摘要

随着互联网和数字多媒体的发展，图像视频数据量爆炸式增长，数据存储和传输面临严峻挑战，因此迫切需要更低码率、更高质量的压缩方法。当前主流的图像压缩标准和基于深度学习的压缩方法采用变换、量化、熵编码构成的技术体系，基于数据统计特性获得信号的紧凑表示，减少冗余信息。然而在极低码率下，从这些表示中保留的少量信息难以恢复出服从自然图像分布的图像，重建的图像有明显的块效应、模糊等问题。随着生成式大模型在文本、图像、视频生成大放异彩，基于一段简单的文本、结构图等即可生成高质量的图像视频，表明生成式大模型学到了从稀疏表示到复杂图像视频信号之间的映射关系。这启示着突破传统压缩框架，建立更加高效、稀疏的压缩表示，发掘并利用海量数据和大规模参数学到的映射关系，在实现极低码率下泛化性较好的高质量图像重构方面具有较大的潜力。本文提出跨模态图像压缩方法，拓展以文本为基础的压缩表示，自底向上增加码率，在提高感知质量的基础上降低像素失真，主要创新点包括以下三个方面：

第一，提出基于文本描述的跨模态图像压缩方法。以构建极低码率图像压缩为动机，考虑到变换系数等常见压缩表示在该码率难以有效编码图像信息，拓展文本作为压缩表示，提出基于文本描述的跨模态图像压缩方法，缓解当前压缩方法在极低码率下存在的块效应、模糊等问题，提高感知质量。考虑到文本是具有语法特性的符号序列，端到端优化将使文本难以保持语法特性和语义信息，因此提出基于多步自评估强化学习的率失真优化方法。发掘图像生成文本的特殊性，重新形式化优势函数，采用自评估方法估计单步和多步优势函数，构建率失真反馈函数实现率失真优化。实验表明，该方法相比基线模型在图像描述相关指标 CIDEr-D 指标上实现 13% 的性能增益，在低于 0.001 bpp 的超低码率下在 MSCOCO 数据集上相比于同时期最佳压缩方法在感知质量评价指标 FID 上提升了 34%。

第二，提出结构纹理紧凑表示的跨模态图像压缩方法。以生成紧凑的结构和纹理两方面入手，生成并压缩边缘图实现紧凑结构表示，生成稠密描述文本作为紧凑的纹理表示，提出联合率失真优化方法，缓解当前压缩方法在极低码率下难以维持结构信息的问题，提高感知质量。具体地，采用边缘图提取网络和矢量量化网络生成和压缩边缘图，将稠密描述分为区域检测和区域描述，统一建模为离散序列预测问题，提出基于 Transformer 的两阶段稠密描述生成模型，提出两阶段率失真优化方法来优化模型。实验表明在低于 0.1 bpp 的极低码率下，在 MSCOCO 数据集上相比于同时期最佳压缩方法在感知质量评价指标 FID 上提升了 27%。

第三，提出多模态一致性约束的跨模态图像压缩方法。以文本表示提供全局语义信息和极低码率下的隐空间表示提供局部纹理结构信息为动机，基于自监督学习范式构建模态间一致性约束，提出交替率失真优化算法实现紧凑文本提取，提出基于文本-隐空间表示的扩散模型，缓解当前压缩方法在极低码率下难以权衡感知失真和像素失真的问题，实现较高的感知质量和较低的像素失真。实验表明该方法在低于 0.1 bpp 的极低码率下相比传统图像压缩方法、端到端图像压缩方法和生成式图像压缩方法实现更好的感知质量和失真权衡，在 Kodak 数据集上相比于当前最佳生成式图像压缩方法在像素失真评价指标 PSNR 上提升了 17%。该模型可与传统压缩标准方法相结合，实现更低码率更高质量重构，相同感知质量下码率节省 50%。该模型获取的压缩表示具有人可理解的文本信息，以较低编码传输代价实现图像检索应用。在自监督的压缩目标优化下，促进多模态生成式大模型的模态转换对齐，其中图文生成性能在测试数据集中可超过 mPLUG-Owl 和 Stable Diffusion 1.5 等多模态大模型。这表明图像压缩可与生成式大模型相辅相成，为提升生成式大模型性能提供一种新颖的视角。

综上所述，本文提出跨模态图像压缩方法，从极低码率的文本表示为基础压缩表示并逐渐拓展，在感知质量和像素失真两方面实现较高的整体性能。所提方法在统一框架中实现高效图像压缩、理解和生成，为构建视觉智能奠定基础。

关键词：图像压缩，跨模态图像压缩，生成式图像压缩，率失真优化

Research on Cross-modal Image Compression Methods

Junlong Gao (Computer Application Technology)
Directed by: Prof. Wen Gao and Prof. Siwei Ma

ABSTRACT

With the development of the Internet and digital multimedia, the amount of image and video data has exploded, and data storage and transmission are facing severe challenges. Therefore, there is an urgent need for compression methods with lower bit rates and higher quality. The current mainstream image compression standards and deep learning-based compression methods use a technical system composed of transformation, quantization, and entropy coding to obtain a compact representation of the signal and reduce redundant information based on the statistical characteristics of the data. However, at extremely low bit rates, it is difficult to recover images that obey natural image distribution from the very small amount of information retained in these representations, and the reconstructed images have obvious block effects, blurring, and other problems. As generative large models shine in text, image, and video generation, high-quality images and videos can be generated conditioned on a simple piece of text or structure map, indicating that generative large models have learned the mapping from sparse representation to complex image and video signals. This enlightens us to break through the traditional compression framework, establish a more efficient and sparse compression representation, discover and utilize the mapping learned with massive data and large-scale parameters, and have great potential to achieve high-quality image reconstruction at extremely low bit rates. Based on this motivation, Cross-Modal Image Compression is proposed, expanding compression representation based on text, increasing bit rate from less to more, and research is carried out from increasing perceptual quality to lowering pixel-level distortion. The main innovation points include the following three aspects:

First, a text-based cross-modal image compression method is proposed. Motivated by the construction of extremely low bit rate image compression, considering that common compression representations such as transformation coefficients are difficult to effectively encode image information at this bit rate, text is extended as a compression representation, and a text-based cross-modal image compression method is proposed to alleviate block effects and blurring problems of current compression methods to improve perceptual quality at extremely low bi-

trates. Considering that text is a sequence of symbols with grammatical characteristics, end-to-end optimization will make it difficult for the text to maintain grammatical characteristics and semantic information. A rate-distortion optimization method based on multi-step self-critical reinforcement learning is proposed. We explore the properties of text generation, reformulate the advantage function, use a self-critical method to estimate single-step and multi-step advantage functions, and construct a rate-distortion reward function to achieve rate-distortion optimization. Experiments show that a performance gain of 13% is achieved on the image captioning-related metric CIDEr-D by the proposed method compared with the baseline model. Compared with the contemporary best image compression method, the proposed method improves FID, a perceptual quality evaluation metric, by 34% on the MSCOCO dataset at an ultra-low bit rate below 0.001 bpp.

Second, a cross-modal image compression method based on the compact representation of structure and texture is proposed. Starting from the two aspects of compact structure and texture generation, we generate and compress edge maps to achieve compact structure representation, generate dense captions as compact texture representation, and use a joint rate-distortion optimization method. As such, the proposed method can alleviate the structure-preserving problem of current compression methods at extremely low bitrates and improve the perceptual quality. Specifically, edge map extraction network and vector quantization network are used to generate and compress edge maps, dense caption generation is divided into region detection and region captioning and unified modeled as a discrete sequence prediction problem. A two-stage dense caption generation model based on Transformer is proposed. A two-stage rate-distortion optimization method is used to optimize the model. Experiments show that compared with the contemporary best image compression method, the proposed method improves FID by 27% on the MSCOCO dataset at extremely low bit rates below 0.1 bpp.

Third, a cross-modal image compression method based on multi-modal consistency constraints is proposed. Motivated by text representation providing glocal semantic information and latent space representation providing local texture and structure information at extremely low bit rates, inter-modal consistency constraints based on a self-supervised learning paradigm are constructed. An alternating rate-distortion optimization algorithm is proposed to achieve compact text extraction, and a diffusion model based on text-latent representation is proposed to alleviate the difficulty in the trade-off between perceptual quality and pixel-level distortion of current compression methods at extremely low bitrates, where perceptual quality gets higher and pixel-level distortion gets lower. Experiments show that a better tradeoff between

the perceptual quality and pixel-level distortion at the extremely low bit rates below 0.1 bpp is achieved by the proposed method, compared to traditional image compression methods, end-to-end image compression methods, and generative image compression methods. Compared with the current best generative image compression method, PSNR, a pixel-level distortion evaluation metric, is improved by 17% on the Kodak dataset. This model can be combined with traditional compression standards and achieve higher quality reconstruction at lower bit rates, saving 50% in bit rate under the same perceived quality. The compressed representation obtained by this model has human-understandable text information and can better support image retrieval applications. Moreover, under self-supervised compression optimization, the cross-modal transformation alignment of multi-modal generative large models is improved. The performance on text and image generation can exceed large models such as mPLUG-Owl and Stable Diffusion 1.5 in the testing dataset. This shows that image compression can complement generative large models, which also provides a novel perspective for improving the performance of generative large models.

To sum up, Cross-Modal Image Compression is proposed. This work starts from text representation as the basic compression representation and gradually expands it to achieve better overall perceptual quality and pixel-level distortion. This work achieves efficient image compression, analysis, and generation in a unified framework, laying the foundation for building visual intelligence.

KEY WORDS: Image compression, cross-modal compression, generative image compression, rate-distortion optimization