基于长短注意力模块的深度学习图像压缩方法

段增辉[1] 胡益诚[1] 张旭博[1]
DUAN Zenghui HU Yicheng ZHAGN Xubo

摘 要 基于超先验自动编码器的潜在表示最近被应用于端到端图像压缩，其性能与最新的通用视频编码（VVC）帧内编码相当。图像压缩的率失真效率很大程度上受到自动编码器提取的潜在表示的影响。为此提出了一种新的注意力机制模块，利用长短注意力（LSA）模块进行深度学习图像压缩，将长短注意力模块引入自动编码器中，分别在编码阶段和解码阶段添加两个长短注意力模块来提高网络的编解码能力，从而使模型获得更准确的图像的潜在特征表示。长短注意力模块提高了自动编码器提取全局和局部图像特征的能力，节省比特率并实现更高的算法压缩性能。在 JPEG-AI 数据集上的实验表明，长短注意力模块成功地重建了图像细节，所提出的方法在多尺度结构相似性（MS-SSIM）方面实现了最先进的性能，并在低比特率下的峰值信噪比（PSNR）方面优于当前最先进的方法。

关键词 深度学习；图像压缩；注意力机制；区域选择；信噪比；率失真

doi：10.3969/j.issn.1672-9528.2024.04.024

0 引言

图像压缩是过去几十年信号处理中的一个基础研究问题。为了图像的高效传输和存储，已经开发了图像编码标准，例如 JPEG[1]、JPEG2000[2]、高效视频编码（HEVC）/H.265[3]、通用视频编码（VVC）。一般来说，它们按如下方式进行图像压缩。

（1）将输入图像划分为块。

（2）将每个块转换到变换域，例如离散余弦变换（DCT）和离散小波变换（DWT），通过帧内预测去除图像中的空间冗余，通过变换域的量化系数去除频率冗余。

（3）使用上下文自适应算术编码器和各种解块或环路滤波器来减少空间冗余并提高编码效率。

（4）通过类似于 CABAC[4] 的熵编码将量化值和预测辅助信息编码到比特流中。

近年来，用于端到端的图像压缩网络被迅速提出，如文献 [5-7] 等。文献 [5] 提出在图像压缩中标量量化通过编码器的非线性分析达到了矢量量化的效果。Theis 等人在文献 [6] 中提出了一种有损压缩方法来处理自动编码器的不可微问题，从而获得比 JPEG2000 更好的性能。Agustsson 等人 [7] 使用软到硬的退火方案进行训练，并提出了第一个端到端的图像压缩框架。同时，Rippel 和 Bourdev[8] 提出了一种金字塔结构的自动编码器来实现实时自适应编码。Liu 等人 [9]、Santurkar 等人 [10] 和 Agustsson 等人 [11] 提出了生成对抗模型，以低比特率实现良好的视觉效果。文献 [12] 提出了一种超先验自动编码器，使端到端图像压缩的性能与 BPG 相当。后来在文献 [13] 和文献 [14] 的研究中，通过在深度学习图像压缩算法中添加上下文模型，将算法的压缩性能提升并超过了 BPG 的压缩效率。文献 [15] 提出了一种高斯混合模型（GMM），通过更精确和灵活的熵模型，使得性能可以与当前最先进的传统方法 VVC 进行比较。在最近的一项研究中 [16]，使用可逆神经网络（INN）来减少信息丢失，从而获得更好的压缩效果。Cui[17] 提出了非对称增益变分自编码器和非对称高斯熵模型，取得了更好的性能。文献 [18] 使用单应矩阵帮助二分支自动编码器压缩立体图像对，最后使用交叉质量增强模块来提高重建图像的质量。

本文提出了一种长短注意力模块，并将其使用在深度学习网络中来实现一种图像压缩算法。通过将具有长短注意力机制的模块加入到编码熵模型，本文发现，当网络中使用长短注意力模块时，可以为熵模型提取准确的图像潜在特征，以减少在压缩图像时所需要的编码位。此外，本文在长短注意力模块中采用了简化版的多头注意力机制，多头注意力机制可以使学习到的模型集中在图像压缩中更需要注意的区域，从而提高编码性能。实验结果表明，与经典图像压缩标准 VVC、JPEG2000 和现有的深度学习压缩方法相比，本文的方法在低比特率下在 PSNR 指标上表现最佳，在 MS-SSIM 指标上，本文的方法在所有比特率下都具有最佳性能。

与现有方法相比，本文的主要贡献如下。

（1）本文设计了长短注意力模块。该模块有两个分支，

1. 航空工业计算机研究所 陕西西安 710068

一个分支用于提取局部特征，另一个分支用于提取全局特征，最后将全局特征和局部特征融合以获得更准确的图像潜在特征。

（2）在长短注意力（LSA）模块的全局特征分支中，为了实现多头注意力机制，网络可以提取多尺度的全局特征。本文使用特征之间的组卷积来实现该功能，增强 LSA 模块提取全局特征的能力。

（3）本文通过将长注意力模块和短注意力模块合并到自动编码器中来设计图像压缩网络。实验表明，在计算 PSNR 时，本文的方法在低比特率下优于其他方法。在高比特率下，本文的方法具有与其他方法相当的性能。在 MS-SSIM 指标上，本文的方法是最好的。

1 深度学习图像压缩算法

1.1 网络结构

本文提出的压缩模型和文献 [15] 具有类似的框架，如图 1 所示，左侧显示自动编码器架构，右侧对应于实现超优先级的自动编码器。其中 Q 代表量化，AE 和 AD 分别代表算术编码器和算术解码器，本文使用 5×5 掩码卷积实现上下文模型，N 根据 λ 选择，前三个低 λ 值为 128，后两个高 λ 值为 192。该网络包含两个主要子网络。第一个是核心自动编码器，它学习图像的量化潜在表示（编码器 g_a 和 g_s）。在 g_a 中，本文使用残差模块 (RB) 来增加大网络的感受野并提高算法的率失真性能，并在下采样 4 倍和 16 倍后添加长短注意力模块来提高网络的潜在特征提取能力，通过添加残差模块和长短注意力模块，可以显著提高编码器学习到的图像的潜在表示。同样，本文的解码器 g_s 使用与 g_a 有类似的结构。

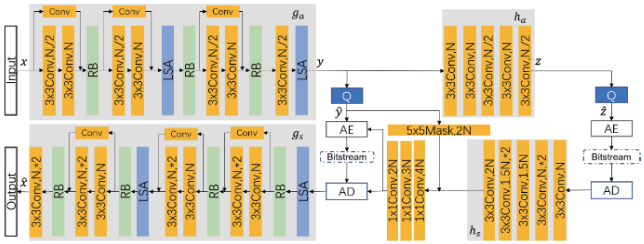

图 1 网络架构

图像的压缩过程可表示为：

$$y = g_a(x;\phi) \tag{1}$$

$$\hat{y} = Q(y) \tag{2}$$

$$\hat{x} = g_s(\hat{y};\theta) \tag{3}$$

式中：x、$\hat{x}$、y 和 $\hat{y}$ 分别表示原始图像、重建图像、量化前的潜在表示和压缩后的编码流。其中 g_a 是编码器，g_s 是解码器，ϕ 和 θ 分别是 g_a 和 g_s 的参数。

第二个子网络负责学习熵编码量化潜在模型的概率模型。它将上下文模型（潜在自回归模型）与超网络（超编码器 h_a 和超解码器 h_s）相结合。当上下文模型由 5×5 掩码卷积实现时，网络主要学习对纠正基于上下文的预测有用的信息。本文将这两部分数据结合起来，通过三个卷积层生成条件高斯熵模型的均值和尺度参数。其过程可表示为：

$$z = h_a(y;\phi_h) \tag{4}$$

$$\hat{z} = Q(z) \tag{5}$$

$$p_{\hat{y}|\hat{z}}(\hat{y}|\hat{z}) \leftarrow h_s(\hat{z};\theta_h) \tag{6}$$

式中：h_a 和 h_s 是超编码器和超解码器，ϕ_h 和 θ_h 是优化参数。$p_{\hat{y}|\hat{z}}(\hat{y}|\hat{z})$是基于$\hat{z}$的估计分布。本文使用非参数、完全分解的密度模型对$\hat{z}$进行建模。

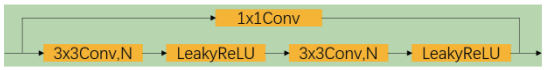

图 2 残差模块框图

$$p_{\hat{z}|\Psi}(\hat{z}|\Psi) = \prod_i \left(p_{z_i|\Psi}(\Psi) \cdot u\left(-\frac{1}{2},\frac{1}{2}\right) \right)(\hat{z}_i) \tag{7}$$

式中：z_i 表示 z 的第 i 个元素，i 指定每个元素或每个信号的位置。

本文使用 MSE 来优化本文的模型，所以完整的速率 - 失真损失函数为：

$$\begin{aligned} L &= \lambda \cdot 255^2 \cdot D_{MSE} + R \\ &= \lambda \cdot 255^2 \cdot MSE(\hat{x},x) \\ &\quad + E\left[-\log_2\left(p_{\hat{y}|\hat{z}}(\hat{y}|\hat{z})\right)\right] \\ &\quad + E\left[\log_2\left(p_{\hat{z}|\Psi}(\hat{z}|\Psi)\right)\right] \end{aligned} \tag{8}$$

式中：λ 控制率失真权衡，不同的 λ 值对应不同的比特率，D_{MSE} 表示畸变项。

1.2 长短注意力模块

本文受文献 [16] 的启发，给全局注意力模块加了一个分支，专门用来提取局部特征，由此设计了一个可以同时提取全局特征和局部特征的长短注意力模块。

如图 3 所示，长短注意力模块（LSA）有两个分支，其中上面的分支由两个残差模块组成，用来提取局部特征，下面的分支先由简化后的注意力模块提取特征，再由组卷积实现多头注意力机制，其中 h 代表控制多头注意力机制的数量。最后，将局部特征与全局特征结合后通过一个卷积层得到融合特征。

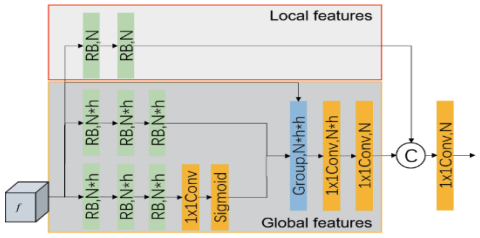

图 3 长短注意力模块

2 实验结果

2.1 模型训练过程

本文使用了 JPEG-AI 提供的数据库 [19]，其中包括 5283 张训练集、300 张验证集和 40 张测试集。本文在一台装了 1080ti 的电脑上训练本文的图像压缩模型， PyTorch 版本为

1.7.1，模型使用 adam 进行优化，Batch Size 大小设置为 8。在训练过程中，本文将图像随机裁剪为 256×256 的图像块作为输入图像训练网络，初始学习率设置为 1e-4，然后使用 ReduceLROnPlateau 来调整学习率，当 loss 不再下降 10 个 epochs 后，学习率下降为原来的二分之一。

本文使用均方误差（MSE）作为损失函数来优化本文提出的模型。根据损失函数，将 λ 设置为集合 {0.001 6, 0.003 2, 0.007 5, 0.015, 0.025, 0.048 3}，将 epochs 参数设置为集合 {200，300，300，300，400，500}。对于三个较低压缩率的模型，N 设置为 128；对于三个较高压缩率的模型，N 设置为 192。

2.2 性能比较

本文用 JPEG-AI 数据集中的 40 张未压缩的图像（8 kB）验证了本文提出的方法的稳健性。本文将所提出的深度学习方法与著名的压缩标准（VVC，JPEG2000）以及一些著名的基于神经网络的学习压缩方法进行了比较，包括 Ball’e 等人、Minnen 等人和 Cheng 的工作。为了评估各方法的率失真性能，速率以每像素比特（bpp）来衡量，质量指标通过 PSNR 和 MS-SSIM 来衡量，其中 bpp 越小，PSNR 值越大，MS-SSIM 越大，则表明算法性能越高。同时，绘制率失真（RD）曲线以证明本文方法和对比方法的编码效率。

2.2.1 视觉比较

如图 4 所示，VVC、JPEG2000 和 bmshj 重建图像时，会出现大量细节损失，例如红框中的皱纹部分边缘模糊，细节缺失，在重建图像的过程中会丢失点状和线状信息。这是因为传统的图像压缩算法是基于分块的图像压缩，在 bpp 比较低时无法保留更多的信息用于图像重建；另外 bmshj 的算法使用的是比较简单的卷积层实现深度学习图像压缩，由于感受野的大小不够，在重建图像时也会丢失图像原本的一些细节。mbt2018 和文献 [15] 是两种带超编解码器的深度学习算法，由于超先验数据的加入，算法在重建图像时，获得了更多的图像潜在特征，可以看出，图像的边缘部分得到了明显改善，点状的图像信息也能得到较好的重建，并且 Cheng 的算法重建结果在视觉上比 mbt2018 的重建结果更好。与其他方法相比，本文使用长和短注意力模块，使网络获取了更多的图像潜在特征，并大大增加了网络的感受野，因此在视觉效果上，本文算法的重建图像具有更清晰的图像边缘和图像细节。

图 4 来自 JPEG AI 数据集的重建图像 00001-TE-960x720 的可视化

如图 5 所示，就峰值信噪比 PSNR 而言，本文提出的方法与 VVC 相比具有竞争力。并且该方法在低比特率下比以前的基于深度学习的方法具有更好的编码性能，这证明了长短注意力模块提取的图像潜在在低比特率要求下压缩图像时可以更好地重建图像。但在高比特率要求下，本文算法的结果在 PSNR 指标上低于文献 [15] 和 mbt2018 方法，这可能是因为长短注意力模块提取的全局特征在高比特率下不利于图像的准确重建。

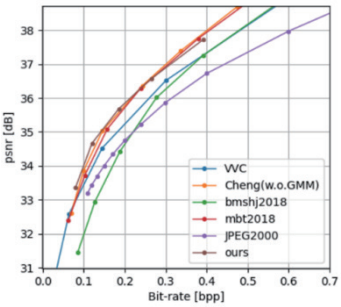

图 5 不同方法之间的 RD 曲线：比特率 (bpp) 与 PSNR(dB)

此外，如图 6 所示，在 MS-SSIM 指标上，本文的方法实现了最先进的压缩性能，这表明本文提出的长短注意力模块可以使重建图像具有更好的结构相似性。如表 1 所示，与传统压缩方法 VVC 相比，本文以 [0.06, 0.50] 的 bpp 对比各方法的平均比特率节省。从结果上看，JPEG2000 的压缩效果在 PSNR 上优于 bmshj，但在 MS-SSIM 指标上效果最差。bmshj 的平均码率节省相比于 VCC 增加超过 20% 开销，表明该方法在 PSNR 和 MS-SSIM 上都要比 VCC 更差。mbt2018 与 VCC 相比在 PSNR 上节省了 9.09% 的平均比特率，在 MS-SSIM 指标上节省了 7.35% 的平均比特率；Cheng 的算法与 VCC 相比在 PSNR 上节省了 12.19% 的比特率，在 MS-SSIM 上节省了 12.46% 的比特率，这两种方法的性能均优于 VVC。本文所提出的方法在 PSNR 中实现了 16.25% 的最高比特率节省，在 MS-SSIM 中实现了 20.59% 的最高比特率节省，表明了本文所提出算法在数据指标上优于传统图像压缩算法和最近的深度学习算法，具有更高的算法压缩和图像重建性能。

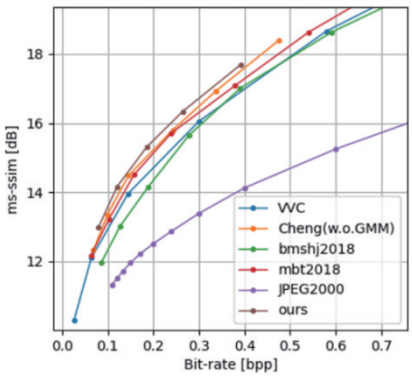

图 6 不同方法之间的 RD 曲线：比特率 (bpp) 与 MS-SSIM

表 1 基于 VVC（VTM15.0）的平均比特率节省

	VVC	JPEG2000	bmshj	mbt2018	Cheng	Ours
PSNR	0	26.91%	28.67%	-9.09%	-12.19%	-16.25%
MS-SSIM	0	162.82%	20.49%	-7.35%	-12.46%	-20.59%

3 结论

本文提出了一种使用长短注意力模块的深度学习图像压缩算法。通过加入长短注意力机制的编码熵模型，本文发现，当在算法网络中使用长短注意力模块时，可以为熵模型提取更准确的图像潜在特征，这直接减少了算法压缩图像所需的编码位。实验结果表明，本文的方法在低比特率下在 PSNR 指标上表现最佳，在 MS-SSIM 指标上，本文的方法在所有比特率下都具有最佳性能。

参考文献：

[1] GREGORY K W. The jpeg still picture compression standard[J]. IEEE transactions on consumer electronics,1992,38(1):57-62.

[2] MAJID R, RAJAN J. An overview of the jpeg 2000 still im age compression standard[J].Signal processing: image communication, 2002,17(1):23-48.

[3] SULLIVAN G J, OHM J-R, 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.

[4] DETLEV M, HEIKO S, THOMAS W. Context-based adaptive binary arithmetic coding in the h.264/AVC video compression standard[J]. IEEE transactions on circuits and systems for video technology, 2003,13(7):620-636 .

[5] JOHANNES B, VALERO L, EERO P S.End-to-end optimized image compression[EB/OL].(2016-11-05)[2024-02-01].https://arxiv.org/abs/1611.01704.

[6] THEIS L, SHI W, CUNNINGHAM A, et al.Lossy image compression with compressive autoencoders[EB/OL].(2017-04-01)[2024-02-15].https://arxiv.org/abs/1703.00395.

[7] EIRIKUR A, FABIAN M, MICHAEL T, et al.Soft-to-hard vec tor quantization for end-to-end learning compressible representations[J].Advances in neural information processing systems, 2017,30:1141-1151.

[8] OREN R, LUBOMIR B. Real-time adaptive image com pression[J].Proceedings of the international conference on machine learning, 2017,70:2922-2930.

[9] LIU B,CAO A,KIM H-S.Unified signal compression using generative adversarial networks[J].In proceedings of the IEEE conference on acoustics, speech and signal processing, 2020,11:3177-3181.

[10] SHIBANI S, DAVID B, NIR S.Generative compression[EB/OL].(2017-03-04)[2024-01-28].https://arxiv.org/abs/1703.01467.

[11] EIRIKUR A, MICHAEL T, FABIAN M, et al.Generative adversarial networks for extreme learned image compression[EB/OL].(2018-04-09)[2024-02-20].https://arxiv.org/abs/1804.02958.

[12] JOHANNES B, DAVID M, SAURABH S, et al.Variational image compression with a scale hyperprior[EB/OL].(2018-04-09)[2024-02-20].https://arxiv.org/abs/1804.02958.

[13] DAVID M, JOHANNES B, GEORGE D T.Joint autoregressive and hierarchical priors for learned image compression[J].Advances in neural information processing systems, 2018,31:10794-10803.

[14] LEE J, CHO S, BEACK S-K. Context adaptive entropy model for end-to-end optimized image compression[EB/OL].(2018-09-27)[2024-02-22].https://arxiv.org/abs/1809.10452.

[15] CHENG Z, SUN H, TAKEUCHI M, et al.Learned image compression with discretized gaussian mixture likeli hoods and attention modules[EB/OL].(2020-01-06)[2024-02-11].https://arxiv.org/abs/2001.01568.

[16] XIE Y, CHENG K L, CHEN Q. Enhanced invertible encoding for learned image compression[J].Proceedings of the 29th ACM International Conference on Multimedia, 2021(10):162-170.

[17] CUI Z, WANG J, GAO S , et al.Asymmetric gained deep image compression with continuous rate adaptation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway:IEEE, 2021:10532-10541.

[18] DENG X, YANG W, YANG R, et al.Deep homography for efficient stereo image compression[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Piscataway: IEEE, 2021:1492-1501.

[19] ISO/IEC, JTC 1/SC29/WG1, ITU-T SG16. Jpeg ai common training and test conditions[EB/OL].(2013-04-24)[2024-02-20].https://ds.jpeg.org/documents/jpegai/wg1n100514-099-CPM-JPEG_AI_Common_Training_&_Test_Conditions.pdf.

【作者简介】

段增辉（1996—），男，陕西韩城人，硕士研究生，助理工程师，研究方向：嵌入式系统、深度学习、图像压缩等。

（收稿日期：2024-03-22）