Learned Dual Perceptual Image Similarity for Compressed Image

Jianzhao Liu [*], Simeng Sun [*], Yiting Lu, Xin Li, Zhibo Chen
University of Science and Technology of China
{jianzhao,smsun20}@mail.ustc.edu.cn, chenzhibo@ustc.edu.cn

Abstract

Designing an image quality assessment (IQA) for compression distortion can not only help to verify the performance of them but also help to guide the optimization. In this paper, we propose a learned dual perceptual image similarity scheme (DPIS) for compressed images, which considering feature similarity on both pixel-wise and gradient-wise. Experimental results show that the proposed metric outperforms both traditional methods and learning-based methods on the CLIC validation set.

1. The proposed framework

In this paper, we propose DPIS metric, which is inspired by structure and texture similarity distance [1] as well as l_2 feature distance [3]. Besides, we also take the gradient map of an image as supplementary input to focus the attention on the regions which have high frequency details. We believe that gradients convey important visual information and can effectively capture the structural and contrast changes which are crucial to compressed image quality assessment. The detailed architecture is shown in Fig.1. In the left half of the figure, we compute each channel's structure and texture similarity of the features extracted from image pair (x, y) and the gradient pair $(Grad(x), Grad(y))$ at six stages of pre-trained VGG (lablled $input$, $conv1\_2$, $conv2\_2$, $conv3\_3$, $conv4\_3$, and $conv5\_3$). Then we employ a set of learnable weights α_{ij} to adpatively weight the texture similarity of each channel and use β_{ij} to weight the structure similarity of each channel, resulting in image similarity and gradient similarity. Then we use a FC layer to fuse the two similarity scores and get the distance $D_1(x, y)$. In the right half of Fig.1, we employ a pre-trained AlexNet and compute each channel's l_2 distance of unit-normalized features extracted from the image pair and the gradient pair at five stages (from $conv1$ - $conv5$), with a vector ψ scaling the features channel-wise. Then we average spatially and sum channel-wise, resulting in image l_2 feature distance and gradient l_2 feature distance. Similarly, we use a FC layer to fuse the two distance and get the distance $D_2(x, y)$. Finally, we weight $D_1(x, y)$ and $D_2(x, y)$ by a learnable weight γ.

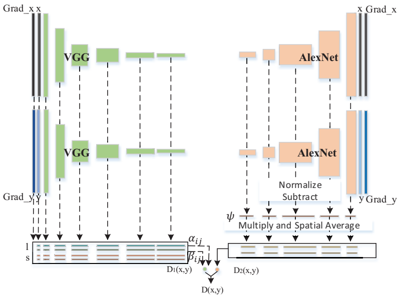

Figure 1: Framework of DPIS.

This paragraph illustrates the details of experiments. The network was trained on BAPPS [3] and PieAPP [2] datasets. We employ a small network to predict perceptual judgment [3] and use BCE loss for training. We randomly cropped the images to $224 \times 224 \times 3$ while training. During testing, we cropped the images into various patches and averaged the predicted distance of all patches to get a more accurate result. To evaluate the performance, given triplets (x,y1,y2), we record the predicted judgment (which distorted image is closer to the reference image x) given by each metric and compute the accuracy. We compare our method with PSNR, MS-SSIM, LPIPS and DISTS on CLIC2021 validation set. Our DPIS metric can achieve 0.792 accuracy on provided validation set, while PSNR, MS-SSIM, LPIPS and DISTS achieve 0.572, 0.612, 0.753 and 0.749 respectively.

References

[1] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *arXiv preprint arXiv:2004.07728*, 2020. 1

[2] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2018. 1

[3] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1