DEEP BLIND IMAGE QUALITY ASSESSMENT USING DUAL-ORDER STATISTICS

Zihan Zhou, Yong Xu, Yuhui Quan, Ruotao Xu

School of Computer Science and Engineering, South China University of Technology, China

ABSTRACT

Deep convolutional neural networks (CNNs) have become a promising approach to blind image quality assessment (BIQA). Existing CNN-based BIQA methods often employ global average pooling (GAP) to aggregate feature maps into a fixed-size representation for regression, so as to handle input images with varying sizes. However, GAP is only capable of extracting the first-order statistics of feature distributions, which is ineffective for distinguishing complex distortions that cause local degradation or preserve global features. To tackle this problem, we introduce the second-order global co-variance pooling (GCP) for aggregating feature maps, leading to a more distortion-sensitive and more discriminative global representation. By incorporating GCP and GAP into a ResNet backbone, we propose an effective deep model for BIQA. The experimental results on five BIQA benchmark datasets, including both the synthetic and authentic ones, have demonstrated the excellent performance of the proposed method.

Index Terms— Blind image quality assessment, Feature pooling, Convolutional neural networks, Deep learning

1. INTRODUCTION

Image quality is the main concern in image processing, and blind image quality assessment (BIQA) techniques are developed to automatically evaluate the perceptual quality of the distorted images. Given an input image, BIQA aims at estimating its perceived quality without accessing any reference information of the image. Such a topic has gained extensive attention from both industries and academics.

In general, there are two phases in a BIQA approach: quality-aware feature extraction and quality score regression. Traditional approaches usually adopt hand-crafted designs for feature extraction, such as natural scene statistics [1, 2] and variants of local binary patterns [3, 4]. Then the score regression is done by some well-established learning-based model, such as support vector regression, random forest, and Gaussian process. Recently, inspired by the success of CNNs in image recognition, a series of CNN-based BIQA approaches have been proposed, where the two phases are jointly optimized in an end-to-end manner. With the data-driven convolutional-layer-based feature extractor and fully-connected-layer-based score regressor, these approaches have shown impressive performance.

Typical CNNs in computer vision often connect convolutional layers with fully-connected (FC) layers. The FC layers encode feature maps into a single quality-aware representation and act as a regressor. Such a structure can only deal with images of a fixed size. Rescaling, one of the most often-used solutions to this problem in image recognition, is not suitable for IQA, as rescaling can obviously change the visual quality of an image. Therefore, global pooling, which aims to aggregate the score maps or feature maps with different spatial sizes into a single score or a fixed-length global vector, is especially critical for an IQA-oriented CNN model. A line of researches [5, 6, 7] proposed to use score pooling strategies, where the pooling module is designed to aggregate the predicted scores on local regions/patches. However, due to the absence of the real supervision to the visual quality of patches, it is hard to guarantee the accuracy of local quality scores, which limits the performance of such score-pooling-based methods. Another pooling strategy that avoids this issue is to operate global pooling on quality-aware feature maps to generate a fixed-length image representation for regression. Several different pooling mechanisms are adopted in this strategy, such as global averaging pooling (GAP) [8, 9] and other average-based pooling [10, 11, 12].

The GAP and average-based pooling mechanisms essentially capture the first-order statistics of a feature distribution. While such average-related statistics can summarize the holistic changes caused by distortions, they are insensitive when handling the distortions that cause little changes to the image features on average. Such distortions, called global-preserving distortions for convenience, are not rare in the real world. For instance, noise degradation which oscillates pixel values is often expected to preserve the average value due to the zero-mean property of the noise. Another often-seen example is image blurring, which also causes little changes to average intensity level due to the sum-to-one property of blur kernels. Since a BIQA-oriented CNN is expected to preserve or propagate the effects caused by distortions for accurate prediction of the image quality score, GAP will be likely to reduce the effectiveness of a BIQA-oriented CNN in
identifying degradation effects.

See Fig. 1 for an illustration, where the feature maps of noisy/blurred images have little changes in terms of global average, compared to those of their clean versions. Furthermore, for some local distortions, the differences in the feature maps only happen in local regions. The changes, though varying in the corresponding small regions, may be diminished by an average operation and produce similar average values. The above analysis motivates us to exploit other statistics for global feature pooling in a BIQA-oriented CNN so as to identify the degradation effects more effectively.

In this paper, we take a step towards addressing CNN-based BIQA by considering the second-order statistics for global pooling mechanisms. The introduction of the second-order statistics (e.g., covariance) to a CNN-based method can lead to a better description for the feature distribution of a distorted image. Consider the average-insensitive distortions such as white noise and Gaussian blur. Although usually preserving the mean, they change the covariance, i.e., noise usually results in higher variances due to its oscillation nature, while blurring often leads to lower variance due to its smoothing effect. Furthermore, second-order statistics are usually more sensitive to outliers. Due to the quadratic calculation, the large changes in a local region can be exposed during the feature pooling. In addition, though not explored in the BIQA realm, the second-order pooling, which captures the self and cross-channel similarities information of feature maps, has proven its ability to generate informative representations and describe complex classification boundary [13] in recognition. Its success in recognition also inspires us to explore the second-order statistics in BIAQ for the complex nature of the human visual system.

To capture the second-order statistics, we introduce the global covariance pooling (GCP) [13] for feature aggregation. Combining GCP with a first-order global pooling implemented by GAP, we propose an end-to-end framework with a dual-order global pooling mechanism for BIQA. The effectiveness of the proposed method is validated on five widely-used IQA benchmark datasets including both synthetic and real-world ones.

2. PROPOSED METHOD

In the proposed method, we construct a two-branch CNN model for blind image quality score prediction. The model accepts an image of an arbitrary size, since a global pooling is employed in both branches. The first branch includes a first-order global pooling module and is expected to process average-sensitive distortions mainly, while the second one, which includes a second-order pooling module, is expected to process average-insensitive distortions. The architecture of the proposed model is illustrated in Fig. 2, whose details are described in the following.

2.1. First-order Global Pooling

Let \( I \in \mathbb{R}^{H_0 \times W_0 \times C_0} \) denote the input image for assessment, and \( X = f(I) \in \mathbb{R}^{H \times W \times C_1} \) denote the extracted features. Given the extracted feature \( X \) with an arbitrary spatial size \( H \times W \), the global pooling mechanism is designed to aggregate \( X \) into a quality-aware global feature \( y \), which is then fed into the regressor for quality prediction. In the pooling scheme, the feature tensor \( X \) is treated as a set of samples \( \{x_i\}_{i=1}^N \) with \( N = HW \) and the sample \( x_i \in \mathbb{R}^{C_1} \) is the feature vector located at \( i \)-th position of \( X \).

The first-order statistic is important to characterize the feature distribution. In a BIQA task, the first-order statistics, often estimated with average or weighted average, are discriminative for the average-sensitive distortions, such as color shifts and brightness changes. To measure the first-order statistics of feature distribution, we introduce the GAP for feature aggregation, which is simply calculated by the average of features:

\[
\mu(X) = \frac{1}{N} \sum_{i=1}^{N} x_i.
\]
2.2. Second-order Global Pooling

Although GAP can handle the average-sensitive distortions, it usually fails to capture the changes of features caused by average-insensitive distortions, such as blur and zero-mean noise. Furthermore, for some spatial distortions, such as color blocks and non-eccentricity patches, the quality drop only happens in a small region. GAP tends to weaken or even eliminate local changes due to the global average.

The second-order statistics can be beneficial in such cases. We introduce the GCP for feature aggregation, which further captures the second-order statistics of feature distribution. Given the feature samples \( \{x_i\}_{i=1}^N \), GCP estimates the covariance of the samples as

\[
\Sigma(X) = \sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})^T,
\]

where \( \bar{x} \) denotes the mean of samples, i.e. \( \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \).

It is known that normalization is beneficial for the robust covariance estimation on high-dimensional low-sample-size data [14]. Encouraged by the promising performance of matrix square-root normalization in classification tasks [13], we adopt it for the GCP:

\[
\Sigma^{\frac{1}{2}}(X) = U \Lambda^{\frac{1}{2}} U^T,
\]

where \( U \) and \( \Lambda \) are the orthogonal matrix of eigenvectors and the diagonal matrix of eigenvalues of \( \Sigma(X) \), respectively.

2.3. Dual-Order BIQA Framework

The GAP and GCP use different feature maps as inputs, as they are designed to process different kinds of distortions, i.e., average-sensitive and average-insensitive distortions, respectively. Given the input image \( I \), the feature extraction procedure can be formulated as follows:

\[
X_1 = f_1 \circ f_0 \circ I,
\]

\[
X_2 = f_2 \circ f_0 \circ I,
\]

where \( \circ \) denotes function composition, \( X_1, X_2 \) denote the input features for GAP and GCP modules respectively, \( f_0 \) denotes the shared feature extractor (i.e. Shared Network in Fig. 2) for low-level features, and \( f_1, f_2 \) are average-sensitive and average-insensitive distortion-aware feature extractors respectively (i.e. Block3 and Block4 in Fig. 2).

The GAP and GCP aggregate the feature maps into global feature vectors respectively, which are then fed to FC layers for score prediction:

\[
s_1 = g_1 \circ \mu \circ X_1, \quad (6)
\]

\[
s_2 = g_2 \circ \Sigma^{\frac{1}{2}} \circ X_2, \quad (7)
\]

where \( g_1, g_2 \) are the FC layer following GAP, GCP respectively, and \( s_1, s_2 \) are the corresponding predicted scores for average-sensitive/insensitive distortions. The final quality score \( s \) is calculated as the weighted sum of \( s_1 \) and \( s_2 \) via a two-neuron FC layer:

\[
s = \omega_1 s_1 + \omega_2 s_2, \quad (8)
\]

where \( \omega_1 \) and \( \omega_2 \) is a pair of learned weights.

Given a set of images \( \{I_i\}_{i=1}^D \) and their subjective scores measured by human \( \{s_i^*\}_{i=1}^D \). Let \( \{s_i\}_{i=1}^D \) denote the scores predicted by the proposed model. We use the Huber loss for training due to its stronger robustness to outliers over the commonly-used mean square error loss, which is defined as

\[
\ell = \sum_{i=1}^D \ell_{\delta}(s_i, s_i^*),
\]

where \( \ell_{\delta} \) is the parameterized Huber loss defined by

\[
\ell_{\delta}(s, s^*) = \begin{cases} \frac{1}{2}(s - s^*)^2, & \text{for } |s - s^*| \leq \delta \\ \delta (|s - s^*| - \frac{1}{2} \delta), & \text{otherwise} \end{cases}.
\]

And \( \delta \) is a parameter to choose the way to penalty outliers. In implementation, we set \( \delta = 1/9 \) as suggested in [11].

3. EXPERIMENTS

3.1. Experimental Setups

Implementation Details. Following [11, 15], we select ResNet-101 [16] as our network backbone due to its effectiveness in feature extraction. In detail, we employ the first
3 building blocks (conv1, conv2_x, conv3_x) in ResNet-101 as the shared feature extraction network \( f_0(\cdot) \), the 4-th building block (conv4_x) for the feature extraction networks \( f_1(\cdot), f_2(\cdot) \) in two branches. These blocks are all initialized with weights of ResNet-101 trained on ImageNet. During training, the weights of \( f_0(\cdot) \) is frozen, while other blocks are optimized with the loss in (9). A two-stage training strategy is adopted. The network without the GCP branch is first trained. Then, the whole network and the paired parameters \( \omega_1, \omega_2 \) are jointly trained with frozen \( f_1(\cdot) \).

Datasets. In order to validate the performance of the proposed method, five publicly available natural image quality databases are employed in our experiments, including three artificially distorted sets (LIVE [17], TID2013 [18] and Kadid-10K [19]) and two realistically distorted sets (LIVE-C [20] and KonIQ-10K [21]). The images in their original resolution are fed into the model as input to test how well the model generalizes to pictures of arbitrary sizes. During training, we follow [8, 11] to randomly sample 80% of the images in each database for training and leave the rest for testing. Regarding the synthetically-distorted datasets, we split training and test sets according to the reference images such that content is not intersected between the two sets. Dynamic data augmentation comprising horizontal flip, vertical flip, and rotation of \( \pm 3^\circ \) is randomly applied to the training images. Since rotation produces extra black borders, we have removed the excess area by cropping.

Evaluation Criteria. Two commonly used evaluation metrics for performance comparison are adopted, including Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Order Correlation Coefficient (SRCC). The SRCC measures the prediction monotonicity, PLCC measures the linear correlation. An effective IQA metric is expected to yield high values of PLCC and SRCC.

Compared Methods. Deep BIQA models have been frequently reported to outperform traditional knowledge-driven BIQA methods, such as NIQE [22] and ILNIQE [23]. Thus we only compared the proposed model against recently State-of-the-Art DNN-based BIQA methods, including PQR [24], deepIQA [5], DBCNN [25], MetaQA [26], CaHDC [12], HyperNet [9], SiamIQA [27] and AIGQA [28]. The experimental results of compared methods are based on implementations obtained from the respective authors or just copied from the original papers.

3.2. Performance Evaluation

The results on synthetically distorted databases are reported in Table 1. It can be seen that the proposed method significantly outperforms the compared methods on datasets TID2013, Kadid-10K, and achieves comparative performance on dataset LIVE. Note that all methods achieve nearly perfect results (over 0.95 on PLCC, SRCC) on LIVE. In comparison, TID2013 and Kadid-10K which contain more samples and complex distortions, are usually more challenging and valuable for performance evaluation.

<table>
<thead>
<tr>
<th></th>
<th>LIVE</th>
<th>TID2013</th>
<th>Kadid-10K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SRCC</td>
<td>PLCC</td>
<td>SRCC</td>
</tr>
<tr>
<td>PQR [24]</td>
<td>0.965</td>
<td>0.971</td>
<td>-</td>
</tr>
<tr>
<td>deepIQA [5]</td>
<td>0.954</td>
<td>0.963</td>
<td>0.761</td>
</tr>
<tr>
<td>DBCNN [25]</td>
<td>0.946</td>
<td>0.959</td>
<td>0.816</td>
</tr>
<tr>
<td>MetaQA [26]</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CaHDC [12]</td>
<td>0.965</td>
<td>0.964</td>
<td>0.862</td>
</tr>
<tr>
<td>HyperNet [9]</td>
<td>0.962</td>
<td>0.966</td>
<td>-</td>
</tr>
<tr>
<td>SiamIQA [27]</td>
<td>0.961</td>
<td>-</td>
<td>0.855</td>
</tr>
<tr>
<td>AIGQA [28]</td>
<td>0.963</td>
<td>0.957</td>
<td>0.871</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td>0.957</td>
<td>0.961</td>
<td><strong>0.946</strong></td>
</tr>
</tbody>
</table>

The results on authentically databases are reported in Table 2. It is shown that our method performs significantly better than all compared methods on the large-scale dataset KonIQ-10K and is comparative on LIVE-C. Note that KonIQ-10K contains over 10 thousand realistically and complexly distorted images, which are usually more challenging for BIQA evaluation.

Table 2: Median SRCC and PLCC results across ten sessions on the test sets of the authentically distorted IQA databases. Bold on digits denote the best result for each criteria.

<table>
<thead>
<tr>
<th></th>
<th>LIVE-C</th>
<th>KonIQ-10K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SRCC</td>
<td>PLCC</td>
</tr>
<tr>
<td>PQR [24]</td>
<td>0.857</td>
<td>0.882</td>
</tr>
<tr>
<td>deepIQA [5]</td>
<td>0.671</td>
<td>0.686</td>
</tr>
<tr>
<td>DBCNN [25]</td>
<td>0.851</td>
<td>0.869</td>
</tr>
<tr>
<td>MetaQA [26]</td>
<td>0.802</td>
<td>0.835</td>
</tr>
<tr>
<td>CaHDC [12]</td>
<td>0.738</td>
<td>0.744</td>
</tr>
<tr>
<td>HyperNet [9]</td>
<td><strong>0.859</strong></td>
<td><strong>0.882</strong></td>
</tr>
<tr>
<td>SiamIQA [27]</td>
<td>0.851</td>
<td>-</td>
</tr>
<tr>
<td>AIGQA [28]</td>
<td>0.751</td>
<td>0.761</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td>0.834</td>
<td>0.852</td>
</tr>
</tbody>
</table>

To evaluate the generalization ability of our method, we conduct a cross-dataset experiment by using the KonIQ-10K as the training set and LIVE-C as the test set. Only compared methods with available results are reported. The results are shown in Table 3, which indicates that our proposed model has a good generalization capability, especially for predicting the quality for pictures of arbitrary resolution and of real-world complex distortions.

3.3. Ablation Study

To analyze the effectiveness of utilizing GCP with GAP module, we conduct an ablation experiment to validate the effec-
In this work, we introduced the global covariance pooling module for the deep BIQA, so as to characterize the second-order statistics of feature distribution and produce a quality-aware global feature to encode the average-insensitive distortions. Incorporating with the global average pooling, we construct a two-branch deep neural network for quality score prediction of degraded images with various distortion types. Experimental results on five benchmark IQA datasets have demonstrated the efficiency of the proposed method as well the dual-order pooling mechanism. In future, we would like to investigate the potentials of higher-order statistics in IQA tasks. Furthermore, inspired by the effectiveness of GCP, we will make a step further by seeking better characterization for the functional relationship that exists between channel consistencies and subjective quality scores.

5. REFERENCES


