DEEP BLIND IMAGE QUALITY ASSESSMENT BY LEARNING SENSITIVITY MAP

Jongyoo Kim, Woojae Kim and Sanghoon Lee*

Yonsei University

ABSTRACT

Applying a deep convolutional neural network CNN to no-reference image quality assessment (NR-IQA) is a challenging task due to the lack of a training database. In this paper, we propose a CNN-based NR-IQA framework that can effectively solve this problem. The proposed method–*the Deep Blind image Quality Assessment predictor (DeepBQA)*–adopts two step training stages to avoid overfitting. In the first stage, a ground-truth objective error map is generated and used as a proxy training target. Then, in the second stage, subjective score is predicted by learning a sensitivity map, which weights each pixel in the predicted objective error map. To compensate the inaccurate prediction of the objective error on the homogeneous regions, we additionally suggest a reliability map. Experiments showed that DeepBQA yields a state-of-the-art correlation with human opinions.

Index Terms— Convolutional neural network, image quality assessment, no-reference image quality assessment.

1. INTRODUCTION

A reliable image quality assessment (IQA) algorithm is needed to quantify the quality of images obtained blindly from the Internet and accurately assess the performance of image processing algorithms. In general, IQA is classified into three categories, depending on whether a reference image (the pristine version of an image) is available: full-reference image quality assessment (FR-IQA), reduced-reference image quality assessment (RR-IQA), and no-reference image quality assessment (NR-IQA). Among them, NR-IQA is the most difficult but the most useful approach, because reference images are not accessible usually. To solve the NR-IQA problem, most of previous methods adopted machine learning techniques, such as support vector machines and neural networks. Research has shown that the accuracy of NR-IQA depends heavily on designing elaborate features.

Recently, convolutional neural networks (CNNs) have become the most popular deep learning model due to their strong representation capability. CNNs have been successfully applied to lots of computer vision problems. However, there is a critical problem that challenges the seamless application of CNNs to NR-IQA [1]. The available training dataset for IQA is insufficient to train deep models. For example, the LIVE Image Quality Assessment Database [2] contains 174 to 233 images according to distortion type, while the most widely used datasets for image recognition contain millions of labeled images [3]. Once can use data augmentation techniques such as rotation, cropping, and horizontal reflection can be used to expand the dataset. However, it is unknown whether any image transformation would alter perceptual quality scores.

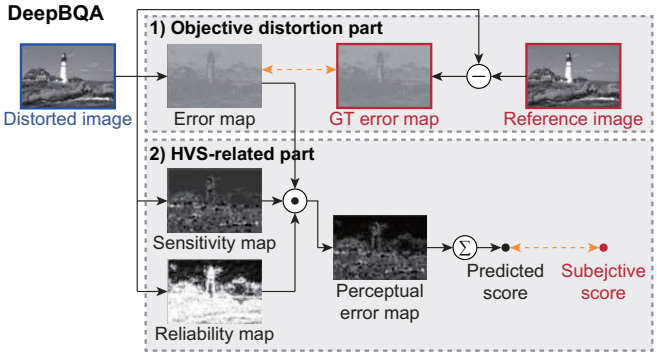


Fig. 1: Conceptual approach of DeepBQA.

We recently proposed a CNN-based FR-IQA method by learning the human visual sensitivity [4]. The proposed model, named DeepQA, seeks the visual weight of each pixel by using a triplet of a distorted image, its objective error map, and its ground-truth subjective score. DeepQA was able to achieve state-of-the-art without overfitting, because the full-reference problem is much easier. In this paper, we propose a novel NR-IQA framework called the *Deep Blind image Quality Assessment predictor* (DeepBQA), which extends thd DeepQA to the no-reference problem. Since the ground-truth objective error map is not available in NR-IQA, DeepBQA also predicts the objective error map using a CNN. The overall diagram of DeepBQA is shown in Fig. 1. In the first stage, DeepBQA is learned to predict the objective error map. Usually, we only have one scalar score for each distorted image. However, since the error map is two-dimensional, it has the same effect of expanding the training dataset by providing more constraints. Therefore, the model

This work is done when Jongyoo Kim is an intern at Microsoft Research Asia

This work was supported by Samsung Research Funding Center of Samsung Electronics under Project Number SRFC-IT1702-08

can be learned without overfitting. Since the objective error is not highly correlated with perceptual quality, we propose an additional CNN path dedicated to learn HVS properties, in the second stage. Based on the predicted objective error map, the model seeks the visual weight of each pixel like [4].

Overall, we attempt to resolve the NR-IQA problem by dividing it into objective distortion and HVS-related parts, as shown in Fig. 1. In the objective distortion part, a pixel-wise objective error map is predicted using the CNN model. In the HVS-related part, the visual sensitivity map that describes the pixel-wise visual importance of the predicted error map is predicted. Note that in the second part, we do not deal with the entire NR-IQA problem, but merely focus on local weight prediction. Since the objective error map is somewhat correlated with the subjective score, the model can be trained successfully by using even a limited dataset.

2. RELATED WORK

Most previously proposed NR-IQA methods were developed based on the machine learning framework [5, 6]. Researchers attempted to design elaborate features that could discriminate distorted images from the pristine images. One popular feature is a family of NSS that assumes that natural scenes contain statistical regularities. Various types of NSS features have been defined in transformation and spatial domains in the literature [7, 8]. Most of these studies were based on conventional machine learning algorithms. Since such models have a limited number of parameters, the size of the dataset was not a significant issue.

Relatively recently, attempts have been made to adopt a deep learning technique for the NR-IQA problem to enhance prediction accuracy. First, deep models were used in place of the conventional regression machine [9, 10, 11]. This involved designing handcrafted features of sufficiently small size such that the neural networks were not sufficiently deep to take full advantage of deep learning. In [12], authors applied a CNN to the NR-IQA problem without handcrafted features. This approach cannot reflect properties of the HVS, since an equal mean opinion score (MOS) was used for all patches in an image. In [13], FR-IQA methods were employed as intermediate training targets of the CNN, and the statistical pooling over minibatch was introduced for end-to-end optimization.

3. PROPOSED METHOD

Following [4], input images are first subtracted from their low-pass filtered images. Let I_r be a reference image and I_d be the corresponding distorted image. The subtracted versions are then denoted by $\hat{I}_r$ and $\hat{I}_d$, respectively. Once an image is normalized, it passes through three paths: 1) objective error map prediction, 2) sensitivity map prediction, and 3) reliability map prediction.

The model consists of two subnetworks. the first subnetwork is trained to predict the objective error map, while the second one learns the sensitivity map. The subnetwokrs uses the same structure: Conv-48, Conv-48 with stride 2, Conv-64, Conv-64 with stride 2, Conv-64, Conv-64, Conv-128, Conv-128 and Conv-1. Here, Conv refers to convolutional layers, and the numbers indicate the number of feature maps. Each layer except for the last has leaky rectified linear units (LReLU) as an activation function, where that of the last layer is a rectified linear unit (ReLU).

3.1. Reliability Map Derivation

When images are severely blurred by distortions, it is difficult to determine whether the blurry region is distorted or not. Furthermore, as severe distortion is applied to an image, its error map includes more high-frequency components. Therefore, the homogeneous regions are not reliable to predict the perceptual quality in NR-IQA.

To avoid this problem, the reliability of the predicted error map is estimated by measuring the texture strength of the distorted image.

$$\mathbf{r} = 2/(1 + exp(-(|\hat{I}_d|))) - 1 \tag{1}$$

The positive half of the sigmoid function is used so that pixels with small values are assigned sufficiently large reliability values. To prevent the reliability map from directly affecting the predicted score, we use a normalized reliability map $\hat{\mathbf{r}}$, where $\mathbf{r}$ is divided by its average.

3.2. Objective Error Map Prediction

The loss function of the first path is defined by the mean squared error between the predicted and ground truth error maps:

$$\mathcal{L}_e(\hat{I}_d, \hat{I}_r; \theta_1) = \left\|(CNN_1(\hat{I}_d; \theta_1) - \mathbf{e}_{gt}) \odot \hat{\mathbf{r}}\right\|_2^2 \tag{2}$$

$$\mathbf{e}_{gt} = |\hat{I}_r - \hat{I}_d|^p \tag{3}$$

where $CNN_1(\cdot)$ is the CNN model in the first subnetwork, θ_1 represents the CNNs parameters. To generate the ground truth error maps $\mathbf{e}_{gt}$, the exponent difference function is used, where p is the exponent number. We chose $p = 0.2$ to spread the distribution of the difference map over the higher values.

3.3. Sensitivity Map Prediction

Because there is no ground truth sensitivity map, the model cannot be trained to directly minimize the pixel-wise difference. Instead, we show a triplet of a distorted image, its objective error map, and its corresponding ground truth subjective score to the deep CNN. Then, the model seeks the optimal weights of the pixels in the error map such that the predicted score approaches the subjective score. The visual sensitivity

map $\mathbf{s}$ is first derived from the CNN model. The perceptual error map $\mathbf{p}$ is then defined by

$$\mathbf{p} = \mathbf{s} \odot \mathbf{e} \odot \hat{\mathbf{r}}. \quad (4)$$

$$\mathbf{s} = CNN_2(\hat{I}_d; \theta_2) \quad (5)$$

where $\odot$ is the Hadamard product, $CNN_2(\cdot)$ indicates the CNN model in the second path with parameter θ_2, and $\mathbf{e}$ is the predicted error map in (2), $\mathbf{e} = CNN_1(\hat{I}_d; \theta_1)$.

Since it cannot be guaranteed that the pooled score from $\mathbf{p}$ has a linear relationship with the ground truth subjective score, we feed the average of $\mathbf{p}$ into the nonlinear regression layers. The objective function of the second path is defined as

$$\mathcal{L}_s(\hat{I}_d; \theta_1, \theta_2) = \|(f(\mu_{\mathbf{p}}) - S)\|_2^2 \quad (6)$$

where $f(\cdot)$ is a nonlinear regression function, $\mu_{\mathbf{p}}$ is the average of $\mathbf{p}$, and S is the ground truth subjective score of the input-distorted image. For the nonlinear regression function, a fully connected neural network with one hidden layer of 4 nodes is used. When the model is optimized to minimize (6) without any constraints, it generates too noisy sensitivity maps, which is not desirable. To avoid this problem, we apply a total variation (TV) $L2$ norm to the sensitivity map to avoid too noisy outputs as proposed in [4].

3.4. Training

We employed the adaptive moment estimation optimizer (ADAM) [14] with Nesterov momentum. The default hyperparameters suggested in the literature [14] were used for ADAM, and the momentum parameter was set to 0.9. The learning rate was set to 5×10^{-4} and 2×10^{-4} for the first and seconds stages, respectively. L2 regularization was applied to all layers (L2 penalty multiplied by 5×10^{-3}).

4. EXPERIEMNT AND ANALYSIS

4.1. Benchmark

Three databases are adopted to evaluate the proposed method: the LIVE IQA database [2], CSIQ [15], and TID2013 [16]. The LIVE IQA database contains 29 reference images and 982 distorted images of five distortion types. The CSIQ database includes 30 reference images and 866 distorted images. TID2013 contains 25 reference images and 3,000 distorted images with 24 different distortions. To train and test DeepBQA, we randomly divided the dataset into two subsets, 80% for training and 20% for testing with respect to reference images. Horizontally flipped images were supplemented to the training set. The training of the error prediction step was iterated for 40 epochs, and the second stage for 60 epochs.

We compared DeepBQA with four FR-IQA methods (PSNR, SSIM [17], FSIMc [18], and DeepQA [4]) and six NR-IQA methods (BLIINDS II [7], BRISQUE [8], IL-NIQE [19], GMLOG [20], and BIECON [13]). Following the recommendation in [21], we evaluated the performance of the IQA algorithms using two metrics: Spearmans rank order correlation coefficient (SRCC) and Pearsons linear correlation coefficient (PLCC).

Table 1: SRCC and PLCC comparison on the 3 databases.

		LIVE IQA		CSIQ		TID2013	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
FR	PSNR	0.876	0.872	0.806	0.800	0.636	0.706
	SSIM	0.948	0.945	0.876	0.861	0.775	0.691
	FSIMc	0.963	0.960	0.931	0.919	0.851	0.877
	DeepQA	0.981	0.982	0.961	0.956	0.939	0.947
NR	BLIINDSII	0.912	0.916	0.780	0.832	0.536	0.628
	BRISQUE	0.939	0.942	0.775	0.817	0.572	0.651
	CORNIA	0.942	0.943	0.714	0.781	0.549	0.613
	IL-NIQE	0.902	0.908	**0.821**	**0.865**	0.521	0.648
	GMLOG	**0.950**	**0.954**	0.803	0.812	**0.675**	**0.683**
	BIECON	**0.958**	**0.962**	**0.825**	**0.838**	**0.721**	**0.765**
	DeepBQA	**0.970**	**0.971**	**0.858**	**0.879**	**0.843**	**0.868**

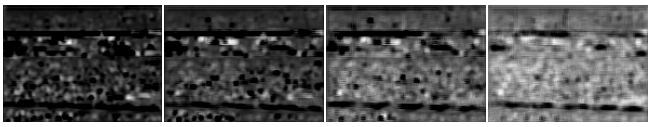
Table 2: SRCC comparison of the models trained using the LIVE IQA database and tested on the TID2013 database.

	JP2K	JPEG	AGN	GB	ALL
PSNR	0.825	0.876	0.918	0.934	0.870
BRISQUE	0.832	0.924	0.829	0.881	0.882
DeepBQA	0.947	0.901	0.827	0.915	0.913

In Table 1, the SRCC and PLCC of the FR- and NR-IQA algorithms compared on the three databases. The correlation coefficients of DeepBQA were averaged after the procedure was repeated 10 times while dividing the training and testing sets randomly. The best three models among the NR-IQA methods for each evaluation criterion are shown in bold. Italics indicate deep learning-based methods. Of the NR-IQA methods, DeepBQA generally achieved the best accuracies in both SRCC and PLCC. On the LIVE IQA database, it is remarkable that even the DeepBQA yielded higher accuracies than previously proposed FR-IQA methods except DeepQA which is a CNN-based model. As a no-reference model, it is difficult to predict the perceptual quality accurately when there is a global change of brightness or contrast in distorted images. In the CSIQ and TID2013 databases, these types of distortions are included, therefore DeepBQA achieved lower accuracies than FSIMc and DeepQA. Compared to previous CNN-based NR-IQA model, BIECON, there were significant increases in the accuracies on all the databases.

4.2. Cross-dataset Test

To evaluate the generalizability of DeepBQA, the model was trained with the LIVE IQA database and tested on the TID2013 database. For testing, four distortion types (JPEG, JP2K, WN, and BLUR) from the TID2013 database were



(a) 10^{-4} (b) 10^{-3} (c) 10^{-2} (d) 10^{-1}

Fig. 2: Examples of predicted sensitivity maps with various TV regularization weights. (a) is distorted image, and (b)-(d) are the predicted sensitivity maps.

Table 3: SRCC and PLCC comparison for each TV regularization weight on the LIVE IQA database.

TV weight	0	10^{-4}	10^{-3}	10^{-2}	10^{-1}
SRCC	0.961	0.967	0.969	0.971	0.969
PLCC	0.963	0.967	0.967	0.972	0.970

Table 4: SRCC and PLCC comparison with and without the reliability map.

Reliability map	LIVE IQA		CSIQ	
	SRCC	PLCC	SRCC	PLCC
w/o	0.969	0.968	0.833	0.859
w/	0.970	0.971	0.858	0.879

selected. The results are shown in Table 2, where DeepBQA achieved a competitive SRCC. It can be concluded that DeepBQA performs well in terms of subjective score prediction, and its performance does not depend on the database.

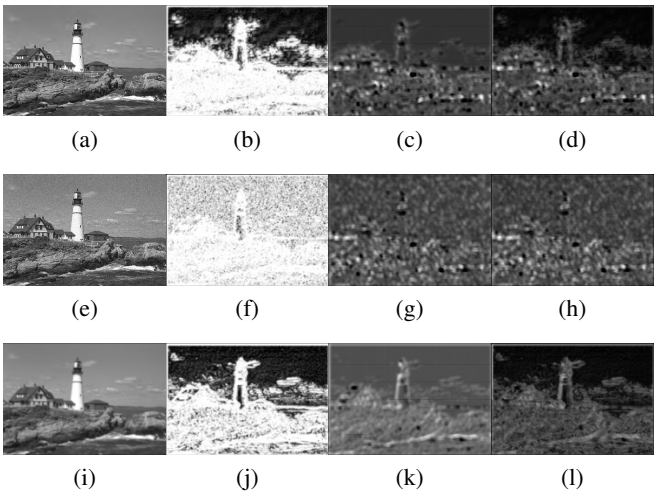
4.3. Ablation Study

We tested four weights (10^{-4}, 10^{-3}, 10^{-2}, and 10^{-1}) of the TV regularization term. Fig. 2 shows the predicted sensitivity maps according to the weight. When the weight was very small (10^{-4}), the overall sensitivity map tended to be zeros, and only small regions had high values. As TV weight increased, the distribution of the sensitivity maps tended to be more uniform, as shown in (e) and (j). This phenomenon also affected the prediction accuracy. The models with sufficient magnitudes of weights (10^{-2} and 10^{-1}) showed higher SRCC, as shown in Table. 3. Too sparse a sensitivity map could not generalize over the various database and distortion types well.

In addition, we study the effect of the reliability map on the LIVE IQA and CSIQ databases. The result is is reported in Table. 4. On the LIVE IQA database, the performance gain by using the reliability map was marginal. However, there was a significant increase in the correlations on CSIQ.

5. PERCEPTUAL ERROR MAP ANALYSIS

The predicted sensitivity and perceptual error maps are shown in Fig. 3. (a), (e), and (i) are distorted by JP2K, white noise, and Gaussian blur, respectively. The reliability maps emphasized high-frequency components, such as edges and complex textures. In case of white noise, the pixels on the whole had a similar reliability as shown in the second column. To analyze DeepBQA, we observed the perceptual error map rather than the sensitivity map. The role of the sensitivity map is tunning the objective error map by multiplication, so that its value does not provide intuitive interpretation. However, it is clear that low values in the perceptual error map can be regarded as perceptually distorted regions. When the image was distorted by white noise, the texture regions were brighter than the homogeneous regions, as shown in (n). For blurred images, textural regions were perceptually more distorted than strong edges, as shown in (l).



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Fig. 3: Examples of predicted sensitivity maps. (a), (e), and (i) are distorted images with JP2K, white noise, and Gaussian blur. (b), (f), and (j) are the reliability maps. (c), (g), and (k) are the predicted sensitivity maps. (d), (h), and (l) are the predicted perceptual error maps.

6. CONCLUSION

In this paper, we described a deep CNN-based NR-IQA framework called DeepBQA. Because of the lack of a training database, applying a deep model to NR-IQA is a challenging issue. In DeepBQA, an objective error map was used as an intermediate regression target to avoid overfitting with the limited database. To reflect the properties of the HVS, an additional path was learned to generate the sensitivity map. As a result, DeepBQA outperformed all benchmark NR-IQA methods. By adding a TV constraint, sensitivity became less sparse and prediction accuracy increased.

7. REFERENCES

[1] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 130–141, Nov. 2017.

[2] H.R. Sheikh, M.F. Sabir, and A.C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[3] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2009, pp. 248–255.

[4] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1676–1684.

[5] Huixuan Tang, N. Joshi, and A. Kapoor, "Learning a blind measure of perceptual image quality," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2011, pp. 305–312.

[6] Peng Ye, J. Kumar, Le Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012, pp. 1098–1105.

[7] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

[8] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[9] W. Hou, X. Gao, D. Tao, and X. Li, "Blind Image Quality Assessment via Deep Learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1275–1286, June 2015.

[10] Yuming Li, Lai-Man Po, Xuyuan Xu, Litong Feng, Fang Yuan, Chun-Ho Cheung, and Kwok-Wai Cheung, "No-reference image quality assessment with Shearlet transform and deep neural networks," *Neurocomputing*, vol. 154, pp. 94–109, Apr. 2015.

[11] Deepti Ghadiyaram and A. C. Bovik, "Feature maps driven no-reference image quality prediction of authentically distorted images," in *Proc. SPIE, Human Vision and Electronic Imaging XX*, 2015, vol. 9394, pp. 93940J–93940J–14.

[12] Le Kang, Peng Ye, Yi Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014, pp. 1733–1740.

[13] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 1, pp. 206–220, Feb. 2017.

[14] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *International Conference for Learning Representations (ICLR)*, 2015.

[15] Eric C. Larson and Damon M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imaging*, vol. 19, no. 1, pp. 19 – 19 – 21, 2010.

[16] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C. C. Jay Kuo, "Image database TID2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57–77, Jan. 2015.

[17] Z. Wang, Alan Conrad Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[18] Lin Zhang, Lei Zhang, Xuanqin Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[19] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.

[20] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and laplacian features," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4850–4862, Nov. 2014.

[21] "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II," *VQEG*, 2003.