帧数，$d$ 表示视频特征的维度。全卷积以 F 为输入生成一个中间特征序列 $P=[P_0,P_1,\cdots,P_N]\in\mathbb{R}^{F*d}$。将原始特征 F 与得到的中间特征序列 P 一同输入到空间注意力与通道注意力中，最终产生序列 $Y=[Y_0,Y_1,\cdots,Y_N]\in\mathbb{R}^{F*d}$，代表每一视频帧的重要性分数预测。DCMAN 模型整个包含两大模块，首先是空洞卷积模块，之后是自注意力模块。下文将详细介绍这两个模块。

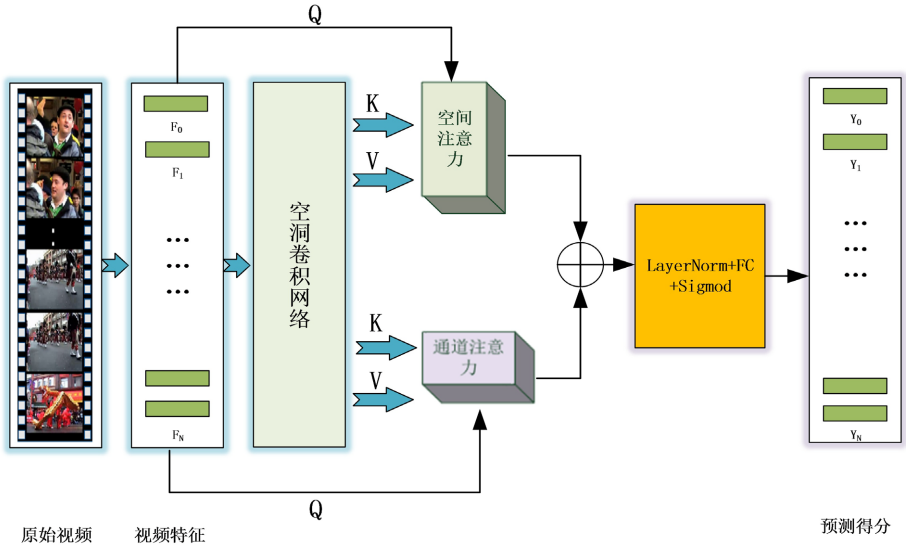

Figure 1. The architecture of DCMAN network
图 1. DCMAN 模型架构

3.2. 级联空洞卷积网络

本文使用级联空洞卷积网络作为视频特征的全局表示提取器，在时间维度上执行一系列卷积操作。与 Rochan [14]中卷积网络结构不同，首先本文不包含重复的下采样与池化操作，因为这将导致视频特征大量细节信息丢失。本文采用空洞卷积，空洞卷积的优势在于既可以通过设置 padding 保证原特征图大小不变，又可以通过设置合适的膨胀系数扩大网络感受野。此外，本文在不同的网络层之间添加了比原文更多的残差连接结构，将浅层的粗糙特征与深层的精细特征结合来获取更多视频的时间信息。级联空洞卷积架构如图 2 所示，首先使用两个三重卷积层对帧特征进行初步提取，三重卷积层由三个 3 × 3 时间卷积层组成，每个时间卷积层后面都会加一个批处理归一化和一个 RELU 激活。然后，使用 4 个三重时间卷积层来扩展网络的感受野。这里的三重卷积层由三个 3 × 3 时间空洞卷积层组成。类似地，每个时间扩张卷积层之后是批处理归一化和 RELU 激活。每三个卷积核膨胀系数依次设置为[1] [6] [12] (下文实验部分将验证该数据设置)，类似于一种锯齿结构，从而能够从不同尺度提取上下文信息，然后通过元素加法的形式融合多尺度上下文信息，得到更丰富的时间信息。

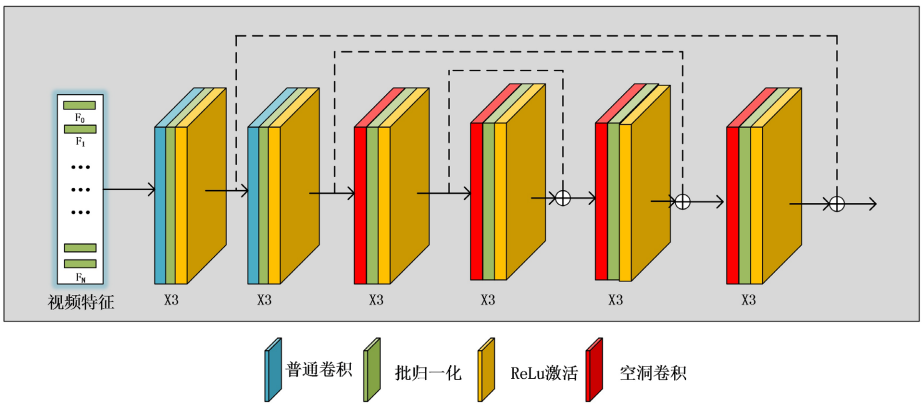

Figure 2. The architecture of cascaded dilated convolutional network
图 2. 级联空洞卷积网络架构

3.3. 多维度自注意力网络

虽然上文中经过改进的全卷积序列网络可以捕获视频的全局表示和短期依赖，但是考虑长期的时间依赖关系对于生成一个高质量摘要来说也是至关重要的。因此，本文在卷积网络后融入自注意力机制来弥补这一缺陷。与前人研究不同，本文不仅融入空间注意力机制，而且将通道注意力也考虑其中。通道注意力在图像领域已经被证明具有重要作用，而在上文中阐明了图像数据与视频数据之间的相似性。因此，我们认为通道注意力对处理视频数据也

应具备相当的益处，下文实验部分也将证明此观点。

自注意力机制计算过程可以分为三大步，如图 3 所示。首先根据网络的输入初始化查询向量，键向量以及值向量，与大多数研究有所不同，本文不直接将卷积网络的输出作为初始化查询向量的前提，而是通过原始特征 F 初始化查询向量，这样可以获取关于整个视频更加完整的信息，这一观点也将在下文实验中证实。接着对于网络的每一个输入，通过查询向量与其余所有键向量做乘法得到注意力权重。最后，把值向量与权重相乘后求和，得到每个向量对应所有向量的权重。

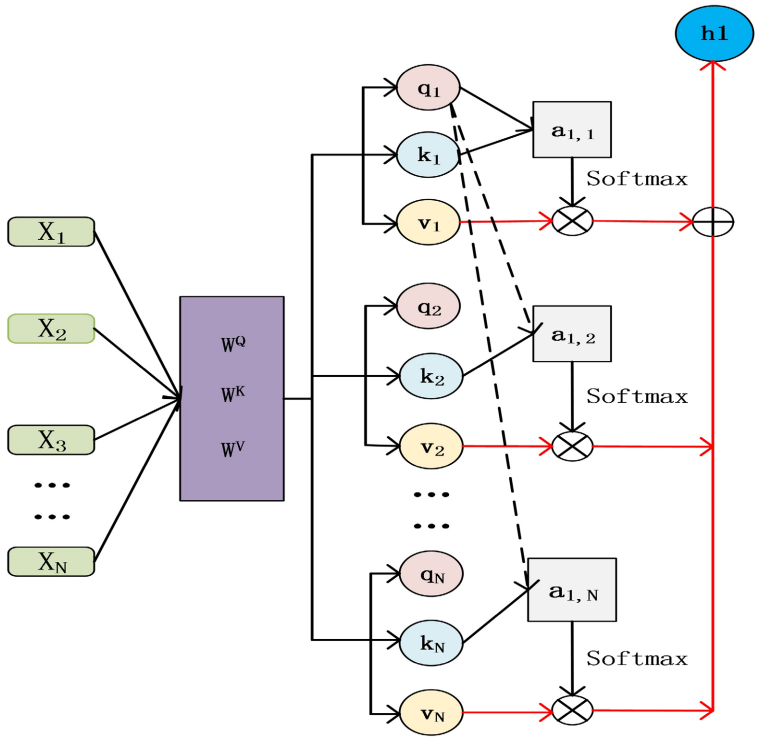

Figure 3. The calculation process of self-attention
图 3. 自注意力计算过程

由前文可知，首先我们需要获取查询，键以及值向量，模型输入为 $F=[F_0,F_1,\cdots,F_N]\in\mathbb{R}^{F*d}$，卷积网络的输出为 $P=[P_0,P_1,\cdots,P_N]\in\mathbb{R}^{F*d}$，则查询向量 Q，键向量 K，值向量 V 可以表示：

$$Q=FW^Q\in\mathbb{R}^{F*d} \tag{1}$$

$$K=XW^K\in\mathbb{R}^{F*d} \tag{2}$$

$$V=XW^V\in\mathbb{R}^{F*d} \tag{3}$$

其中 $W^Q \in \mathbb{R}^{d*d}$，$W^K \in \mathbb{R}^{d*d}$，$W^V \in \mathbb{R}^{d*d}$ 分别是网络优化时待学习的权重矩阵，F 表示视频的帧数，d 表示视频特征的维度。

接着需要计算注意力分数。第一步利用得到 Q 和 K 以及缩小的点积计算任意某个时刻 t 的输入特征 Fi 与整个序列之间的相关性，第二步将得到的相关性乘以 o，o 是一个缩放参数，目的是削弱查询向量与键向量之间的点积值，这样做利于网络的反向传播。空间与通道注意力分数的计算过程分别如式(4)，式(5)所示，其中<.>表示点积计算。

$$E_{spa}^{t}\left(Q_i,K_i\right)=<Q_i,K_i^T>*o,\forall i,t\in[N] \tag{4}$$

$$E_{cha}^{t}\left(Q_i,K_i\right)=<Q_i^T,K_i>*o,\forall i,t\in[N] \tag{5}$$

然后根据得到的注意力分数计算注意力权重，空间与通道注意力权重分别如式(6)，式(7)所示，其中都使用 softmax 函数进行归一化操作。

$$Spa_{t,i}=\text{softmax}\left(E_{spa}^{t}\left(Q_i,K_i\right)\right)V \tag{6}$$

$$Cha_{t,i}=\text{softmax}\left(E_{cha}^{t}\left(Q_i,K_i\right)\right)V \tag{7}$$

将得到的权重进行求和之后，使用两层神经网络对求和结果进行回归操作，计算如式(8)所示。与原始的 transformer 结构一样，第一层神经网络对求和的结果施加层归一化(Layer Normalization)，避免模型过拟合。接着使用两个线性层分别是 1024 个神经元和 1 个神经元从而可以得到我们想要的维度。第二层使用 Sigmoid 预测最终的视频帧得分。

$$Y=sigmoid\left(linear\left(norm\left(Spa_{t,i}+Cha_{t,i}\right)\right)\right) \tag{8}$$

最后，我们选择均方误差(MSE, Mean Squared Error)损失函数作为网络的目标函数。MSE 通过计算网络预测分数与人工注释分数的差值的平方和后最后再求平均，如式(9)所示，通过反向传播算法使得目标函数损失值达到

最小，得到最优解。

$$MSELoss\left(S,Y\right)=\frac{1}{n}\sum_{i=1}^{n}\left(S_i-Y_i\right)^2 \tag{9}$$

4. 实验设计与验证

4.1. 视频摘要数据集与数据集设置

我们在两个常见的公共数据集上进行了实验，即 SumMe 和 TVSum。SumMe 由 25 个视频组成，涵盖如体育和假期等主题。每个视频持续时间为 1.5 到 6 分，这些视频由 15 至 18 名人类以帧级重要性得分形式进行注释。TVSum 包含 50 个不同主题的视频。每个单独视频的持续时间为 1 至 5 分钟，注释形式为帧级别重要性得分，鉴于这两个数据集的小规模不足以训练神经网络，我们还使用 OVP (50 个视频)和 YouTube (39 个视频)来加强训练数据集。OVP 数据集包含纪录片和其他不同类型的视频。YouTube 的数据集视频有不同的主题，如新闻和体育。这两个数据集以关键帧进行注释。四种数据集信息如表 1 所示。

根据数据集介绍可知有两种数据集设置：未增强与增强。未增强模式下仅对 SumMe 与 TvSum 数据集进行训练，将其中 80%数据集作为训练集，其余作为测试集。增强模式下是在未增强的基础上增加了其余三个数据集扩充了原先的训练集。本文按照 80%训练集，20%测试集比例随机划分 5 次，并将 5 次结果的平均作为最终的模型的表现分数。具体设置信息如表 2 所示。

Table 1. Dataset description
表 1. 数据集信息

数据集名称	视频个数	内容	时长(min)	用户注释数	注释形式
SumMe	25	人工拍摄	1.5~6	18	帧级得分
TvSum	50	网站视频	1~10	20	帧级得分
OVP	50	纪录片等	1~4	5	关键帧
YouTube	39	网站视频	1~10	5	关键帧

Table 2. Dataset settings
表 2. 数据集设置

数据集名称	模式	训练集	测试集
SumMe	未增强	80% SumMe	20% SumMe
SumMe	增强	80% SumMe + TvSum + OVP + YouTube	20% SumMe
TvSum	未增强	80%TvSum	20% TvSum
TvSum	增强	80% TvSum + SumMe + OVP + YouTube	20% TvSum

4.2. 评价指标与环境

为了与其他方法公平比较，本文采用其他大多视频摘要方法采用的指标 F_Score 作为评判标准，同时，F_Score 能够评估模型摘要与用户摘要的相似性。假设 S_0 是模型预测摘要，S_1 是用户打分摘要，然后通过 S_0 和 S_1 的时间重叠部分计算精确率 P 和召回率 R。其中 P 表示预测正确的摘要占整个算法预测摘要长度的比例，R 表示预测正确的摘要占整个注释摘要的长度的比例。P 和 R 的计算方法如下：

$$P=\frac{S_0\text{与}S_1\text{时间重叠部分}}{\text{算法输出摘要长度}S_0} \tag{10}$$

$$R=\frac{S_0\text{与}S_1\text{时间重叠部分}}{\text{注释摘要长度}S_1} \tag{11}$$

则 F_Score 可以定位为：

$$F=2\times\frac{P\times R}{P+R}\times 100\% \tag{12}$$

与其他研究一样，本文采用五折交叉验证的测试方法。并且对每个视频的下采样率设置为 2 fps，下采样后每一视频帧维度为 1024 维，该过程是通过从 ImageNet 训练的 GoogLeNet 网络的倒数第二层提取出来的。训练过程中，采用 ADAM 优化器以及 L2 的正则化，epoch 设置为 300，学习率 rate 设置为 5×10^{-5}。本章实验的软硬件配置为：PyTorch 1.10.1，python3.6 内存为 16 GB 的 NVIDIA GeForce GTX 3090 GPU 的计算机。

此外，视频经模型处理后产生序列化的重要性得分。在评估时，本文采用 KTS [3]算法把这些帧级别的分数转换为镜头级别的分数，再由这些镜头级别的分数作为依据挑选出重要镜头形成最终的摘要视频。

4.3. 实验结果与分析

为了验证本文 DCMAN 模型的有效性，将 DCMAN 模型与最新的 5 个模型进行比较，包括基于注意力机制的，基于 LSTM 以及改进 RNN 的，以及基于全卷积序列网络的。因 DCMAN 模型是基于有监督

的，而一般无监督的方法的结果会比有监督方法差，所以本章只挑选有监督的方法作为对比对象。1) Zhang [8]等人是第一个运用 LSTM 捕获视频中前向和后向信息，并提出 DPP 结构提升关键帧之间的差异性。2) Fajtl [15]提出一种纯注意力机制，序列到序列的网络——VANSNet。3) Ji 等人[11]提出 A-AVS 和 M-AVS，都是基于注意力编码解码，其中 M-AVS 是乘法形式注意力，A-AVS 则是加法形式注意力。4) Rochan [14]阐述了语义分割任务与视频摘要之间的联系，并将主流的语义分割网络改造后应用于视频摘要中，取得了不错的表现。5) Mahasseni [18]提出基于生成对抗网络(Generative Adversarial Network, GAN)的方法，目标是最小化重建摘要于人工注释之间的距离，本文只引用其中有监督的方法。这 5 种方法的实验结果均取自于原文。

表 3 展示了 DCMAN 模型与其他方法在 SumMe 和 TvSum 数据集上的结果对比，涵盖了增强与未增强两种数据集设置。显而易见，DCMAN 模型在两个数据集上的表现领先其他方法。M-AVS 略领先 A-AVS，说明乘法形式的注意力更为有效，且运算速度方面更有优势。VANSET 略优于 FCSN。FCSN 虽然获取到视频的全局表示与短期依赖，但视频的长期依赖性的捕捉没有 VANSET 完善。VANSET 明显优于 DPP-LSTM 与 GAN (sup)，因为 LSTM 也无法捕获视频的长期依赖，且运算速度慢，生成对抗方法会给摘要结果带来巨大的不稳定性。在 SumMe 数据集上，本文提出的 DCMAN 优于 VANSET，在未增强设置下比 VANSET 提升了 1.87%，在增强模式下提升了 1.38%，在 TvSum 数据集上，在未增强设置下比 VANSET 提升了 0.88%，在增强模式下提升了 1.14%，这证明通道注意力对结果提升是有益处的。此外，与 FCSN 比较，DMCAN 提升效果更加明显，因为 DMCAN 中级联空洞卷积网络保留了视频大量的细节信息，而且还融合了注意力模块以及更多的跳连接结构，前者捕获了视频的长期依赖性，后者融合了更多的上下尺度信息。

综合表 3 的结果，DCMAN 优于其他方法的原因可以总结为以下几点：

1) 一维卷积对处理序列问题更具优势，提取视频帧之间的时间信息，而且无需关注相关的位置信息；2) 一维卷积网络中丢弃了最大池化和反卷积操作，保留了更多的细节信息，同时空洞卷积网络保证感受野不受影响，此外，更多的跳连接结构提供更多不同尺度的信息，有效提升了分类正确率；3) 注意力机制保证了网络不受序列长度的影响，并且通道注意力的加入对建模视频长期依赖性又提升了一个高度。

Table 3. Comparison of DCMAN model with other methods
表 3. DCMAN 模型与其他方法对比

Method	SumMe		TVSum	
	未增强	增强	未增强	增强
DPP-LSTM	38.6	42.9	54.7	59.9
VANSNet	49.71	51.09	61.42	62.37
A-AVS	43.9	44.6	59.4	60.8
M-AVS	44.4	46.1	61.0	61.8
FCSN	48.8	50.2	58.4	59.1
GAN (sup)	41.7	43.6	56.3	59.6
DCMAN (ours)	**51.58**	**52.47**	**62.30**	**63.51**

4.4. 消融实验

本节评估 DCMAN 模型相关的变种模型的表现，以证明相关组件的影响。为了简捷起见，只在未增

强数据设置下进行实验。探究了三种变体模型 DCMAN (FSCN)，DCMAN (w/o-CA)，DCMAN (Q = P)。第一个代表使用文献[14]中的 FCSN 模型替代 DCMAN 中的空洞卷积结构。第二个代表去除 DCMAN 中通道注意力结构。第三个代表使用空洞卷积网络的输出 P 作为初始化查询向量 Q 的网络结构。三个变种模型的效果如图 4 所示。DCMAN (FSCN)，DCMAN (w/o-CA)的表现都没有 DCMAN 好，证明了 DCMAN 模型空洞卷积网络结构和通道注意力能更好提升模型性能。DCMAN (Q = P)的表现不如 DCMAN 说明本文用原始特征 F 初始化查询向量 Q 确实能获得更完整的视频信息。

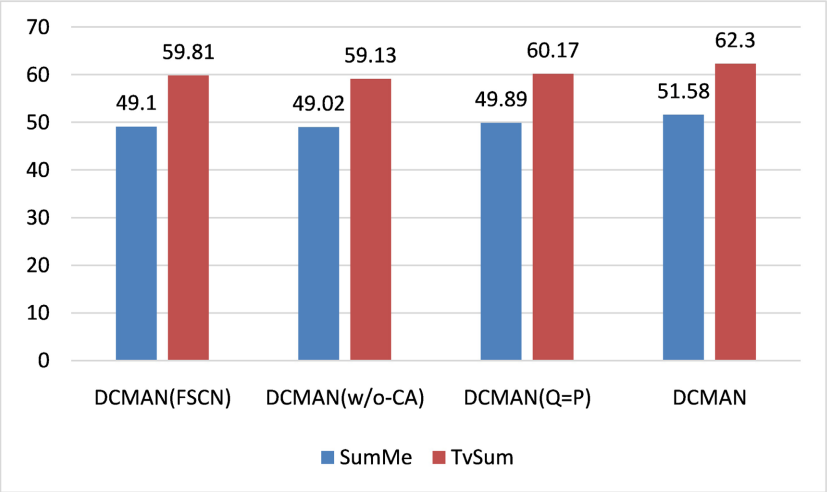

Figure 4. Comparison of the effect of DCMAN's variant models
图 4. DCMAN 的变种模型效果对比

4.5. 参数分析

本节针对 DCMAN 设置不同膨胀系数的情况下，对模型的表现的影响进行了实验。为了简捷起见，只在未增强数据设置下进行实验。结果如表 4 所示，刚开始随着膨胀系数增大，效果也越好，这得益于模型能提取到更丰富的上下不同尺度的信息，但当膨胀系数进一步增大时，如表 4 第四行所示，模型表现下降，本文分析这是由于随着膨胀系数增大，为保证输入特征图大小尺寸不变，padding 也要随之增大，导致模型提取到许多无效的信息。最终实验结果表明当膨胀系数为[1] [6] [12]时，取得最佳表现。

Table 4. Effect of different expansion coefficients on the model
表 4. 不同膨胀系数对模型的影响

膨胀系数	SumMe	TvSum
[1] [2] [4]	46.70	57.21
[1] [6] [12]	51.58	62.30
[1] [12] [24]	49.72	60.11

5. 结束语

本文提出一种级联空洞卷积的多维度注意力视频摘要网络 DCMAN。其中级联空洞卷积结构提取视频的全局表示以及捕获了视频数据的短期依赖，该网络不包含了重复的下采样以及最大池化操作，最大程度保证了视频的细节信息的保留，同时通过设置合适的膨胀系数扩大网络的感受野。在级联空洞卷积

网络后融入了空间与通道注意力机制，捕获了视频信息的长期依赖。最后在四个公共数据集上验证DCMAN模型的性能，实验结果表明，优于其他最新方法以及基线模型。

本文设计的模型仍有不足之处，模型结构过于庞大，导致训练时间长，在未来的工作中需要考虑在不影响模型性能的前提下，缩小模型的结构。同时考虑应用更多先进的语义分割模型在视频摘要领域。

参考文献

[1] Li, P., Ye, Q.H., Zhang, L.M., Yuan, L., Xu, X. and Shao, L. (2021) Exploring Global Diverse Attention via Pairwise Temporal Relation for Video Summarization. *Pattern Recognition*, **111**, Article ID: 107677. https://doi.org/10.1016/j.patcog.2020.107677

[2] Pfeiffer, S., Lienhart, R., Fischer, S. and Effelsberg, W. (1996) Abstracting Digital Movies Automatically. *Journal of Visual Communication and Image Representation*, **7**, 345-353. https://doi.org/10.1006/jvci.1996.0030

[3] Potapov, D., Douze, M., Harchaoui, Z. and Schmid, C. (2014) Category-Specific Video Summarization. In: Fleet, D., Pajdla, T., Schiele, B. and Tuytelaars, T., Eds., *Computer Vision—ECCV* 2014, Springer, Cham, 540-555. https://doi.org/10.1007/978-3-319-10599-4_35

[4] Liu, T. and Kender, J.R. (2002) Optimization Algorithms for the Selection of Key Frame Sequences of Variable Length. *Computer Vision—ECCV* 2002: 7*th European Conference on Computer Vision Copenhagen*, Denmark, 28-31 May 2002, 403-417. https://doi.org/10.1007/3-540-47979-1_27

[5] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[6] Li, Z., Liu, F., Yang, W., Peng, S. and Zhou, J. (2021) A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems*, **33**, 6999-7019. https://doi.org/10.1109/TNNLS.2021.3084827

[7] Zhao, H., Jia, J. and Koltun, V. (2020) Exploring Self-Attention for Image Recognition. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 10073-10082. https://doi.org/10.1109/CVPR42600.2020.01009

[8] Zhang, K., Chao, W.L., Sha, F. and Grauman, K. (2016) Video Summarization with Long Short-Term Memory. In: Leibe, B., Matas, J., Sebe, N. and Welling, M., Eds., *Computer Vision—ECCV* 2016, Springer, Cham, 766-782. https://doi.org/10.1007/3-540-47979-1_27

[9] Zhao, B., Li, X. and Lu, X. (2017) Hierarchical Recurrent Neural Network for Video Summarization. *Proceedings of the* 25*th ACM International Conference on Multimedia*, Los Cabos, 23-27 October 2017, 863-871. https://doi.org/10.1145/3123266.3123328

[10] Zhao, B., Li, X. and Lu, X. (2018) HSA-RNN: Hierarchical Structure-Adaptive RNN for Video Summarization. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake, 18-23 June 2018, 7405-7414. https://doi.org/10.1109/CVPR.2018.00773

[11] Ji, Z., Xiong, K., Pang, Y., Member, S. and Li, X. (2020) Video Summarization with Attention-Based Encoder—Decoder Networks. *IEEE Transactions on Circuits and Systems for Video Technology*, **30**, 1709-1717. https://doi.org/10.1109/TCSVT.2019.2904996

[12] Ji, Z., Jiao, F., Pang, Y. and Shao, L. (2020) Deep Attentive and Semantic Preserving Video Summarization. *Neurocomputing*, **405**, 200-207. https://doi.org/10.1007/3-540-47979-1_27

[13] Lal, S., Duggal, S. and Sreedevi, I. (2019) Online Video Summarization: Predicting Future to Better Summarize Present. 2019 *IEEE Winter Conference on Applications of Computer Vision* (*WACV*), Waikoloa, 7-11 January 2019, 471-480.

[14] Rochan, M., Ye, L. and Wang, Y. (2018) Video Summarization Using Fully Convolutional Sequence Networks. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *Computer Vision—ECCV* 2018, Springer, Cham, 347-363. https://doi.org/10.1007/978-3-030-01258-8_22

[15] Fajtl, J., Sokeh, H.S., Argyriou, V., Monekosso, D. and Remagnino, P. (2019) Summarizing Videos with Attention. In: Carneiro, G. and You, S., Eds., *Computer Vision—ACCV* 2018 *Workshops*, Springer, Cham, 39-54. https://doi.org/10.1007/978-3-030-21074-8_4

[16] Gupta, D. and Sharma, A. (2021) Attentive Convolution Network-Based Video Summarization. In: Choudhary, A., Agrawal, A.P., Logeswaran, R. and Unhelkar, B., Eds., *Applications of Artificial Intelligence and Machine Learning*, Springer, Singapore, 333-346. https://doi.org/10.1007/978-981-16-3067-5_25

[17] Liang, G., Lv, Y., Li, S., Zhang, S. and Zhang, Y. (2021) Unsupervised Video Summarization with a Convolutional Attentive Adversarial Network. http://arxiv.org/abs/2105.11131

[18] Mahasseni, B., Lam, M. and Todorovic, S. (2017) Unsupervised Video Summarization with Adversarial LSTM Networks. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 202-211. https://doi.org/10.1109/CVPR.2017.318