

ResHiFiC: High-Fidelity Generative Image Compression with ResNet

LeXing Zhang
Peking University

1900012963@pku.edu.cn

Xiao Lin
Peking University

1900011027@pku.edu.cn

YuHan Zhou
Peking University

zhouyuhuan@pku.edu.cn

Abstract

HiFiC(High-Fidelity Generative Image Compression) 是Google 在2020年提出的最高水准的有损图像压缩模型。此模型使用生成对抗网络 (GAN) 并修改了Normalization Layer, Generator 和Discriminator 但保留了传统的Encoder架构。我们使用残差网络 (ResNet) 改进了Encoder得到了新的模型 (ResHiFiC) 并测试了两种模型对低、高分辨率图像的压缩效果。

1. 简介

随着信息时代的到来，信息量展现出爆炸性增长的趋势，一度超越基础设施的发展速度。这使得数据的压缩变得至关重要。现在，信息压缩已经遍布我们生活的角角落落，无论是在网络传输，抑或是数据存储上，都能窥见它的身影。而随着现代视频通信技术的发展，网络直播行业的兴起，作为信息的一大组成部分的图像，其传输、储存、处理问题逐渐进入人们的视野。图像作为信息的聚合体，其包含的庞大信息量若直接进行传输、储存，将会产生极大的负荷以及资源占用，对实际的应用造成巨大的影响。因此，图像的压缩是一个有着实际应用价值的课题。

2. 相关工作

图像压缩的研究工作开始于电视信号数字化的1948年。1969年，第一届“图像编码会议”的召开，标志着图像编码正式成为一门独立的学科。在之后的研究中，我们往往会根据压缩后的图像能否完全恢复，将图像压缩算法分两类。首先是无损压缩。无损压缩是一种没有信息损耗的压缩方式。其通过消除或减少数据中的冗余度进行压缩。该方法消除或减少的各种形式的冗余可以重新插入到数据中，因此，无损压缩是一个可逆过程，也称无失真压缩。与无损压缩相对的是有损压缩。有损压缩会压缩熵，从而使得信息量减少。因而其信息是不能再恢复的，是一个不可逆的过程。

谈及图像编码方法的发展，我们大致可将其分为两个阶段。第一个阶段中采用的“第一代”图像编

码方法，以信息论和数字信号处理为理论基础，旨在去除图像数据中的线性相关性。这类技术去除客观和视觉的冗余信息的能力已经接近极限，其压缩比并不高。“第二代”图像压缩编码技术则不局限于信息论的框架，它充分利用人的视觉生理、心理和图像信源的各种特征，获得了较高压缩比。

至今我们仍在大量使用的最早的有损压缩算法应该当属JPEG [10]。在其之后依靠视频编解码器HEVC [7] 的BPG方法出现，它在不同码率下实现了非常高的PSNR。随着研究的深入，神经压缩方法出现，并取得了卓越的成效。神经压缩方法最初的工作依赖于RNNs [8, 9]，而后续则转向基于自编码器的工作。研究者们为了降低所需的码率，已经使用了各种方法来更准确地建模自编码器潜在的概率密度，从而实现更高效的算术编码，这些方法包括使用层次先验、具有不同上下文形状的自回归或它们的组合。在这之后，GANs 的出现，更是带来了全新的突破。自Goodfellow等人 [2] 介绍GANs以来，GANs在无条件和条件图像生成方面取得了快速进展。现在，最先进的GANs可以产生高分辨率的逼真图像。谈及这一进展的重要驱动因素，离不开与日俱增的训练数据、模型规模 [1]，网络架构的创新 [3]，以及稳定训练的新的标准化技术 [6]。同时，还有研究者证明了在非常低的码率下，使用基于GANs的压缩系统可以比最先进压缩算法节省2倍的码率。

3. 方法

我们的算法基于Google在2020年10月提出的一篇论文 [4]。我们在这篇论文的基础上，对其方法进行了一定的改进。在这里，我先对论文中涉及到算法模型进行简单的介绍。

3.1. 背景

3.1.1 条件生成对抗网络

条件生成对抗网络 [5] 是一种学习在附加信息 s 下数据条件分布 $p_{X|S}$ 生成模型的方法。这里 x 表示我们的数据点，具体到图像压缩任务，也就是原始图片；而 s 则代表附加信息，比如标签、语义地图。 x, s 通过一个未知的联合分布 $p_{X,S}$ 相关联。GANs 主要包含两个互相竞争的网络结构。其一是生成器G，它以 s 为

条件，将输入的样本 y 从一个固定的已知分布映射到一个条件分布。而另一个则是鉴别器D，其起到了判断一个输入是来自生成器G还是原始数据条件分布的作用。形式化的说，它将输入的数据点和附加信息数据对映射成了样本来自条件分布而不是生成器G输出的概率。我们的目标是要让生成器G去欺骗鉴别器D，让D相信G生成的样本是真实的。而我们要让D尽可能的去识别出输入的真正来源。这个思想体现在了损失函数的设计当中。

$$\mathcal{L}_G = \mathbb{E}_{y \sim p_Y} [-\log(D(G(y, s), s))] \quad (1)$$

$$\begin{aligned} \mathcal{L}_D = & \mathbb{E}_{y \sim p_Y} [-\log(1 - D(G(y, s), s))] \\ & + \mathbb{E}_{x \sim p_{X|s}} [-\log(D(x, s))]. \end{aligned} \quad (2)$$

我们令生成器G的损失函数与鉴别器D识别错误的概率呈相反趋势。我们在生成器中输入我们的样本 y ，把得到的输出通过鉴别器D来判别，D给出的概率越大，也就是识别错误的概率越大，我们得到的损失函数值越小，这也正是我们想要的。而对鉴别器D损失函数的设计中，我们考察了鉴别器D对生成器G输出以及原始数据的辨别结果，使得对生成器G输出的误判概率越低，对原始数据判断正确的概率越高，损失函数的值越小。

3.1.2 自动编码器

自动编码器包括一个编码器E与解码器G。我们输入数据 x ，这个 x 可以为图片，其通过编码器E得到表示 y ，然后 y 再通过解码器G得到了压缩后的图片 x' 。我们根据压缩后图片的质量来评价编码器和解码器的优劣。在这里，我们取图片的码率 $r(y)$ ，以及其失真 $d(x, x')$ 作为描述图片质量的指标。同时我们引入一个参数 λ 来调节两者之间的着重程度， λ 越大，表示越重视码率。因此我们选取以下损失函数，码率越小（压缩程度越大），以及失真度越小，则Loss就越小。

$$\mathcal{L}_{EG} = \mathbb{E}_{x \sim p_X} [\lambda r(y) + d(x, x')] \quad (3)$$

3.1.3 合并架构

我们将条件生成对抗网络和自动编码器结合在了一起，令编码器的输出 y 为生成器的附加信息 s ，此时生成器成为了解码器。此外，我们取失真函数为MSE以及LPIPS的结合 $d = k_M \text{MSE} + k_P \text{LPIPS}$ ，MSE为均方误差，而LPIPS为知觉失真的一种衡量。 k_M, k_P 是用来调节两者平衡的超参数。此时，我们将前面的损失函数 $\mathcal{L}_G(1)$ 与 $\mathcal{L}_{EG}(3)$ 结合，得到损失函数 $\mathcal{L}_{EGP}(4)$ ；再将 s 替换成 y ，将 $G(y, s)$ 替换成 x' ，我们就得到了以下的两个损失函数。

$$\mathcal{L}_{EGP} = \mathbb{E}_{x \sim p_X} [\lambda r(y) + d(x, x') - \beta \log(D(x', y))] \quad (4)$$

$$\mathcal{L}_D = \mathbb{E}_{x \sim p_X} [-\log(1 - D(x', y))] + \mathbb{E}_{x \sim p_X} [-\log(D(x, y))] \quad (5)$$

我们来看下整体的运行流程。输入图像 x ，通过编码器E得到其表示，分别通过概率模型P存储以及通过Q量化得到表示 y 。我们将表示 y 输入解码器G，也是我们的生成器，得到压缩后的图像 x' 。最后，我们把 x' 输入辨别器D，让其来判断 x' 的来源。在运行的途中，我们通过计算损失函数，条件网络中的权重，就此完成了整个模型的训练。

3.2. 改进和优化

3.2.1 对原始架构的分析

在原始架构中为了提高Encoder的压缩效率，采取了一种简单但却高效的降采样方式。原始论文直接使用了6个卷积层来实现Encoder的功能（见Fig.1）。其中，第一层和第六层的卷积层用于匹配图像通道数。而剩下的2至4层的步长均被设置为2，以实现降采样；与此同时，每经过一次降采样，图像的通道数加倍以避免信息丢失。这样的架构能够在尽可能少的网络层数上实现较好的压缩性能。考虑到原始架构在通过Encoder之后，还需要经过网络层数较高的Decoder和生成器Generator，故采用这种简单但却高效的降采样方式来减少神经网络层数，以此提高网络收敛速度是非常有必要的。

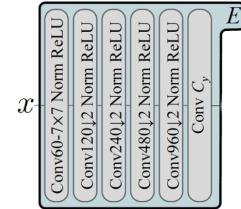


Fig. 1.

但是，与此同时，这种架构设计的缺点也是显而易见的。由于原始架构的Encoder基本上每一层都对图像进行了降采样，因此，尽管通道数加倍，但是高维信息的丢失可以说是无法避免的。从某种层面上说，降低高维信息的损失与降低Encoder的网络层数是相互龃龉的。因此，如何在保证整体的网络依然容易收敛的前提下，尽可能的降低高维信息的损失成为了架构改进和优化的重难点所在。

3.2.2 架构改进

通过进一步的分析，我们可以发现，若能够找到一种新型的网络架构，使得随着网络层数的加深，其收敛能力并不会有明显下降，那么所有的问题就迎刃而解。在这种想法的指导下，我们开始寻找在这个问题上表现更优异的网络架构，以期实现更好的压缩性能。

为了尝试解决网络层数过深产生的退化问题, *ResNet*的研究员选择使用学习残差的方式来构建网络模型。实验证明, 即便在网络层数很深的情况下, *ResNet*也能够保证不错的收敛能力。因此, 我们尝试基于*ResNet*架构, 修改原始论文中的*Encoder*模型。

我们将原始论文中*Encoder*的每一层卷积层都替换成相对应的残差块。每一个残差块中都含有2层的卷积层: 其中一层用于降采样, 与原论文类似的, 其步长被设置为2; 另外一层则使用小卷积核, 并且步长被设置为1, 来实现对图像高维信息的抽取和保存。相比于原始的*Encoder*, 改进后的*Encoder*的参数基本增加了一倍, 因此合理猜想, 其保留图像高维信息的能力也会随之增加。而且, 由于采取了残差学习的方式, 我们认为, *Encoder*收敛能力将不会明显下降, 亦即, 对于*Encoder*的改进工作将不会对后续的*Decoder*和*Generator*的性能有很明显的负面效果。

4. 评测

对于图像压缩任务, 我们从比特率 (Bits Per Pixel, BPP), 峰值讯噪比 (Peak Signal-to-Noise Ratio, PSNR), 结构相似性指标 (Structural Similarity index, SSIM index) 三个指标评测其压缩效果。

4.1. 比特率BPP

BPP评估图片压缩效率

$$BPP = \frac{\text{压缩后图片总比特数}}{\text{图片像素数}} \quad (6)$$

4.2. 峰值讯噪比PSNR

PSNR 衡量两幅图像的相似程度, 基于MSE 误差直接计算, 记 I, K 为两张 $m \times n$ 图像

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i, j) - K(i, j)]^2 \quad (7)$$

$$PSNR = 10 \log_{10} \left(\frac{MAX_I^2}{MSE} \right) = 20 \log_{10} \left(\frac{MAX_I}{\sqrt{MSE}} \right) \quad (8)$$

PSNR 指标越高表明两张图片越相似, 但其直接比较两张图片对应像素的差异, 对图片视觉上的相似性评估效果较差。

4.3. 结构相似性指标SSIM

自然影像是高度结构化的, 即在自然影像中相邻像素之间有很强的关联性, 因此结构相似性在影像品质的衡量上更能符合人眼对影像品质的判断。给定信号 x, y , 定义结构相似性指标为

$$SSIM(x, y) = [l(x, y)]^\alpha [c(x, y)]^\beta [s(x, y)]^\gamma \quad (9)$$

$$\begin{aligned} \text{其中 } l(x, y) &= \frac{2\mu_x\mu_y+C_1}{\mu_x^2+\mu_y^2+C_1}, \quad c(x, y) = \frac{2\sigma_x\sigma_y+C_2}{\sigma_x^2+\sigma_y^2+C_2}, \\ s(x, y) &= \frac{\sigma_{xy}+C_3}{\sigma_x\sigma_y+C_3}, \quad \text{一般取 } \alpha = \beta = \gamma = 1 \end{aligned}$$

5. 实验结果

5.1. 数据集

由于华为云资源上传和下载数据速率过慢而无法接受, 本实验全过程在Google Colab 平台完成。Google Colab 是虚拟机环境云端平台, 限制了虚拟机硬盘的大小, 因此不能使用过大数据集 (如原HiFiC 论文使用的coco2014 数据集和CLIC 2021 数据集), 因此我们采用较小数据集cifar10, 其包含60000张 32×32 大小图片, 其中50000张为训练集, 10000张为数据集。

5.2. 训练

Google Colab 限制了连续连接时长, 因此不能进行过长时间的训练, 本实验训练参数如表1

参数	数值	备注
batch size	10	对MSE + LPIPS 模型的预训练
pre-train iteration	1k	训练GAN 模型
train iteration	4k	

我们使用同样参数训练原HiFiC 模型和改进后的ResHiFiC 模型, 并分别测试了二者对低分辨率和高分辨率图像的压缩效果。

5.3. 低分辨率图像压缩

我们使用测试集的 32×32 图片作为低分辨率图像, 得到模型视觉效果如图2。可以看出压缩效果很差, 这是因为原图片像素数本身过少, 压缩后图片单个像素点的畸变对总体效果都有较大影响。

5.4. 高分辨率图像压缩

尽管模型采用训练集质量不高且训练迭代次数较少, 但对于高分辨率图片有可以接受的压缩效果, 且对分辨率越高的图片压缩效果越好。图4、图5给出改进前后的模型对一张 1280×720 图片的压缩效果, 可以看出较之低分辨率图像压缩效果已经有明显改善; 图6、图7给出改进前后的模型对一张 1500×2228 图片的压缩效果, 此效果已经能够为人眼所接受。

5.5. 改进效果

对于低分辨率图像, 增加了残差块的ResHiFiC 牺牲了部分BPP 而获得了PSNR, SSIM 的提升 (图3)。这表明改进方法在测试集上已经取得成效。

但对于高分辨率图像, 改进措施没有取得更好效果, 甚至使压缩效果变差。这可能是因为Encoder 模块卷积层数本身并不够多, 残差块的引入没有起到较好的帮助收敛作用, 反而为卷积层引入新的误差。而增大卷积层数会导致训练参数过多, 在Google Colab 平台上训练时间难以接受。

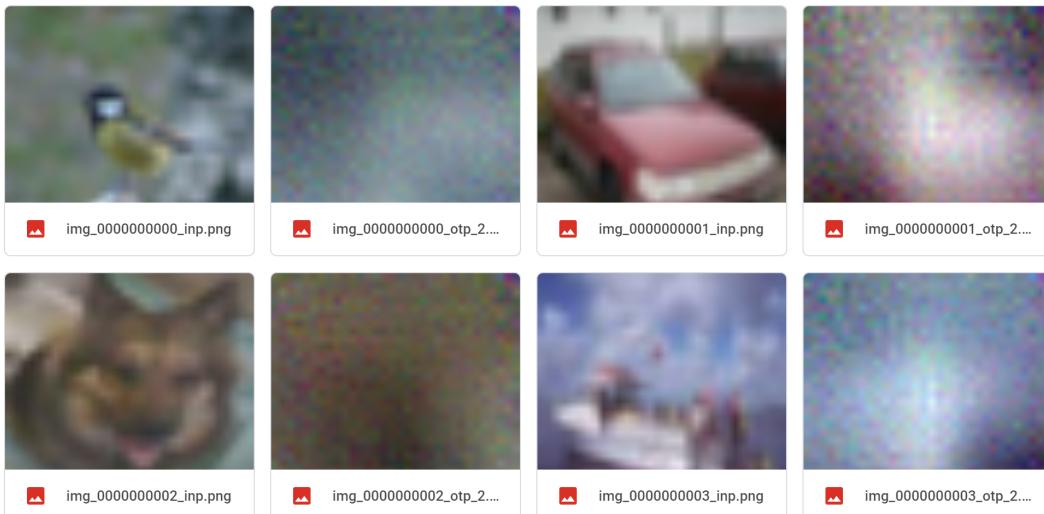


Fig. 2. 低分辨率图像压缩视觉效果

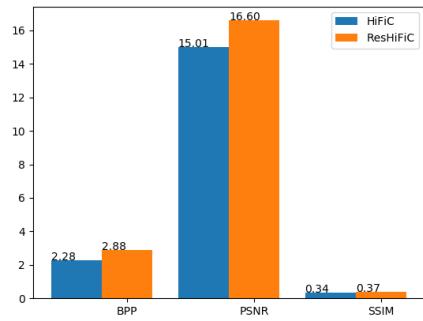


Fig. 3. HiFiC与ResHiFiC在低分辨率图像上的压缩效果

6. 模型使用

本项目所有内容已经开源，任何人可在 <https://github.com/HumphreyChou/CLIC-2021-Learned-Image-Compression> 找到。其中，我们已在 Google Colab 平台创建一个使用 ResHiFiC 模型的图像压缩文件，你可以使用它训练模型并压缩图像。

References

- [1] A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.
- [2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.
- [3] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.
- [4] F. Mentzer, G. Toderici, M. Tschanne, and E. Agustsson. High-fidelity generative image compression. *arXiv preprint arXiv:2006.09965*, 2020.
- [5] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- [6] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.
- [7] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand. Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on circuits and systems for video technology*, 22(12):1649–1668, 2012.
- [8] G. Toderici, S. M. O’Malley, S. J. Hwang, D. Vincent, D. Minnen, S. Baluja, M. Covell, and R. Sukthankar. Variable rate image compression with recurrent neural networks. *arXiv preprint arXiv:1511.06085*, 2015.
- [9] G. Toderici, D. Vincent, N. Johnston, S. Jin Hwang, D. Minnen, J. Shor, and M. Covell. Full resolution image compression with recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5306–5314, 2017.
- [10] G. K. Wallace. The jpeg still picture compression standard. *IEEE transactions on consumer electronics*, 38(1):xviii–xxxiv, 1992.



bpp: 0.2349 PSNR: 22.60 MS-SSIM: 0.741

Fig. 4. HiFiC 对 1280×720 图片压缩效果



bpp: 0.49 PSNR: 21.33 SSIM: 0.64

Fig. 5. ResHiFiC 对 1280×720 图片压缩效果

Original Image (Left) vs. Decompressed Image (Right)



bpp: 0.2165 PSNR: 26.78 MS-SSIM: 0.881

Fig. 6. HiFiC 对 1500×2228 图片压缩效果

Original Image (Left) vs. Decompressed Image (Right)



bpp: 0.35 PSNR: 26.82 SSIM: 0.76

Fig. 7. ResHiFiC 对 1500×2228 图片压缩效果