Entended Variable Rate Image Compression With a Context Entropy Model

Yichen Qian, Zhiyu Tan, Hesen Chen, Ming Lin, Xiuyu Sun, Hao Li
Alibaba, Group
{yichen.qyc,zhiyu.tzy,hesen.chs,ming.l,xiuyu.sxy,lihao.lh}@alibaba-inc.com

Abstract

In this paper, we propose an extended variable-rate image compression method using a conditional autoencoder[2]. We deploy one variable-rate image compression network with a conditional autoencoder that can yield compressed images of varying quality. Based on this effective framework, we use a non-local module to train model with attention mechanism. The probability estimation of latent representation is jointly modeled by a hyperprior autoencoder and autoregressive context module. The Encoder uses pyramidal feature maps to improve the compression performance.

1. The Proposed framework

Fig.1 depicts the proposed image compression framework. To avoid training and deploying multiple networks, we use a conditional autoencoder [2] to yield varying quality compression image. The condition audoencoder takes λ as a conditioning input parameter, along with the input image, and produces a compressed image with different rate and distortion.

$$R_{\phi,\theta}(\lambda) = \mathbb{E}_{p(x)p_\phi(y|x,\lambda)}[-\log_2 q_\theta(y|\lambda)]$$

And then we trainthe model with the loss function,

$$\min_{\phi,\theta} \sum_{\lambda\in\Lambda}(R_{\phi,\theta}(\lambda) + \lambda D_{\phi,\theta}(\lambda))$$

where λ is random selected from a pre-defined finite set Λ, and we use MS-SSIM as distortion metric. We use a conditional convolution to implement a conditional autoencoder. Through a one-hot encoding, λ is used to channel-wise scale and bias feature maps of each layer.

We also use Non-local Attention Modules (NLAM)[3] to capture global connections between features in different channels and spatial locations. The NLAM module is shown in fig.1, which has three branch.

In hyperprior probabilistic model, two different way is to capture context information. First, we use additional side information [1] to capture context by a hyper-latents $\hat{z}$. Second, we utilize the current spatial feature of $\hat{y}$ to predict its bias and scale autoregressively. The autoregressive model is based on [2].

This paragraph illustrates the details of experiments. We trained 8 models with the same framework but different parameters. The difference lays in the channel of convolution layer and latent $\hat{y}$. We trained all the models using 5123 images and random crop then to 256×256 in the training. ADAM optimizer is used with learning rate of 1×10^{-4}, and then reduced half in the last 20 epochs. For the low-rate image compression track (0.15 bits per pixel constraint) in CLIC-2020, we obtain 0.9774 and 0.9811 performance in terms of multi-scale structural similarity (MS-SSIM) on validation and test dataset respectively.

References

[1] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. *ICLR*, 2018.

[2] Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. Variable rate deep image compression with a conditional autoencoder. In *ICCV*, 2019.

[3] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019.

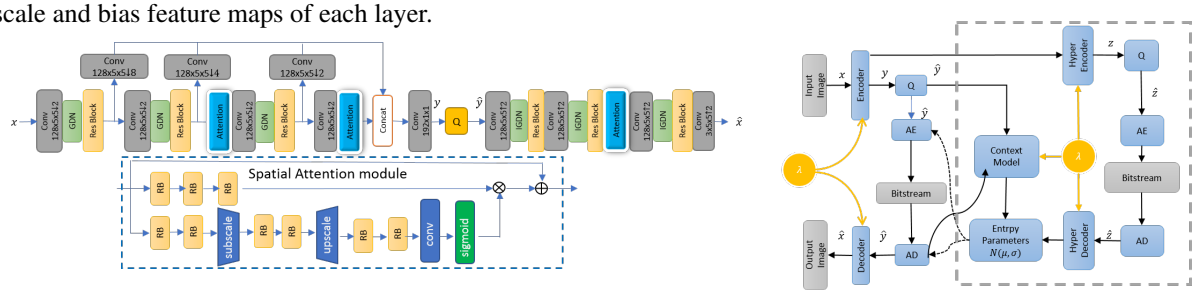

Figure 1. Proposed image compression framework.