DEEP PERCEPTUAL IMAGE QUALITY ASSESSMENT FOR COMPRESSION

Juan Carlos Mier, Eddie Huang*, Hossein Talebi, Feng Yang, and Peyman Milanfar*

Google Research

ABSTRACT

Lossy Image compression is necessary for efficient storage and transfer of data. Typically the trade-off between bit-rate and quality determines the optimal compression level. This makes the image quality metric an integral part of any imaging system. While the existing full-reference metrics such as PSNR and SSIM may be less sensitive to perceptual quality, the recently introduced learning methods may fail to generalize to unseen data. In this paper we propose the largest image compression quality dataset to date with human perceptual preferences, enabling the use of deep learning, and we develop a full reference perceptual quality assessment metric for lossy image compression that outperforms the existing state-of-the-art methods. We show that the proposed model can effectively learn from thousands of examples available in the new dataset, and consequently it generalizes better to other unseen datasets of human perceptual preference.

***Index Terms*—** Full-reference Image Quality Assessment, Image Compression, Compression Quality Metric, Perceptual Quality Dataset

1. INTRODUCTION

Compressed image data make up a very large portion of data stored in data centers around the world. At the rapidly increasing scale of data storage, compression techniques are more important now than ever. Even a marginal change in efficiency can have large impact on data storage. While image compression techniques are vital for efficiently storing enormous quantities of data, they have remained relatively static over the years [1].

Lossy image compression is a family of techniques for digital images, where bits are discarded when compression takes place resulting in a possible degradation in quality perceived by a human. The degree of compression or bit rate is directly proportional to file size. There is an inherent trade-off between the bit-rate and distortion of a compressed image [2]. Currently, some of the the main metrics used for adjusting bit-rate of an encoder is MSE (mean squared error), PSNR (peak-signal-to-noise-ratio), or other similar quality metrics however, it has been shown that large PSNR values do not necessarily mean a high perceptual quality. For instance, the

* equal contribution

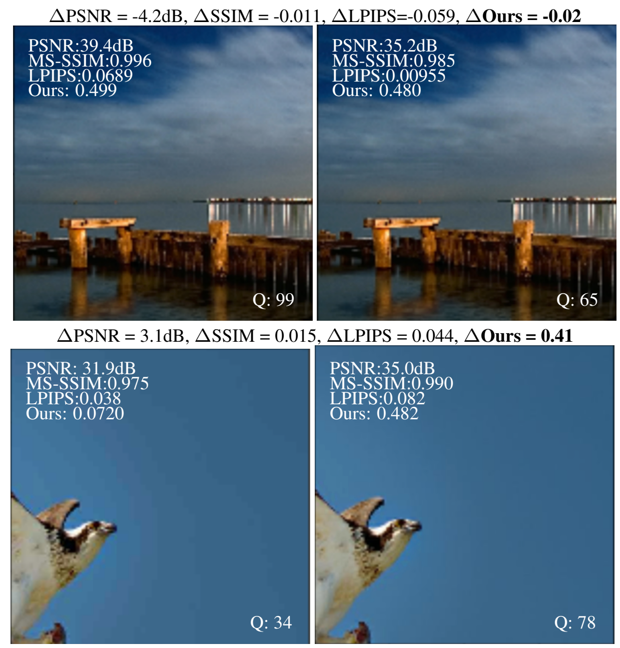

Fig. 1. Examples of CIQA pairwise comparisons with labeled JPEG Q factor, PSNR, SSIM, LPIPS and our model output. PSNR and SSIM do not adequately correlate with visible distortions in the compressed image. For each metric the Δ is taken as $Metric(Image_2, Reference) - Metric(Image_1, Reference)$.

JPEG Encoder can generate artifacts such as color banding and blockiness, that are not well-captured by PSNR [3]. In addition, other popular baseline metrics such as SSIM [4] and MS-SSIM [5] do not show a consistently high correlation with human perceptual quality preferences. Assessing compression quality based on these metrics can lead to sub-par performances. For example, using SSIM as a loss function for deep learning can sometimes guide neural network training in the wrong direction [6]. This leads to low confidence in the results of compression with adjusted bit rates or to the use of a catch-all bit-rate for the encoder to minimize the perceptual quality loss for the vast majority of images [6].

Learning based methods are the logical next step. These methods do show a significant improvement on correlation with human preferences [7, 8]. The LPIPS metric [7] has very promising results correlating with human perceptual preferences in measuring similarity of images that have very different and obvious perturbations. However, since most of these models are developed to address generic quality assessment, they may not perform well outside of their main objective and do not generalize consistently to unseen data.

This creates need for a perceptual quality metric that can be used for image compression. Such a metric should not only correlate well with human perceptual preferences but also generalize well to new datasets, robust to unseen data in the wild. To this aim, we first introduce the largest set of human perceptual preference labels for compressed image pairwise comparisons that is better suited for the creation of deep learning networks for compression image quality assessment (IQA). Since 72.4%of all websites on the internet use images with JPEG standard [9], we focus on this compression format. Our proposed CIQA is $3\times$ larger than previous datasets enabling semantic information to be more general with more examples and allows the model to focus on compression artifacts during training. Secondly, this paper introduces a deep learning full-reference, perceptual quality metric, trained on CIQA for image compression IQA that correlates strongly with human perception and preference. This model not only outperforms current state of the art metrics on the dataset used for training but also generalizes well to SOTA performance with unseen datasets, namely LIVE-Wild [10] and FG-IQA [11].

2. RELATED RESEARCH

In recent years, there has been a resurgence in interest for image compression techniques particularly in IoT and robotics applications [12, 13], particularly with the advance in machine learning, which introduces the capability for personalized or smart compression as a way to increase efficiency while minimizing perceptual quality loss.

2.1. Compression

Storing digital images is subject to an important tradeoff between compression file size (bit-rate), perceptual quality (distortion) [2]. Different domains emphasize the tension between different pairs of these goals.

The most well studied obvious is the bit-rate distortion tradeoff. MSE, PSNR, SSIM, MS-SSIM [5] etc., are the dominant metrics used for measuring distortion resulting from compression. Recent work in this domain advances the search for a metric and method for maintaining a high quality after compression [14, 15]. In [16] discusses the importance of a perceptual quality metric that can be used to train a deep compression model.

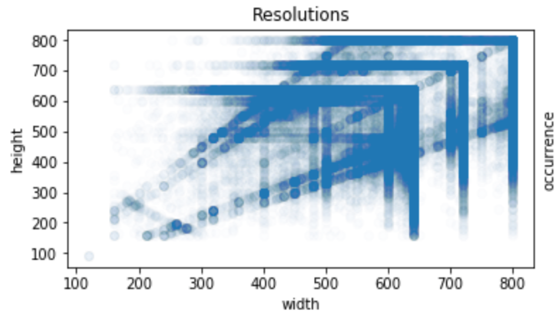

Fig. 2. Resolution of images sampled for the CIQA dataset showing diversity in heights and widths

2.2. Compressed Image Quality Assessment

Image quality assessment broadly has been accelerated by the release of large scale datasets that include image labels of human perceptual ratings. Datasets such as AVA [17], TID2013 [3], KonIQ-10k [18] and WILD [10] have enabled the use of deep learning for this task. Image compression perceptual quality datasets, however, are still somewhat limited in size limiting the use of deep learning due to likely overfitting.

Two prominent datasets [10, 11] and associated analysis have been performed in image quality assessment (IQA) for compressed images. [10] created the LIVE-Wild in which 100 reference images were compressed to different degrees using different encoders and the compressed images were rated using single stimulus ranking. [11] introduces the fine-grained image quality assessment dataset, a much larger-scale image database used for fine grained quality assessment noting the preference of human raters to more subtle changes in compression.

3. METHODS

The current perceptual quality datasets for image compression, FG-IQA and LIVE-Wild [10, 11], are well suited for the evaluation of baseline methods, but fall short when training deep learning models due to the limitations in the number of training examples resulting in over-fitting of the network. Any attempt to create a neural network to address this domain will face the problem of overfitting.

The Aesthetic Visual Analysis dataset (AVA) is well suited for deep learning applied to perceptual IQA proven by its successful implementation in the NIMA NR-IQA model [8] and others [19, 20]. AVA contains $\sim 255,000$ images-rated based on aesthetic qualities by amateur photographers [17], and its size and semantic content distribution may be attributed to its success in deep learning applications. In order to mirror this success in the field of compression perceptual IQA, we introduce the Compression Image Quality Assessment (CIQA) dataset, whose new labels are used to compare quality between images compressed to different Q values

Database	No. Ref Images	No. Distortions	No. Distinct Comparisons	Number of Ratings
LIVE-Wild	29	779	780	single stimulus (20-29 per image)
FGIQA	100	1,200	3600	54,000 (15 per pair)
CIQA	6,667	13,868	7808	249,856 (32 per pair)

Table 1. Summary of current Compression IQA Datsets

from 6,667 unique images were uniformly sampled(6400/267 train/test split) from AVA.

Because these images are stored using JPEG compression, only the subset of images with a JPEG Q factor of 99 or more (essentially uncompressed) were sampled and subsequently compressed at random JPEG quality factors. The original dataset contained semantic labels in addition to perceptual quality labels, and the distribution images from each of the semantic groups was preserved. Semantics labels for each of animal, architecture, cityscape, floral, food/drink, landscape, portrait, and still life each have almost equal occurrence ($\sim 6.25\%$) in all images references and the category of 'generic' has an occurrence of 50%. This was done to ensure the diversity of sampled images and to be a more accurate representation of general perceptual image quality.

Similarly, we analyzed the distribution of resolution from the sampled images which includes ranges from 200×200 to 800×800, not necessarily always of equal height and width. The distribution of image area $(height \times width)$ followed close to a normal distribution with a mean of $\sim 300,000$ pixels and a standard deviation of $\sim 100,000$ pixels.

The aforementioned study was a forced choice pairwise comparison study [21, 22] and consisted of human raters who were presented with two images derived from the same reference, compressed to different degrees, and were asked to select which image has the better perceptual quality. The rater was asked to select an image at random when unsure.

7808 pairwise comparisons were generated and each was rated by 32 individual participants. Pairs were chosen by compressing the reference image to two random quality factors in [10,100]. 13,868 compressed images were generated from the 6,667 reference images, 2 for each image in the train set and 4 for each image in the test set.

4. PERCEPTUAL QUALITY METRIC FOR COMPRESSED IMAGES

As shown in figure 3, We propose a full-reference CNN model architecture that leverages a partial weight matrix pretrained on ImageNet [23]. The base CNN corresponds to the first N layers from one of, EfficientNet-b0 [24] (73 layers), EfficientNet-b7 (157 layers), DenseNet [25] (121 layers), ResNet50-v2 [26] (32 or 44 layers) or VGG16 (7 layers). Both images are fed through a base CNN model to extract features separately up to some some hidden layer l. Both features are then fed through a 1×1 convolution filter with

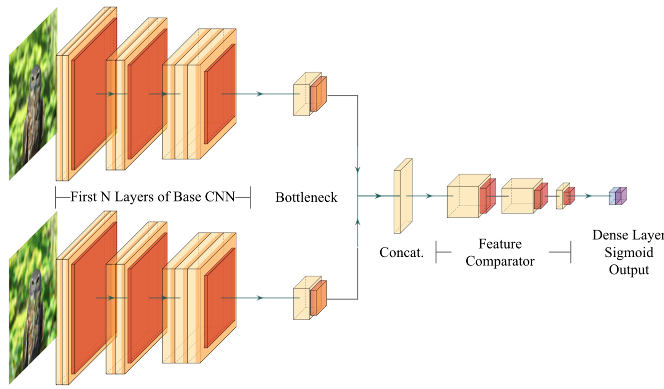

Fig. 3. Architecture used for our proposed deep learning model to predict perceptual IQA preference. The output corresponds to the proportion of responders prefer JPEG2

2 features maps to reduce the dimensional of the original features, namely the bottleneck. Next, we concatenate these features, and feed it into the feature comparator, a two layer 1x1 convolutional layer with 256 hidden feature maps and a 1 feature map output. Finally, we apply a global average pooling, an affine transform, and a sigmoid activation, which represents our final prediction. The output is a number between 0 and 1 corresponding to the predicted preference of JPEG 2. A value of 0.5 corresponds to an indistinguishable change in perceptual quality between the two images, and a prediction approaching 0 or 1 corresponds to the respective image being of higher quality.

The neural network based models this architecture described above are trained on our CIQA labels and then evaluated on FG-IQA and LIVE-Wild datasets. Then the current best methods used pervasively including PSNR, MS-SSIM and Q2StepQA are used on the same evaluation set of the three datasets.

Implementations of PSNR, MS-SSIM and NIQE differ slightly with their implementations in their respective papers as we used the readily available TensorFlow implementation where available. This table includes Q2StepQA which is equal to MS-SSIM multiplied with the negative inverse of NIQE [10].

5. EXPERIMENTAL RESULTS

The proposed CNN architecture was trained on the new augmented CIQA dataset using multiple base models and subsequently evaluated on FG-IQA and LIVE-Wild to test gener-

Models	LIVE-WILD	FG-IQA	CIQA
PSNR	0.51	-0.29	-0.35
MS-SSIM	0.76	0.81	0.44
Q2StepQA	**0.79**	0.53	0.41
LPIPS	0.70	0.55	0.81
Ours	0.76	**0.89**	**0.91**

Table 2. Correlation of popular and best performing baseline models on Major compression IQA datasets. Ours corresponds to the neural network model trained on CIQA and evaluated on all three datasets.

Base CNN	Train correlation	Test correlation
EfficientNet-b7	0.95	0.92
EfficientNet-b0	0.90	0.88
DenseNet121	0.85	0.85
ResNet50-v2	0.87	0.82
VGG16	0.88	0.87

Table 3. Correlation results from various base cnn models

alizability in the wild. The results of the CNN model comparison are presented in Table 3. N layers of the base CNN model were integrated into our architecture. EfficientNet was the best performing CNN by a considerable margin, and the EfficientNet-b0's first 73 layers was chosen due to its performance and lesser complexity when compared to EfficientNet-b7's first 159 layers, also tested.

The two most important criteria for a compression IQA metric are high correlation with human labels (Accurate) and consistent results across a diverse set of unseen examples (Reliable). The baseline models show high variability in correlation between datasets and, in some cases, low correlation with labels. Both these qualities indicate a poor real correlation with human perception and a real opportunity for improvement. Table 2 shows the correlation of our proposed model trained on CIQA and subsequently evaluated on both CIQA, FG-IQA, and LIVE-Wild compared with the other baseline models tested.

Our proposed model outperforms SOTA for the FG-IQA and our the CIQA datasets while remaining competitive when evaluated on the LIVE-Wild dataset achieving similar to MS-SSIM and Q2stepQA.

The LIVE-Wild dataset used single stimulus scores and the correlation is computed from the difference in the mean opinion scores (DMOS) for the images in the pairwise comparisons.

The LIVE-wild images are much more noticeably compressed than FG-IQA and CIQA and the labels have different meaning. The model as well as FGIQA and CIQA have labels that represent which is the higher quality image and the degree to which the distinction is clear. The LIVE-Wild labels use a ranking system such that even if the degradation in one image is clear and there is a clear "higher quality image" the magnitude of of the DMOS score will depend on the number of images with the same reference ranked higher and lower with respect to the two images being compared.

With this knowledge, a full hyper parameter sweep was performed on the model trained on the CIQA labels. Our model's input image size of 400×400 was produced using random cropping with the possibility for zero padding along dimensions smaller than 400 pixels. Training was conducted for 15 epochs of 150 steps each with a batch size of 64 and at a learning rate of 0.001. L_2 regularization, weight decay, and dropout were not successful in increasing generalizability and were not used.

6. DISCUSSION

In this work we used LPIPS, a SOTA perceptual distance metric, and we evaluate it as a perceptual quality metric, which may not be its intended use. However, the variability seen between datasets is noteworthy. This model is based on human perception and its variance across three datasets supports the need for more training data.

The proposed model performs at or above state of the art for the three evaluated datasets. This performance suggests that it may be practical to use this model in a production setting. Although the labels of LIVE-Wild are on a slightly different scale, the performance of the model trained on the new CIQA data fulfils the two requirements maximizing correlation with perceptual preference and reducing variance across unseen data.

Our proposed dataset and model is not necessarily a compression artifact detection algorithm and dataset. The model performs well but not at SOTA when detecting compression artifacts that are difficult for humans to detect. This fact may be a benefit rather than a limitation. Figure 1 shows in its first example how a large Q value difference can result in an indistinguishable perceptual result (~ 0.5) even though the two images differ greatly in file size. This may lead to the potential to use our model as a loss function for the degree of compression. The model will not be able to detect artifacts invisible to the naked eye. The same may be overly penalized when using an artifact detection model. This leads to the potential use of this model in the rate/distortion trade-off.

7. CONCLUSION

The new CIQA dataset is the largest available dataset currently available, opening the possibility for further improvements in deep learning applications to compression IQA. This new augmented database performs significantly better than the current standards when training deep neural networks, leading to less substantial over-fitting and greater generalization. Using this data we were able to train a deep model that outperformed state-of-the-art models across 3 datasets.

8. REFERENCES

[1] Muzhir Al-Ani and Fouad Awad, “The jpeg image compression algorithm,” *Int. Journal of Advances in Engineering And Technology*, vol. 6, pp. 1055–1062, 05 2013.

[2] Xiyang Luo, H. Talebi, Feng Yang, Michael Elad, and P. Milanfar, “The rate-distortion-accuracy tradeoff: Jpeg case study,” *ArXiv*, vol. abs/2008.00605, 2020.

[3] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al., “Image database tid2013: Peculiarities, results and perspectives,” *Sig. proces.: Image comm.*, vol. 30, pp. 57–77, 2015.

[4] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE TIP*, vol. 13, no. 4, pp. 600–612, 2004.

[5] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, “Multiscale structural similarity for image quality assessment,” in *ACSSC*. Ieee, 2003, vol. 2, pp. 1398–1402.

[6] Jim Nilsson and Tomas Akenine-Möller, “Understanding ssim,” *arXiv preprint arXiv:2006.13846*, 2020.

[7] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *CVPR*, 2018, pp. 586–595.

[8] Hossein Talebi and Peyman Milanfar, “Nima: Neural image assessment,” *IEEE TIP*, vol. 27, no. 8, pp. 3998–4011, 2018.

[9] “Usage statistics of jpeg for websites,” https://w3techs.com/technologies/details/im-jpeg, Accessed: 2021-01-25.

[10] L. Cormack H.R. Sheikh, Z.Wang and A.C. Bovik, “Live image quality assessment database release 2,” 2006, Accessed: 2020-07-25.

[11] X. Zhang, W. Lin, S. Wang, J. Liu, S. Ma, and W. Gao, “Fine-grained quality assessment for compressed images,” *IEEE TIP*, vol. 28, no. 3, pp. 1163–1175, 2019.

[12] N Krishnaraj, Mohamed Elhoseny, M Thenmozhi, Mahmoud M Selim, and K Shankar, “Deep learning model for real-time image compression in internet of underwater things (iout),” *J. of Real-Time Image Proc.*, vol. 17, no. 6, pp. 2097–2111, 2020.

[13] J. Uthayakumar, M. Elhoseny, and K. Shankar, “Highly reliable and low-complexity image compression scheme using neighborhood correlation sequence algorithm in wsn,” *IEEE Trans. on Reliability*, vol. 69, no. 4, pp. 1398–1423, 2020.

[14] Siu-Wai Wu and Allen Gersho, “Rate-constrained picture-adaptive quantization for jpeg baseline coders,” in *ICASSP*. IEEE, 1993, vol. 5, pp. 389–392.

[15] Ties van Rozendaal, Guillaume Sautiere, and Taco S. Cohen, “Lossy compression with distortion constrained optimization,” in *CVPR*, June 2020.

[16] Li-Heng Chen, Christos G Bampis, Zhi Li, Andrey Norkin, and Alan C Bovik, “Perceptually optimizing deep image compression,” *arXiv preprint arXiv:2007.02711*, 2020.

[17] Naila Murray, Luca Marchesotti, and Florent Perronnin, “Ava: A large-scale database for aesthetic visual analysis,” in *CVPR*. IEEE, 2012, pp. 2408–2415.

[18] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe, “Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment,” *IEEE TIP*, vol. 29, pp. 4041–4056, 2020.

[19] Chen Kang, Giuseppe Valenzise, and Frédéric Dufaux, “Predicting subjectivity in image aesthetics assessment,” in *MMSP*. IEEE, 2019, pp. 1–6.

[20] Yan-Han Chew, Lai-Kuan Wong, John See, Huai-Qian Khor, and Balasubramanian Abivishaq, “Liteemo: Lightweight deep neural networks for image emotion recognition,” in *MMSP*. IEEE, 2019, pp. 1–6.

[21] Hossein Talebi, Ehsan Amid, Peyman Milanfar, and Manfred K Warmuth, “Rank-smoothed pairwise learning in perceptual quality assessment,” in *ICIP*. IEEE, 2020, pp. 3413–3417.

[22] Rafal Mantiuk, Anna Tomaszewska, and Radoslaw Mantiuk, “Comparison of four subjective methods for image quality assessment,” *Comp. Graph. Forum*, vol. 31, 11 2012.

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *CVPR*. Ieee, 2009, pp. 248–255.

[24] Mingxing Tan and Quoc Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *ICML*. PMLR, 2019, pp. 6105–6114.

[25] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, “Densely connected convolutional networks,” in *CVPR*, 2017, pp. 4700–4708.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks," in *ECCV*. Springer, 2016, pp. 630–645.