基于深度残差网络的 HEVC 压缩视频增强

何晓艺[1] 段凌宇[2] 林巍峣[1]
(上海交通大学电子信息与电气工程学院 上海 200240)[1]
(北京大学数字视频编解码技术国家工程实验室 北京 100871)[2]

摘 要 文中提出了一种基于深度残差网络的 HEVC 压缩视频增强方法。该方法利用一系列级联的残差模块来完成特征提取,然后基于这些特征进行视频的质量增强。与现有的方法相比,所提方法能够捕捉到压缩视频帧更清晰和泛化的特征。实验结果表明,所提方法在 20 个通用的测试视频序列上能够实现平均 6.92% 的 BD-rate 增益,是所有参与比较的方法中效果最好的。

关键词 高效率视频编码,卷积神经网络,残差网络,压缩视频增强
中图法分类号 TP751 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.03.011

Deep Residual Network Based HEVC Compressed Videos Enhancement

HE Xiao-yi[1] DUAN Ling-yu[2] LIN Wei-yao[1]
(School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)[1]
(National Engineering Laboratory for Video Technology, Peking University, Beijing 100871, China)[2]

Abstract This paper proposed a HEVC-compressed videos enhancement method based on deep residual network. This method utilizes several stacked residual blocks to achieve feature extraction, followed by feature enhancement and reconstruction. Compared with the existing methods which only use a few convolutional layers, the proposed method can capture the feature of input compressed frames in a more distinctive and stable way. Experimental results show that the proposed method leads to over 6.92% BD-rate saving on 20 benchmark sequences and achieves the best performance among the compared methods.

Keywords High efficiency video coding, Convolutional neural network, Residual network, Compressed videos enhancement

1 引言

随着视频拍摄技术和显示设备的不断发展,高清甚至超高清视频出现得越来越频繁。与上一代视频编码标准 H.264/AVC 相比,新一代的视频编码标准[1](High Efficiency Video Coding, HEVC)的编码效率提升了大约 50%[2]。但是作为一种有损视频压缩算法,HEVC 在压缩视频的同时不可避免地会造成一些图像的失真,例如出现块效应和振铃效应等。因此,在保证压缩率的前提下,压缩视频质量的提升成为了一项重要任务。

目前,深度学习不仅在图像分类、物体检测和识别等高级计算机视觉任务中表现出色,而且逐渐被应用在图像超分辨率、图像压缩和增强等领域,并取得了良好的效果。本文着力于研究和改进利用深度学习实现的压缩视频质量增强方法。

为了进一步提升压缩视频的质量,在总结了现有基于卷积神经网络的质量增强方法的基础上,提出了一种新的基于残差网络的压缩视频的质量增强方法。现有方法所使用的网络的结构比较简单,且能够保持输入视频帧的分辨率,但对输入视频帧的特征泛化能力不够,且随着网络深度的增加,训练难度也增大。本文首次将批归一化层和残差结构用于压缩视频的质量增强,在增加网络深度以提高模型的特征表达能力的同时,降低了模型的训练难度。基于卷积神经网络的压缩视频增强方法的整体框架如图 1 所示。

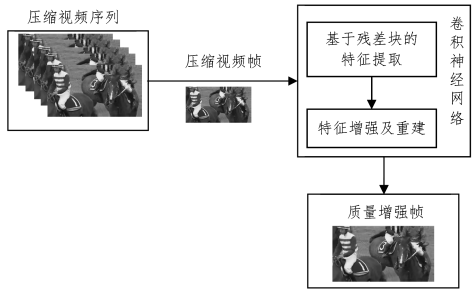

图 1 基于卷积神经网络的压缩视频增强方法

Fig. 1 Compressed videos enhancement method based on convolutional neural network

到稿日期:2018-07-05 返修日期:2018-09-21 本文受国家自然科学基金项目(61471235),上海市"一带一路"青年科学家交流国际合作项目(17510740100),Ng Teng Fong 慈善基金 PKU-NTU 联合研究中心项目(JRI)资助。

何晓艺(1995—),男,硕士生,主要研究方向为视频压缩编码、计算机视觉,E-mail:xiaoyi.he@outlook.com;**段凌宇**(1975—),男,教授,主要研究方向为图像识别与多媒体大数据分析;**林巍峣**(1980—),男,博士,教授,主要研究方向为视频压缩编码、计算机视觉,E-mail:wylin@sjtu.edu.cn(通信作者)。

2　相关工作

深度学习在多个领域的任务中都取得了显著的效果，最近国内外学者又提出了基于深度学习的压缩图像和视频增强方法[3]。

Park 等[4]使用卷积神经网络 SRCNN[5]替代了 HEVC 中的样点自适应补偿(SAO)模块，用于重建视频帧的质量增强。由于该模块处于编解码环路内，因此该方法在提升视频帧质量的同时节省了码率。但由于网络过于简单，该方法的效果非常有限。

Dai 等[6]和 Yang 等[7]分别提出了 VRCNN 和 QECNN，用于 HEVC 压缩视频的质量增强，在保持码率不变的情况下改善了压缩视频帧的图像质量，提升了编码效率。VRCNN 和 QECNN 均主要由卷积层和激活函数层组成。由于两者均使用了残差学习的方法，即网络学习的目标是残差图像，因此网络最终的输出须加上输入图像。这样，网络的优化目标将更简单，且更容易收敛。VRCNN 采用了不同大小的卷积核，在帧内编码模式下节省了约 4.6%的码率。而 QECNN 则主要针对视频编码中的两种不同帧，即 I 帧和 P 帧，分别设计了 QECNN-I 和 QECNN-P 两个网络；QECNN-P 基于 P 帧内同时存在帧内编码和帧间编码两个模式，设计了网络结构。

现有的基于卷积神经网络的压缩视频增强方法均是基于简单的卷积层和激活函数层的有限叠加(深度不超过 5)，模型的特征提取和表征能力有限，且得到的特征可能存在一定的噪声。

Wang 等[8]提出了 Deep CNN-based Auto Decoder(DCAD)作为解码端的后处理步骤，用于 HEVC 压缩视频的块效应去除和细节增强。DCAD 由 10 个卷积核大小为 3×3 的卷积层级联构成，网络较深。为了有效训练该网络，作者选择了多样性较强的训练数据，同时应用了学习率调整和迁移学习的方法来避免过拟合。区别于现有方法，本文提出了一种基于残差块的深度神经网络用于压缩视频的质量增强。该方法具有更强的特征提取和繁华能力，能够进一步改善压缩视频的质量。

3　基于深度残差网络的质量增强方法

3.1　基于残差块的质量增强网络

文献[6-8]中设计的网络仅由几个卷积层及其激活函数构成，非常简单且深度较小，模型的性能具有一定的局限性。文献[9]中的实验表明增加网络的深度能够提升其性能，但随着网络深度的增加，其训练的难度也逐渐加大，导致性能出现饱和甚至下降[10]。因此 He 等[11]提出了残差学习方法，以解决网络深度增加时性能下降的问题。一般的残差模块可以表示为：

$$y=F(x;\theta)+x$$

其中，x 表示输入，y 表示输出，θ 表示非线性映射 F 的参数。实验证明，这样的残差结构更容易优化和学习。

图 2 为基于残差块的质量增强网络结构的示意图，该网络以压缩后的视频帧作为输入，直接预测输出质量增强后的视频帧。其中主要包含了特征提取、特征增强、映射和重建 4 个过程。所有卷积层均由 3×3 的卷积核组成，除最靠近输出的一个卷积层外，输出的特征图数量均为 64。特征提取由 32 个级联的残差模块完成。

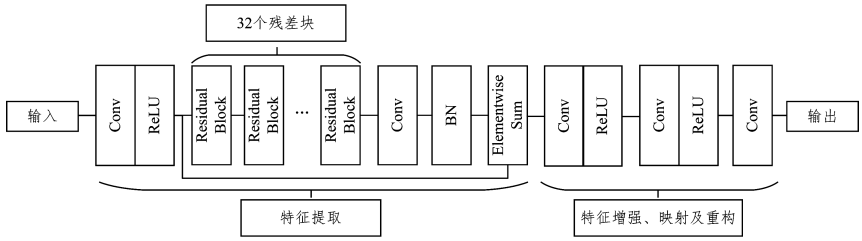

图 2　基于残差块的质量增强网络结构示意图

Fig. 2　Network architecture of proposed deep residual network for compressed videos enhancement

我们选择使用了文献[12]中描述的一种残差模块，其结构如图 3 所示。一个残差模块主要包含 2 个卷积层、1 个批归一化层[13]和 1 个修正线性单元。卷积层的卷积核大小均为 3×3，且被设置成输入与输出大小相同。

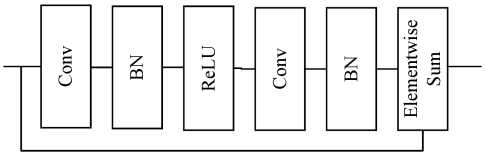

图 3　残差模块结构示意图

Fig. 3　Architecture of residual block

在得到输入视频帧的特征后，基于这些特征网络进行特征映射及重建等。特征增强、映射和重建均由一个卷积层及其激活函数完成，可以表示为：

$$Y_0=\max(0,W_0*F+b_0)$$

$$Y_1=\max(0,W_1*Y_0+b_1)$$

$$Y=W_2*Y_1+b_2$$

其中，W_i，b_i ($i=0,1,2$)为对应卷积层的参数；F 表示输入的特征图；Y_0 表示特征增强后的输出；Y_1 表示映射后的输出；而 Y 为重建后的输出，即网络最终输出的增强后的视频帧。

3.2　训练

如图 4 所示，受 GPU 内存大小和训练数据集规模的限制，为了有效地进行训练，从每一个视频中随机挑选 3 帧，然后将每一个视频帧及其标注帧划分成多个不相交的 64×64 的子图，然后进行训练。在视频压缩算法中，不同的量化参数决定了视频的压缩率，而不同压缩率下压缩视频的失真程度和特性各不相同，因此我们针对每一个量化参数(QP)训练一个单独的模型。

给定包含 N 个训练样本的数据集$\{(X^i,Y^i)\mid i=1,\cdots,N\}$，其中 X^i 表示压缩视频帧，Y^i 为对应的标注帧(原始视频帧)，残差网络的目标是学习一个非线性模型 M，最小化均方误差函数 L：

$$L=\frac{1}{N}\sum_{i=1}^{N}(Y^i-M(X^i))^2$$

训练时，采用 Adam[14] 作为优化方法（参数 $\beta_1=0.9$，其余参数为默认值），批尺寸为 32，初始学习速率为 1×10^{-4}。经过 20 个训练周期（epoch）后学习速率衰减为 1×10^{-5}，训练过程总共包含 40 个训练周期。

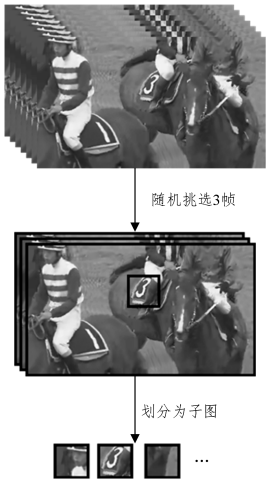

图 4 训练数据的生成

Fig. 4 Generation of training samples

与现有工作需要在训练时做额外的处理以解决网络难收敛的问题（如文献[6-7]中均使用了梯度裁剪的技术）不同，本文中所有的模型均采用一样的训练参数和普通的优化方式，并没有进行任何额外的处理或调参。

3.3 测试

与训练阶段类似，每一个压缩视频帧同样被划分为多个不相交的 64×64 大小的子图，然后将其分别输入对应 QP 的预训练模型中，得到预测的质量增强帧。在 LDP 编码模式下，压缩视频帧的类型有 I 帧和 P 帧两种。虽然在训练阶段只使用 P 帧，但在测试时对 I 帧和 P 帧均进行处理。

4 实验结果

4.1 实验条件

为了训练该网络，建立一个压缩视频数据集。首先收集 600 个不同分辨率的视频片段，然后使用 HEVC 的参考软件 HM-16.0 在默认的 LDP（Low DelayP）[15] 编码条件下对这些视频进行压缩。在每个压缩视频中随机地选取 3 帧（只有图像的 Y 通道被使用），并将其对应的原始视频帧作为标注帧，得到 1800 个数据对（一个数据对包含压缩视频帧及其标注帧）。TensorFlow[16] 被用于模型的训练和测试。为了与现有的性能较好的方法进行比较，使用同样的训练数据和超参数对 VRCNN[6] 和 QECNN[7] 进行了训练，并在同样的测试条件下对得到的模型进行了测试。注意，本文实验正是在 LDP 编码条件下进行（只有第一帧为 I 帧）的，使用针对 P 帧而设计的 QECNN-P 而非 QECNN-I 来进行比较。

为了评估模型的性能，在文献[17]中给出的 20 个 HEVC 标准测试序列上对其进行了测试。在测试过程中，选定量化参数 QP，使用 HM-16.0 在 LDP 的条件下压缩测试序列，然后将其输入质量增强网络进行测试。文中采用峰值信噪比 PSNR（Peak Signal to Noise Ratio）作为评价标准来评价视频帧质量增强的效果。

另外，本文方法在保持比特率不变的情况下提升了压缩视频的图像质量，从而等效地提升了视频压缩编码的效率。对此，使用文献[18]中描述的 BD-rate（Bjøntegaard delta rate）来评估本方法对编码效率的提升。文献[18]中提出的 Bjøntegaard model 主要通过计算两条 RD 曲线（rate-distortion curves）之间的差异来计算平均 PSNR 和比特率的差异，主要包含两个指标：1）BD-PSNR（Bjøntegaard delta PSNR），用来反映在相同码率下的平均 PSNR 差值（单位为 dB）；2）BD-rate，用来反映在相同 PSNR 下的比特率差异（单位为%）。若 BD-rate 为正则表示性能下降；若 BD-rate 为负，则表示性能上升。

4.2 PSNR 提升

表 1 列出了本文方法、VRCNN[6] 和 QECNN-P[7] 在 4.1 节描述的测试条件下与 HM-16.0 相比 PSNR 的提升。可以看出，本文方法是所有参与比较的方法中性能最好的，所有测试序列的平均 PSNR 提升为 0.25 dB。其中，类别 A 中的 PeopleOnStreet 视频序列的 PSNR 提升最高，为 0.56 dB，而 VRCNN 和 QECNN-P 的平均 PSNR 提升分别为 0.38 dB 和 0.48 dB，也就是说，本文的方法与之相比提升了 0.18 dB 和 0.08 dB，这得益于本文方法强大的特征提取和泛化能力。

表 1 不同方法在 $QP=37$ 时相比于 HM-16.0 的 PSNR 提升

Table 1 ΔPSNR of different methods over HM-16.0 at $QP=37$

（单位：dB）

类别	序列名称	VRCNN	QECNN-P	Ours
A	Traffic	0.22	0.28	0.31
	PeopleOnStreet	0.38	0.48	0.56
	Nebuta	0.19	0.22	0.27
	SteamLocomotive	0.13	0.16	0.19
B	Kimono	0.27	0.33	0.36
	ParkScene	0.13	0.15	0.17
	Cactus	0.16	0.22	0.23
	BQTerrace	0.05	0.16	0.18
	BasketballDrive	0.13	0.16	0.19
C	RaceHorses	0.18	0.23	0.26
	BQMall	0.02	0.14	0.10
	PartyScene	−0.04	0.07	0.11
	BasketballDrill	0.09	0.18	0.22
D	RaceHorces	0.23	0.29	0.31
	BQSquare	−0.20	0.04	0.05
	BlowingBubbles	0.05	0.11	0.12
	BasketballPass	0.06	0.15	0.19
E	FourPeople	0.30	0.40	0.44
	Johnny	0.23	0.31	0.35
	KristenAndSara	0.30	0.37	0.39
Average		0.14	0.22	0.25

4.3 图像主观质量评价

除了使用 PSNR 等来评价图像的客观质量外，我们还可以通过图像的主观质量来评估质量增强的效果。图 5 显示了使用本文方法以及 VRCNN 产生的质量增强帧。从图中可以看出，由视频压缩算法产生的视频帧存在明显的块效应和失真（眼睛、鼻子等地方）；但经过处理后的视频帧的块效应明显减轻，且本文方法增强后的视频帧效果更好，这一观察与 4.2 节中得出的结论一致。

(a)原始视频帧

(b)HM-16.0 压缩视频帧 (*PSNR*:34.2)

(c)VRCNN(*PSNR*:34.50)

(d)Ours(*PSNR*:34.56)

图 5　*QP* 为 37 时本文所提方法与 VRCNN 的主观效果图

Fig. 5　Subjective results of our network and VRCNN at *QP*=37

4.4　码率节省

表 2 列出了不同方法在 4.1 节描述的测试条件下以 BD-rate 为衡量指标的编码效率提升情况(与 HM-16.0 基准相比)。从中看出，本文方法取得了 6.92%的增益，以较大的幅超过 VRCNN(3.81%)和 QECNN-P(5.71%)。针对不同类别的视频序列，本文方法在类别 E 中取得了最高的编码效率，分别提升了 7.12%，7.00%和 7.13%，这说明了基于深度卷积网络的质量增强算法具有重大的意义和潜力。

表 2　不同方法相比于 HM-16.0 的码率节省(BD-rate，Y 通道)

Table 2　BD-rate(Y，%) saving of different methods over HM-16.0

类别	序列名称	VRCNN	QECNN-P	Ours
A	Traffic	−6.84	−8.28	−9.27
	PeopleOnStreet	−7.41	−8.66	−9.84
	Nebuta	−5.65	−7.56	−6.23
	SteamLocomotive	−7.71	−9.18	−10.22
B	Kimono	−7.39	−8.70	−9.49
	ParkScene	−3.97	−4.73	−5.40
	Cactus	−5.86	−7.39	−8.13
	BQTerrace	−1.73	−4.87	−7.25
	BasketballDrive	−3.75	−5.91	−6.42
C	RaceHorses	−3.60	−4.78	−5.57
	BQMall	0.11	−2.91	−4.01
	PartyScene	2.72	−1.03	−2.48
	BasketballDrill	−0.08	−2.36	−5.71
D	RaceHorces	−4.05	−5.03	−6.66
	BQSquare	−0.57	−0.11	−2.48
	BlowingBubbles	−0.15	−2.07	−4.12
	BasketballPass	−0.15	−2.37	−4.49
E	FourPeople	−7.12	−9.27	−10.69
	Johnny	−7.00	−9.78	−10.40
	KristenAndSara	−7.13	−9.21	−9.50
Average		−3.81	−5.71	−6.92

结束语　为了提高基于卷积神经网络的压缩视频质量增强方法的性能，本文提出了一种深度残差网络用于压缩视频的质量增强。与现有工作只使用少数卷积层不同，本文首次使用了批归一化层和残差结构来实现质量增强中的特征提取。级联的残差块增加了网络深度，提升了网络的特征表达能力，能够学习更复杂的非线性模型。实验结果表明，本文方法在不同指标下均超过了现有的 VRCNN 和 QECNN-P，以 HM-16.0 作为基准，实现了 6.92% 的 BD-rate 增益。

参 考 文 献

[1] SULLIVAN G J，OHM J，HAN W J，et al. Overview of the high efficiency video coding (HEVC) standard[J]. IEEE Transactions on Circuits and Systems for Video Technology，2012，22(12)：1649-1668.

[2] OHM J R，SULLIVAN G J，SCHWARZ H，et al. Comparison of the coding efficiency of video coding standards—including high efficiency video coding (HEVC)[J]. IEEE Transactions on circuits and systems for video technology，2012，22(12)：1669-1684.

[3] DONG C，DENG Y，CHANGE LOY C，et al. Compression artifacts reduction by a deep convolutional network[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015：576-584.

[4] PARK W S，KIM M. CNN-based in-loop filtering for coding efficiency improvement[C]// Image，Video，and Multidimensional Signal Processing Workshop (IVMSP). IEEE，2016：1-5.

[5] DONG C，LOY C C，HE K，et al. Learning a deep convolutional network for image super-resolution[C]// European Conference on Computer Vision. Cham：Springer，2014：184-199.

[6] DAI Y，LIU D，WU F. A convolutional neural network approach for post-processing in HEVC intra coding[C]// International Conference on Multimedia Modeling. Cham：Springer，2017：28-39.

[7] YANG R，XU M，WANG Z，et al. Enhancing Quality for HEVC Compressed Videos[J]. arXiv preprint arXiv：1709.06734，2017.

[8] WANG T，CHEN M，CHAO H. A novel deep learning-based method of improving coding efficiency from the decoder-end for hevc[C]// Data Compression Conference (DCC)，2017. IEEE，2017：410-419.

[9] SIMONYAN K，ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv：1409.1556，2014.

[10] HE K，SUN J. Convolutional neural networks at constrained time cost[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE，2015：5353-5360.

[11] HE K，ZHANG X，REN S，et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016：770-778.

[12] GROSS S，WILBER M. Training and investigating residual nets [EB/OL]. http：//torch.ch/blog/2016/02/04/resnets.html 2016.

[13] SCHWARZ H，MARPE D，WIEGAND T. Analysis of hierarchical B pictures and MCTF[C]// 2006 IEEE International Conference on Multimedia and Expo. IEEE，2006：1929-1932.

[14] KINGMA D P，BA J. Adam：A method for stochastic optimization[J]. arXiv preprint arXiv：1412.6980，2014.

[15] IOFFE S，SZEGEDY C. Batch normalization：Accelerating deep network training by reducing internal covariate shift[J]. arXiv preprint arXiv：1502.03167，2015.

[16] ABADI M，BARHAM P，CHEN J，et al. TensorFlow：A System for Large-Scale Machine Learning[C]// Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation. 2016：265-283.

[17] BOSSEN F. Common test conditions and software reference configurations[C]// 12th Joint Collaborative Team on Video Coding Meeting. 2011.

[18] BJONTEGARRD G. Calculation of average PSNR differences between RD-curves：ITU-T SG16/Q6[R]. VCEG-M33，Austin，US，2001.