SIViP (2018) 12:355–362
https://doi.org/10.1007/s11760-017-1166-8

ORIGINAL PAPER

On the use of deep learning for blind image quality assessment

Simone Bianco[1] · **Luigi Celona**[1] · **Paolo Napoletano**[1] · **Raimondo Schettini**[1]

Received: 10 January 2017 / Revised: 23 June 2017 / Accepted: 19 August 2017 / Published online: 31 August 2017

Abstract In this work, we investigate the use of deep learning for distortion-generic blind image quality assessment. We report on different design choices, ranging from the use of features extracted from pre-trained convolutional neural networks (CNNs) as a generic image description, to the use of features extracted from a CNN fine-tuned for the image quality task. Our best proposal, named DeepBIQ, estimates the image quality by average-pooling the scores predicted on multiple subregions of the original image. Experimental results on the LIVE In the Wild Image Quality Challenge Database show that DeepBIQ outperforms the state-of-the-art methods compared, having a linear correlation coefficient with human subjective scores of almost 0.91. These results are further confirmed also on four benchmark databases of synthetically distorted images: LIVE, CSIQ, TID2008, and TID2013.

Keywords Deep learning · Convolutional neural networks · Transfer learning · Blind image quality assessment · Perceptual image quality

✉ Paolo Napoletano
napoletano@disco.unimib.it

Simone Bianco
bianco@disco.unimib.it

Luigi Celona
celona@disco.unimib.it

Raimondo Schettini
schettini@disco.unimib.it

[1] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Viale Sarca 336, 20126 Milan, Italy

1 Introduction

Digital pictures may have a low perceived visual quality. Capture settings, such as lighting, exposure, aperture, sensitivity to noise, and lens limitations, if not properly handled could cause annoying image artifacts that lead to an unsatisfactory perceived visual quality. Being able to automatically predict the quality of digital pictures can help to handle low-quality images or to correct their quality during the capture process [5]. An automatic image quality assessment (IQA) algorithm, given an input image, tries to predict its perceptual quality. The perceptual quality of an image is usually defined as the mean of the individual ratings of perceived quality assigned by human subjects (mean opinion score—MOS).

In recent years, many IQA approaches have been proposed [29,43]. They can be divided into three groups, depending on the additional information needed: full-reference image quality assessment (FR-IQA) algorithms, e.g., [1,4,9,14,34,48], reduced-reference image quality assessment (RR-IQA) algorithms, and no-reference/blind image quality assessment (NR-IQA) algorithms, e.g., [25,26,28,30,31,33]. FR-IQA algorithms perform a direct comparison between the image under test and a reference or original in a properly defined image space [7]. RR-IQA algorithms are designed to predict image quality with only partial information about the reference image [7]. In their general form, these methods extract a number of features from both the reference and the image under test, and image quality is assessed only by the similarity of these features. NR-IQA algorithms assume that image quality can be determined without a direct comparison between the original and the image under test [7]. Thus, they can be used whenever the original image is unavailable. NR-IQA algorithms can be further classified into two main subgroups: to the first group belong those targeted to estimate the presence of a specific image artifact (i.e., blur,

blocking, grain) [6,8]; to the second group the ones that estimate the overall image quality and thus are distortion generic [5,7,32,40]. In this work, we focus on distortion-generic NR-IQA.

Most of the distortion-generic methods estimate the image quality by measuring deviations from natural scene statistics (NSS) models [5] that capture the statistical "naturalness" of non-distorted images. The natural image quality evaluator (NIQE) [31] is based on the construction of a quality-aware collection of statistical features based on a space domain NSS model. The Distortion Identification-based Image Verity and INtegrity Evaluation (DIIVINE) index [33] is based on a two-stage framework for estimating quality based on NSS models, involving distortion identification and distortion-specific quality assessment. C-DIIVINE [51] is an extension of the DIIVINE algorithm in the complex domain. The BLIINDS-II [39] method, given an input image, computes a set of features and then uses a Bayesian approach to predict quality scores.

The Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) [30] operates in the spatial domain and is also based on a NSS model.

The use of a database of images along with their subjective scores is fundamental for both the design and the evaluation of IQA algorithms [13,41]. Recent approaches to the blind image quality assessment problem use these images coupled with the corresponding human provided quality scores within machine learning frameworks to learn directly from the data a quality measure. The Feature maps based Referenceless Image QUality Evaluation Engine (FRIQUEE) [11,13] combines a deep belief net and a SVM to predict image quality. Tang et al. [45] define a simple radial basis function on the output of a deep belief network to predict the perceived image quality. Hou et al. [15] propose to represent images by NSS features and to train a discriminative deep model to classify the features into five grades (i.e., excellent, good, fair, poor, and bad). Quality pooling is then applied to convert the qualitative labels into scores. In [27] a model is proposed which uses local normalized multiscale difference-of-Gaussian (DoG) response as feature vectors. Then, a three-step framework based on a deep neural network is designed and employed as pooling strategy. Ye et al. [50] presented a supervised filter learning-based algorithm that uses a small set of supervised learned filters and operates directly on raw image patches. Later they extended their work using a shallow convolutional neural network [18]. The same CNN architecture has been then used to simultaneously estimate image quality and identify the distortion type [19] on a single-type distortion dataset [41]. Kottayil et al. [20] used a hybrid approach composed by a shallow CNN architecture followed by a regressor to refine the quality score prediction.

Features extracted from CNN pre-trained for object and scene recognition tasks have been shown to provide image representations that are rich and highly effective for various computer vision tasks. This paper investigates their use for multiple generic distortions NR-IQA and their capability to model the complex dependency between image content and subjective image quality [2,8,46]. The hypothesis motivating our research is that the presence of image distortion, such as JPEG compression, noise, blur, is captured and modeled by these features as well. Furthermore, the more concepts the CNN has been trained to recognize, the better are the extracted features. We evaluate the effect of several design choices:

(i) the use of different features extracted from CNNs that are pre-trained on different image classification tasks for an increasing variety and number of concepts to recognize;
(ii) the use of a number of different image subregions (opposed to the use of the whole image) to better capture image artifacts that may be local or partially masked by specific image content;
(iii) the use of different strategies for feature and score predictions pooling.

We then propose a novel procedure for the fine-tuning of a CNN for multiple generic distortions NR-IQA, which consists in discriminatively fine-tuning the CNN to classify image crops into five distortion classes (i.e., bad, poor, fair, good, and excellent) and then using it as feature extractor. Whatever is the feature extraction strategy and the related CNN, we finally exploit a Support Vector Regression (SVR) machine to learn the mapping function from the CNN features to the perceived quality scores [25].

The experiments are conducted on the *LIVE In the Wild Image Quality Challenge Database* which contains widely diverse authentic image distortions on a large number of images captured using a representative variety of modern mobile devices [12]. The result of this study is a CNN suitably adapted to the blind quality assessment task that accurately predicts the quality of images with a high agreement with respect to human subjective scores. Furthermore, we show the applicability of our method to the legacy LIVE Image Quality Assessment Database [41], CSIQ [22], TID2008 [36] and TID2013 [35].

2 Deep learning for BIQ assessment

Deep convolutional neural networks (CNNs) are a class of learnable architectures used in many image domains [23,37] such as recognition, annotation, retrieval, object detection. CNNs are usually composed of several layers of processing, each involving linear as well as nonlinear operators that are jointly learned in an end-to-end manner to solve a particular task.

Table 1 Architecture of Caffe network

	conv1	*pool1*	*norm1*	*conv2*	*pool2*	*norm2*	*conv3*	*conv4*	*conv5*	*pool5*	*fc6*	*fc7*	*fc8*
Type	Conv	MaxPool	LRN	Conv	MaxPool	LRN	Conv	Conv	Conv	MaxPool	FC	FC	FC
Kernel size	11×11	3×3		5×5	3×3		3×3	3×3	3×3	3×3			
Depth	96			256			384	384	256		4096	4096	
Stride	4	2		1	2		1	1	1	2			
Padding	0			2			1	1	1				

A typical CNN architecture consists of a set of stacked layers: convolutional layers to extract local features; pointwise nonlinear mappings; pooling layers, which aggregates the statistics of the features at nearby locations; and fully connected layers. The result of the last fully connected layer is the CNN output. CNN architectures vary in the number of layers, the number of outputs per layer, the size of the convolutional filters, and the size and type of spatial pooling. CNNs are usually trained in a supervised manner by means of standard back-propagation [24].

In practice, very few people train an entire CNN from scratch, because it is relatively rare to have a dataset of sufficient size. Instead, it is common to take a CNN that is pre-trained on a different large dataset (e.g., ImageNet [38]), and then use it either as a feature extractor or as an initialization for a further learning process (i.e., transfer learning, known also as fine-tuning [3]). Among possible CNN architectures [17,42,44], after preliminary investigations, we have chosen the Caffe network architecture [17] (inspired by the AlexNet [21]) as a feature extractor on top of which we exploit a support vector regression (SVR) machine with a linear kernel to learn a mapping function from the CNN features to the perceived quality scores (i.e., MOS). The detailed architecture of the CNN used is reported in Table 1.

Given an input image, the CNN performs all the multilayered operations, and the corresponding feature vector is obtained by removing the final softmax nonlinearity and the last fully connected layer. The length of the feature vector is 4096.

In this work, we evaluate the effect of several design choices for feature extraction, such as: (1) the use of different CNNs that are pre-trained on different image classification tasks; (2) the use of a number of different image subregions (opposed to the use of the whole image) as well as the use of different strategies for feature and score prediction pooling; (3) the use of a CNN that is fine-tuned for category-based image quality assessment.

2.1 Image description using pre-trained CNNs

Razavian et al. [37] showed that the generic descriptors extracted from convolutional neural networks are very powerful and their use outperforms hand-crafted, state-of-the-art systems in many visual classification tasks. Within the approach previously described, our baseline consists in the use of off-the-shelf CNNs as feature extractors. Features are computed by feeding the CNN with the whole image, that must be resized to fit its predefined input size. We experiment with the use of three CNNs sharing the same architecture that have been pre-trained on three different image classification tasks:

- ImageNet-CNN, which has been trained on 1.2 million images of ImageNet (ILSVRC 2012) for object recognition belonging to 1000 categories;
- Places-CNN, which has been trained on 2.5 million images of the Places Database for scene recognition belonging to 205 categories;
- ImageNet+Places-CNN, which has been trained using 3.5 million images from 1183 categories, obtained by merging the scene categories from Places Database and the object categories from ImageNet.

2.2 Feature and prediction pooling strategies

In the design choice described in Sect. 2.1, we resized the image to match the predefined CNN input size. Since the resizing operation can mask some image artifacts, we consider here a different design choice in which CNN features are computed on multiple subregions (i.e., crops) of the input image. Crops dimensions are chosen to be equal to the CNN input size so that no scaling operation is involved. Each crop covers almost 21% of the original image (227×227 out of 500×500 pixels); thus, the use of multiple crops permits to evaluate the local quality. The final image quality is then computed by pooling the evaluation of each single crop. This permits, for instance, to distinguish between a globally blurred image and a high-quality depth-of-field image.

We experiment the use of a different number randomly selected subregions [21], ranging from 5 to 50. The information coming from the multiple crops has to be fused to predict a single quality score for the whole image. The different fusion strategies are here reported:

– *feature pooling* information fusion is performed element by element on the subregion feature vectors to generate a single feature vector for each image minimum, average, and maximum feature pooling are considered.
– *feature concatenation* information fusion is performed by concatenating the subregion feature vectors in a single longer feature vector.
– *prediction pooling* information fusion is performed on the predicted quality scores. The SVR predicts a quality score for each image crop, and these scores are then fused using a minimum, average, or maximum pooling operators.

Fig. 1 Examples from the LIVE In the Wild IQ Chall. DB

2.3 Image description using a fine-tuned CNN

Convolutional neural networks usually require millions of training samples in order to avoid overfitting. Since in the blind image quality assessment domain the amount of data available is not so large, we investigate the fine-tuning of a pre-trained CNN exploiting the available NR-IQA data. When the amount of data is small, it is likely best to keep some of the earlier layers fixed and only fine-tune some higher-level portion of the network. This procedure, which is also called transfer learning [3], is feasible since the first layers of CNNs learn features similar to Gabor filters and color blobs that appear not to be specific to a particular image domain, while the following layers of CNNs become progressively more specific to the given domain [3].

We start the fine-tuning procedure to the image quality assessment task by substituting the last fully connected layer of a pre-trained CNN with a new one initialized with random values. The new layer is trained from scratch, and the weights of the other layers are updated using the back-propagation algorithm [24] with the available data for image quality assessment. In this work, image quality data are a set of images having human average quality scores (i.e., MOS). The CNN is discriminatively fine-tuned to classify image subregions into five classes according to the FIVE-points MOS scale. The five classes are obtained by a crisp partition of the MOS: bad (score $\in$ [0, 20]), poor (score $\in$]20, 40]), fair (score $\in$]40, 60]), good (score $\in$]60, 80]), and excellent (score $\in$]80, 100]). Once the CNN is trained, it is used for feature extraction, just like one of the pre-trained CNNs.

3 Experimental results

Different standard databases are available to test the algorithms performance with respect to the human subjective judgments. Most of them have been created starting from high-quality images and adding synthetic distortions. However, as pointed out by Ghadiyaram and Bovik [12]: "images captured using typical real-world mobile camera devices are usually afflicted by complex mixtures of multiple distortions, which are not necessarily well modeled by the synthetic distortions found in existing databases."

We evaluate the different design choices within the proposed approach on the LIVE In the Wild IQ Challenge DB [12,13]. It contains 1162 images with resolution equal to 500×500 pixels affected by diverse authentic distortions and genuine artifacts such as low-light noise and blur, motion-induced blur, over and underexposure, compression errors. Figure 1 shows some database samples. Database images have been rated by many thousands of subjects via an online crowdsourcing system designed for subjective quality assessment. Over 350,000 opinion scores from over 8100 unique human observers have been gathered. The mean opinion score (MOS) of each image is computed by averaging the individual ratings across subjects and used as ground truth quality score. The MOS values are in the [1, 100] range.

We compared the different design choices within the proposed approach with a number of leading blind IQA algorithms. Since most of these algorithms are machine learning-based training procedures, following [13] in all the experiments, we randomly split the data into 80% training and 20% testing sets, using the training data to learn the model and validating its performance on the test data. To mitigate any bias due to the division of data, the random split of the dataset is repeated 10 times. For each repetition, we compute the Pearson's linear correlation coefficient (LCC), the Spearman's rank-ordered correlation coefficient (SROCC), and the normalized mean absolute error (nMAE) between the predicted and the ground truth quality scores, reporting the median of these measures across the 10 splits. The nMAE is obtained by normalizing the MAE with respect to the upper limit of the MOS range in order to make it more easy to be compared across datasets.

In all the experiments, we use the Caffe open-source framework [17] for CNN training and feature extraction and the LIBLINEAR library [10] for SVR training.

Experiment I, pre-trained CNNs We extract the 4096-dimensional features from the *fc7* layer of the pre-trained ImageNet-CNN, Places-CNN, and ImageNet+Places-CNN. Since these CNNs require an input with a dimensionality equal to 227×227 pixels, we rescale the original 500×500

Table 2 Median LCC, SROCC, and nMAE across 10 train–test random splits of the LIVE In the Wild IQ Challenge Database considering only the central crop of the subsampled image as input for the pre-trained CNNs considered

	LCC	SROCC	nMAE
Imagenet-CNN	0.6782	0.6381	0.12
Places-CNN	0.6267	0.6055	0.12
ImageNet+Places-CNN	**0.7215**	**0.7021**	**0.11**

The best method is reported in bold

Table 3 Median LCC, SROCC, and nMAE across 10 train–test random splits of the LIVE In the Wild IQ Challenge DB considering randomly selected crops as input for the ImageNet+Places-CNN and three different fusion approaches

	LCC	SROCC	nMAE
Feature pool. (avg@30crops)	**0.7938**	**0.7828**	**0.09**
Feature concat. (@35crops)	0.7864	0.7724	0.10
Prediction pool.(avg@20crops)	0.7873	0.7685	0.10

The best method is reported in bold

images to 256 × 256 keeping aspect ratio, and then we crop out the central 227 × 227 subregion from the resulting image. All the images are pre-processed by subtracting the mean image, that is computed by averaging all the images in the training set on which the CNN was pre-trained. The median LCC, SROCC, and nMAE over the 10 train–test splits are reported in Table 2. From the results, it is possible to see that ImageNet+Places-CNN outperforms both Imagenet-CNN and Places-CNN, with Places-CNN giving the worst performance confirming our original hypothesis that the more concept the CNN has been trained to recognize, the more effective are its features for modeling generic image content.

Experiment II, feature, and prediction pooling In the previous experiment the resize operation could have reduced the effect of some artifacts, e.g., noise. In order to keep unchanged the distortion level, we evaluate the performances of features extracted from a variable number of randomly cropped 227 × 227 subregions from the original image. This choice is confirmed in preliminary experiments (not reported here due to lack of space) where taking crops at different scales demonstrated to perform worse than taking them at the original image scale.

Given the results in the previous section, the only features considered here are those extracted using the ImageNet+Places-CNN. We evaluate three different fusion schemes for combining the information generated by the multiple subregions to obtain a single score prediction for the whole image. The first scheme is feature pooling that can be seen as an early fusion approach, performing element-wise fusion on the feature vectors. The second scheme is feature concatenation, performing information fusion by concatenating the multiple feature vectors into a single feature vector. The third scheme is prediction pooling that can be seen as a late fusion approach, where information fusion is performed on the predicted quality scores.

In all the experiments, the number of random crops is varied between 5 and 50 in steps of 5. The values of LCC, SROCC, and nMAE for the best configurations of each fusion scheme (across pooling operators and number of crops) are reported in Table 3. The optimal number of crops has been selected by running the two-sample t test whose results are reported as additional material. Concerning the best configurations reported in Table 3, the output of the two-sample t test shows that the results obtained by feature average-pooling are statistically better than both those obtained by feature concatenation (p value equal to 3.4×10^{-9}) and prediction average-pooling (p value equal to 8.8×10^{-5}). The difference between feature concatenation and prediction average-pooling is not significative instead (p value equal to 0.23).

Experiment III, fine-tuned CNN In all previous experiments, we use pre-trained CNNs for feature extraction. In this experiment instead, we fine-tune the ImageNet+Places-CNN for the NR-IQA task. The CNN is discriminatively fine-tuned to classify image crops into five distortion classes (i.e., bad, poor, fair, good, and excellent) obtained by crisp partitioning the MOS into five disjoint sets. Since the number of images belonging to the five sets is uneven , during training we use a sample weighting approach [16] giving larger weights to images belonging to less represented distortion classes [52]. Weights are computed as the ratio between the frequency of the most represented class and the frequency of the class to which the image belongs.

On the NR-IQA task, this weighting scheme gives better results compared to batch-balancing (i.e., assuring that in each batch all the classes are evenly sampled) since it guarantees more heterogeneous batches.

Given the results of the previous experiments, we only evaluate the performance of the fine-tuned CNN with feature pooling and prediction pooling with the average operator. We fine-tune the network for 5000 iterations using Caffe framework [17] on a NVIDIA K80 GPU. The total training time was about 2 h, while predicting the MOS for a single image at test time requires about 20 ms.

The numerical values of LCC, SROCC, and nMAE for the best configurations are reported in Table 4. As for the previous experiment, the optimal number of crops has been selected by running the two-sample t test whose results are reported as additional material. Concerning the best configurations reported in Table 4, the output of the two-sample t test shows that the results obtained by prediction average-pooling are statistically better than those obtained by feature average-pooling (p value equal to 4.7×10^{-4}).

Table 4 Median LCC, SROCC, and nMAE across 10 train–test random splits of the LIVE In the Wild IQ Challenge Database considering randomly selected crops as input for the fine-tuned CNN and two different fusion approaches

	LCC	SROCC	nMAE
Feature pool. (avg@20crops)	0.9026	0.8851	**0.06**
Prediction pool. (avg@25crops)	**0.9082**	**0.8894**	**0.06**

The best method is reported in bold

Table 5 Median LCC, SROCC, and nMAE across 10 train–test random splits of the LIVE In the Wild IQ Chall. DB

	Sub-reg	LCC	SROCC	nMAE
DIIVINE [33]		0.57	0.52	0.14
BRISQUE [30]		0.61	0.61	0.13
BLIINDS-II [39]	✓	0.46	0.41	0.16
S3 index [47]		0.32	0.30	0.18
NIQE [31]		0.48	0.43	0.15
C-DIIVINE [51]	✓	0.67	0.65	0.12
FRIQUEE [11,13]		0.71	0.68	0.12
HOSA [49]	✓	0.74	0.72	0.10
DeepBIQ (Exp. I)		0.72	0.70	0.11
DeepBIQ (Exp. II)	✓	0.79	0.79	0.09
DeepBIQ (Exp. III)	✓	**0.91**	**0.89**	**0.06**

The best method is reported in bold

In Table 5 we compare the results of the different instances of the proposed approach, that we name DeepBIQ, with those of some NR-IQA algorithms in the state of the art. Each method is tested on both the original image and with the sub-region approaches proposed in Sect. 2.2, and the best result among the two is reported. The column *Sub-reg* indicates which version obtained the best result. From the results it is possible to see that the use of a pre-trained CNN on the whole image is able to give slightly better results than the best in the state of the art. The use of multiple crops with average-pooled features is able to improve LCC and SROCC with respect to the best method in the state of the art by 0.08 and 0.11, respectively. Finally, the use of the fine-tuned CNN with multiple image crops and average-pooled predictions is able to improve LCC and SROCC by 0.20 and 0.21 respectively. Moreover, this last solution is also able to almost halve the nMAE with respect to the best solution in the state of the art lowering it from 0.10 to 0.06.

Error statistics may not give an intuitive idea of how well a NR-IQA algorithm performs. On the other hand, individual human scores can be rather noisy. Taking into account that the LIVE In the Wild Image Quality Challenge Database gives for each image the MOS as well as the standard deviation of the human subjective scores, to have an intuitive assessment of DeepBIQ performance we proceed as follows: We divide the absolute prediction error of each image by the standard deviation of the subjective scores for that particular image. We then build a cumulative histogram and collect statistics at one, two, and three standard deviations. Results indicate that 97.2% of our predictions are below σ, 99.4% below 2σ and 99.8% below 3σ. Assuming a normal error distribution, this means that in most of the cases the image quality predictions made by DeepBIQ are closer to the average observer than those of a generic human observer.

Table 6 Median LCC, SROCC, and nMAE across 100 random splits of the legacy LIVE Image Quality Assessment DB

Method	Sub-reg	LCC	SROCC	nMAE
DIIVINE [33]		0.93	0.92	0.10
BRISQUE [30]		0.94	0.94	0.08
BLIINDS-II [39]	✓	0.92	0.91	0.11
NIQE [31]		0.92	0.91	0.10
C-DIIVINE [51]	✓	0.95	0.95	0.07
FRIQUEE [11,13]	✓	0.95	0.94	0.07
ShearletIQM [28]		0.94	0.93	0.10
MGMSD [1]		0.97	**0.97**	**0.05**
Low Level Feat. [20]		0.95	0.94	0.08
Rectifier NN [45]		0.95	0.96	0.06
Multitask CNN [19]		0.95	0.95	0.07
Shallow CNN [18]	✓	0.95	0.96	0.07
DLIQA [15]		0.93	0.93	0.09
HOSA [49]	✓	0.95	0.95	0.07
CNN-Prewitt [26]	✓	0.97	0.96	**0.05**
CNN-SVR [25]	✓	0.97	0.96	0.06
DeepBIQ	✓	**0.98**	**0.97**	**0.05**

The best method is reported in bold

For sake of comparison with other methods in the state of the art, as an additional experiment, we evaluate our method on the older but widely used benchmark databases of synthetically distorted images: LIVE Image Quality Assessment Database [41], Categorical Subjective Image Quality (CSIQ) Database [22], TID2008 [36], TID2013 [35].

We evaluate our method on these datasets dealing with the different human judgments and distortion ranges by only re-training the SVR, while keeping the CNN unchanged. We follow the experimental protocol used in [18,19]. This protocol consists in running 100 iterations, where in each iteration 60% of the reference images and their distorted versions is randomly select as the training set, 20% as the validation set, and the remaining 20% as the test set. The experimental results in terms of average LCC and SROCC values on LIVE are reported in Table 6, on CSIQ in Table 7, on TID2008 in Table 8, and on TID2013 in Table 9. From these results, it is possible to see that our method, DeepBIQ, is able to obtain the best performance in terms of LCC, SROCC, and nMAE notwithstanding that differently from all the other methods reported, the features have been learned on a different dataset

Table 7 Median LCC, SROCC, and nMAE across 100 train–val–test random splits of the CSIQ

Method	Sub-reg	LCC	SROCC	nMAE
DIIVINE [33]		0.90	0.88	0.16
BRISQUE [30]		0.93	0.91	0.12
BLIINDS-II [39]	✓	0.93	0.91	0.11
Low Level Feat. [20]		0.94	0.94	0.09
Multitask CNN [19]		0.93	0.94	0.09
HOSA [49]	✓	0.97	0.96	**0.05**
DeepBIQ	✓	**0.97**	**0.96**	0.06

The best method is reported in bold

Table 8 Median LCC, SROCC, and nMAE across 100 train–val–test random splits of the TID2008

Method	Sub-reg	LCC	SROCC	nMAE
DIIVINE [33]		0.90	0.88	0.09
BRISQUE [30]		0.93	0.91	0.04
BLIINDS-II [39]	✓	0.92	0.90	0.05
MGMSD [1]		0.88	0.89	0.06
Low Level Feat. [20]		0.89	0.88	0.08
Multitask CNN [19]		0.90	0.91	0.04
Shallow CNN [18]	✓	0.90	0.92	0.05
DeepBIQ	✓	**0.95**	**0.95**	**0.03**

The best method is reported in bold

Table 9 Median LCC, SROCC, and nMAE across 100 train–val–test random splits of the TID2013

Method	Sub-reg	LCC	SROCC	nMAE
DIIVINE [33]		0.89	0.88	0.07
BRISQUE [30]		0.92	0.89	0.06
BLIINDS-II [39]	✓	0.91	0.88	0.06
Low Level Feat. [20]		0.89	0.88	0.07
HOSA [49]	✓	**0.96**	0.95	**0.03**
DeepBIQ	✓	**0.96**	**0.96**	**0.03**

The best method is reported in bold

containing images with real distortions and not on a portion of the test database itself. Therefore, the results confirm the effectiveness of our approach for no-reference IQ assessment.

4 Conclusions

In this work, we have investigated the use of deep learning for distortion-generic blind image quality assessment. We report on different design choices in three different experiments, ranging from the use of features extracted from pre-trained convolutional neural networks (CNNs) as a generic image description, to the use of features extracted from a CNN fine-tuned for the image quality task.

Our best proposal, named DeepBIQ, consists of a CNN originally trained to discriminate 1,183 visual categories that is fine-tuned for category-based image quality assessment. This CNN is then used to extract features that are then fed to a SVR to predict the image quality score. By considering multiple image crops and exploiting the prediction pooling fusion scheme with the average operator, DeepBIQ reaches a LCC of almost 0.91, that is 0.20 higher than the best solution in the state of the art [13]. Furthermore, in many cases, the quality score predictions of our method are closer to the average observer than those of a generic human observer.

DeepBIQ is then further tested on four benchmark databases of synthetically distorted images: LIVE, CSIQ, TID2008, and TID2013. To deal with the different types of human opinion scores and distortion ranges, we only re-trained the SVR, while keeping the CNN unchanged. Experimental results show that DeepBIQ is able to outperform all the methods in the state of the art also on these datasets, even if the features have been learned on a different dataset, confirming the effectiveness of our approach for no-reference image quality assessment.

A Web demo of the DeepBIQ network and additional materials are available at http://www.ivl.disco.unimib.it/activities/deep-image-quality/.

References

1. Alaei, A., Raveaux, R., Conte, D.: Image quality assessment based on regions of interest. Signal Image Video Process. **11**(4), 673–680 (2017)
2. Allen, E., Triantaphillidou, S., Jacobson, R.: Image quality comparison between jpeg and jpeg2000. I. Psychophysical investigation. J. Imaging Sci. Technol. **51**(3), 248–258 (2007)
3. Bengio, Y.: Deep learning of representations for unsupervised and transfer learning. In: Unsupervised and Transfer Learning Challenges in Mach. Learn. vol. 7, p. 19 (2012)
4. Bianco, S., Ciocca, G., Marini, F., Schettini, R.: Image quality assessment by preprocessing and full reference model combination. In: IS&T/SPIE Electronic Imaging, pp. 72,420O (2009)
5. Bovik, A.C.: Automatic prediction of perceptual image and video quality. Proc. IEEE **101**(9), 2008–2024 (2013)
6. Ciancio, A., Da Costa, A.L.N.T., da Silva, E.A., Said, A., Samadani, R., Obrador, P.: No-reference blur assessment of digital pictures based on multifeature classifiers. IEEE Trans. Image Process. **20**(1), 64–75 (2011)
7. Ciocca, G., Corchs, S., Gasparini, F., Schettini, R.: How to assess image quality within a workflow chain: an overview. Int. J. Digit. Libr. **15**(1), 1–25 (2014)
8. Corchs, S., Gasparini, F., Schettini, R.: No reference image quality classification for jpeg-distorted images. Digit. Signal Process. **30**, 86–100 (2014)
9. Eckert, M.P., Bradley, A.P.: Perceptual quality metrics applied to still image compression. Signal Process. **70**(3), 177–200 (1998)
10. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: a library for large linear classification. J. Mach. Learn. Res. **9**, 1871–1874 (2008)

11. Ghadiyaram, D., Bovik, A.C.: Blind image quality assessment on real distorted images using deep belief nets. In: Global Conference on Signal and Information Processing (GlobalSIP), pp. 946–950. IEEE (2014)
12. Ghadiyaram, D., Bovik, A.C.: Crowdsourced study of subjective image quality. In: Asilomar Conference on Signals, Systems and Computers (2014)
13. Ghadiyaram, D., Bovik, A.C.: Massive online crowdsourced study of subjective and objective picture quality. IEEE Trans. Image Process. **25**(1), 372–387 (2016)
14. He, L., Gao, X., Lu, W., Li, X., Tao, D.: Image quality assessment based on S-CIELAB model. Signal Image Video Process. **5**(3), 283–290 (2011)
15. Hou, W., Gao, X., Tao, D., Li, X.: Blind image quality assessment via deep learning. IEEE Trans. Neural Netw Learn. Syst. **26**(6), 1275–1286 (2015)
16. Huang, Y.M., Du, S.X.: Weighted support vector machine for classification with uneven training class sizes. In: 2005 International Conference on Mach. Learn. and Cybernetics, vol. 7, pp. 4365–4369. IEEE (2005)
17. Jia, Y., et al.: Caffe: convolutional architecture for fast feature embedding. In: ACM MM, pp. 675–678. ACM (2014)
18. Kang, L., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for no-reference image quality assessment. In: CVPR, pp. 1733–1740 (2014)
19. Kang, L., Ye, P., Li, Y., Doermann, D.: Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In: ICIP, pp. 2791–2795. IEEE (2015)
20. Kottayil, N.K., Cheng, I., Dufaux, F., Basu, A.: A color intensity invariant low-level feature optimization framework for image quality assessment. Signal Image Video Process. **10**(6), 1169–1176 (2016)
21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS, pp. 1097–1105 (2012)
22. Larson, E.C., Chandler, D.M.: Most apparent distortion: full-reference image quality assessment and the role of strategy. JEI **19**(1), 011,006 (2010)
23. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
24. LeCun, Y.A., Bottou, L., Orr, G.B., Müller, K-R.: Efficient Back Prop. In: Montavon, G., Orr, G.B., Müller, K-R. (eds) Neural Networks: Tricks of the Trade, 2nd edn, pp. 9–48, Springer, Berlin, Heidelberg (2012)
25. Li, J., Yan, J., Deng, D., Shi, W., Deng, S.: No-reference image quality assessment based on hybrid model. Signal Image Video Process. **11**(6), 985–992 (2017)
26. Li, J., Zou, L., Yan, J., Deng, D., Qu, T., Xie, G.: No-reference image quality assessment using Prewitt magnitude based on convolutional neural networks. Signal Image Video Process. **10**(4), 609–616 (2016)
27. Lv, Y., Jiang, G., Yu, M., Xu, H., Shao, F., Liu, S.: Difference of Gaussian statistical features based blind image quality assessment: a deep learning approach. In: ICIP, pp. 2344–2348. IEEE (2015)
28. Mahmoudpour, S., Kim, M.: No-reference image quality assessment in complex-shearlet domain. Signal Image Video Process. **10**(8), 1465–1472 (2016)
29. Manap, R.A., Shao, L.: Non-distortion-specific no-reference image quality assessment: a survey. Inf. Sci. **301**, 141–160 (2015)
30. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Trans Image Process. **21**(12), 4695–4708 (2012)
31. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a completely blind image quality analyzer. SPL **20**(3), 209–212 (2013)
32. Mittal, A., Moorthy, A.K., Bovik, A.C., Chen, C.W., Chatzimisios, P., Dagiuklas, T., Atzori, L.: No-reference approaches to image and video quality assessment. In: Multimedia Quality of Experience (QoE): Current Status and Future Requirements, vol. 99. Wiley (2015)
33. Moorthy, A.K., Bovik, A.C.: Blind image quality assessment: from natural scene statistics to perceptual quality. IEEE Trans Image Process. **20**(12), 3350–3364 (2011)
34. Pappas, T.N., Safranek, R.J., Chen, J.: Perceptual criteria for image quality evaluation. In: Handbook of Image and Video Processing, pp. 669–684 (2000)
35. Ponomarenko, N., Ieremeiev, O., Lukin, V., Egiazarian, K., Jin, L., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., et al.: Color image database tid2013: peculiarities and preliminary results. In: Visual Information Processing (EUVIP), 2013 4th European Workshop on, pp. 106–111. IEEE (2013)
36. Ponomarenko, N., et al.: Tid 2008-a database for evaluation of full-reference visual quality assessment metrics. Adv. Mod. Radioelectron. **10**(4), 30–45 (2009)
37. Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. In: CVPR Workshops, pp. 806–813 (2014)
38. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet Large Scale Visual Recognition Challenge. Int J Comput Vis. **115**(3), 211–252 (2015)
39. Saad, M.A., Bovik, A.C., Charrier, C.: Blind image quality assessment: a natural scene statistics approach in the DCT domain. Trans. Image Process. **21**(8), 3339–3352 (2012)
40. Seshadrinathan, K., Bovik, A.C.: Automatic prediction of perceptual quality of multimedia signals—a survey. Multimed. Tools Appl. **51**(1), 163–186 (2011)
41. Sheikh, H.R., Wang, Z., Cormack, L., Bovik, A.C.: Live image quality assessment database release 2 (2005). http://live.ece.utexas.edu/research/quality/subjective.htm. Accessed 29 Aug 2017
42. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
43. Soundararajan, R., Bovik, A.C.: Survey of information theory in visual quality assessment. Signal Image Video Process. **7**(3), 391–401 (2013)
44. Szegedy, C., et al.: Going deeper with convolutions. In: IEEE CVPR, pp. 1–9 (2015)
45. Tang, H., Joshi, N., Kapoor, A.: Blind image quality assessment using semi-supervised rectifier networks. In: CVPR, pp. 2877–2884 (2014)
46. Triantaphillidou, S., Allen, E., Jacobson, R.: Image quality comparison between jpeg and jpeg2000. II. Scene dependency, scene analysis, and classification. J. Imaging Sci. Technol. **51**(3), 259–270 (2007)
47. Vu, C.T., Phan, T.D., Chandler, D.M.: S3: a spectral and spatial measure of local perceived sharpness in natural images. Trans. Image Process. **21**(3), 934–945 (2012)
48. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
49. Xu, J., Ye, P., Li, Q., Du, H., Liu, Y., Doermann, D.: Blind image quality assessment based on high order statistics aggregation. IEEE Trans. Image Process. **25**(9), 4444–4457 (2016)
50. Ye, P., Kumar, J., Kang, L., Doermann, D.: Real-time no-reference image quality assessment based on filter learning. In: CVPR, pp. 987–994 (2013)
51. Zhang, Y., Moorthy, A.K., Chandler, D.M., Bovik, A.C.: C-DIIVINE: no-reference image quality assessment based on local magnitude and phase statistics of natural scenes. Signal Process. Image Commun. **29**(7), 725–747 (2014)
52. Zhou, Z.H., Liu, X.Y.: Training cost-sensitive neural networks with methods addressing the class imbalance problem. Trans. Knowl. Data Eng. **18**(1), 63–77 (2006)