DEEP CNN BASED BLIND IMAGE QUALITY PREDICTOR

Seminar Report

Submitted in partial fulfillment of the requirements for the award of degree of

BACHELOR OF TECHNOLOGY

In

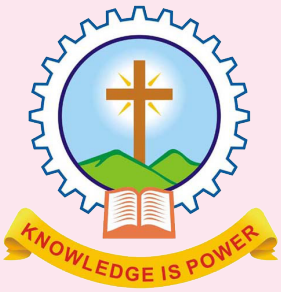
COMPUTER SCIENCE AND ENGINEERING

of

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Submitted By

AKHILA DINESH.R



Department of Computer Science & Engineering
Mar Athanasius College Of Engineering Kothamangalam

DEEP CNN BASED BLIND IMAGE QUALITY PREDICTOR

Seminar Report

Submitted in partial fulfillment of the requirements for the award of degree of

BACHELOR OF TECHNOLOGY

In

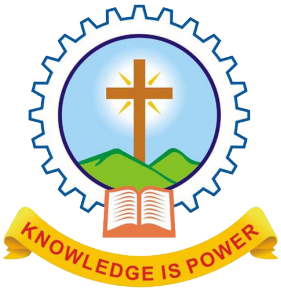
COMPUTER SCIENCE AND ENGINEERING

of

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Submitted By

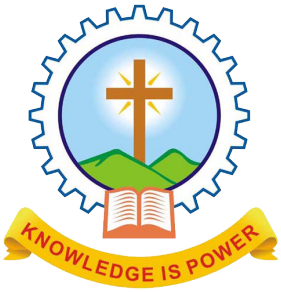
AKHILA DINESH.R



Department of Computer Science & Engineering

Mar Athanasius College Of Engineering Kothamangalam

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MAR ATHANASIUS COLLEGE OF ENGINEERING
KOTHAMANGALAM



CERTIFICATE

This is to certify that the report entitled **Deep CNN Based Blind Image Quality Predictor** *submitted by Ms.* **AKHILA DINESH.R** *, Reg. No.* **MAC15CS008** *towards partial fulfillment of the requirement for the award of Degree of Bachelor of Technology in Computer science and Engineering from APJ Abdul Kalam Technological University for December 2018 is a bonafide record of the seminar carried out by her under our supervision and guidance.*

.....	.....	.....
Prof. Joby George	**Prof. Neethu Subash**	**Dr. Surekha Mariam Varghese**
Faculty Guide	***Faculty Guide***	***Head of the Department***

Date: Dept. Seal

ACKNOWLEDGEMENT

First and foremost, I sincerely thank the God Almighty for his grace for the successful and timely completion of the seminar.

I express my sincere gratitude and thanks to Dr. Solly George, Principal and Dr. Surekha Mariam Varghese, Head Of the Department for providing the necessary facilities and their encouragement and support.

I owe special thanks to the staff-in-charge Prof. Joby george, Prof. Neethu Subash and Prof. Joby Anu Mathew for their corrections, suggestions and sincere efforts to co-ordinate the seminar under a tight schedule.

I express my sincere thanks to staff members in the Department of Computer Science and Engineering who have taken sincere efforts in helping me to conduct this seminar.

Finally, I would like to acknowledge the heartfelt efforts, comments, criticisms, co-operation and tremendous support given to me by my dear friends during the preparation of the seminar and also during the presentation without whose support this work would have been all the more difficult to accomplish.

ABSTRACT

Image recognition based on convolutional neural networks (CNNs) has recently been shown to deliver the state of the-art performance in various areas of computer vision and image processing. Nevertheless, applying a deep CNN to no reference image quality assessment (NR-IQA) remains a challenging task due to critical obstacles, i.e., the lack of a training database. CNN-based NR-IQA framework that can effectively solve this problem is hence proposed. The proposed method - deep image quality assessor (DIQA) - separates the training of NR-IQA into two stages: 1) An objective distortion part and 2) A human visual system-related part. In the first stage, the CNN learns to predict the objective error map, and then the model learns to predict subjective score in the second stage. To complement the inaccuracy of the objective error map prediction on the homogeneous region, a reliability map is implemented. Two simple handcrafted features were additionally employed to further enhance the accuracy. In addition, a way to visualize perceptual error maps is implemented.

Contents

List of Figures

List of abbreviation

IQA	Image Quality Assessment
NR-IQA	No-Reference Image Quality Assessment
FR-IQA	Full Reference Image Quality Assessment
DIQA	Deep Image Quality Assessment
SVM	Support Vector Machine
DBN	Deep Belief Network
CNN	Convolutional Neural Network
HVS	Human Visual System
NSS	Natural Scene Statistics
RNN	Recurrent neural network

Introduction

The goal of image quality assessment (IQA) is to predict the perceptual quality of digital images in a quantitative manner. Digital images are likely to be inevitably degraded in the process from content generation to consumption. The acquisition, transmission, storage, post-processing, or compression of images introduces various distortions, such as Gaussian white noise, Gaussian blur (GB), or blocking artifacts.

A reliable IQA algorithm can help quantify the quality of images obtained blindly from the Internet and accurately assess the performance of image processing algorithms, such as image compression and super-resolution, from the perspective of a human observer. IQA is classified in general into three categories, depending on whether a reference image (the pristine version of an image) is available: full-reference IQA (FR-IQA), reduced-reference IQA (RR-IQA), and noreference IQA (NR-IQA). In general, the performance of these techniques, in order of decreasing accuracy, is FR-IQA, RR-IQA, and NR-IQA. However, since reference images are not accessible in a number of practical scenarios, NR-IQA is most appropriate as the most general method.

The bit rate of computer networks has continued to increase in recent years and has enabled the provision of high-quality entertainment to end users who do not have reference images; hence, significant research efforts have been made to enhance the accuracy of NR-IQA from the perspective of the end user. Many recently proposed NR-IQA algorithms involve the use of machine learning, such as support vector machines (SVMs) and neural networks (NN), to blindly predict image quality scores. Research has shown that the accuracy of NR-IQA depends heavily on designing elaborate features.

Natural scene statistics (NSS) is one of the most successful features under the assumption that natural images have statistical regularity that is altered when distortions are introduced. Due to the difficulties involved in obtaining reliable features, research on NR-IQA has progressed significantly since NSS. Deep learning has lately been adopted in a few NR-IQA studies as a different method from conventional approaches based on NSS. However, most such studies have continued to use handcrafted features, and deep models, such as deep belief

networks (DBNs) and stacked autoencoders, have been used in place of conventional regression machines.

System work in three main steps: collecting brain signals, interpreting them and outputting commands to a connected machine according to the brain signal received. It can be applied to a variety of tasks, including but not limited to neurofeedback, restoring motor function to paralyzed patients, allowing communication with locked in patients and improving sensory processing.

Convolutional neural networks (CNNs) form the most popular deep learning model nowadays due to their strong representation capability and impressive performance. CNNs have been successfully applied to various computer vision and image processing problems.

Related works

Convolutional neural networks (CNNs) form the most popular deep learning model nowadays due to their strong representation capability and impressive performance. CNNs have been successfully applied to various computer vision and image processing problems.

The performance of deep neural networks heavily depends on the number of training data. However, the currently available IQA databases are much smaller compared to the typical computer vision data set for deep learning.

Moreover, obtaining large-scale reliable human subjective labels is very difficult. Unlike classification labels, constructing an IQA database requires a complex and timeconsuming psychometric experiment. To expand the training data set, one can use data augmentation techniques such as rotation, cropping, and horizontal reflection. Unfortunately, any transformation of images would affect perceptual quality scores.

2.1 Natural scene statistics

Natural scene statistics (NSS) [1], [2] is one of the most successful features under the assumption that natural images have statistical regularity that is altered when distortions are introduced. Due to the difficulties involved in obtaining reliable features, research on NR-IQA has progressed significantly since NSS. Deep learning has lately been adopted in a few NR-IQA studies as a different method from conventional approaches based on NSS [3], [4]. However, most such studies have continued to use handcrafted features, and deep models, such as deep belief networks (DBNs) and stacked autoencoders, have been used in place of conventional regression machines.

Researchers in visual quality assessment have endeavored to understand how the presence of these distortions affects the viewing experience. The ideal approach to measure the effect of distortions on the quality of viewing experience is to solicit opinions from a sufficiently large sample of the human populace. Averaging across these opinions produces a mean opinion score (MOS) which is considered to be the perceived quality of the stimulus. Such subjective assessment of visual quality is the best indicator of how distortions affect perceived

quality; however, they are time-consuming, cumbersome, and impractical. Hence, one seeks to develop algorithms that produce quality estimates of these distorted visual stimuli with high correlation with MOS. Such objective image quality assessment (IQA) is the focus of this paper.

Objective quality assessment can be divided into three categories depending on the amount of information provided to the algorithm [1]. Full-reference (FR) algorithms are provided with the original undistorted visual stimulus along with the distorted stimulus whose quality is to be assessed. Reduced-reference (RR) approaches are those in which the algorithm is provided with the distorted stimulus and some additional information about the original stimulus, either by using a auxiliary channel or by incorporating some information in the distorted stimulus (such as a watermark). Finally, no-reference (NR)/blind approaches to quality assessment are those in which the algorithm is provided only with the distorted stimulus. In this paper, we propose a method for NR/blind IQA.

2.2 Deep belief network

Deep Belief Network form the most popular deep learning model nowadays due to their strong representation capability and impressive performance. CNNs have been successfully applied to various computer vision and image processing problems.

The performance of deep neural networks heavily depends on the number of training data. However, the currently available IQA databases are much smaller compared to the typical computer vision data set for deep learning. For example, the LIVE IQA database [5] contains 174233 images for each distortion type, while the widely used data set for image recognition contain more than 1.2 million pieces of labeled data [6]. Moreover, obtaining large-scale reliable human subjective labels is very difficult. Unlike classification labels, constructing an IQA database requires a complex and timeconsuming psychometric experiment. To expand the training data set, one can use data augmentation techniques such as rotation, cropping, and horizontal reflection. Unfortunately, any transformation of images would affect perceptual quality scores.

Even though NR QA is potentially the most useful goal, the difficulty of creating algorithms that accurately predict visual quality, especially without any information about the

original image, has led to greater activity in the FR QA area [2]. These studies have yielded considerable insights into the perception of image distortions and image quality that may prove useful in creating algorithms that assess quality without need for a reference. RR QA remains attractive, not only as a solution to the QA problem on its own, but also as a stepping stone towards solving the NR QA problem [3], [4]. Indeed, our approach to NR QA finds inspiration from previously proposed models for FR QA [5], [6] as well as for RR QA [3], [4].

Most present-day NR IQA algorithms assume that the distorting medium is knownfor example, compression, loss induced due to noisy channel, etc. Based on this assumption, distortions specific to the medium are modeled and quality is assessed. By far the most popular distorting medium is compression, which implies that blockiness and bluriness should be evaluated. In the following, we study blind QA algorithms that target three common distortion categories: JPEG compression, JPEG2000 compression, and blur. We also survey blind QA algorithms that operate holistically in fig.2.1.

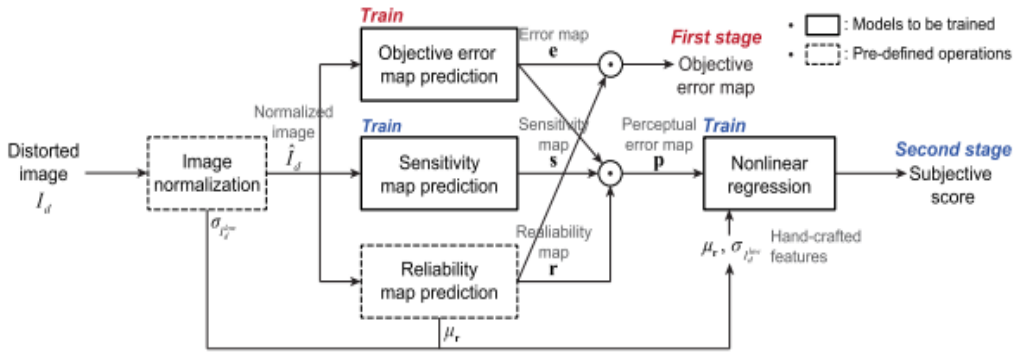


Fig. 2.1: Flowchart of Deep Image Quality Assessor-sense

Moreover, the perceptual process of the human visual system (HVS) includes multiple complex processes, which makes training of a deep model with limited data set even harder. For example, the visual sensitivity of the HVS varies according to spatial frequency of stimuli [7], [8], and the presence of texture hinders other spatially coincident image changes [9]. In addition, the perceived signals go through bandpass, multiscale, and directional decompositions in the visual cortex [10]. Such complex behaviors need to be embedded in the data set with

human subjective labels. However, it is difficult to claim that a small data set can represent general visual stimuli, which results in an overfitting problem.

To generate a signal that is detectable, approximately 50,000 active neurons are needed. Since current dipoles must have similar orientations to generate magnetic fields that reinforce each other, it is often the layer of pyramidal cells, which are situated perpendicular to the cortical surface, that gives rise to measurable magnetic fields. Bundles of these neurons that are orientated tangentially to the scalp surface project measurable portions of their magnetic fields outside of the head, and these bundles are typically located in the sulci. Researchers are experimenting with various signal processing methods in the search for methods that detect deep brain (i.e., non-cortical) signal, but no clinically useful method is currently available.

The proposed method

To tackle this problem, we propose a novel NR-IQA framework called deep blind image quality assessor (DIQA). The DIQA is trained in two separated stages as shown in Fig. 3.1. In the first stage, an objective error map is used as a proxy training target to expand the data set labels. The existing database provides a subjective score for each distorted image. In other words, one training data item includes a mapping from a 3-D tensor (width, height, and channel) to a scalar value.

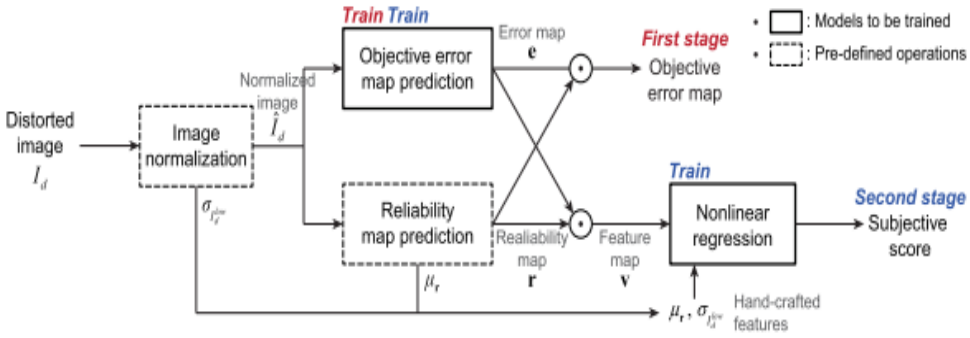


Fig. 3.1: Overall flowchart of Deep Image Quality Assessor

In contrast, the DIQA utilizes reference images during training and generates a 2-D intermediate target called the objective error map. Please note that the reference images are accessible during training as long as the database provides them, and the ground-truth objective error map can be easily derived by comparing the reference and distorted images.

Our NR IQA model utilizes a 2-stage framework for blind IQA that we introduced in [12]. In this framework, scene statistics extracted from a distorted natural image are used to first explicitly classify the distorted image into one of distortions (distortion identificationstage 1). Then, the same set of statistics are used to evaluate the distortion-specific quality (distortion-specific QAstage 2) of the image. A combination of the two stages leads to a quality score for the image which, as we shall soon demonstrate, correlates quite well with human perception and is competitive with leading FR IQA algorithms. The proposed approach we call Distortion

Identification-based Image Verity and INtegrity Evaluation (DIIVINE). The name is appropriate as the algorithm resulting from the modeling framework succeeds at divining image quality without any reference information or the benefit of distortion models.

Before proceeding, we state some salient aspects of DIIVINE. Present-day NR IQA algorithms are distortion-specific, i.e., the algorithm is capable of assessing the quality of images distorted by a particular distortion type. For example, the algorithm in [13] is for JPEG compressed images, that in [14] is for JPEG2000 compressed images, and that in [15] is for blur. DIIVINE, however, is not bound by the distortion-type affecting the image since we do not seek distortion-specific indicators of quality (such as edge strength at block boundaries) but provide a modular strategy that adapts itself to the distortion in question. Indeed, our framework is ostensibly distortion-agnostic.

Li proposed a series of heuristic measures to characterize image quality based on three quantitiesedge sharpness, random noise level (impulse/additive white Gaussian noise), and structural noise [29]. Edge sharpness is measured using an edge-detection approach, while the random noise level is measured using a local smoothness approach (impulse noise) and PDE-based model (Gaussian noise). Structural noise as defined by Li relates to blocking and ringing from compression techniques such as JPEG and JPEG2000. Unfortunately, the author does not analyze the performance of the proposed measures, nor propose a technique to combine the measures to produce a general-purpose quality assessment algorithm.

3.1 Control essentials

Once the deep neural network is trained with sufficient training data set, the model is fine tuned to predict the subjective scores. Since the objective error map is somewhat correlated with the subjective score, the second stage can be trained without great difficulty by using even a limited data set. In the end, our model can predict the subjective scores without accessing the groundtruth objective error maps during testing.

3.1.1 Image quality assessment

The outer-loop control system, which is typically located in the ground station, gives the pilot the ability to deliver a desired mission, such as following a trajectory. This can be done through transmitting a simple set of navigational commands via a telemetry link. Employing a high-level control system is indispensable in aircraft that operate in the first or second autonomous mode, as these systems rely on external instructions to function. Therefore, designing a robust outer-loop control system is essential.

Once a network has been structured for a particular application, that network is ready to be trained. To start this process, the initial weights (described in the next section) are chosen randomly. Then the training (learning) begins.The network processes the records in the Training Set one at a time, using the weights and functions in the hidden layers, then compares the resulting outputs against the desired outputs. Errors are then propagated back through the system, causing the system to adjust the weights for application to the next record. This process occurs repeatedly as the weights are tweaked. During the training of a network, the same set of data is processed many times as the connection weights are continually refined.

The network processes the records in the Training Set one at a time, using the weights and functions in the hidden layers, then compares the resulting outputs against the desired outputs. Errors are then propagated back through the system, causing the system to adjust the weights for application to the next record. This process occurs repeatedly as the weights are tweaked. During the training of a network, the same set of data is processed many times as the connection weights are continually refined.

3.2 Training and feedback

Training is a sine qua non for BCI users, and in particular for controlled operators. The crucial importance of training could be viewed from two aspects. First, users can learn how to modulate their activity patterns, based on the system feedback, in order to gain a dexterous control over the aircraft. Second, the system can learn how to avoid future errors by co-adaptation via machine learning strategies. In a typical system, users go through an offline training phase

prior to running the experiment; however, coadaptive algorithms, offered by state-of-the-art classification, make it possible to fulfill the training online. Another determining factor in the training phase is paradigm for systems based on induced potentials that requires a more rigorous training compared to those using evoked potentials.

Overall, we resolve the NR-IQA problem by dividing it into the objective distortion and the HVS-related parts. In the objective distortion part, a pixelwise objective error map is predicted using the CNN model. In the HVS-related part, the model further learns the human visual perception behavior.

Gabrada and Cristobal proposed an innovative strategy for blind IQA which utilized the Renyi entropy measure [30] along various orientations to measure anisotropy. The proposed approach is attractive since natural images are anisotropic in nature and possesses statistical structure that distortions destroy. They measure mean, standard deviation, and range of the Renyi entropy along four pre-defined orientations in the spatial domain and demonstrate their correlation with perceived quality. Unfortunately, a thorough evaluation of the proposed measure is again lacking.

3.3 Image normalisation

As a preprocessing, the input images are first converted to grayscale, and they are subtracted from their low-pass filtered images. Let Ir be a reference image and Id be the corresponding distorted image. The normalized versions are then denoted by Ir and Id , respectively. The low-frequency image is obtained by downscaling the input image to 1/4 and upscaling it again to the original size, which is denoted by Ilowr and Ilowd . A Gaussian low-pass filter and subsampling were used to resize the images.There are two reasons for this

The network processes the records in the Training Set one at a time, using the weights and functions in the hidden layers, then compares the resulting outputs against the desired outputs. Errors are then propagated back through the system, causing the system to adjust the weights for application to the next record. This process occurs repeatedly as the weights are tweaked. During the training of a network, the same set of data is processed many times as the connection weights are continually refined.

Blind/NR video quality assessment (VQA) is an important problem that has followed a similar trajectory. Some authors have proposed techniques which measure blockiness, blur, corner outliers, and noise separately, and use a Minkowski sum to pool the measures of quality together [32], [33]. In both these approaches, distortion-specific indicators of quality are computed and pooled using a variety of pre-fixed thresholds and training, as against our approach that uses concepts from NSS to produce a modular and easily extensible approach that can be modified to include other distortions than those discussed here. We anticipate that the approach taken here could be eventually extended to video to achieve good results.

The obtained subband coefficients are then utilized to extract a series of statistical features. These statistical features are stacked to form a vector which is a statistical description of the distortion in the image. Our goal is to utilize these feature vectors across images to perform two tasks in sequence: 1) identify the probability that the image is afflicted by one of the multiple distortion categories, then 2) map the feature vector onto a quality score for each distortion category, i.e., build a regression model for each distortion category to map the features onto quality, conditioned on the fact that the image is impaired by that particular distortion category (i.e., distortion-specific QA). The probabilistic distortion identification estimate is then combined with the distortion-specific quality score to produce a final quality value for the image.

3.4 Reliability map prediction

When severe distortion is applied to an image and its high-frequency detail is lost, its error map obtains more high-frequency components. Meanwhile the distorted image does not have high-frequency details. Therefore, without the reference image, it is difficult to predict an accurate error map from the distorted image, in particular, on homogeneous regions. To avoid this problem, we propose deriving a reliability map by measuring textural strength to compensate for the inaccuracy of the error map.

Pre-processing techniques help to remove unwanted artifacts from the EEG signal and hence improve the signal to noise ratio. A pre-processing block aids in improving the performance of the system by separating the noise from the actual signal.preprocessing might

vary case by case. The simplest and most widely used method to remove artifacts and extract the desired frequency band is filtering. The most common filters include high-pass, low-pass, band-pass, and notch filter, which can be combined with other computational techniques such as independent component analysis (ICA) or common average reference to remove artifacts and improve signal to noice ratio.

3.4.1 Low pass filter

A low-pass filter is a filter that passes signals with a frequency lower than a selected cutoff frequency and attenuates signals with frequencies higher than the cutoff frequency. The exact frequency response of the filter depends on the filter design. The filter is sometimes called a high-cut filter, or treble-cut filter in audio applications. A low-pass filter is the complement of a high-pass filter.

Low-pass filters exist in many different forms, including electronic circuits such as a hiss filter used in audio, anti-aliasing filters for conditioning signals prior to analog-to-digital conversion, digital filters for smoothing sets of data, acoustic barriers, blurring of images, and so on. The moving average operation used in fields such as finance is a particular kind of low-pass filter, and can be analyzed with the same signal processing techniques as are used for other low-pass filters. Low-pass filters provide a smoother form of a signal, removing the short-term fluctuations and leaving the longer-term trend.

3.4.2 High pass filter

A high-pass filter is an electronic filter that passes signals with a frequency higher than a certain cutoff frequency and attenuates signals with frequencies lower than the cutoff frequency. The amount of attenuation for each frequency depends on the filter design. A high-pass filter is usually modeled as a linear time-invariant system. It is sometimes called a low-cut filter or bass-cut filter. High-pass filters have many uses, such as blocking DC from circuitry sensitive to non-zero average voltages or radio frequency devices. They can also be used in conjunction with a low-pass filter to produce a bandpass filter.

In signal processing, a band-stop filter or band-rejection filter is a filter that passes most

frequencies unaltered, but attenuates those in a specific range to very low levels.[1] It is the opposite of a band-pass filter. A notch filter is a band-stop filter with a narrow stopband (high Q factor).

Narrow notch filters (optical) are used in Raman spectroscopy, live sound reproduction (public address systems, or PA systems) and in instrument amplifiers (especially amplifiers or preamplifiers for acoustic instruments such as acoustic guitar, mandolin, bass instrument amplifier, etc.) to reduce or prevent audio feedback, while having little noticeable effect on the rest of the frequency spectrum (electronic or software filters). Other names include 'band limit filter', 'T-notch filter', 'band-elimination filter', and 'band-reject filter'.Typically, the width of the stopband is 1 to 2 decades (that is, the highest frequency attenuated is 10 to 100 times the lowest frequency attenuated). However, in the audio band, a notch filter has high and low frequencies that may be only semitones apart.

3.5 Feature extraction

feature extraction block helps to retrieve the most relevant features from the signal. These features will aid the decision making mechanism in giving the desired output.In machine learning feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is a dimensionality reduction process, where an initial set of raw variables is reduced to more manageable groups (features) for processing, while still accurately and completely describing the original data set.

When the input data to an algorithm is too large to be processed and it is suspected to be redundant (e.g. the same measurement in both feet and meters, or the repetitiveness of images presented as pixels), then it can be transformed into a reduced set of features (also named a feature vector). Determining a subset of the initial features is called feature selection. The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data.

The training process normally uses some variant of the Delta Rule, which starts with the calculated difference between the actual outputs and the desired outputs. Using this error, connection weights are increased in proportion to the error times, which are a scaling factor for global accuracy. This means that the inputs, the output, and the desired output all must be present at the same processing element. The most complex part of this algorithm is determining which input contributed the most to an incorrect output and how must the input be modified to correct the error. (An inactive node would not contribute to the error and would have no need to change its weights.)

To solve this problem, training inputs are applied to the input layer of the network, and desired outputs are compared at the output layer. During the learning process, a forward sweep is made through the network, and the output of each element is computed by layer. The difference between the output of the final layer and the desired output is back-propagated to the previous layer(s), usually modified by the derivative of the transfer function. The connection weights are normally adjusted using the Delta Rule. This process proceeds for the previous layer(s) until the input layer is reached.

3.6 Learning objective error map

In the first stage of training, the objective error maps are used as proxy regression targets to get the effect of increasing data. The loss function is defined by the mean squared error between the predicted and ground-truth error maps.

When the energy of the signal is concentrated around a finite time interval, especially if its total energy is finite, one may compute the energy spectral density. More commonly used is the power spectral density (or simply power spectrum), which applies to signals existing over all time, or over a time period large enough (especially in relation to the duration of a measurement) that it could as well have been over an infinite time interval. The spectral density then refers to the spectral energy distribution that would be found per unit time, since the total energy of such a signal over all time would generally be infinite.

3.6.1 Phase coherency

The spectral coherence is a statistic that can be used to examine the relation between two signals or data sets. It is commonly used to estimate the power transfer between input and output of a linear system. If the signals are ergodic, and the system function linear, it can be used to estimate the causality between the input and output.

3.6.2 Common spatial pattern

Common spatial pattern (CSP) is a mathematical procedure used in signal processing for separating a multivariate signal into additive subcomponents which have maximum differences in variance between two windows

3.7 Subjective error map prediction

Once the model is trained to predict the objective error maps, we move to the next training stage, where DIQA is trained to predict subjective scores. To achieve this, the trained subnetwork f () is connected to a global average pooling layer followed by the fully connected layers. The feature map is averaged over spatial domain leading to a 128-D feature vector.

Since zeros are padded before each convolution, the feature maps near the borders tend to be zeros. Therefore, during the minimization of the loss functions in (3) and (6), we ignored pixels near borders around the error and the perceptual error maps. Each of four rows or columns for each border was excluded in the experiment, which compensated for information loss in the last two convolutional layers.

In the field of machine learning, the goal of statistical classification is to use an object's characteristics to identify which class (or group) it belongs to. A linear classifier achieves this by making a classification decision based on the value of a linear combination of the characteristics. An object's characteristics are also known as feature values and are typically presented to the machine in a vector called a feature vector. Such classifiers work well for practical problems such as document classification, and more generally for problems with many variables (features), reaching accuracy levels comparable to non-linear classifiers while taking less time

to train and use. linear methods can only solve problems that are linearly separable (usually via a hyperplane).

3.7.1 Neural networks

Artificial neural networks are relatively crude electronic networks of neurons based on the neural structure of the brain. They process records one at a time, and learn by comparing their classification of the record (i.e., largely arbitrary) with the known actual classification of the record. The errors from the initial classification of the first record is fed back into the network, and used to modify the networks algorithm for further iterations. A neuron in an artificial neural network is

1. A set of input values (xi) and associated weights (wi).
2. A function (g) that sums the weights and maps the results to an output (y).

Neurons are organized into layers: input, hidden and output. The input layer is composed not of full neurons, but rather consists simply of the record's values that are inputs to the next layer of neurons. The next layer is the hidden layer. Several hidden layers can exist in one neural network. The final layer is the output layer, where there is one node for each class. A single sweep forward through the network results in the assignment of a value to each output node, and the record is assigned to the class node with the highest value.

Training an artificial neural network

In the training phase, the correct class for each record is known (termed supervised training), and the output nodes can be assigned correct values – 1 for the node corresponding to the correct class, and 0 for the others. (In practice, better results have been found using values of 0.9 and 0.1, respectively.) It is thus possible to compare the network's calculated values for the output nodes to these correct values, and calculate an error term for each node (the Delta rule). These error terms are then used to adjust the weights in the hidden layers so that, hopefully, during the next iteration the output values will be closer to the correct values.

the kernel is chosen be centered on a unit (odd size), to have sufficient overlap to not lose information (3 would be too small with only one unit overlap), but yet to not have redundant

computation (7 would be too large, with 5 units or over 70 percent overlap). A convolution kernel of size 5 is shown in Figure 4. The empty circle units correspond to the subsampling and do not need to be computed. Padding the input (making it larger so that there are feature units centered on the border) did not improve performance significantly. With no padding, a subsampling of 2, and a kernel size of 5, each convolution layer reduces the feature size from n to (n-3)/2. Since the initial MNIST input size 28x28, the nearest value which generates an integer size after 2 layers of convolution is 29x29. After 2 layers of convolution, the feature size of 5x5 is too small for a third layer of convolution. The first feature layer extracts very simple features, which after training look like edge, ink, or intersection detectors. We found that using fewer than 5 different features decreased performance, while using more than 5 did not improve it. Similarly, on the second layer, we found that fewer than 50 features (we tried 25) decreased performance while more (we tried 100) did not improve it. These numbers are not critical as long as there are enough features to carry the information to the classification layers (since the kernels are 5x5, we chose to keep the numbers of features multiples of 5).

The iterative learning process

A key feature of neural networks is an iterative learning process in which records (rows) are presented to the network one at a time, and the weights associated with the input values are adjusted each time. After all cases are presented, the process is often repeated. During this learning phase, the network trains by adjusting the weights to predict the correct class label of input samples. Advantages of neural networks include their high tolerance to noisy data, as well as their ability to classify patterns on which they have not been trained. The most popular neural network algorithm is the back-propagation algorithm proposed in the 1980s.

Once a network has been structured for a particular application, that network is ready to be trained. To start this process, the initial weights (described in the next section) are chosen randomly. Then the training (learning) begins.

The network processes the records in the Training Set one at a time, using the weights and functions in the hidden layers, then compares the resulting outputs against the desired outputs. Errors are then propagated back through the system, causing the system to adjust the

weights for application to the next record. This process occurs repeatedly as the weights are tweaked. During the training of a network, the same set of data is processed many times as the connection weights are continually refined.

Note that some networks never learn. This could be because the input data does not contain the specific information from which the desired output is derived. Networks also will not converge if there is not enough data to enable complete learning. Ideally, there should be enough data available to create a Validation Set.

Feedforward, back-propagation

The feedforward, back-propagation architecture was developed in the early 1970s by several independent sources (Werbor; Parker; Rumelhart, Hinton, and Williams). This independent co-development was the result of a proliferation of articles and talks at various conferences that stimulated the entire industry. Currently, this synergistically developed back-propagation architecture is the most popular model for complex, multi-layered networks. Its greatest strength is in non-linear solutions to ill-defined problems.

The typical back-propagation network has an input layer, an output layer, and at least one hidden layer. There is no theoretical limit on the number of hidden layers but typically there are just one or two. Some studies have shown that the total number of layers needed to solve problems of any complexity is five (one input layer, three hidden layers and an output layer). Each layer is fully connected to the succeeding layer.

The training process normally uses some variant of the Delta Rule, which starts with the calculated difference between the actual outputs and the desired outputs. Using this error, connection weights are increased in proportion to the error times, which are a scaling factor for global accuracy. This means that the inputs, the output, and the desired output all must be present at the same processing element. The most complex part of this algorithm is determining which input contributed the most to an incorrect output and how must the input be modified to correct the error. (An inactive node would not contribute to the error and would have no need to change its weights.)

To solve this problem, training inputs are applied to the input layer of the network, and

desired outputs are compared at the output layer. During the learning process, a forward sweep is made through the network, and the output of each element is computed by layer. The difference between the output of the final layer and the desired output is back-propagated to the previous layer(s), usually modified by the derivative of the transfer function. The connection weights are normally adjusted using the Delta Rule. This process proceeds for the previous layer(s) until the input layer is reached.

3.7.2 Convolutional neural network

The netcontains eight layers with weights; the first five are convolutional and the remaining three are fullyconnected. The output of the last fully-connected layer is fed to a 1000-way softmax which producesa distribution over the 1000 class labels. Our network maximizes the multinomial logistic regression objective, which is equivalent to maximizing the average across training cases of the log-probability of the correct label under the prediction distribution.

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.Both for classification and regression, a useful technique can be used to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor.The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.A peculiarity of the k-NN algorithm is that it is sensitive to the local structure of the data

Generally speaking, the deep network facilitates the proposed method in two ways. On one hand, deep network is an efficient way to represent highly varying functions. It can mine the inherent structure of data without labels, which is inspired by the fact that humans heavily use unsupervised learning. Therefore, it would be an excellent model for learning the highly varied mapping between visual stimuli and quality, because the human perception of quality has an extremely strong nonlinearity, and researchers still have inadequate insight into

its mechanism. On the other hand, the deep network has a stronger power of generalization than shallow methods, especially when training samples are limited. In this case, depth and pre-training act as a smart regularization choice to help the model prevent overfitting. When the training set is small, even shallow machine learning methods can fit the training set perfectly, but they generalize poorly.

3.8 Non linear regression

To study and analyze what was learned by a deep model, we additionally propose a variant version of the DIQA, named DIQA-SENS. Different from the normal model, the DIQA-SENS contains three paths: 1) objective error map prediction, 2) sensitivity map prediction, and 3) reliability map prediction. The overview of DIQA-SENS is shown in Fig. 7. The same architecture of the DIQA (from Conv1 to Conv9 in Fig. 2) is used for the objective error map prediction and the sensitivity map prediction subnetworks, respectively. The hidden layer of the fully connected layer has 20 perceptrons. Its training scheme is similar to that of the DIQA. However, the objective error map prediction subnetwork is frozen while the sensitivity map prediction subnetwork is updated. In addition, the perceptual error map is obtained by multiplying the predicted error and sensitivity maps, and then directly regressed onto the subjective score.

The kernels of the second, fourth, and fifth convolutional layers are connected only to those kernel maps in the previous layer which reside on the same GPU (see Figure 2). The kernels of the third convolutional layer are connected to all kernel maps in the second layer. The neurons in the fullyconnected layers are connected to all neurons in the previous layer. Response-normalization layers follow the first and second convolutional layers. Max-pooling layers, of the kind described in Section 3.4, follow both response-normalization layers as well as the fifth convolutional layer. The ReLU non-linearity is applied to the output of every convolutional and fully-connected layer.

The first form of data augmentation consists of generating image translations and horizontal reflections. We do this by extracting random 224 224 patches (and their horizontal reflections) from the 256256 images and training our network on these extracted patches4 .

This increases the size of our training set by a factor of 2048, though the resulting training examples are, of course, highly interdependent. Without this scheme, our network suffers from substantial overfitting, which would have forced us to use much smaller networks. At test time, the network makes a prediction by extracting five 224 224 patches (the four corner patches and the center patch) as well as their horizontal reflections (hence ten patches in all), and averaging the predictions made by the networks softmax layer on the ten patches.

3.9 Handcrafted features

To investigate the contribution of each module and training scheme, we utilized different combinations of them. In other words, each ablated model needs to be repeatedly trained and tested while dividing training and testing sets randomly, which leads to a huge training time of deep CNN models. Thus, learning and testing were conducted on LIVE IQA and CSIQ, which are reasonable with respect to the data size.

Several ad hoc strategies have been used in previous studies to evaluate the performance of developed brain-control UAVs. LaFleur et al. used ITR metric based on Shanons work, customized for asynchronous BCI. In many of these studies, performance assessment is task specific and is measured intrinsically for each subject. A universal framework for assessing the performance of brain-controlled UAVs is yet to be established

Regardless of their underlying configuration, the basic principle that governs different hybrid BCIs remains the same; one of the engaged modalities is counterbalancing limitations of the other and vice versa. In these systems, the grand hierarchy determines whether different components are engaged concurrently or sequentially. In the first case, recorded signals from different sources are processed in parallel and subsequently mapped to control commands. The simultaneous use of different input modalities could significantly improve the BCI bit-rate and the information throughput. On the other hand, in sequential hybrid BCIs, the output of one modality serves as the input for others. The practical merit of sequential hybrid BCIs is twofold. First, they could be designed in a way that enables the operator to selectively initiate/terminate other control processes, e.g., a sequential ERS-based brain switch could be used to turn ON/OFF an SSVEP BCI. Second, by the same token, a sequential hybrid BCI could

enable the operator to rectify, reinforce, or cancel the issued control commands and, therefore, reduce the overall false positive rate of the system. Numerous studies were dedicated to review the design and application of the hybrid BCI.

the kernel is chosen be centered on a unit (odd size), to have sufficient overlap to not lose information (3 would be too small with only one unit overlap), but yet to not have redundant computation (7 would be too large, with 5 units or over 70of size 5 is shown in Figure 4. The empty circle units correspond to the subsampling and do not need to be computed. Padding the input (making it larger so that there are feature units centered on the border) did not improve performance significantly. With no padding, a subsampling of 2, and a kernel size of 5, each convolution layer reduces the feature size from n to (n-3)/2. Since the initial MNIST input size 28x28, the nearest value which generates an integer size after 2 layers of convolution is 29x29. After 2 layers of convolution, the feature size of 5x5 is too small for a third layer of convolution. The first feature layer extracts very simple features, which after training look like edge, ink, or intersection detectors. We found that using fewer than 5 different features decreased performance, while using more than 5 did not improve it. Similarly, on the second layer, we found that fewer than 50 features (we tried 25) decreased performance while more (we tried 100) did not improve it. These numbers are not critical as long as there are enough features to carry the information to the classification layers (since the kernels are 5x5, we chose to keep the numbers of features multiples of 5).

3.10 Performance evaluation

. Since regions with low reliability were ignored during the training of the error map, the prediction was inaccurate on homogeneous regions. In the fourth and fifth columns, the ground-truth and predicted error maps are multiplied by the reliability maps as in (4), such that the inaccurate regions were ignored. The images in the last column are images actually used for NR-IQA.

The predicted perceptual error maps are shown in Fig. 10. The reliability maps [Fig. 10(b), (g), and (l)] emphasized highfrequency components, such as edges and complex textures. To analyze human visual sensitivity, we observed the perceptual error map rather than the

sensitivity map. The role of the sensitivity map is tuning the objective error map by weighting. It is clear that low values in the perceptual error map can be regarded as perceptually distorted regions. However, it is difficult to ensure that low values in the sensitivity map indicate less important pixels, since the SNES subnetwork was trained based on the prepredicted error map.

The obtained subband coefficients are then utilized to extract a series of statistical features. These statistical features are stacked to form a vector which is a statistical description of the distortion in the image. Our goal is to utilize these feature vectors across images to perform two tasks in sequence: 1) identify the probability that the image is afflicted by one of the multiple distortion categories, then 2) map the feature vector onto a quality score for each distortion category, i.e., build a regression model for each distortion category to map the features onto quality, conditioned on the fact that the image is impaired by that particular distortion category (i.e., distortion-specific QA). The probabilistic distortion identification estimate is then combined with the distortion-specific quality score to produce a final quality value for the image.

The distorted image is first decomposed using a scale-space-orientation decomposition (loosely, a wavelet transform) to form oriented band-pass responses. The obtained subband coefficients are then utilized to extract a series of statistical features. These statistical features are stacked to form a vector which is a statistical description of the distortion in the image. Our goal is to utilize these feature vectors across images to perform two tasks in sequence: 1) identify the probability that the image is afflicted by one of the multiple distortion categories, then 2) map the feature vector onto a quality score for each distortion category, i.e., build a regression model for each distortion category to map the features onto quality, conditioned on the fact that the image is impaired by that particular distortion category (i.e., distortion-specific QA). The probabilistic distortion identification estimate is then combined with the distortion-specific quality score to produce a final quality value for the image in fig 3.2.

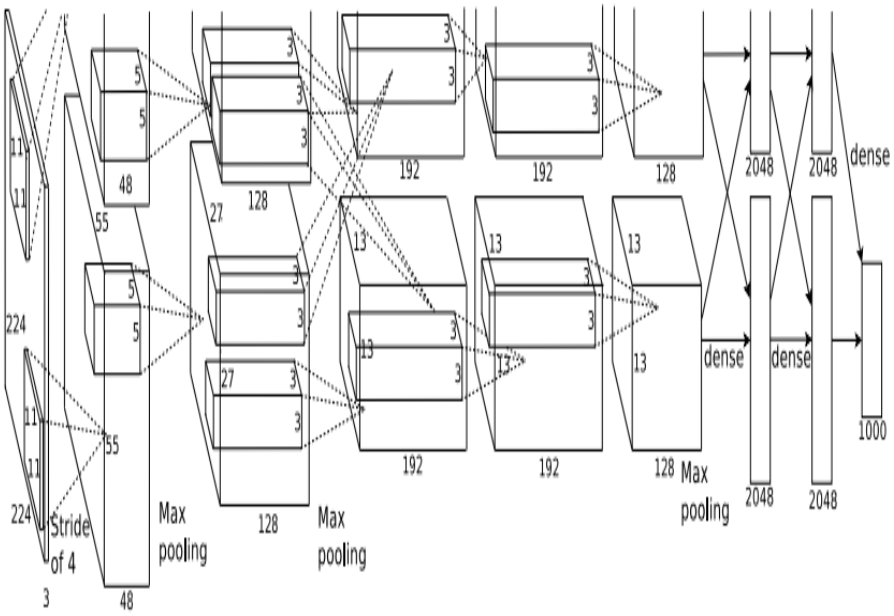


Fig. 3.2: Neural network

Conclusion

We described a deep CNN-based NR-IQA framework. Applying a CNN to NR-IQA is a challenging issue, because there are critical obstacles. In the DIQA, an objective error map was used as an intermediate regression target to avoid overfitting with the limited database.

When the first training stage is not run enough, the DIQA suffers from the overfitting problem leading to a degradation of performance. The inputnormalization and the reliability map increased the accuracy.

The final DIQA model outperformed all the benchmarked full-reference methods as well as noreference methods. We further showed that the performance of the DIQA is independent of the selection of the database. We additionally proposed the DIQA-SENS to visualize and analyze the learned perceptual error maps. The perceptual error maps followed the behavior of the HVS. In the future, we will investigate a new way to obtain more meaningful sensitivity maps that can provide a more interpretable analysis with respect to the HVS.

REFERENCES

[1] A. K. Moorthy and A. C. Bovik, Blind image quality assessment: From natural scene statistics to perceptual quality, IEEE Trans. Image Process., vol. 20, no. 12, pp. 33503364, Dec. 2011.

[2] M. A. Saad, A. C. Bovik, and C. Charrier, Blind image quality assessment: A natural scene statistics approach in the DCT domain, IEEE Trans. Image Process., vol. 21, no. 8, pp. 33393352, Aug. 2012.

[3] W. Hou, X. Gao, D. Tao, and X. Li, Blind image quality assessment via deep learning, IEEE Trans. Neural Netw. Learn. Syst., vol. 26, no. 6, pp. 12751286, Jun. 2015.

[4] H. Cecotti, Spelling with non-invasive braincomputer interfaces Current and future trends, J. Physiol, Paris, vol. 105, no. 1, pp. 106114, 2011.

[5] Y. Li et al., No-reference image quality assessment with shearlet transform and deep neural networks, Neurocomputing, vol. 154, pp. 94109, Apr. 2015.

[6] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, IEEE Trans. Image Process., vol. 15, no. 11, pp. 34403451, Nov. 2006.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2009, pp. 248255.

[8] S. J. Daly, The visible differences predictor: An algorithm for the assessment of image fidelity, Proc. SPIE, vol. 1666, pp. 179206, Jan. 1992