软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(Suppl.(2)):103−112

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于 3D 卷积神经网络的无参考视频质量评价*

王春峰[1], 苏 荔[1,2], 张维刚[1,3], 黄庆明[1,2]

[1](中国科学院大学 数据挖掘与知识管理重点实验室,北京 100049)

[2](中国科学院 计算技术研究所 智能信息处理重点实验室,北京 100190)

[3](哈尔滨工业大学(威海) 计算机与科学与技术学院,山东 威海 264209)

通讯作者: 苏荔, E-mail: suli@ucas.ac.cn

摘 要: 无参考视频质量评价(NR-VQA)在无法获得原始高质量视频参照的前提下,对失真视频的视觉质量进行定量度量.常规 NR-VQA 方法通常针对特定失真类型设计,或者与人的主观感受存在偏差.首次将 3D 深度卷积神经网络(3D-CNN)引入到了视频质量评价中,提出了一种基于 3D-CNN 的无参考视频质量评价方法,可以适用于非特定失真类型的 NR-VQA.首先,通过 3D 块来有效学习和表征视频内容的时空特征.其次,对常规的 3D 卷积网络模型进行改进,使其适用于视频质量评价的任务.实验结果表明,所提出的方法在多种失真类型和多个测试指标上,与人的主观感知一致性较高.作为无参考视频质量评价方法,其性能与许多全参考评价方法具有可比性,同时比主流的 NR-VQA 方法具有更快的运行速度,这使得所提模型在实际中具有更好的应用前景.

关键词: 视频质量评价;3D;深度卷积神经网络;无参考;全参考

中文引用格式: 王春峰,苏荔,张维刚,黄庆明.基于 3D 卷积神经网络的无参考视频质量评价.软件学报,2016,27(Suppl.(2)):103−112. http://www.jos.org.cn/1000-9825/16025.htm

英文引用格式: Wang CF, Su L, Zhang WG, Huang QM. No reference video quality assessment based on 3D convolutional neural network. Ruan Jian Xue Bao/Journal of Software, 2016,27(Suppl.(2)):103−112 (in Chinese). http://www.jos.org.cn/1000-9825/16025.htm

No Reference Video Quality Assessment Based on 3D Convolutional Neural Network

WANG Chun-Feng[1], SU Li[1,2], ZHANG Wei-Gang[1,3], HUANG Qing-Ming[1,2]

[1](Key Laboratory on Big Data Mining and Knowledge Management, University of Chinese Academy of Sciences, Beijing 100049, China)

[2](Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100190, China)

[3](Institute of Computer Science and Technology, Harbin Institute of Technology (Weihai), Weihai 264209, China)

Abstract: No reference video quality assessment (NR-VQA) measures distorted videos quantitatively without the reference of original high quality videos. Conventional NR-VQA methods are generally designed for specific types of distortions, or not consistent with human's perception. This paper innovatively introduces 3D deep convolutional neural network (3D-CNN) into VQA and proposes a 3D-CNN based NR-VQA method, which is universal for non-specific types of distortions. First, the proposed method utilizes 3D patches to learn spatio-temporal features that represent video content effectively. Second, the original 3D-CNN model is modified which is used to classify videos to make it adapt to VQA task. Experiments demonstrate that the proposed method is highly consistent with human's perception across numerous distortions and metrics. Compared with other state-of-the-art no-reference VQA methods, the proposed method runs much faster while keeping the similar performance. As a no-reference VQA method, it is even comparable with many of the state-of-the-art full-reference VQA methods, which provides the proposed method with better application prospects.

* 基金项目: 国家自然科学基金(61025011, 61332016, 61472389); 国家重点基础研究发展计划(973)(2015CB351802)

Foundation item: National Natural Science Foundation of China (61025011, 61332016, 61472389); National High Technology Research and Development Program of China (973) (2015CB351802)

收稿时间: 2016-05-01; 采用时间: 2016-10-18

Key words: video quality assessment; 3D; convolutional neural network; no reference; full reference

伴随着智能手机、平板电脑和掌上电脑等各种多媒体终端设备的快速发展与普及,视频资源的数量呈现出爆炸式增长的趋势.但是,受限于视频采集与传输过程中的一些客观条件,最终呈现在用户面前的视频往往都伴随着不同程度的失真.视频质量评价在很多视频处理应用中都起着非常重要的作用,诸如视频增强、视频压缩与重建、视频水印添加等.因此,近年来,视频质量评价这个研究方向吸引了越来越多的研究者投身其中.

现有的视频质量评价方法大致可以划分为两类:主观评价和客观评价.主观评价是由观测者根据评测规范进行人工评分,且需要通过大量观测者的测评分来计算平均意见得分(MOS).除此之外,对于不同的测评者,其测试结果可能偏差较大.因此,主观评价通常需要对测评者经过特定的训练.然而,即便是同一观测者对同一观测对象的评分,其结果也是不稳定的.综上,主观视频评价费时费力,且评分不稳定,很难在线使用.

因此,构建一种自动的视频质量客观评价模型就变得非常必要.客观评价方法是基于自动测量标准和指标的,可提供与主观质量评价近似的结果.根据使用原始无损视频信息的程度,视频质量评价方法可以划分为 3 类:全参考(FR)、半参考(RR)和无参考(NR)评价方法.

主流的全参考评价方法有 STMAD[1],ViS3[2],MOVIE[3]等,主流的半参考评价方法有 RRED[4]等.尽管上述全参考和半参考评价方法已经可以取得较好的性能,但是无参考视频质量评价方法有着更广的应用前景.在很多情况下,我们并不能获得足够的参考信息,甚至是无从获得参考信息.比如,当需要评价某段由数码摄相机获取的压缩后的视频图像时,我们无法获得传感器或录制系统的参数,需仅凭视频图像内容本身给予质量评价;又如,由于传输带宽的限制,接收端用户只能根据压缩码流重建存在失真的视频,无法获得原始的无损视频作为参考比对.因此我们不得不把评价系统看作一个黑盒子,应用无参考视频质量评价方法来对失真视频进行评价.很显然,设计一种可以集成至实时应用系统的无参考视频质量评价方法是一项非常有意义和挑战性的任务.然而,设计一种与人主观感知一致性强的无参考质量评价方法要比全参考和半参考评价方法困难得多.

近年来,一些无参考图像评价方法[5−8]相继涌现出来.针对无参考视频质量评价任务,一种很直接的方法就是应用无参考图像评价方法来进行逐帧打分,然后最终将各帧的得分聚合得到视频的评分.但是这种方式忽略了视频中的运动信息以及视频中更加复杂的失真现象.目前针对视频设计的无参考评价方法还相对较少,并且现有的无参考评价方法大多是针对特定的失真类型进行设计的.这些无参考方法应用起来有很大的局限性,因为它们需要预先知道视频的失真类型,或者使用多种类型失真评测模型分别进行评测后再对最终评分进行融合.

失真视频中通常存在多种不同的失真类型,显然,简单地将这些失真类型的结果叠加并不是一种明智的做法.目前仅有少量无参考评价方法是针对非特定失真类型的,即广义无参考视频质量评价方法.这些广义无参考评价方法中的多数需要预先提取大量的人为分析设计的特征,这些特征的分析与提取过程比较复杂并且耗时.

通过大量观测实验我们发现,视频中的一些失真类型具有相似的模式,所以我们认为无参考视频质量评价模型需要自学习能力.卷积神经网络(CNN)是人工神经网络的一种,已成为当前图像识别领域的研究热点,在图像分类[9−12]、物体检测[13−15]以及图像语义分割[16,17]等领域都取得令人瞩目的成果,但在无参考视频质量评价领域的应用较少.我们认为,3D-CNN 非常适用于视频质量评价任务,因为:

从原理上看,卷积神经网络最贴近人眼视觉系统的工作机制.1962 年 Hubel 和 Wiesel 通过对猫的视觉系统的研究[18],提出了感受野(receptive field)的概念.后续研究表明,人的视网膜、视皮层及更加深入的大脑皮层中,神经元细胞之间组成了层次结构,多个次级神经元共同决定一个上级神经元的反应,构成该上级神经元的感受野.局部感受野的示意图如图 1 所示.当人在观看图像或视频时,是借用感受野来进行视觉信息处理的.不同的感受野结构对应不同的视觉特征,如边缘、方向、空间频率、运动等.这些都是影响我们判断图像质量好坏的重要因素.

卷积神经网络最早就是基于生物神经网络中的感受野的工作原理提出的.1984 年日本学者 Fukushima 基于感受野概念提出的神经认知机(neocognitron)[19],可以看作是卷积神经网络的第一个实现网络.从卷积神经网

络可视化的角度来讲,感受野就是输出特征图中某个节点的响应所对应的输入图像的区域.

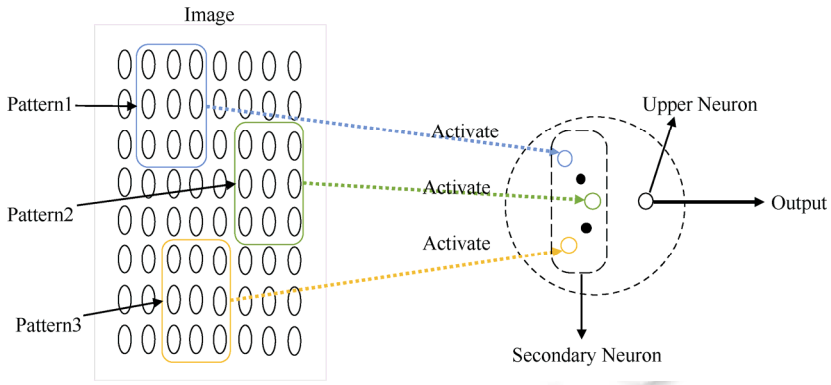

Fig.1 Local receptive field of human's vision system

图 1 人类视觉系统的局部感受野示意图

而视频相较于图像,增加了时域信息.因此 3D 卷积神经网络对于视频质量评价而言,是更合适的一种选择.近年来,一些针对 3D 卷积神经网络的工作也相继在一些国际顶级会议中被提出.在当前具有代表性的工作中[20−22],本文所使用的基础模型是目前性能最好的 3D-CNN 模型[22],如图 2(a)所示.

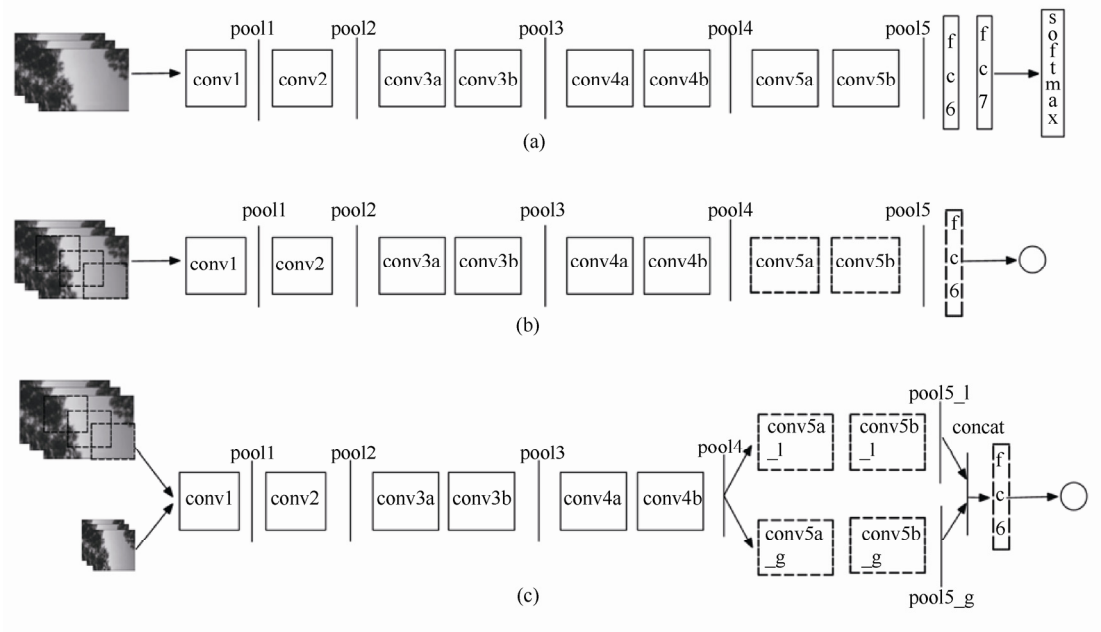

Fig.2 3D-CNN architechture for video classification (a) and the two modified architechures (b)(c) used for VQA in this paper

图 2 用于视频分类的 3D 卷积网络结构(a)与本文用于无参考视频质量评价的两种网络结构(b)(c)对比图

基于上述问题和发现,本文首次将 3D 深度卷积神经网络(3D-CNN)引入到了视频质量评价模型中,提出了一种基于 3D-CNN 的无参考视频质量评价方法,可以适用于非特定失真类型的无参考视频质量评价.该方法是一种基于学习的算法,通过 3D 块来有效学习和表征视频内容的空域与时域特征,并对常规的 3D 卷积网络模型[22]进行了改进,使其适用于视频质量评价的任务.整个网络是一个端到端的网络,评价过程相较于传统的手工分析提取特征的方法,处理速度更快,使用更为简便且与人工主观评价结果更为一致.

本文所提出的方法框架如图 2(b)和图 2(c)所示,其主要创新点包括:

(1) 本文中所提出的无参考评价方法首次在视频评价领域引入了 3D 卷积神经网络.并且在此基础上我们对于原始的 3D 卷积网络[22]进行了修改与适配,通过大量实验找到了一种对于视频评价任务有效的网络框架结构.

(2) 本文提出的方法,相较于传统的基于统计分析、手工筛选提取特征的方法以及目前已有的基于特征学习的方法,首次将空域特征学习与时域特征学习集成进一个框架内,实现了一种端到端的评价预测方法.

(3) 通过在公开的 LIVE[23,24]视频评价数据集上的大量实验,其结果表明,本文所提出的无参考视频评价方法的性能甚至超过了某些经典的全参考评价方法,并且本文所提出的方法在运行效率上比主流的全参考以及无参考评价方法快 2~8 倍,具备很好的实时应用的前景.

1 无参考视频质量评价相关工作

近年来,一些研究者致力于无参考视频质量评价领域,并且取得了一定的进展,我们将在本节对已有的一些无参考视频质量评价方法进行简要的介绍.现有的绝大多数无参考评价方法都只针对特定的失真类型.

文献[25]中提出了一种通过计算一组针对块效应、模糊效应以及清晰度的特征来对视频失真进行建模预测.文献[26]主要针对块效应和网络丢包产生的失真进行建模.文献[27]则针对块效应、模糊效应以及白噪声 3 种失真类型进行建模预测.文献[28]提出了一种在检测出的感兴趣区域上针对块效应与模糊效应进行建模的方法.文献[29]中的方法使用拉普拉斯金字塔特征来度量视频的压缩失真.文献[30]提出了一种针对视频时域失真的无参考评价方法,该方法通过计算视频中相邻帧对应块的运动补偿信息来进行时域失真建模.

目前也有少数不针对特定失真类型的无参考视频质量评价方法,即广义无参考评价方法.文献[31]提出了一种基于空域自然视频统计的方法.文献[32]提出了一种在 DCT 域自然视频统计的方法,该方法还同时结合了视频的时域运动信息.

最近发表的文献[33]和文献[34]证明了基于机器学习方法的视频质量评价方法可以得到不错的效果,并且这也逐渐成为一种研究趋势.文献[33]提出了一种词袋模型,通过支持向量机回归(SVR)模型来获得每一帧的得分,时域上使用磁滞效应来融合各帧的得分从而得到最终整个视频的得分.文献[34]提出了一种新型的基于 1D 卷积神经网络和 Logistic 回归的模型.该模型使用了剪切波变换,导致其特征提取过程比较费时.尽管这些方法已被证明在无参考视频评价领域中比较有效,但是其中的很多方法对视频的时域特性的考虑仍有欠缺,其性能仍有很大的提升空间.

2 3D 卷积神经网络

2.1 基于分类的 3D 卷积模型

文献[22]提出了一种 3D 深度卷积网络的框架,该网络在原文中是用于学习对运动视频有较强判别力的一种时空特征.其网络结构如图 2(a)所示,具体包含 8 个卷积层以及 2 个全连接层,激活函数为 ReLU 函数,前 2 个卷积层中每个卷积层后接入一个最大池化层,后 6 个卷积层中每 2 个卷积层之后接入一个最大池化层,每个全连接层后都有一个 dropout 层来减轻网络过拟合,最后通过 softmax 来对视频进行分类.网络的输入视频尺寸为 171×128×3×16,通过对输入进行中心裁剪得到尺寸为 112×112×3×16 的输入,整个网络中卷积核的大小均为 3×3×3,卷积步长为 1×1×1,第 1 个最大池化层窗口大小为 2×2×1,步长为 2×2×1,其他的最大池化层窗口大小均为 2×2×2,步长为 2×2×2.如此设置的目的在于可以保证时域信息的渐进融合.

文献[22]中的模型是通过在百万级别的视频库上进行训练得到的,并且考虑到目前公开的视频质量评价数据库的资源仍然有限,因此我们决定利用原模型的先验,在原模型基础上利用质量评价的视频对网络进行微调.由于视频评价任务不同于分类任务,因此我们首先把原网络中的 softmax 层去掉,然后替换为回归预测节点,采用欧几里德损失函数,如下所示:

$$loss=\frac{1}{2N}\sum_{i=1}^{N}\| f_i(w,b)-y_i\|_2 \tag{1}$$

其中,N 表示样本数,$f_i(w,b)$表示第 i 个样本预测的得分,y_i 表示视频的真实得分.

由于本文所提方法在测试时是将一段测试视频分为多个视频片段或多个视频小块,因此,会有多个预测结果,本文所提方法通过对多个预测结果进行求均值来获得最终的视频得分,即

$$score=\frac{1}{M}\sum_{j=1}^{M}f_j(w,b) \tag{2}$$

其中,M 表示待测视频划分的片段数或块数.

2.2 针对视频质量评价的 3D 卷积模型

为了确保输入视频尺寸与预训练的网络保持一致,我们有两种策略可供选择.

一种方案是直接将视频进行等比例缩放,缩放至 171×128 的尺寸,然后将一段视频以 16 帧无重合帧的步长划分为视频片段集,每个视频片段的得分等同于这个视频的得分.这种方案可以将原始的视频训练集扩充 13~31 倍,但是数据量相对仍较小.

另一种方案是将视频切分为 171×128×16 的视频小块,小块之间在序列级别无重合,在帧级别可以重合也可以不重合,这里我们先以小块不重合的方式来进行测试.以小块不重合的方式对一段视频进行切分,由于小块足够大,基本保证了每个视频小块的失真分布相对均匀,所以可以近似地将每个视频小块的得分等同于这个视频的得分,这样可以将原始的视频训练集扩充至 156~372 倍,数据量相对较为充足.

通过实验我们发现,采用以视频小块作为输入的方式要比以视频等比例缩放作为输入的方式性能要好,当采取小块间有重叠的方式时,性能仍略有提升.在微调网络的过程中我们发现,当直接逐层加深微调的网络层数时,会出现过拟合的现象.去掉一层全连接层后,可以缓解这一现象.因此最终我们的针对视频质量评价任务的微调网络最优结构为,去掉最后一层卷积层,然后微调后两层卷积层以及一层全连接层.使用该微调结构,获得了最优的性能.具体的实验及分析过程可见实验部分.

3 实 验

在主流的视频质量评价数据库 LIVE 库上对我们的方法进行了测试.LIVE 库总共包括 160 段视频,其中包括 10 段无损视频,每段无损视频又对应着 15 段失真类型与失真程度均不同的视频,一共包括 4 种失真类型,分别是 MPEG-2 压缩失真、H.264 压缩失真、IP 网络失真以及无线网络失真.每段视频对应一个主观评分,即不同平均意见得分(DMOS),该得分范围为 0~100,分数越高,代表视频的质量越差.LIVE 库中视频的基本信息见表 1.

Table 1 Videos' information in LIVE

表 1 LIVE 库中视频的基本信息

序列名	分辨率	帧率(fps)	时长(s)	失真类型
bs	768×432	25	8.68	wireless,ip,H.264,MPEG-2
pa,rb,rn,sf,sh,st,tr	768×432	25	10	wireless,ip,H.264,MPEG-2
mc,pr,sh	768×432	50	10	wireless,ip,H.264,MPEG-2

我们在测试性能时采用了两个评测指标,分别是斯皮尔曼等级相关系数(SROCC)和皮尔逊线性相关系数(LCC),前者用来衡量两组变量间的单调关系,后者用于测量两个变量之间线性关系的强度和方向,两个指标值均是越高越好.

实验过程中,只使用库中的失真视频,总共有 150 段视频,选取其中的 80%作为训练集,剩下的 20%作为测试集.做了 5 组不同拆分方式的实验,然后选取 5 组测试结果的中值作为最终的结果.而在确定网络微调结构的实验中,我们针对 1 组训练测试集进行了测试,用以辅助微调结构的确定.

3.1 网络输入与微调结构相关实验

针对第 2.2 节中的两种输入方案,我们进行了测试实验,实验中,固定卷积层参数,然后对全连接层的参数进

行微调.实验结果见表 2.从实验结果可以看出,使用拆分小块的输入方式其性能要优于直接将视频等比例缩放作为输入的方式.不难解释,因为对于视频质量评价任务,其局部的细节信息是较为敏感的,等比例缩放的方式模糊了局部信息,对最终性能产生了不良影响.

为了进一步增强网络对于视频质量评价任务的学习表达能力,需要微调更多的网络层,而不只局限于全连接层,所以接下来,我们使用以视频小块为输入的方式试图微调卷积层,但是,当加入 conv5b 时,性能发生了骤降,见表 3.

Table 2 Performance of different inputs

表 2 不同输入方式性能对比

输入方式	SROCC	LCC
拆分视频小块	**0.759**	**0.764**
视频等比例缩放	0.538	0.689

Table 3 Performance of fine-tuning and not fine-tuning conv5b

表 3 微调 conv5b 前后性能对比

微调方式	SROCC	LCC
微调 conv5b 前	**0.759**	**0.764**
微调 conv5b 后	0.565	0.671

由于受限于网络的训练时间以及设备,我们并没有在此基础上进行进一步向更前面的网络层进行微调的实验,但是我们猜想微调更多前面的层应该会导致结果变差.我们分析出现表 3 中所示的现象是因为参数增多导致出现了过拟合的现象.因此,基于此假设,我们对网络进行了修改.考虑到网络参数大多集中在全连接层,参数量约占了总参数量的 76%.因此,我们决定去掉最后一层全连接层,增强网络学习能力,使其可以微调更多的卷积层.基于此,针对网络的微调层数进行了对比实验,对比实验的结果见表 4.

Table 4 Performance of fine-tuning different layers without FC7

表 4 去掉 FC7 微调不同层数性能对比

微调方式	SROCC	LCC
微调 conv5b-fc6	0.754	0.767
微调 conv5a-fc6	0.782	**0.809**
微调 conv4b-fc6	**0.790**	0.778
微调 conv4a-fc6	0.752	0.735

通过表 4 的对比实验可以看出,去掉 FC7 层的方法改善了过拟合的现象,证明了改进的有效性.通过微调不同深度的卷积层发现,微调 conv5a 层后的网络综合来看可以得到最好的结果,因此我们决定采取这种微调方式作为视频评价任务的网络基础框架.在确定了网络框架后,又尝试将输入的方式改为小块之间有重叠的方式,性能整体仍有所提升,实验结果见表 5,实验中块与块之间的步长为 50 像素.至此,我们得到了一个单路的针对视频质量评价任务的 3D 网络,如图 2(b)所示.

Table 5 Performance of no-overlap patches and overlap patches adopting the best fine-tuning framework

表 5 采用最佳微调网络框架块间有无重叠性能对比

输入方式	SROCC	LCC
块间无重叠	0.782	**0.809**
块间重叠	**0.798**	0.798

3.2 局部与全局网络结合实验

通过上述实验讨论,我们得到了一个单路的视频质量评价 3D 网络,该网络基于视频小块,对于局部细节信息描述力更强.在此基础上,我们希望讨论在这种局部信息上添加一些全局的粗糙信息是否会提升最终的性能.因此我们又有针对性地提出了一个双路的网络框架,如图 2(c)所示.该网络由两路组成,其中一路以第 3.1 节中确定的网络组成,作为局部细节信息的考虑,一路和前述网络相似,只是将输入改为对视频下采样后的缩放视频,

作为一种全局粗糙信息的考量,配对输入.在该网络结构中,我们并未进行进一步的对网络结构进行调试确定的实验,因为我们假设上节确定的最佳微调结构具有较好的迁移能力,因此两路输入均采用第 3.1 节中的最优微调结构,conv5a 前面的卷积层共享权重且参数保持不变,在 conv5a 和 conv5b 层分开学习各自层的参数,然后在 conv5b 层以后汇合连接至全连接层.实验结果见表 6.实验过程中,我们发现结果并没有如预料的那样,反而性能下降得很严重,出现了严重的过拟合现象.对于出现过拟合,我们分析其可能的原因是,因为两路输入的数据量相差过于悬殊所致,局部输入是全局输入数据量的 84 倍,为了逐对匹配,因此全局输入通路的数据在训练过程中经过的迭代次数也是局部输入的 84 倍,因而导致过训练现象的发生,最终全局输入占据了主导地位,进而诱发了过拟合的现象.

Table 6 Performance of network intergrating local and global information

表 6 局部与全局结合网络性能

SROCC	LCC
0.207	0.757

3.3 主流无参考视频质量评价方法性能对比

基于上述两种不同的网络结构的尝试,最终我们决定采用单路的只使用局部信息的网络.实验过程中,我们与 5 种主流的全参考方法和 2 种主流的无参考方法进行了实验对比.其中,PSNR、SSIM[35]、VIF[36]是用于图像质量评价的全参考方法,首先用其对视频进行逐帧打分,然后通过对所有的分数取平均来获得最终的视频得分.STMAD[1]和 ViS3[2]是较新的两种性能较好的全参考视频质量评价方法.V-Bliinds[32]和 V-CORNIA[33]是两种主流的无参考视频质量评价方法,由于无法获得其他的无参考视频质量评价方法的代码,所以本文中只和这两种无参考评价方法进行了对比实验,实验结果见表 7 和表 8.

Table 7 Performance of different VQA methods on LIVE database

表 7 不同评价方法在 LIVE 数据库上的性能对比

	方法名称	SROCC	LCC
全参考	PSNR	0.552	0.623
	SSIM	0.653	0.655
	VIF	0.641	0.726
	STMAD	**0.884**	**0.878**
	ViS3	0.873	0.870
无参考	V-Bliinds	0.580	0.663
	V-CORNIA	**0.769**	0.792
	本文所提方法	0.763	**0.795**

Table 8 Results of the two sample T-test performed between SROCC values obtained by different measures. 1 (−1) indicates the algorithm in the row is statistically superior (inferior) than the algorithm in the column. 0 indicates the algorithm in the row is statistically equivalent to the algorithm in the column

表 8 不同评价方法的 SROCC 系数 T 检验结果.1(−1)表示当前行所表示的方法性能优于(差于)对应列所表示方法的性能.0 表示当前行所表示的方法性能与对应列所表示方法的性能相当

方法	PSNR	SSIM	VIF	STMAD	ViS3	V-Bliinds	V-CORNIA	本文方法
PSNR	0	−1	−1	−1	−1	−1	−1	−1
SSIM	1	0	1	−1	−1	1	−1	−1
VIF	1	−1	0	−1	−1	1	−1	−1
STMAD	1	1	1	0	1	1	1	1
ViS3	1	1	1	−1	0	1	1	1
V-Bliinds	1	−1	−1	−1	−1	0	−1	−1
V-CORNIA	1	1	1	−1	−1	1	0	1
本文方法	1	1	1	−1	−1	1	−1	0

从实验结果可以看出,本文所提方法达到了与主流无参考评价方法性能相当的水平,性能甚至优于全参考评价方法 PSNR、SSIM 和 VIF.本文所提方法的性能与全参考方法 STMAD 和 ViS3 相比略差,这一点在我们的预料之内,因为前两种方法用到了原始视频的信息,可以对视频质量进行更精确的建模与评估.除此之外,我们

还为这些对比方法绘制了盒图,用于对比各种方法预测性能的稳定性,SROCC 指标和 LCC 指标对应的盒图如图 3 和图 4 所示.从图中的结果可以看出,本文所提出的方法不仅性能较好,而且方法的预测稳定性较强.

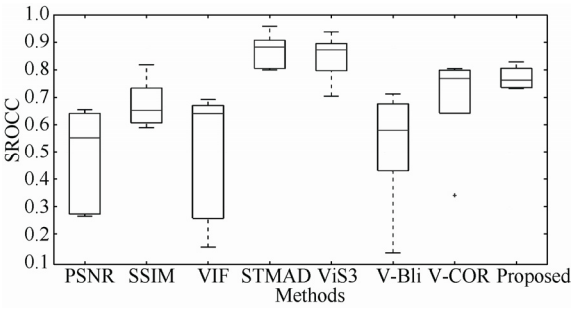

Fig.3 Box plot of SROCC distributions of algorithms over 5 iterations on LIVE

图 3 LIVE 库上 5 次实验结果的 SROCC 指标的盒图分布

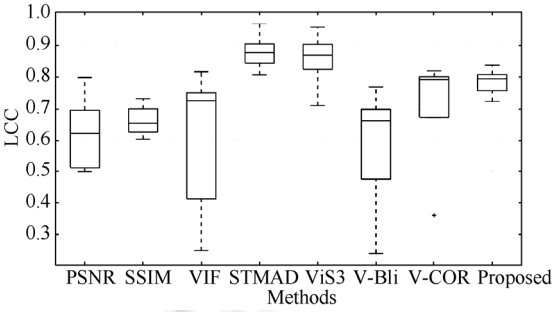

Fig.4 Box plot of LCC distributions of algorithms over 5 iterations on LIVE

图 4 LIVE 库上 5 次实验结果的 LCC 指标的盒图分布

此外,我们还针对性能较好的几种主流全参考及无参考评价方法进行了运行效率的实验对比,具体实验在一台 GPU 型号为 Tesla K40、主频为 2.6GHz 的机器上进行测试,本文方法的实现框架为 CAFFE[37].实验方案为在 LIVE 库上人为选取 10 段不同场景的视频,对每种方法计算其预测 10 段视频质量的平均时间,实验结果如图 5 所示.值得一提的是,我们的方法在运行过程中使用了 GPU 加速,这对于本文所提方法的运行效率有很大的助益,考虑到其他对比方法并没有提供 GPU 版本的代码,因此本文中只得和其 CPU 版本程序进行效率对比,虽然可能有失公正,但还是可以从图 5 中可以看出,本文所提方法与主流的全参考和无参考方法相比,在运行效率上有了很大的提升,这不仅归功于 GPU 加速,还主要归功于本文方法不需要复杂的变换与繁复的特征提取过程,而是直接由尺寸适中的视频块直接进行端到端的测试,并行度高且需要的运算不是很多.这使得我们的方法在实际应用中有更好的前景.

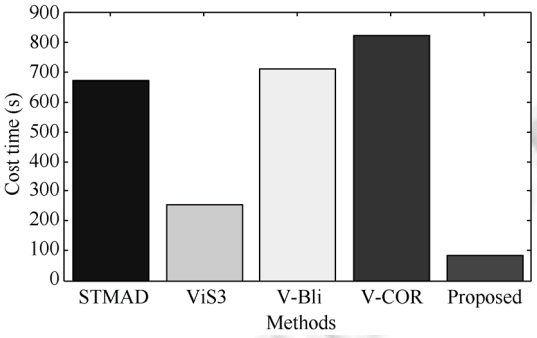

Fig.5 Runtime of state-of-the-art VQA methods

图 5 主流评价方法运行时间对比

4 总 结

本文提出一种基于 3D 卷积神经网络的无参考视频质量评价方法,通过对视频小块进行 3D 卷积来同时学习视频的时域和空域特征,整个学习预测框架是一个端到端的网络.与主流的无参考方法相比,性能相当且运行速度提升很大,性能上甚至超过了一些主流的全参考评价方法,有很好的应用前景.未来工作将对本文中全局与局部结合的双路网络进行进一步的改进与研究,尝试引入 triplet 损失来适当增大数据集,以期实现更好的预测

性能.

References:

[1] Vu PV, Vu CT, Chandler DM. A spatiotemporal most-apparent-distortion model for video quality assessment. In: Proc. of the 18th IEEE Int'l Conf. On Image Processing. Brussels: IEEE, 2011. 2505–2508.

[2] Vu PV, Chandler DM. Vis3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. Journal of Electronic Imaging, 2014,23(1):013016.

[3] Seshadrinathan K, Bovik AC. Motion tuned spatio-temporal quality assessment of natural videos. IEEE Trans. on Image Processing, 2010,19(2):335–350.

[4] Soundararajan R, Bovik AC. Video quality assessment by reduced reference spatio-temporal entropic differencing. IEEE Trans. on Circuits and Systems for Video Technology, 2013,23(4):684–694.

[5] Ye P, Kumar J, Kang L, Doermann D. Unsupervised feature learning framework for no-reference image quality assessment. In: Proc. of the 2012 IEEE Int'l Conf. on Computer Vision and Pattern Recognition. Providence: IEEE, 2012. 1098–1105.

[6] Mittal A, Moorthy AK, Bovik AC. No-Reference image quality assessment in the spatial domain. IEEE Trans. on Image Processing, 2012,21(12):4695–4708.

[7] Xue W, Zhang L, Mou X. Learning without human scores for blind image quality assessment. In: Proc. of the 2013 IEEE Int'l Conf. on Computer Vision and Pattern Recognition. Portland: IEEE, 2013. 995–1002.

[8] Kang L, Ye P, Li Y, Doermann D. Convolutional neural networks for no-reference image quality assessment. In: Proc. of the 2014 IEEE Int'l Conf. On Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 1733–1740.

[9] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: 2012 Advances in Neural Information Processing Systems. Barcelona: MIT Press, 2012. 1097–1105.

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

[11] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 1–9.

[12] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385.

[13] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proc. of the 2014 IEEE Int'l Conf. on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 580–587.

[14] Girshick R. Fast R-CNN. In: Proc. of the 2015 IEEE Int'l Conf. On Computer Vision. Santiago: IEEE, 2015. 1440–1448.

[15] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In: 2015 Advances in Neural Information Processing Systems. Montréal: MIT Press, 2015. 91–99.

[16] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 3431–3440.

[17] Chen LC, Yang Y, Wang J, Xu W, Yuille AL. Attention to scale: Scale-Aware semantic image segmentation. arXiv preprint arXiv:1511.03339.

[18] Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology, 1962,160(1):106–154.

[19] Fukushima K, Miyake S. Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. Pattern Recognition, 1982,15(6):455–469.

[20] Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L. Large-Scale video classification with convolutional neural networks. In: Proc. of the 2014 IEEE Int'l Conf. on Computer Vision and Pattern Recognition. Columbs: IEEE, 2014. 1725–1732.

[21] Ng JYH, Hausknecht M, Vijayanarasimhan S. Beyond short snippets: Deep networks for video classification. In: Proc. of the 2015 IEEE Int'l Conf. On Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 4694–4702.

[22] Tran D, Bourdev LD, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3D convolutional networks. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 4489–4497.

[23] Seshadrinathan K, Soundararajan R, Bovik AC, Cormack LK. A subjective study to evaluate video quality assessment algorithms. In: Proc. of the IS&T/SPIE Electronic Imaging. San Jose: SPIE, 2010. 75270H–75270H.

[24] Seshadrinathan K, Soundararajan R, Bovik AC, Cormack LK. Study of subjective and objective quality assessment of video. IEEE Trans. on Image Processing. 2010,19(6):1427−1441.

[25] Caviedes JE, Oberti F. No-Reference quality metric for degraded and enhanced video. In: Visual Communications and Image Processing. Lugano: Int'l Society for Optics and Photonics, 2003. 621−632.

[26] Babu RV, Bopardikar AS, Perkis A, Hillestad OI. No-Reference metrics for video streaming applications. In: Proc of the Int'l Workshop on Packet Video. 2004.

[27] Farias MC, Mitra SK. No-Reference video quality metric based on artifact measurements. In: Proc. of the 2005 IEEE Int'l Conf. on Image Processing. Genoa: IEEE, 2005. III−141.

[28] Lin XY, Tian X, Chen YW. No-Reference video quality assessment based on region of interest. In: Proc. of the 2nd Int'l Conf. on Consumer Electronics, Communications and Networks. Yichang: IEEE, 2012. 1924−1927.

[29] Zhu KF, Keisuke Hirakawa, Asari V, Saupe D. A no-reference video quality assessment based on laplacian pyramids. In: Proc. of the 2013 IEEE Int'l Conf. on Image Processing. Melbourne: IEEE, 2013. 49−53.

[30] Yang F, Wan S, Chang Y, Wu HR. A novel objective no-reference metric for digital video quality assessment. IEEE Signal Processing Letters, 2005,12(10):685−688.

[31] Mittal A, Saad M, Bovik AC. Assessment of video naturalness using time-frequency statistics. In: Proc. of the 2014 IEEE Int'l Conf. on Image Processing. Paris: IEEE, 2014. 571−574.

[32] Saad MA, Bovik AC, Charrier C. Blind prediction of natural video quality. IEEE Trans. on Image Processing, 2014,23(3): 1352−1365.

[33] Xu JT, Ye P, Liu Y, Doermann D. No-Reference video quality assessment via feature learning. In: Proc. of the 2014 IEEE Int'l Conf. on Image Processing. Paris: IEEE, 2014. 491−495.

[34] Li YM, Po LM, Cheung CH, Xu XY, Feng LT, Yuan F, Cheung KW. No-Reference video quality assessment with 3D shearlet transform and convolutional neural networks. IEEE Trans. on Circuits and Systems for Video Technology, 2016,26(6): 1044−1057.

[35] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. IEEE Trans. on Image Processing, 2004,13(4):600−612.

[36] Sheikh HR, Bovik AC, Veciana GD. An information fidelity criterion for image quality assessment using natural scene statistics. IEEE Trans. on Image Processing, 2005,14(12):2117−2128.

[37] Jia YQ, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T. Caffe: Convolutional architecture for fast feature embedding. In: Proc. of the 22nd ACM Int'l Conf. on Multimedia. Orlando: ACM, 2014. 675−678.

王春峰(1990－),男,河北保定人,硕士,CCF 学生会员,主要研究领域为图像,视频质量评价,深度学习.

张维刚(1980－),男,博士,副教授,CCF 专业会员,主要研究领域为多媒体分析,视频处理,跨媒体计算.

苏荔(1980－),女,博士,副教授,CCF 专业会员,主要研究领域为多媒体技术,模式识别,计算机视觉.

黄庆明(1965－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为多媒体技术,模式识别,计算机视觉.