On the Use of Deep Learning for Blind Image Quality Assessment

S. Bianco, L. Celona, P. Napoletano, R. Schettini

[a]*DISCo, Department of Informatics, Systems and Communication, University of Milano-Bicocca, Viale Sarca 336, Milan 20126, Italy*
(e-mail: {bianco,celona,napoletano,schettini}@disco.unimib.it)

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 sub-regions of the original image. The score of each sub-region is computed using a Support Vector Regression (SVR) machine taking as input features extracted using a CNN fine-tuned for category-based image quality assessment. Experimental results on the LIVE In the Wild Image Quality Challenge Database and on the LIVE Image Quality Assessment Database show that DeepBIQ outperforms the state-of-the-art methods compared, having a Linear Correlation Coefficient (LCC) with human subjective scores of almost 0.91 and 0.98 respectively. Furthermore, in most of the cases, the quality score predictions of DeepBIQ are closer to the average observer than those of a generic human observer.

Keywords: Deep learning, Convolutional neural networks, Transfer learning, Blind image quality assessment, Perceptual image quality.

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 [1]. 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. They can be divided into three groups, depending on the additional information needed: full-reference image quality assessment (FR-IQA) algorithms e.g. [2, 3, 4, 5, 6, 7], reduced-reference image quality assessment (RR-IQA) algorithms, and no-reference/blind image quality assessment (NR-IQA) algorithms e.g. [8, 9, 10]. FR-IQA algorithms perform a direct comparison between the image under test and a reference or original in a properly defined image space [11]. Having access to an original is a requirement of the usability of such metrics. RR-IQA algorithms are designed to predict image quality with only partial information about the reference image [11]. 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 [11]. Thus, they can be used whenever the original image is unavailable. NR-IQA algorithms can be further classified into two main sub-groups: to the first group belong those targeted to estimate the presence of a specific image artifact (i.e. blur, blocking, grain, etc.) [12, 13]; to the second group the ones that estimate the overall image quality and thus are distortion generic [14, 15, 1, 11]. 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 Statistic (NSS) models [1] that capture the statistical "naturalness" of non-distorted images. These models are based on the two following principles: i) good quality real-world photographic images obey certain perceptually relevant statistical laws; ii) common image distortions alter such statistical laws. The Natural Image Quality Evaluator (NIQE) [10] 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 [8] is based on a two-stage framework for estimating quality based on NSS models, involving distortion identification and distortion-specific quality assessment. The core of the method uses a Gaussian scale mixture to model neighboring wavelet coefficients. C-DIIVINE [16] is an extension of the DIIVINE algorithm in the complex domain, and blindly assesses image quality based on the complex Gaussian scale mixture model corresponding to the complex version of the steerable pyramid wavelet transform. The BLIINDS-II [17] method, given an input image, computes a set of features and then uses a Bayesian approach to predict quality scores. Such features are obtained

by transforming the model parameters of a generalized NSS-based model of local Discrete Cosine Transform coefficients into a vector of features.

The Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) [9] operates in the spatial domain and is also based on a NSS model. The algorithm quantifies possible losses of naturalness in the image due to the presence of distortions.

The use of a database of images along with their subjective scores is fundamental for both the design and the evaluation of IQA algorithms [18, 19]. 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) [20, 19] combines a deep belief net and a SVM to predict image quality. Tang et al. [21] define a simple radial basis function on the output of a deep belief network to predict the perceived image quality. They first pre-train the network in an unsupervised manner and then fine-tune it with labeled data. Finally they model the quality of images exploiting a Gaussian Process regression. Hou et al. [22] 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 [23] a model is proposed which uses local normalized multi-scale difference of Gaussian (DoG) response as feature vectors. Then, a three-steps framework based deep neural network is designed and employed as pooling strategy. Ye et al. [24] 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 [25]. The same CNN architecture has been then used to simultaneously estimate image quality and identify the distortion type [26] on a single-type distortion dataset [18].

This paper investigates the use of *deep learning* for multiple generic distortions NR-IQA. More precisely, we use CNN as feature extractor on top of which we exploit a Support Vector Regression (SVR) machine [27, 28] to learn the mapping function from the image features to the perceived quality scores. We evaluate the effect of several design choices: i) the use of different CNNs that are pre-trained on different image classification tasks; ii) the use of a number of different image sub-regions (opposed to the use of the whole image) as well as the use of different strategies for feature and score predictions pooling; iii) the use of a CNN that is fine-tuned for image quality assessment.

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 [29]. 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 [18].

The rest of the paper is organized as follows: Section II introduces the proposed approach and the different design choices considered; Section III describes the data, evaluation metrics, and the experimental setup. Sections IV and V analyze the experimental results. Finally, Section VI presents our final considerations.

2. Deep Learning for blind image quality assessment

Deep Convolutional Neural Networks (CNNs) are a class of learnable architectures used in many image domains [30, 31] such as recognition, annotation, retrieval, object detection, etc. CNNs are usually composed of several layers of processing, each involving linear as well as non-linear operators that are jointly learned in an end-to-end manner to solve a particular task.

A typical CNN architecture consists of a set of stacked layers: convolutional layers to extract local features; point-wise non-linear 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 [32].

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 [33]), 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 [34, 35]). In this work, we use the Caffe network architecture [36] (inspired by the AlexNet [37]) as a feature extractor on top of which we exploit a Support Vector Regression (SVR) machine [27, 28] 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. A graphical representation of the described approach is reported in Figure 1.

In this work we evaluate the effect of several design choices for feature extraction, such as: i) the use of different CNNs that are pre-trained on different image classification tasks; ii) the use of a number of different image sub-regions (opposed to the use of the whole image) as well as the use of different strategies for feature and score prediction pooling; iii) 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. [31] showed that the generic descriptors extracted from convolutional neural networks are very powerful

Table 1: Architecture of Caffe network. It consists in 8 weight layers. The ReLU activation layers after each weight layer (except for *fc8*) are not shown for brevity. FC denotes fully connected layer type, while LRN represents the Local Response Normalization layer type.

	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				

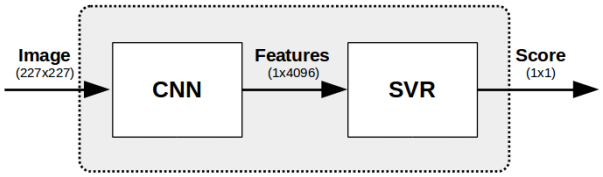


Figure 1: Graphical representation of the main steps of the proposed approach: the input image is fed to a CNN which performs all the multilayered operations and extracts a feature vector. Then, an SVR maps the extracted features to the perceived quality scores (i.e. MOS).

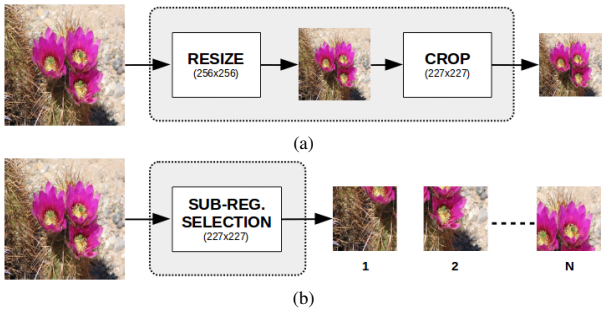


Figure 2: Graphical representation of different design choices: use of the whole image resized to fit the CNN input size (a); and use of multiple image sub-regions taken from the fullsize image (b).

and their use outperforms hand crafted, state-of-the-art systems in many visual classification tasks. Within the approach depicted in Figure 1, 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 (see Figure 2.a).

We experiment with the use of three CNNs 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 978 categories;
- Places-CNN, which has been trained on 2.5 million images of the Places Database for scene recognition belonging to 205 categories;
- Hybrid-CNN [38], which has been trained using 3.5 million images from 1,183 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 previous design choice, 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 sub-regions (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 (see Figure 2.b).

We experiment the use of a different number randomly selected sub-regions, 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. Different fusion strategies are experimented:

- *feature pooling*: information fusion is performed element by element on the sub-region feature vectors to generate a single feature vector for each image (see Figure 3.a). Minimum, average, and maximum feature pooling are considered.
- *feature concatenation*: information fusion is performed by concatenating the sub-region feature vectors in a single longer feature vector (see Figure 3.b).
- *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 (see Figure 3.c).

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 [34, 35], 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 [34, 35].

We start the fine-tuning procedure to the image quality assesment task by substituting the last fully connected layer of a

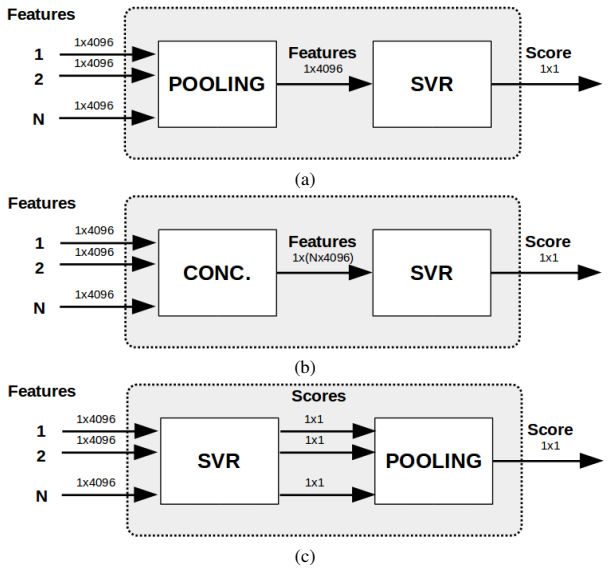


Figure 3: Graphical representation of different design choices to pool information coming from multiple image sub-regions: feature pooling (a), feature concatenation (b), and prediction pooling (c).

pre-trained CNN with a new one initialized with random values. The new layer is trained from scratch using the back-propagation algorithm [32] 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 sub-regions into five classes corresponding to five image quality grades. 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 within the approach depicted in Figure 1, just like one of the pre-trained CNNs.

3. Image Database and evaluation criterions

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 [29]: "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 Image Quality Challenge Database [29, 19]. It contains 1,162 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, etc. Figure 4 shows some database samples. Database images have been rated by many thousands of subjects via an online crowdsourcing system designed for subjective quality assessment. About 280,000 opinion scores from over 5,000 subjects 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.

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

	LCC	SROCC
Imagenet-CNN	0.6782	0.6381
Places-CNN	0.6267	0.6055
Hybrid-CNN	**0.7215**	**0.7021**

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 [19] 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 Pearsons Linear Correlation Coefficient (LCC) and the Spearmans Rank Ordered Correlation Coefficient (SROCC) between the predicted and the ground truth quality scores, reporting the median of these correlation coefficients across the 10 splits. In all the experiments we use the Caffe open-source framework [36] for CNN training and feature extraction, and the LIBLINEAR library [39] for SVR training.

4. Experimental results

In this section we evaluate the performance of each design choice introduced in Section II.

4.1. Experiment I: Image description using pre-trained CNNs

We extract the 4096-dimensional features from the *fc7* layer of the pre-trained ImageNet-CNN, Places-CNN and Hybrid-CNN. Since these CNNs require an input with a dimensionality equal to 227×227 pixels, we rescale the original 500×500 images to 256×256 keeping aspect ratio, and then we crop out the central 227×227 sub-region from the resulting image. All the images are pre-processed by subtracting the mean image. The median LCC and SROCC over the 10 train-test splits are reported in Table 2. From the results it is possible to see that Hybrid-CNN outperforms both Imagenet-CNN and Places-CNN, with Places-CNN giving the worst performance.

4.2. Experiment II: feature and prediction pooling strategies

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 sub-regions from the original image. Given the results of the previous experiment, the only features considered here are those extracted using the Hybrid-CNN.

Figure 4: Sample images from the LIVE In the Wild Image Quality Challenge Database.

We evaluate three different fusion schemes for combining the information generated by the multiple sub-regions to obtain a single score prediction for the whole image. The first scheme is feature pooling, where information fusion is performed element-wise on the feature vectors: this can be seen as an early fusion approach, in which a single feature vector is generated and given as input to the SVR to predict a single quality score for the whole image. Three different pooling operators are considered: minimum, average, and maximum.

The second scheme is feature concatenation, where information fusion is achieved by concatenating the multiple feature vectors into a single feature vector with higher dimensionality.

The third scheme is prediction pooling, where information fusion is performed on the predicted quality scores: this can be seen as a late fusion approach, in which a score is predicted from each crop. Three different pooling operators are considered to combine the multiple scores: minimum, average, and maximum pooling.

In all the experiments the number of random crops is varied between 5 and 50 in steps of 5. Figure 5 shows the plots for LCC and SROCC with respect to the number of crops considered, while the numerical values 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 in Appendix. From the plots it is possible to see that feature pooling conveys the best results. Prediction pooling is able to give comparable results with those of feature pooling only when a small number of crops is considered. Finally, feature concatenation gives the worst results, giving comparable results with those of prediction pooling only when a large number of crops is considered. 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 $3.4{\cdot}10^{-9}$) and prediction average-pooling ($p-$value equal to $8.8{\cdot}10^{-5}$). The difference between feature concatenation and prediction average-pooling is not significative instead ($p-$value equal to 0.23).

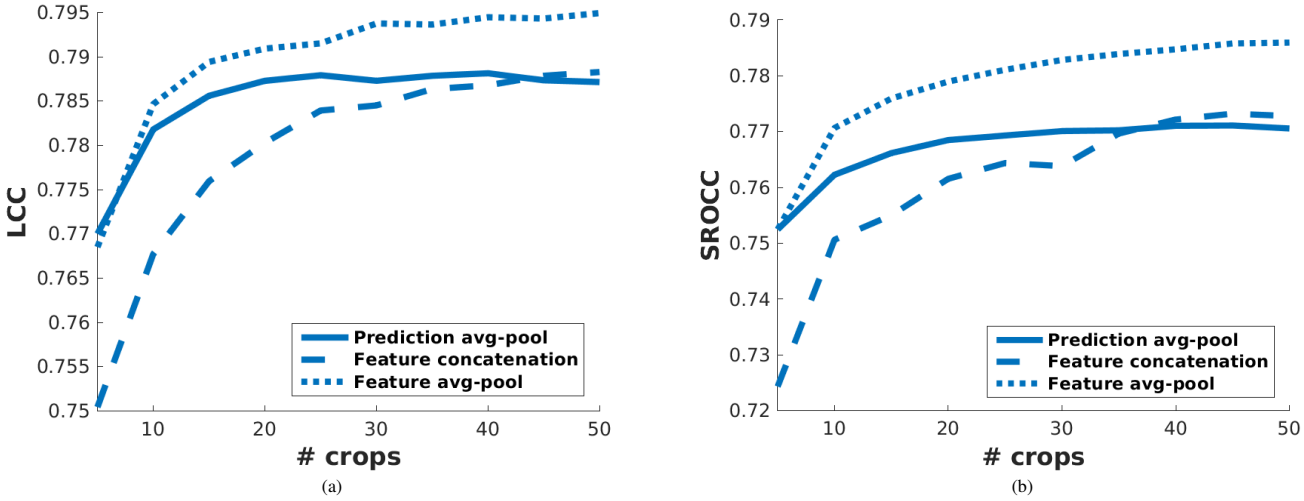
Table 3: Median LCC and SROCC across 10 train-test random splits of the LIVE In the Wild Image Quality Challenge Database considering randomly selected crops as input for the Hybrid-CNN and three different fusion approaches: feature pooling, feature concatenation and prediction pooling.

	LCC	SROCC
Feature pooling (avg-pool,@30crops)	**0.7938**	**0.7828**
Feature concatenation (@35crops)	0.7864	0.7724
Prediction pooling (avg-pool,@20crops)	0.7873	0.7685

4.3. Experiment III: Image description using a fine-tuned CNN

In all previous experiments we use pre-trained CNNs for feature extraction. In this experiment instead, we fine-tune the Hybrid-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 (see Figure 7), during training we give larger weights to images belonging to less represented distortion classes [40, 41]. 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. Figure 6 shows the plots for LCC and SROCC with respect to the number of crops considered, while the numerical values 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 test whose results are reported in Appendix. From the plots it is possible to notice that prediction pooling conveys the best results whatever is the number of crops considered. 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



(a) (b)

Figure 5: Median LCC (a) and SROCC (b) across 10 train-test random splits of the LIVE In the Wild Image Quality Challenge Database, with respect to the number of image crops given in input to the pre-trained Hybrid-CNN. Three fusion schemes are considered (feature pooling, feature concatenation and prediction pooling), and for each of them only the best configuration over the pooling operators considered is reported.

Table 4: Median LCC and SROCC across 10 train-test random splits of the LIVE In the Wild Image Quality Challenge Database considering randomly selected crops as input for the fine-tuned CNN and two different fusion approaches (feature pooling and prediction pooling).

	LCC	SROCC
Feature pooling (avg-pool,@20crops)	0.9026	0.8851
Prediction pooling (avg-pool,@25crops)	**0.9082**	**0.8894**

better than those obtained by feature average-pooling (p-value equal to $4.7{\cdot}10^{-4}$).

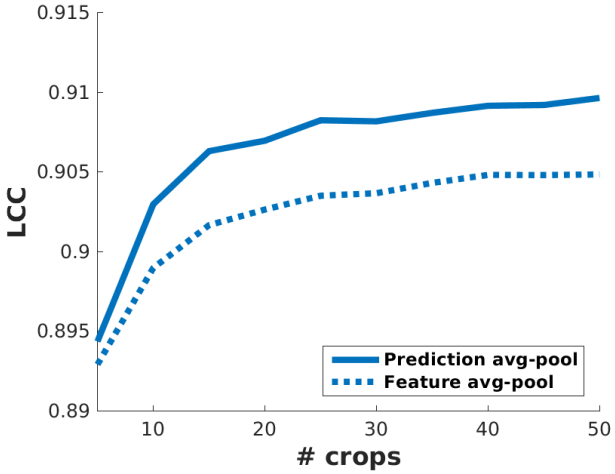
5. Discussion

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. 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.

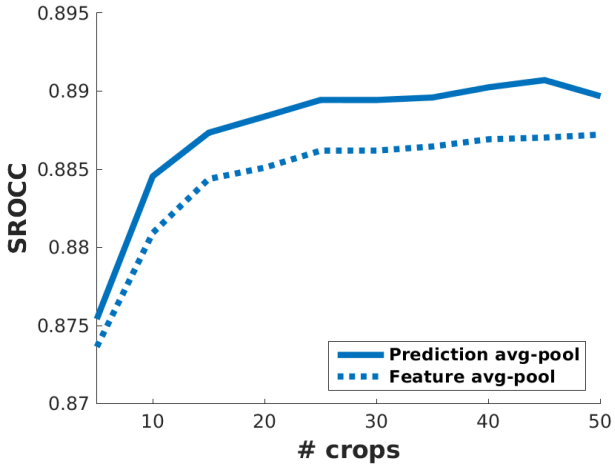
Error statics may not give an intuitive idea of how well a NR-IQA algorithm performs. On the other hand, individual human scores can be rather noisy. Taking into account that the LIVE In the Wild Image Quality Challenge Database gives for each image the MOS as well as the standard deviation of the human 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.

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 LIVE Image Quality Assessment Database [18]. It contains a total of 779 distorted images with five different single distortions: JPEG2000 compression, JPEG compression, White Gaussian noise, Gaussian blur and Fast Fading at 7-8 synthetic degradation levels derived from 29 reference images. Differential Mean Opinion Scores (DMOS) are provided for each image in the range $[0, 100]$, where higher DMOS indicates lower quality.

We evaluate our method on this dataset dealing with the different human judgements and distortion ranges by only retraining the SVR, while keeping the CNN unchanged. We follow the experimetal protocol used in [25, 26]. 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 are reported in Table 6 in terms of average LCC and SROCC values. From these results it is possible to see that our method, DeepBIQ, is able to obtain the best performance in terms of both LCC and SROCC notwithstanding that differently from the all the other methods reported, the features have been learned on a different dataset and not on a portion of the LIVE Image Quality Assessment Database itself. Therefore, the results confirm the effectiveness of our approach for no-reference image quality assessment.



(a)



(b)

Figure 6: Median LCC (a) and SROCC (b) across 10 train-test random splits of the LIVE In the Wild Image Quality Challenge Database, with respect to the number of image crops given in input to the fine-tuned CNN. Two fusion schemes are considered (feature average pooling and prediction average pooling).

Table 5: Median LCC and median SROCC across 10 train-test random splits of the LIVE In the Wild Image Quality Challenge Database.

	LCC	SROCC
DIIVINE [8]	0.56	0.51
BRISQUE [9]	0.61	0.60
BLIINDS-II [17]	0.45	0.40
S3 index [42]	0.32	0.31
NIQE [10]	0.48	0.42
C-DIIVINE [16]	0.66	0.63
FRIQUEE [20, 19]	0.71	0.68
DeepBIQ (Exp. I: pre-trained CNN, whole image)	0.72	0.70
DeepBIQ (Exp. II: pre-trained CNN, image sub-regions, feat. avg-pool)	0.79	0.79
DeepBIQ (Exp. III: fine-tuned CNN, image sub-regions, pred. avg-pool)	**0.91**	**0.89**

Table 6: Median LCC and median SROCC across 100 train-val-test random splits of the legacy LIVE Image Quality Assessment Database.

Method	LCC	SROCC
DIIVINE [8]	0.93	0.92
BRISQUE [9]	0.94	0.94
BLIINDS-II [17]	0.92	0.91
NIQE [10]	0.92	0.91
C-DIIVINE [16]	0.95	0.94
FRIQUEE [20, 19]	0.95	0.93
Rectifier Neural Network [21]	–	0.96
Multi-task CNN [26]	0.95	0.95
Shallow CNN [25]	0.95	0.96
DeepBIQ	**0.98**	**0.97**

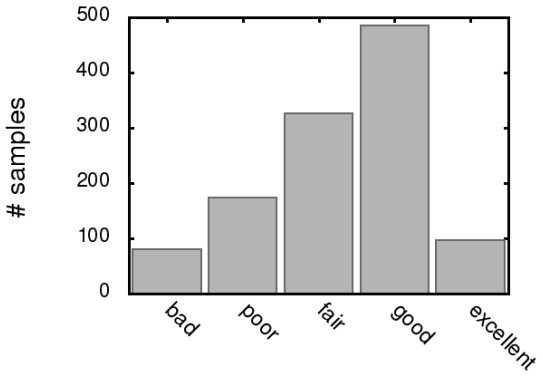


Figure 7: Sample distribution over the five quality grades considered for the LIVE In the Wild Image Quality Challenge Database.

6. 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 Convolu-

tional Neural Networks (CNNs) as a generic image description, to the use of features extracted from a CNN fine-tuned for the image quality task.

Results in the first experiment on the LIVE In the Wild Image Quality Challenge Database [29, 19] showed that using a pre-trained CNN as a generic image description and using a Support Vector Regression (SVR) machine to map these features to quality scores, is able to slightly outperform all the methods in the state of the art. In particular, the more concepts the original CNN is able to discriminate, the more accurate are the predicted scores. In fact, the best results are obtained by the Hybrid-CNN, that is able to classify both scene and object categories, with a Linear Correlation Coefficient (LCC) between predicted and human subjective scores of more than 0.72.

In the second experiment we have shown that the performance can be increased by considering multiple sub-regions. Three different fusion schemes have been tested, i.e. feature pooling, feature concatenation and prediction pooling, coupled with three pooling operators: minimum, average and maximum. The best results are obtained using feature average-pooling with a LCC of more than 0.79.

In the third experiment we have shown that a further improvement in performance can be obtained by fine-tuning the Hybrid-CNN for category-based image quality assessment. The use of the prediction pooling fusion scheme with the average operator reaches a LCC of almost 0.91, that is 0.20 higher than the best solution in the state of the art [19]. Furthermore, in many cases, the quality score predictions of our method are closer to the average observer than those of a generic human observer.

Our best proposal, named DeepBIQ, is then further tested on the legacy LIVE Image Quality Assessment Database [18]. To deal with the different type 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 this dataset, even if the features have been learned on a different dataset, confirming the effectiveness of our approach for no-reference image quality assessment.

References

[1] A. C. Bovik, "Automatic prediction of perceptual image and video quality," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 2008–2024, 2013.

[2] M. P. Eckert and A. P. Bradley, "Perceptual quality metrics applied to still image compression," *Signal processing*, vol. 70, no. 3, pp. 177–200, 1998.

[3] S. Winkler, "Issues in vision modeling for perceptual video quality assessment," *Signal Processing*, vol. 78, no. 2, pp. 231–252, 1999.

[4] T. N. Pappas, R. J. Safranek, and J. Chen, "Perceptual criteria for image quality evaluation," *Handbook of image and video processing*, pp. 669–684, 2000.

[5] Z. Wang, H. R. Sheikh, and A. C. Bovik, "Objective video quality assessment," *The handbook of video databases: design and applications*, pp. 1041–1078, 2003.

[6] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, 2004.

[7] S. Bianco, G. Ciocca, F. Marini, and R. Schettini, "Image quality assessment by preprocessing and full reference model combination," in *IS&T/SPIE Electronic Imaging*, pp. 72420O–72420O, International Society for Optics and Photonics, 2009.

[8] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *Image Processing, IEEE Transactions on*, vol. 20, no. 12, pp. 3350–3364, 2011.

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

[10] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," *Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

[11] G. Ciocca, S. Corchs, F. Gasparini, and R. Schettini, "How to assess image quality within a workflow chain: an overview," *International Journal on Digital Libraries*, vol. 15, no. 1, pp. 1–25, 2014.

[12] A. Ciancio, A. L. N. T. Da Costa, E. A. da Silva, A. Said, R. Samadani, and P. Obrador, "No-reference blur assessment of digital pictures based on multifeature classifiers," *Image Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 64–75, 2011.

[13] S. Corchs, F. Gasparini, and R. Schettini, "No reference image quality classification for jpeg-distorted images," *Digital Signal Processing*, vol. 30, pp. 86–100, 2014.

[14] A. Mittal, A. K. Moorthy, A. C. Bovik, C. W. Chen, P. Chatzimisios, T. Dagiuklas, and L. Atzori, "No-reference approaches to image and video quality assessment," *Multimedia Quality of Experience (QoE): Current Status and Future Requirements*, p. 99, 2015.

[15] K. Seshadrinathan and A. C. Bovik, "Automatic prediction of perceptual quality of multimedia signalsa survey," *Multimedia Tools and Applications*, vol. 51, no. 1, pp. 163–186, 2011.

[16] Y. Zhang, A. K. Moorthy, D. M. Chandler, and A. C. Bovik, "C-diivine: No-reference image quality assessment based on local magnitude and phase statistics of natural scenes," *Signal Processing: Image Communication*, vol. 29, no. 7, pp. 725–747, 2014.

[17] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the dct domain," *Transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, 2012.

[18] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "Live image quality assessment database release 2," 2005.

[19] D. Ghadiyaram and A. C. Bovik, "Massive Online Crowdsourced Study of Subjective and Objective Picture Quality," *IEEE Transactions on Image Processing*, vol. 25, pp. 372–387, jan 2016.

[20] D. Ghadiyaram and A. C. Bovik, "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, dec 2014.

[21] H. Tang, N. Joshi, and A. Kapoor, "Blind image quality assessment using semi-supervised rectifier networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2877–2884, 2014.

[22] W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment via deep learning," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 26, no. 6, pp. 1275–1286, 2015.

[23] Y. Lv, G. Jiang, M. Yu, H. Xu, F. Shao, and S. Liu, "Difference of gaussian statistical features based blind image quality assessment: A deep learning approach," in *Image Processing (ICIP), 2015 IEEE International Conference on*, pp. 2344–2348, IEEE, 2015.

[24] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Real-time no-reference image quality assessment based on filter learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 987–994, 2013.

[25] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1733–1740, 2014.

[26] L. Kang, P. Ye, Y. Li, and D. Doermann, "Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks," in *Image Processing (ICIP), 2015 IEEE International Conference on*, pp. 2791–2795, IEEE, 2015.

[27] V. N. Vapnik and V. Vapnik, *Statistical learning theory*, vol. 1. Wiley New York, 1998.

[28] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[29] D. Ghadiyaram and A. C. Bovik, "Crowdsourced study of subjective im-

age quality," in *Asilomar Conf. Signals, Syst. Comput*, 2014.
[30] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[31] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 806–813, 2014.
[32] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural networks: Tricks of the trade*, pp. 9–48, Springer, 2012.
[33] J. Deng, A. Berg, S. Satheesh, H. Su, A. Khosla, and L. Fei-Fei, "Imagenet large scale visual recognition competition 2012 (ilsvrc2012)," 2012.
[34] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Advances in Neural Information Processing Systems*, pp. 3320–3328, 2014.
[35] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," *Unsupervised and Transfer Learning Challenges in Machine Learning*, vol. 7, p. 19, 2012.
[36] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the ACM International Conference on Multimedia*, pp. 675–678, ACM, 2014.
[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.
[38] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in Neural Information Processing Systems*, pp. 487–495, 2014.
[39] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.
[40] K. M. Ting, "An instance-weighting method to induce cost-sensitive trees," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 14, no. 3, pp. 659–665, 2002.
[41] Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *Transactions on Knowledge and Data Engineering*, vol. 18, no. 1, pp. 63–77, 2006.
[42] C. T. Vu, T. D. Phan, and D. M. Chandler, "S3: A spectral and spatial measure of local perceived sharpness in natural images," *Transactions on Image Processing*, vol. 21, no. 3, pp. 934–945, 2012.

Appendix

In this section we report the results of the statistical significance two-sample t-test to select the best number of crops for each design choice investigated. The p-values for Experiment II and III are respectively reported in Figure 8 and 9. The optimal number or crops is the lowest number after which any increase does not give an improvement at the 5% significance level.

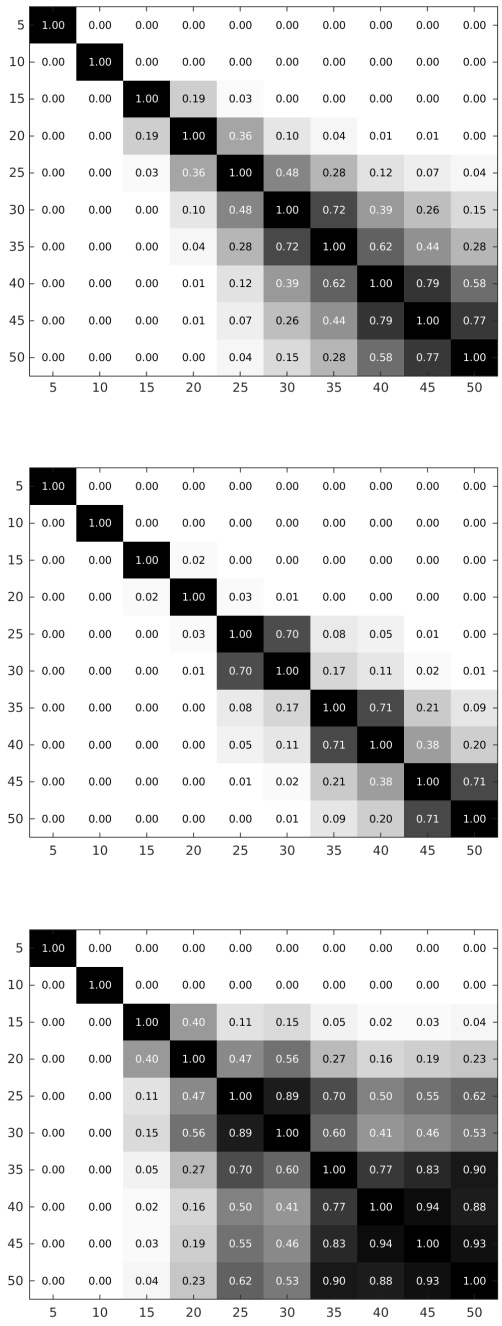


Figure 8: p-values of the two-sample t-test in Experiment II for the different design choices: feature pooling (top), feature concatenation (middle), and prediction pooling (bottom).

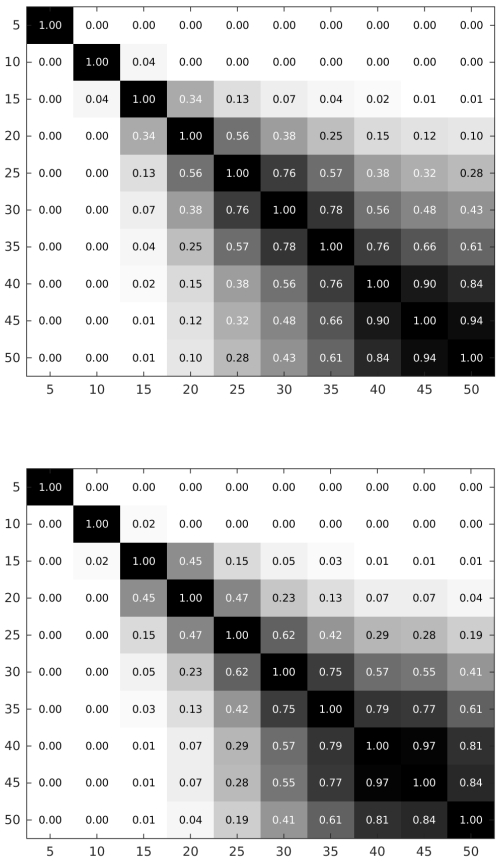


Figure 9: p-values of the two-sample t-test in Experiment III for the different design choices: feature pooling (top), and prediction pooling (bottom).