LETTER

Evaluating the Stability of Deep Image Quality Assessment With Respect to Image Scaling

Koki TSUBOTA\(^{a}\), Nonmember, Hiroaki AKUTSU\(^{†\dagger}\), Member, and Kiyoharu AIZAWA\(^{c}\), Fellow

SUMMARY Image quality assessment (IQA) is a fundamental metric for image processing tasks (e.g., compression). With full-reference IQAs, traditional IQAs, such as PSNR and SSIM, have been used. Recently, IQAs based on deep neural networks (deep IQAs), such as LPIPS and DISTS, have also been used. It is known that image scaling is inconsistent among deep IQAs, as some perform down-scaling as pre-processing, whereas others instead use the original image size. In this paper, we show that the image scale is an influential factor that affects deep IQA performance. We comprehensively evaluate four deep IQAs on the same five datasets, and the experimental results show that image scale significantly influences IQA performance. We found that the most appropriate image scale is often neither the default nor the original size, and the choice differs depending on the methods and datasets used. We visualized the stability and found that PieAPP is the most stable among the four deep IQAs.

key words: Full-reference image quality assessment, image scale, neural networks.

1. Introduction

This paper discusses full-reference image quality assessment (IQA), which aims to predict how humans perceive the visual quality of distorted images by comparing them to non-distorted references. Full-reference IQA provides evaluation criteria for image processing tasks, such as image compression and image restoration. Recently, the full-reference IQA has also been used as a loss function for such tasks \([1,2]\).

Peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) index metrics \([3]\) are widely used full-reference IQAs. Despite their popularity, they do not accurately model human perception. To better model our complex visual system, IQAs using deep neural networks (deep IQAs) have recently been investigated \([4,6]\), achieving a better correlation to human perception. Specifically, learned perceptual image patch similarity (LPIPS) \([4]\), a deep IQA, is widely used as the evaluation metric in computer vision tasks \([1]\).

This study examines the image scale used when predicting image quality. Many traditional IQAs \([3,7,8]\) down-scale images to a specific size for pre-processing. However, the process is inconsistent among deep IQAs, as some \([6,9]\) perform down-scaling and others \([4,5,10]\) use the original size.

In this paper, we consider the image scale as the independent variable and empirically evaluate its effect on IQA performance. Notably, a few other researchers have investigated this effect. Joy et al. \([11]\), for example, studied the effect of image scale on SSIM \([3]\) and feature similarity index (FSIM) metrics \([7]\). Gao et al. \([9]\) used the distance between intermediate representations extracted by a VGG network \([12]\) pre-trained on ImageNet \([13]\) for IQA. They investigated two image sizes: original (e.g., 384 × 512 pixels in the TID2013 \([14]\) dataset) and resized (224 × 224 pixels), empirically demonstrating that using the resized version improved performance. However, their investigation was limited to these two sizes.

In our experiments, we evaluate four deep IQAs: Alex-lin and VGG-lin of LPIPS \([4]\), the perceptual image-error assessment through pairwise preference (PieAPP) \([5]\), and deep image structure and texture similarity (DISTS) \([6]\). We apply each of these IQAs to the same five datasets. Through comprehensive experimentation, we show the effect of image scale on IQA performance.

Our contributions are summarized as follows:

- Using the image scale as the independent variable, we evaluate deep IQAs with different image scales and comprehensively measure their effect on performance. To the best of our knowledge, this is the first work to do so.
- We empirically show that the image scale greatly affects deep IQA performance and that the most desirable image scale is often neither the default nor the original size; it differs depending on the deep IQAs and the datasets used.
- We visualize the stability of the deep IQAs with respect to their scales and performance scores and found that PieAPP is the most stable.

2. Related Works

IQA is a fundamental image processing task that has been studied for several decades. The simplest IQA is the mean squared error (MSE). It averages the squared errors of pixel values to obtain pixel-level errors. PSNR is a different representation of MSE. SSIM \([3]\) is also widely used. It captures the structural similarity of images instead of their pixel-level errors. FSIM \([7]\) uses phase congruency and gradient magnitude as features to calculate similarity. Multi-scale
SSIM (MS-SSIM) \[15\] calculates SSIM at multiple scales by down-scaling the image.

With the development of deep IQAs, several related studies have been published. Zhang et al. \[4\] constructed a large-scale patch dataset for examining perceptual loss \[16\], which is the computed distance between intermediate representations of models pre-trained on ImageNet \[13\]. They also proposed LPIPS, which performs a weighted summation of the distance across channels. Prashnani et al. \[5\] presented a pairwise-learning framework using their PieAPP dataset. Ding et al. \[6\] focused on image textures and trained an invariant model under texture shifts. Their model computes similarity not only in the feature space, but also in the image space to obtain an injective function. Notably, the image scaling used for these IQAs was inconsistent.

3. Method

We illustrate the flow of our evaluation procedure in Fig. 1. Given a reference and distorted image pair, we perform down-scaling on both using various scales for preprocessing. We then apply IQAs to the pair to obtain the IQA score.

We use four deep IQAs (i.e., Alex-lin and VGG-lin of LPIPS \[4\], PieAPP \[5\], and DISTS \[6\]) for evaluation. We list the training data and initial weights of these IQAs in Table 1. PieAPP is initialized with random weights, whereas the others are initialized with ImageNet \[13\] pretrained weights that are fixed during training.

LPIPS \[4\] extracts intermediate representations from an image pair using a backbone network and computes their weighted distance per channel. Alex-lin and VGG-lin respectively use AlexNet \[17\] and VGG \[12\] as the backbone network. LPIPS does not perform down-scaling during training and inference.

PieAPP \[5\] leverages a VGG-based network architecture on the PieAPP dataset \[5\] using multiple random 64×64 cropped patches. During inference, PieAPP is applied to the image patches cropped using sliding window. PieAPP does not perform down-scaling during training and inference.

DISTS \[6\] uses a VGG network as the backbone network. It calculates the global similarity between intermediate representations of the backbone network extracted from an image pair. It is applied to scaled image pairs whose shorter side is set to 256 pixels during training and inference.

4. Experiments

4.1 Experimental Settings

We evaluated the four deep IQAs using the same five datasets: TID2013 \[14\], PieAPP \[5\], KADID-10k \[18\], PIPAL \[20\], and CLIC2021 validation dataset for perceptual metrics (CLIC2021) \[21\]. For reference, we evaluated three traditional methods: PSNR, SSIM \[3\], and MS-SSIM \[15\]. In Table 2, we list the properties of these datasets (i.e., image size, number of images, and metric used). PieAPP and PIPAL datasets comprise patch images, whereas TID2013, KADID-10k, and CLIC2021 datasets comprise whole images. The CLIC2021 dataset contains higher-resolution images than the other datasets. We used images of uniform size from the CLIC2021 dataset by removing five reference images and their corresponding distorted images, whose sizes were not 768×768 pixels.

The TID2013 and KADID-10k datasets contain distortions applied by traditional operations, such as color change and JPEG compression. PieAPP and PIPAL datasets contain distortions applied by traditional operations as well as obtained by algorithm outputs such as super-resolution. The PIPAL dataset includes the output of a trained model with adversarial loss.

The TID2013, PieAPP, KADID-10k, and PIPAL datasets comprise pairs of reference and distorted images. Because each pair is annotated with its relative image quality in the dataset, we used Spearman’s rank correlation coefficients (SRCC) as the evaluation metric for these datasets. The CLIC2021 dataset comprises sets containing a reference image and two distorted images. Because each set was annotated to identify the distorted image with better quality, we used accuracy (ACC) as the evaluation metric for CLIC2021 dataset.
We down-scaled the images using 10 different ratios: \(\{x/9, x/2/9, x/3/9, x/4/9, x/1/2, x/5/9, x/7/9, x/8/9, x/9\}\). \(x/9\) is the original size. These scales included the default scale for each method in each dataset. We used bilinear down-scaling and evaluated the IQA performance when the image size was larger than the minimum size for each IQA model.

### 4.2 Experimental Results

We present the experimental results in Table 3. The results show that the appropriate image scale differs depending on the IQA and evaluation dataset.

Interestingly, the best image scale, where the SRCC or ACC are the best, differs from the default scale and the original size in most cases. For example, the best image scale for LPIPS is \(x/1/2, x/7/9, x/6/9, x/1/2,\) and \(x/5/9\) for TID2013, PieAPP, KADID-10k, PIPAL, and CLIC2021, respectively. These scales differ from the original image size and the default scale.

Down-scaling improves the score in most cases, but discrepancies tend to depend on the method and dataset. Specifically, the difference in LPIPS is large. The SRCC of LPIPS (VGG-lin) when using the TID2013 dataset differs from the default scale by 0.172.

We evaluate the stability of the deep IQAs. We evaluate the stability in the image scale by computing the absolute differences between the default scale and the best scale. The absolute differences are normalized by the best scale and represented by percentage. We also evaluate the stability in the performance score by computing the absolute differences between the scores at the default scale and the best scale. We show the results in Fig. 3. PieAPP is the most stable among the deep IQAs.

We consider the stability depends on the initial weights and the training data. It is because these conditions are the main differences except the network architecture between PieAPP and the other three deep IQAs as shown in Table 4. To validate this, we add experiments of...
LPIPS-RND-BAPPS

LPIPS (VGG-lin) default / best\[4\]

0.034 (1) / 0.004 (1)

0.106 (2) / 0.008 (1)

0.451 (1) / 0.304 (1)

0.024 (1) / 0.014 (1)

LPIPS (Alex-lin) default / best\[4\]

0.189 (3) / 0.030 (2)

0.285 (3) / 0.037 (2)

1.436 (3) / 0.891 (2)

1.012 (2) / 1.250 (3)

Fig. 2: Qualitative results. The images in the far-left column are reference images. The remaining columns contain samples of distorted images sorted by annotated IQA scores. “Default” denotes the default scale, and “best” denotes the best scale for each IQA. The value in parentheses denotes the ranking of the annotated or predicted image quality. The underline denotes that the predicted IQA scores are in the correct order.

Fig. 3: The stability of each IQA. The difference is shown in relative percentage. The IQA is stable if the differences for scales and scores are small.

LPIPS under different conditions. We train LPIPS with a backbone of AlexNet initialized with random weights on the PieAPP dataset (LPIPS-RND-PieAPP) and the BAPPS dataset (LPIPS-RND-BAPPS), respectively. We show the results in Fig. 4. LPIPS-RND-PieAPP is more stable than LPIPS-RND-BAPPS and LPIPS (Alex-lin) and competitive with PieAPP. These results suggest that training deep IQAs on the PieAPP dataset from random initial weights contributes to the stability.

We show the qualitative results of the four IQAs on the TID2013 dataset in Fig. 4. The results demonstrate that the order of IQAs with the best scale is correct, whereas the order of IQAs with the default scale is incorrect. This may be because the IQAs are sensitive to fine and sharp distortions, such as multiplicative Gaussian noise.

Fig. 4: The stability of LPIPS\[4\] with a backbone of AlexNet under different conditions. The difference is shown in relative percentage. The IQA is stable if the differences for scales and scores are small.

5. Conclusions

In this paper, we have analyzed the effect of image scale on performance. Experimental results demonstrated that downscaling images has a large impact on performance and that the best image scale often differs from the default scale and the original size. The best image scale differs depending on the model and the evaluation dataset. We also found that PieAPP was the most stable among the deep IQAs. Further investigation of the factor of stability is the future work.

Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Number 22J13735.

References

[1] F. Mentzer, G. Toderici, M. Tschannen, and E. Agustsson, “High-fidelity generative image compression,” NeurIPS, Virtual, pp.11913–11924, Dec. 2020.
[2] K. Ding, K. Ma, S. Wang, and E.P. Simoncelli, “Comparison of full-reference image quality models for optimization of image processing systems,” IJCV, vol.129, pp.1258–1281, Jan. 2021.
[3] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, “Image
quality assessment: from error visibility to structural similarity,” TIP, vol.13, no.4, pp.600–612, Apr. 2004.
[4] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” CVPR, USA, pp.586–595, June 2018.
[5] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, “Pingapp: Perceptual image-error assessment through pairwise preference,” CVPR, USA, pp.1808–1817, June 2018.
[6] S.W. Keyan Ding, Kede Ma and E.P. Simoncelli, “Image quality assessment: Unifying structure and texture similarity,” TPAMI, 2020.
[7] L. Zhang, L. Zhang, X. Mou, and D. Zhang, “Fsim: A feature similarity index for image quality assessment,” TIP, vol.20, no.8, pp.2378–2386, Jan. 2011.
[8] B. Zhang, P.V. Sander, and A. Bermak, “Gradient magnitude similarity deviation on multiple scales for color image quality assessment,” ICASSP, USA, pp.1253–1257, Mar. 2017.
[9] F. Gao, Y. Wang, P. Li, M. Tan, J. Yu, and Y. Zhu, “Deepsim: Deep similarity for image quality assessment,” Neurocomputing, vol.257, pp.104–114, Sep. 2017.
[10] S. Bosse, D. Maniry, K.R. Müller, T. Wiegand, and W. Samek, “Deep neural networks for no-reference and full-reference image quality assessment,” TIP, vol.27, no.1, pp.206–219, Jan. 2018.
[11] K. Joy and E.G. Sarma, “An evaluation of the effect of image down-sampling on performance indicators of iqa algorithms,” ARPN Journal of Engineering and Applied Sciences, vol.10, no.17, pp.7507–7513, Sep. 2015.
[12] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” ICLR, May 2015.
[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M.S. Bernstein, A.C. Berg, and F.F. Li, “Imagenet large scale visual recognition challenge,” IJCV, vol.115, no.3, pp.211–252, Apr. 2015.
[14] N.N. Ponomarenko, L. Jin, O. Ieremeiev, V.V. Lukin, K.O. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.C.J. Kuo, “Image database tid2013: Peculiarities, results and perspectives,” SPIC, vol.30, pp.57–77, Jan. 2015.
[15] Z. Wang, E.P. Simoncelli, and A.C. Bovik, “Multiscale structural similarity for image quality assessment,” ACSSC, USA, pp.1398–1402, Nov. 2003.
[16] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” ECCV, pp.694–711, Oct. 2016.
[17] A. Krizhevsky, I. Sutskever, and G.E. Hinton, “Imagenet classification with deep convolutional neural networks,” NeurIPS, USA, pp.1106–1114, Dec. 2012.
[18] H. Lin, V. Hosu, and D. Sauge, “Kadid-10k: A large-scale artificially distorted iqa database,” QoMEX, pp.1–3, June 2019.
[19] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, “Describing textures in the wild,” CVPR, USA, pp.3606–3613, IEEE Computer Society, June 2014.
[20] J. Gu, H. Cai, H. Chen, X. Ye, J.S. Ren, and C. Dong, “Pipal: A large-scale image quality assessment dataset for perceptual image restoration,” ECCV, Virtual, pp.633–651, Aug. 2020.
[21] G. Toderici, W. Shi, R. Timofte, L. Theis, J. Ballé, E. Agustsson, N. Johnston, F. Mentzer, Z. Simon, A. Norkin, K. Rapaka, and E. Noury, “Clic 2021: Challenge on learned image compression.” http://compression.cc 2021.