Computing and Informatics, Vol. 39, 2020, 385–409, doi: 10.31577/cai_2020_3_385

PERCEPTUAL QUALITY ASSESSMENT OF DIGITAL IMAGES USING DEEP FEATURES

Nisar Ahmed

Department of Computer Engineering
University of Engineering and Technology, Lahore
Pakistan
e-mail: nisarahmedrana@yahoo.com

Hafiz Muhammad Shahzad Asif

Department of Computer Science
University of Engineering and Technology, Lahore
Pakistan
e-mail: shehzad@uet.edu.pk

Abstract. Perceptual quality assessment is a tough task especially in the absence of reference information. No-reference image quality assessment is more challenging than full-reference or reduced reference methods, as the system has to model the different image distortions in the form of a quality score. Most of the approaches are based on handcrafted features which are based on natural scene statistics and are specific to some distortion types. These approaches provide high correlation with human opinion score for datasets containing specific distortions, but they fail to generalize well in scenarios were multiple distortions or real-time distortions are present in images. Deep learning algorithms, on the other hand, demonstrated their abilities to learn expert features with better discriminatory power for various classification and regression tasks. It is a big challenge to use those deep learning methods for image quality assessment as the image datasets with human opinion score are very small and cannot be used effectively to train a deep learning algorithm. We experimented with activations of different deep layers of thirteen pre-trained models and checked for their suitability for the task of no-reference quality assessment. Fine-tuning of these models on quality assessment datasets provided even better performance. A Gaussian process regression model is trained on these activations to perform the quality assessment and it provided state-of-the-art performance. Cross-dataset validation demonstrated its performance further and also provided further prospects of research in this direction.

Keywords: Image quality, image quality assessment, IQA, deep features, perceptual quality assessment

Mathematics Subject Classification 2010: 60Gxx

1 INTRODUCTION

Assessment of perceptual quality of digital images can be very useful in several image processing applications. Image quality assessment algorithms can be used to monitor the video or image quality to optimize the parameters of image-processing algorithms. Such system can be used to adjust compression ratio, amount of color saturation, contrast adjustment, etc. It can also be used to evaluate the performance of image acquisition hardware and amount of perceptual distortions occurring during transmission. Usually the above applications are based on full-reference methods which require the original image to compare the extracted features with distorted image for assessment of amount of distortion, but this approach has practical limitations. Full-reference methods can be used to assess the compression performance, but they cannot be used to assess the perceptual distortion in transmitted video when the original one is not available (e.g. broadcasting) or in case of evaluation of image acquisition/enhancement. Moreover, full-reference approaches measure the amount of change in original and distorted image but in case of image enhancement applications such as contrast enhancement the perceptual quality of reproduced image is better than the original image and full-reference approaches fail to provide quality score for these applications.

No-reference image quality assessment is therefore crucial for several image processing systems for evaluation of perceptual image quality. These approaches do not require any reference image for comparison but they extract or learn discriminatory features from images which can be used to assess the perceptual quality. On the contrary, due to lack of information, it is harder for no-reference image quality assessment algorithms to assess the perceptual quality better than full-reference approaches. It is therefore more difficult for no-reference approaches to adapt to the behavior of human visual system and result in decreased prediction performance in terms of correlation with Mean Opinion Score (MOS).

No-reference quality assessment is usually performed by extracting handcrafted features such as natural scene statistics and then training a regression algorithm to obtain the quality score. Many no-reference image quality assessment algorithms are proposed in the literature and development of a robust quality assessment algorithm is dependent on quality discrimination ability of the features. Mittal et al. [1], Liu et al. [2] and Sazzad et al. [3] extracted features in the spatial domain. Saad et al. [4], Ma et al. [5] and Liu et al. [6] worked on transform domain features. He at el. [7] and Chang et al. [8] used sparse representation for image quality assessment. These approaches can predict perceptual quality in high correlation to the human judgments.

Deep features, on the other hand, are the activations of convolutional neural networks which are extracted to perform different classification and regression tasks [9, 10, 7, 11]. These features demonstrated very powerful capabilities for image quality assessment task as well [12, 13, 14, 15]. There are two approaches to extract deep features, one is to use the pre-trained CNN model and extract deep features [9, 15] and the other is to fine-tune the model on your problem set and then perform the feature extraction [10]. The second approach is required on the problem of image quality as it is of different nature than the original dataset (ImageNet) which is used in the pre-trained model. Some researchers designed their own architecture [16, 17, 18] or used a previously designed architecture to train from scratch for feature extraction. The deep feature based approaches are much better at predicting perceptual image quality than the handcrafted feature based. However, there is no clarity as how to obtain most representative feature set for perceptual image quality assessment.

In this work, we have performed an analysis of different deep features to identify the most representative feature set for image quality assessment. Our contributions are twofold, the first involves identification of most suitable way of deep feature extraction and the second is construction of a quality prediction model by using Gaussian process regression (GPR). The identification of deep features extraction method is performed by first selecting thirteen popular CNN architectures, pre-trained on ImageNet, and performing feature extraction at different bottleneck layers. Eight of these pre-trained models are fine-tuned on image quality database and then deep features are extracted in a similar manner. The best performing feature set is used to train a GPR as it has been demonstrated that the GPR is the most suitable regression algorithm for the task of image quality assessment. The specific contributions are highlighted below:

- We have provided a comparison of deep features performance for image quality using several popular pre-trained models with and without fine-tuning.
- We have extracted deep features at several bottleneck points and provided three best points to extract features in these architectures for image quality.
- We have highlighted and used NASNet after fine-tunning for feature extraction as it provided most quality aware features.
- We have proposed a Gaussian process regression based model trained using deep features to obtain state-of-the-art performance on several benchmark databases.

2 RELATED WORK

Bosse et al. [19] presented a deep neural network based quality assessment approach. They follow configuration of a Siamese network in which the differences of extracted features for original and distorted images are taken and features are fused to perform regression with fully connected neural network. The whole process is applied in patch-wise fashion and weighted averaging is used for final quality score. As their

approach requires both pristine and distorted image, their approach is only useful for full-reference quality assessment.

Zhang et al. [15] presented a study of effectiveness of deep features for image quality assessment. They described the use of VGG, AlexNet and SqueezeNet for extraction of deep features and demonstrated that they are very good at prediction of image quality. They presented a new dataset of images with 'just noticeable difference' images to demonstrate the performance of deep networks. They demonstrated with VGG pre-trained model on their dataset that deep features are superior to almost all of the models utilizing handcrafted features with full-reference or reduced reference. It is to highlight that only VGG network provides a rich feature representation among the three used architectures but the more advanced architectures with deeper representation may prove more useful for the task of image quality assessment.

Gao et al. [9] has presented a deep CNN based image quality predictor. They have used VGG model pre-trained on 1 000 ImageNet categories. They extracted the deep features at each layer and trained a Support Vector Regression (SVR) on each of the deep features. These SVR are combined to form an ensemble to predict the image quality. They tested their model on several benchmark datasets and reported comparable performance. The idea behind their approach is very naïve. But the issue is that combining the deep features from all the layers results in a very large feature set size which is computationally expensive at one end and has a very large feature space at the other end. This large feature space will easily overfit the model rather than learning a more generalized form because the training database used in their experiment is relatively small.

Bianco et al. [17] proposed the use of convolutional neural networks for the task of image quality assessment. They used features extracted from the layers of CNN and also proposed their own architecture for quality prediction. Their final proposal is a deep feature extractor and these features are pooled and provided to an SVR for quality assessment. Multiple crops of an image are used and their estimated scores are averaged to provide the final quality score. They have demonstrated their model on five benchmark datasets and claimed comparable performance. Their architecture is very primitive, but they have used different learning strategies to improve the prediction performance. These learning strategies combined with a deeper and representative architecture may prove helpful in obtaining better prediction performance.

Fan et al. [18] proposed a CNN based two stage image quality assessment approach. The first phase identifies the type of distortion present in the image and the second stage contains multiple image quality assessment modules trained for each distortion type. The quality score is provided based on the distortion type identified in the previous stage. This approach can be used with success for some specific distortion types only and cannot be applied to naturally distorted images which contain number of image distortions occurring simultaneously.

Guan et al. [10] proposed a deep features based image quality assessment approach. The first step in their approach performs spatial sampling and the next

stage performs the feature extraction through CNN. There are two configurations, one performs 5×5 and then 7×7 convolution and the other performs 7×7 convolutions and their activations are concatenated to obtain the final feature vector. The next layer performs patch-wise quality assessment and weight learning for the specific patch. The last layer finds the global image quality by weighted addition of patch-wise quality. Their approach uses bilinear pooling by extracting features with different sized filters, but the depth of CNN architecture is not very deep and better representations cannot be obtained.

Bosse et al. [20] proposed a deep neural network based image quality assessment approach. The input image is divided into patches and quality is estimated for each patch and final score is obtained by averaging these quality scores. They have provided another architecture for patch-wise weighted aggregation which uses an additional fully-connected layer to learn the weight for each patch and then patch-wise weighted averaging is used for score calculation, and it has shown superior performance over the other. The patch based CNN models are easier to train and can be used on variable sized inputs by combining the scores of multiple image crops.

Bare et al. [16] proposed a specialized CNN for image quality assessment. They have used six convolutional layers along with skip connections and sum layers in their architecture. The output of last sum layer is provided to a fully connected layer of 1 024 and regression score is obtained to indicate quality for a single patch. The overall quality can be obtained by averaging over the patch estimates. The drawback of these approaches is that the proposed architecture cannot be completely trained using small image database and there is high probability of overfitting.

Hou et al. [11] proposed a deep image quality assessment approach based on qualitative scoring. Their work is based on the premise that humans prefer to provide quality judgment qualitatively rather than quantitatively, so following the qualitative approach would be more beneficial. Natural Scene Statistics (NSS) features are provided to deep belief network to learn qualitative representations which are then converted to quantitative scores for further utilization. They have presented a new direction of research in the area of image quality assessment and more work is required to improve its efficacy.

There are two approaches to perform image quality assessment,

1. handcrafted features based and
2. deep features based.

It has been demonstrated experimentally that deep features based approaches are better at quality assessment keeping in view the complexity of factors affecting the perceptual quality of an image. It can be observed from the literature that there is no consensus in the use of a pre-trained model for deep features extraction. Most of the researchers has used primeval architectures such as VGG and AlexNet, but it is not clear whether it is the best pre-trained model or some other pre-trained

model can perform better. Moreover, some authors has presented their own architectures which are inspired from AlexNet or have entirely different architecture. It is highlighted that these architectures cannot be optimally trained keeping in view the size of database they used for its training (typically containing up to 3 000 images). Therefore, we have found the need to highlight the efficacy of deep features extracted from the popular pre-trained models. We have extracted deep features at several bottleneck points and presented the three best layers for their extraction, their relative performance and the size of feature set to highlight the computational complexity. Moreover, the deep features are extracted with and without fine-tuning and important observations are highlighted to guide the reader about the architecture which provide the most quality aware features.

3 METHODOLOGY

Assessment of image quality in the absence of reference information is a complex task. No-reference quality assessment of digital images is a subjective task and there is a slight variation in quality score provided by different humans based on content and type of distortions. Therefore, subjective opinion of a number of humans is obtained and averaged to obtain MOS. Development of a machine learning model to perform quality assessment is therefore modeling of Human Visual System (HVS). The conventional approach towards this task is extraction of Natural Scene Statistics (NSS) in either spatial domain [1, 2, 3] or spectral domain [2, 4, 5, 6, 21] and then training of a regression algorithm such as SVR. The performance of the trained model is therefore based on feature set obtained through NSS. A quality aware feature set will provide better estimate of perceptual quality. HVS is a naïvely understood system and therefore its modeling through extraction of NSS is a tough task and there is always room for improvement.

Convolutional Neural Networks demonstrated expertise in the area of visual recognition in the past few years. They automatically learn the discriminatory features and perform the task of visual recognition with high accuracy and achieve/beat the human level accuracy. Some researchers [19, 17, 18, 10, 16, 11] have tried to utilize CNN for the task of image quality assessment. Their works use expertly curated deep learning architectures as well as pre-trained CNN models which are originally trained for the task of visual recognition. As the datasets with subjective score (i.e. MOS) are limited due to involvement of subjective scoring nature and the largest dataset with MOS contains 3 000 images [22]. Training from scratch with 3 000 images cannot be performed successfully on deeper models therefore shallow CNN are designed for this task which lack the impressive performance provided by CNN in visual recognition. Moreover, the pre-trained networks are originally designed for object recognition and their fine-tuning provides better performance than NSS based methods but it still needs improvement.

Recently some researchers have explored the use of activations of deep CNN layers [19, 15, 9, 17, 10] for training of a SVR to perform the quality assessment

task and achieved an impressive performance. These activations of deep layers are also referred to as deep features and have shown a good performance in several tasks which have a small dataset size and are different in nature than the pre-trained model itself. One of these researchers explored AlexNet, VGG and SqueezeNet and extracted the activations of its fully connected layers to estimate quality. Similarly, the other works principally focus on the pre-trained VGG model, as it has been demonstrated that it provides a rich representation of image content. Gao et al. [9] on the other hand extracted deep features from each layer of VGG and trained an SVR for each of them and constructed an ensemble to perform image quality estimation.

3.1 Deep Features Extraction from Pre-Trained CNNs

We have opted the approach of deep features to train an image quality assessment model. In contrast to the previous work, we have explored a number of pre-trained models to select the one with most quality aware features. We have performed an extensive experimentation on thirteen pre-trained models and extracted the activations at several deep layers of these CNN models instead of the last fully-connected layer. Table 1 provides the list of pre-trained CNN models along with three best performing layers of each model with Root Mean Squared Error (RMSE), Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Order Correlation Coefficient (SROCC) and Kendall Rank Order Correlation Coefficient (KROCC). Size of feature vector for each layer is also provided to indicate the corresponding complexity of feature space.

Some observations based on Table 1 are highlighted below:

1. Final activations did not provide the most quality aware features as these are more focused on visual recognition. However, activations of the layers earlier in the CNN are better suited for the task of image quality assessment.
2. Deeper networks with more number of filters provided better image quality assessment performance as compared to networks with lesser number of filters.
3. Networks with better visual recognition performance are also better for image quality assessment task, especially the networks which have larger number of filters.

3.2 Deep Features Extraction after Fine-Tuning the Pre-Trained CNNs

Pre-trained CNNs are trained on ImageNet visual recognition dataset which has images of objects falling in 1 000 different categories. CNNs trained on these images have learned the features which are most suitable to the task of visual recognition but may not perform very well on image quality assessment task. It is therefore attempted to fine-tune these CNNs on image quality assessment datasets and then extract the activations. The pre-trained CNN architecture is modified by removing

#	Architecture	RMSE	PLCC	SROCC	KROCC	Features	Layer Name
1	AlexNet	0.7729	0.7679	0.7435	0.5512	1 000	fc8
2	AlexNet	0.7784	0.7865	0.7700	0.5760	4 096	fc7
3	AlexNet	0.7648	0.7904	0.7761	0.5802	4 096	fc6
4	Vgg16	0.8148	0.8343	0.8214	0.6309	100 352	conv5_3
5	Vgg16	0.7879	0.8433	0.8354	0.6471	100 352	conv5_2
6	Vgg16	0.7274	0.8377	0.8327	0.6478	100 352	conv5_1
7	GoogleNet	0.7966	0.7483	0.7259	0.5359	1 000	loss3-classifier
8	GoogleNet	0.6680	0.8266	0.8176	0.6232	50 176	Inception5bOutput
9	GoogleNet	0.9313	0.6877	0.8241	0.6364	163 072	Inception4eOutput
10	SqueezeNet	0.8588	0.7956	0.7835	0.5878	100 352	fire8-concat
11	SqueezeNet	0.7875	0.8139	0.8065	0.6101	75 264	fire7-concat
12	SqueezeNet	0.7655	0.8107	0.8115	0.6181	75 264	fire6-connect
13	ShuffleNet	0.9566	0.6856	0.6597	0.4724	1 000	'node_202'
14	ShuffleNet	0.9127	0.6946	0.6477	0.4717	26 656	'node_198'
15	ShuffleNet	0.8708	0.7091	0.6642	0.4820	26 656	'node_174'
16	InceptionV3	0.7198	0.8036	0.7551	0.5706	131 072	mixed9
17	InceptionV3	0.6802	0.8286	0.7921	0.6122	81 920	mixed8
18	InceptionV3	0.7190	0.8082	0.7718	0.5911	221 952	mixed6
19	DenseNet201	0.6645	0.8375	0.8127	0.6328	94 080	conv5_block32 _concat
20	DenseNet201	0.6494	0.8413	0.8165	0.6362	92 512	conv5_block31 _concat
21	DenseNet201	0.6690	0.8393	0.8115	0.6289	90 944	conv5_block30 _concat
22	MobileNetV2	0.7317	0.8050	0.7717	0.5840	7 840	block_15_add
23	MobileNetV2	0.7416	0.7897	0.7654	0.5757	62 720	Conv_1
24	MobileNetV2	0.7283	0.7923	0.7680	0.5717	15 680	block_16_project
25	ResNet50	0.6422	0.8455	0.8317	0.6458	100 352	add_15
26	ResNet50	0.6449	0.8470	0.8320	0.6455	200 704	add_14
27	ResNet50	0.6608	0.8376	0.8310	0.6405	200 704	add_12
28	ResNet101	0.6758	0.8312	0.8156	0.6273	200 704	res5b
29	ResNet101	0.6980	0.8362	0.8265	0.6416	200 704	res5a
30	ResNet101	0.6819	0.8305	0.8102	0.6274	200 704	res4b21
31	Inception-ResNet-V2	0.6851	0.8216	0.7873	0.6006	133 120	block8_9
32	Inception-ResNet-V2	0.6871	0.8199	0.7856	0.6059	133 120	block8_8
33	Inception-ResNet-V2	0.6945	0.8171	0.7776	0.5953	133 120	block8_7
34	Xception	0.7786	0.7590	0.7094	0.5251	1 000	add_12
35	Xception	0.7443	0.7855	0.7500	0.5607	262 808	add_11
36	Xception	0.7502	0.7882	0.7462	0.5616	262 808	add_10
37	NASNet	0.8858	0.7347	0.7288	0.5385	1 000	predictions
38	**NASNet**	**0.5735**	**0.8909**	**0.8982**	**0.7160**	**487 872**	**normal_concat_13**
39	NASNet	0.7095	0.8438	0.8555	0.6721	325 248	reduction_concat _reduce12

Note: Best performing layer with its corresponding scores is in bold.

Table 1. Quality assessment performance using pre-trained CNN models

the 'softmax' and 'classification' layers and adding a fully-connected layer with one neuron and a regression layer. The training is performed with a low learning rate of 0.0001. As the training dataset contains different image size than the pre-trained models, we have used random crop from the image with size complying with the pre-trained model. This will provide regularization by not letting the training architecture to learn the content of the images. Resizing is not used as the image scale is observed to affect the perception of quality. The training is performed for a total of 30 epochs with Adam optimizer. The training progress for MobileNet-v2 for 20 epochs is depicted in Figure 1.

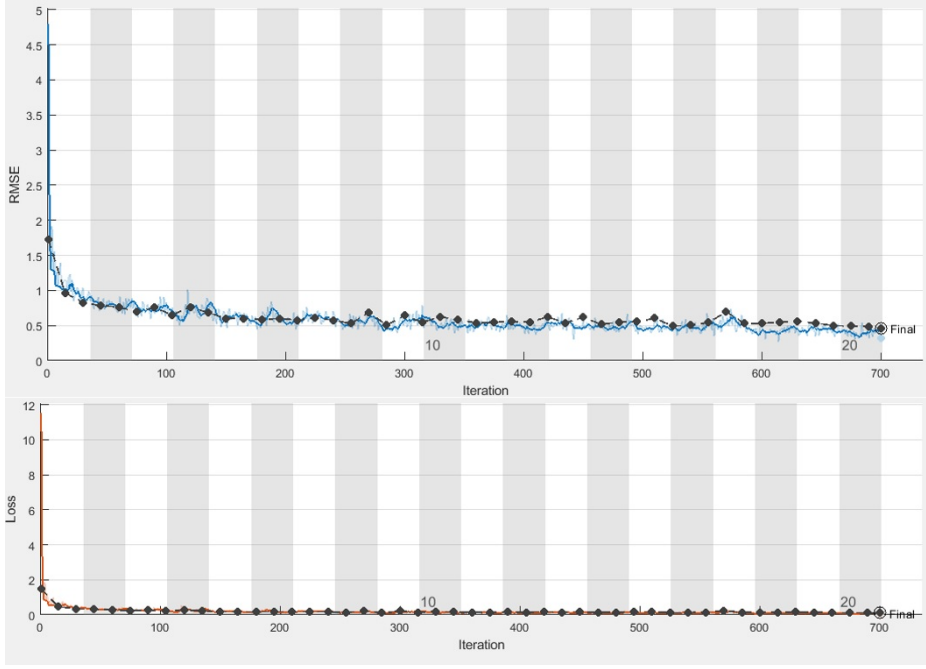


Figure 1. Training progress of MobileNet-v2

We have combined five benchmark datasets for image quality assessment to perform the fine-tuning of these pre-trained CNN models. The fine-tuning is performed with eight pre-trained architectures and the activations of different layers of trained models are obtained and trained with Support Vector Regression (SVR). The predictions from the trained model are evaluated by calculating RMSE, PLCC, SROCC and KROCC and reported in Table 2.

Observations based on results of Table 2 are highlighted below:

1. The fine-tuning of CNN models has adjusted the pre-trained models weights in a way to make it predict the image quality, so the extracted activations are quality aware features and provide better estimates.

#	Architecture	RMSE	PLCC	SROCC	KROCC	Layer Name	Features
1	VGG16	0.6407	0.8474	0.8323	0.6426	fc8	1 000
2	VGG16	0.7594	0.8121	0.7944	0.5994	fc7	4 096
3	VGG16	1.0467	0.6827	0.6672	0.4834	fc6	4 096
4	ShuffleNet	0.7391	0.8115	0.7934	0.5993	node_202	1 000
5	ShuffleNet	0.7541	0.792	0.7751	0.5805	node_198	26 656
6	ShuffleNet	0.7612	0.7912	0.758	0.568	node_186	26 656
7	DenseNet201	0.535	0.8966	0.8826	0.7022	fc1000	1 000
8	DenseNet201	0.5439	0.8943	0.8818	0.7024	conv5_block32 _concat	94 080
9	DenseNet201	0.5061	0.9085	0.896	0.7228	conv5_block30 _concat	90 944
10	MobileNet-V2	0.5784	0.8736	0.8583	0.6704	Logits	1 000
11	MobileNet-V2	0.605	0.8616	0.8438	0.6531	Conv_1	62 720
12	MobileNet-V2	0.6454	0.8425	0.8211	0.6326	block_16_project	15 680
13	ResNet50	0.5129	0.9065	0.897	0.7211	fc1000	1 000
14	ResNet50	0.5037	0.9066	0.8946	0.7167	add_16	100 352
15	ResNet50	0.4978	0.9094	0.9039	0.7307	add_15	100 352
16	Inception-ResNet-V2	0.4676	0.925	0.9187	0.7586	block8_9	133 120
17	Inception-ResNet-V2	0.4729	0.9268	0.9202	0.7596	block8_7	133 120
18	Inception-ResNet-V2	0.4684	0.9282	0.923	0.7654	block8_8	133 120
19	Xception	0.5704	0.8946	0.8833	0.6989	predictions	1 000
20	Xception	0.5712	0.8867	0.8775	0.6946	add_12	1 000
21	Xception	0.7044	0.8148	0.7958	0.6003	add_10	262 808
22	**NASNet-Large**	**0.4432**	**0.9357**	**0.9327**	**0.7768**	**predictions**	**1 000**
23	NASNet-Large	0.5117	0.9129	0.9006	0.732	normal_concat_17	487 872
24	NASNet-Large	0.5242	0.9041	0.8879	0.7139	normal_concat_16	487 872

Note: Best performing layer with its corresponding scores is in bold.

Table 2. Quality assessment performance using fine-tuned CNN models

2. More the training data, better are the deep features for quality estimation as experimented with different combinations of image quality assessment datasets and augmentation methods.
3. After fine-tuning, the last layers started providing higher performance as the model is learning the quality aware features.

3.3 Proposed Approach

Tables 1 and 2 provide the quality estimation performance of deep features using SVR with linear kernel with single image crop only. These results demon-

strate the effectiveness of deep features for the problem of image quality estimation. Some of these models have performed better or comparable to the top performing models described in literature. We have chosen pre-trained NASNet and trained it for 1 000 iterations using Adam optimizer with 0.003 learning rate. Random cropping is used during training as the NASNet input image size is different from the images in benchmark databases. Random cropping serves two purposes, firstly, it converts image size equal to the NASNet input size, and secondly, it serves as the regularization method by varying the cropping region of the image.

The feature extraction phase on the other hand provides the cropped region of the input image to generate probabilities in the last fully connected layer of 1 000 neurons which are used as features. It is to be noted that during training random image regions are used for training in each epoch and therefore performing the quality estimate at several different crops of the image will provide a better estimate. We, therefore, performed random cropping ad feature extraction for 10 random crops and an averaging ensemble is constructed to make the final prediction. The predicted image quality is therefore an average of 10 predictions obtained from different cropped regions of the image under test.

These features can be used to train a regression algorithm to provide quality estimates. It has been observed that the subjective quality scores are the mean opinion scores and therefore they tend to follow normal distribution as shown in Figure 2.

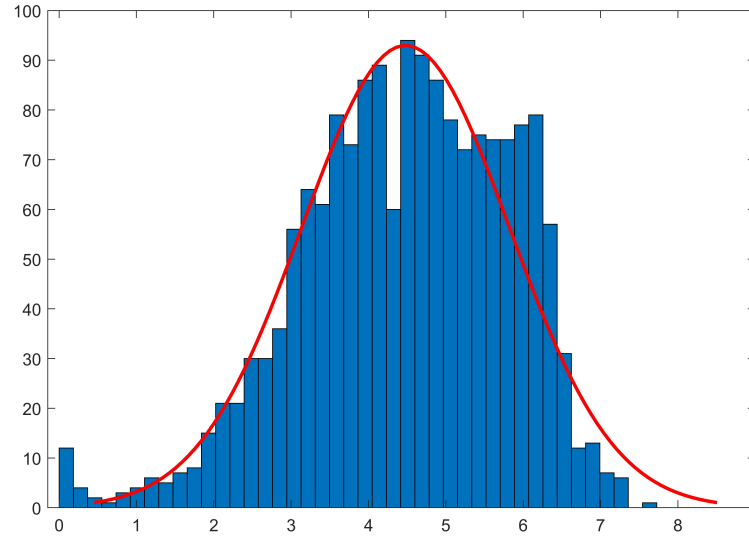


Figure 2. Histogram of mean opinion scores for TID2008 with Gaussian curve

Therefore, we decided to model the problem using Gaussian Process Regression (GPR) which is a stochastic Gaussian process-based algorithm. A Gaussian process is based on random variables and any finite set of these variables follows a joint Gaussian distribution. The standard form of Gaussian process can be defined by it mean $\mu(x)$ and covariance $C_v(x, y)$ functions and is provided in Equation (1).

$$f(x) = G(\mu(x), C_v(x, y) \tag{1}$$

where x and y are the random variables

- $\mu(x) = E[f(x)]$ and
- $C_v(x, y) = E\left[(f(x) - \mu(x))(f(y) - \mu(y))\right]$.

A number of different methods can be used to train a Gaussian process [23]. The function to make mean predictions for the GPR is provided in Equation (2) which is defined for a single test point only.

$$(\overline{\rho_x}) = \kappa_*^T \left(C_v + \sigma_n^2 I\right) y \tag{2}$$

where $\kappa_* = \kappa(x_*)$. There are different options of covariance functions for GPR, but we have used the Matérn covariance function with parameter 5/2 which is defined in Equation (3). This covariance function provided best modeling for our scenario.

$$C_v(x, y) = \sigma_m^2 \left(1 + \frac{\sqrt{5}r}{\sigma_l} + \frac{5r^2}{(3\sigma_l)^2}\right) \exp\left(\frac{-\sqrt{5}r}{\sigma_l}\right) \tag{3}$$

where, r is the distance function and σ^2 is the maximum allowed variance.

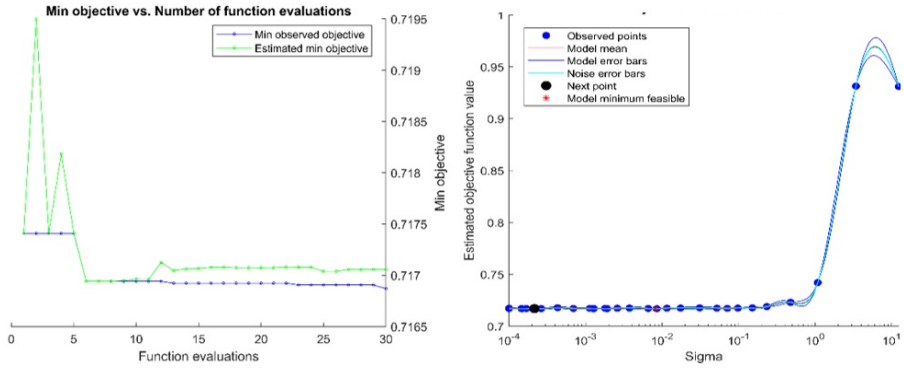


Figure 3. Left: Number of objective function evaluation vs the minimum objective function value. Right: Objective function value for corresponding value of sigma.

Optimization of hyperparameters is performed to find the optimal value of sigma, as it is crucial for the estimation of the covariance function. Bayesian optimization

is used to perform hyperparameter search and the curve for number of function evaluation is plotted against objective function value and provided in Figure 3.

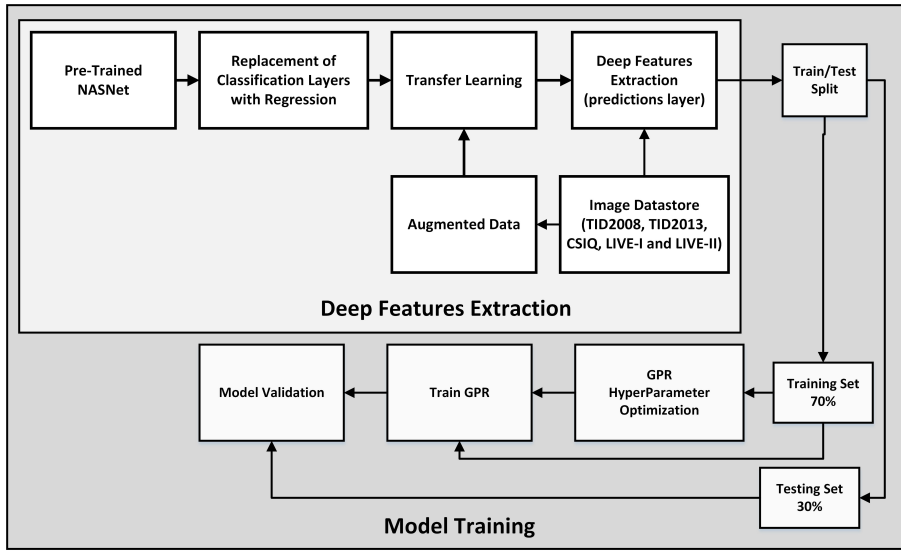


Figure 4. Model training work flow

The flow chart of feature extraction and model training is provided in Figure 4. Five datasets are used in the process of fine-tuning and model training and validation is performed for each dataset individually. Ablation study is conducted to check if the final CNN architecture used is optimal for the image quality assessment task and its results are reported in the next section.

4 RESULTS AND DISCUSSION

Some of the benchmark datasets for image quality assessment are provided in Table 3. These datasets have different number of distortion types and scoring method is either MOS or Differential MOS (DMOS). The scores are standardized using the below formula, so they fall in the same range and the cross-dataset evaluation becomes possible.

$$x_1 = \frac{x - x_2}{\sigma} \tag{4}$$

where x is the original score, x_2 is the mean of subjective scores and σ is the standard deviation of subjective scores.

The performance of the final method is measured by finding the correlation between the predicted quality score and the human subjective evaluation. Three correlation measures: Pearson Linear Correlation Coefficient (PLCC) and Spearman's Rank-Order Correlation Coefficient (SROCC) and Kendall Rank-Oder Correlation

Dataset Name	Number of Reference Images	Number of Distorted Images	Scoring Method	Range
LIVE-I	29	460	DMOS	0–100
LIVE-II	29	982	DMOS	0–100
TID2008	25	1 700	MOS	1–10
TID2013	25	3 000	MOS	1–10
CSIQ	30	900	DMOS	0–1

Table 3. Benchmark datasets for image quality assessment

Coefficient (KROCC) along with RMSE are reported for each test dataset. Table 4 provides the results of experimental testing on five benchmark datasets.

Datasets	RMSE	PLCC	SROCC	KROCC
TID2013	0.4300	0.9685	0.9717	0.8636
TID2008	0.4316	0.9487	0.9504	0.8161
CSIQ	0.0552	0.9802	0.9776	0.8696
LIVE-I	4.5840	0.9779	0.9754	0.8267
LIVE-II	3.5696	0.9752	0.9741	0.8597

Table 4. Correlation and RMSE of the proposed scheme on five benchmark datasets

The bar-chart in Figure 5 provides ground-truth values in the form of bar (green) and the predicted values in the form of stem (red) for 20 random values. Whereas Figure 6 provides the scatter plot between ground-truth and predicted values along with fitting of regression line and value of R-squared. These two plots are provided for TID2013 database and similar plots can be obtained for other benchmark database. It can be noted that the proposed model provided a good quality predicted performance and can be used as a representative model for the objective quality assessment.

4.1 Residual Analysis

The residuals are the difference between the ground-truth and predicted values and are normally plotted in the form of a bar chart. As the value of residual can be negative or positive so the bar-chart is pivoted on the x-axes with the y-axes providing the magnitude of the residuals. Figure 7 provides the bar chart of residuals for 750 (20 %) testing values of TID2013 database. The residual analysis is important in the identification of model's behavior. The residuals are checked for their normal distribution and two tests are conducted for this purpose. The histogram of residuals is plotted with Gaussian fitting in Figure 8 a) and probability plot of the residual is provided in Figure 8 b) indicating the residuals are very close to a normal distribution. The histogram is showing a symmetric distribution around zero and follows a close trend with Gaussian curve plotted for comparison. The validation of

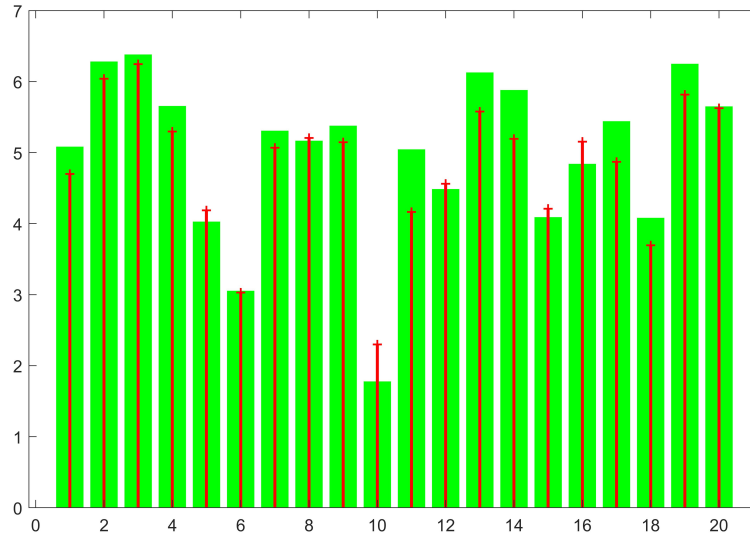


Figure 5. Model training work flow

normality test of residuals indicates that the underlying assumptions of the model are true.

4.2 Cross-Dataset Evaluation

Generalization is a major challenge in the no-reference image quality assessment. A model trained on one dataset usually performs poor on some other dataset which has a different type of distortions and uses a different experimental setup. We have therefore evaluated the performance of the proposed model by training it on one type of dataset and testing on other type of datasets. There are three categories of datasets in our experiment:

1. TID2008 and TID2013,
2. CSIQ, and
3. LIVE-I and LIVE-II.

Three experiments are conducted and reported in Tables 5, 6 and 7.

The generalizability of the proposed method can be explained due to use of a deeper architecture which provides more abstract representations of the learned features. The selected feature set is therefore the representative of image quality and provides features which are quality aware rather than content aware. Moreover, we have incorporated random cropping and other image augmentation strategies such as rotation, scaling and translation to make it robust to small variations. The scores

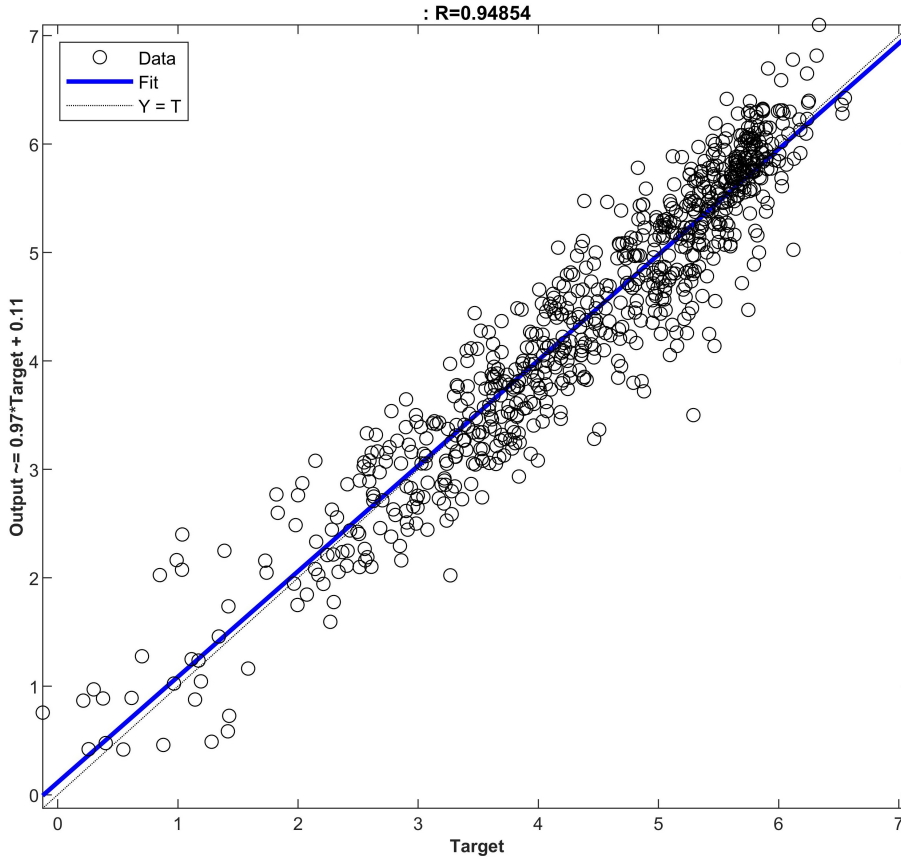


Figure 6. Model training work flow

of the different databases are standardized so the model trained on one database can predict the other database.

4.3 Comparison with Existing Methods

The performance of the proposed scheme is demonstrated in comparison to the existing methods. Ten top performing deep learning based methods are incorporated in the comparison. We have used two performance metrics PLCC and SROCC as they are the widely reported metrics and comparison is performed against three widely used datasets. The results of the comparison are reported in Table 8. It can be noted that LIVE is the most widely used dataset whereas few of the authors has reported performances for other datasets. Two best performing methods on each dataset are in bold face. It can be noted that the proposed approach has

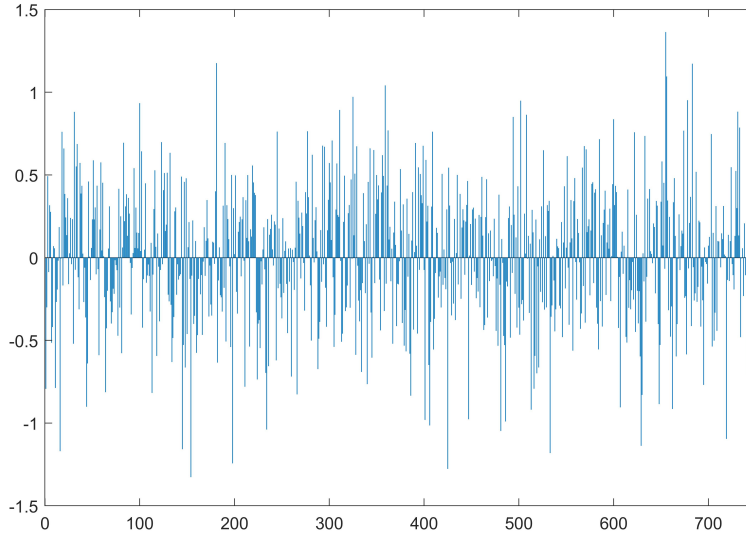


Figure 7. Bar-chart of residuals for test set of TID2013

Evaluation Measure	LIVE-I	LIVE-II	CSIQ
RMSE	5.1542	4.7514	1.2172
PLCC	0.8917	0.8815	0.8912
SROCC	0.8801	0.8798	0.8814
KROCC	0.8204	0.8102	0.8204

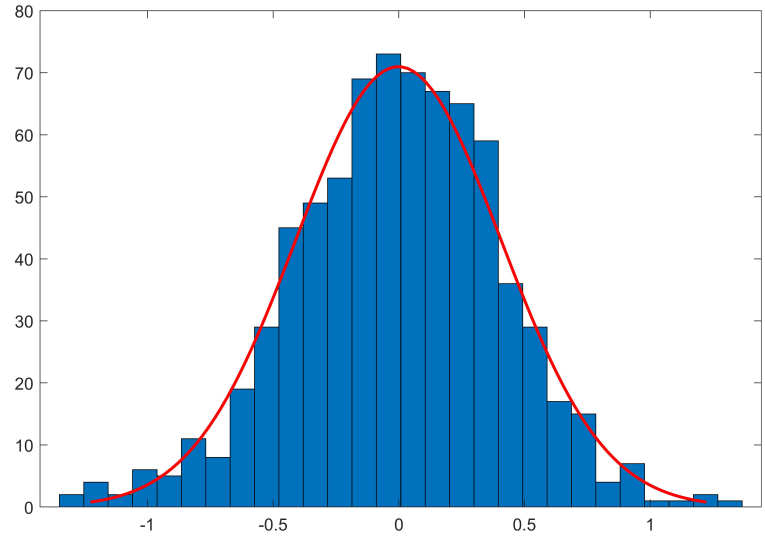
Table 5. Training on category-I dataset and testing on LIVE-I, LIVE-II and CSIQ datasets

Evaluation Measure	LIVE-I	LIVE-II	TID2008	TID2013
RMSE	7.2174	4.1572	0.7524	2.1872
PLCC	0.8617	0.8421	0.8157	0.7214
SROCC	0.8531	0.8681	0.8624	0.7189
KROCC	0.8278	0.7907	0.7124	0.5891

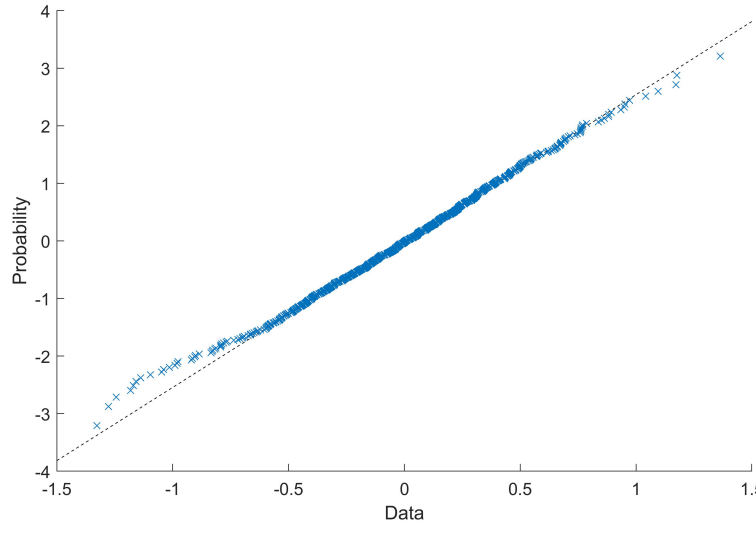
Table 6. Training on category-II dataset and testing on LIVE-I, LIVE-II, TID2008 and TID2013 datasets

Evaluation Measure	TID2008	TID2013	CSIQ
RMSE	0.7813	1.2415	2.1571
PLCC	0.8354	0.7354	0.8872
SROCC	0.8781	0.7257	0.8798
KROCC	0.7254	0.5914	0.8012

Table 7. Training on category-III datasets and testing on TID2008, TID2013 and CSIQ datasets



a) Histogram of residuals with Gaussian fitting



b) Probability plot of normal distribution

Figure 8. Normality tests using histogram of residuals and probability plots

provided the highest performance by using a single learning algorithm with multiple crops.

Dataset	LIVE	LIVE	TID2013	TID2013	CSIQ	CSIQ
Metric	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
[19]a	0.972	0.96	0.855	0.835	–	–
[19]b	0.963	0.954	0.787	0.761	–	–
[9]	0.959	0.966	0.838	0.819	0.968	0.961
[17]	0.98	0.97	0.96	0.96	0.97	0.96
[18]	0.957	0.953	–	–	0.894	0.877
[24]	0.952	0.95	–	–	–	–
[10]	0.973	0.969	–	–	–	–
[20]	0.972	0.96	–	–	–	–
[16]	0.974	0.971	–	–	–	–
[11]	0.93	0.927	–	–	–	–
[26]	0.958	0.957	0.894	0.877	0.949	0.93
[27]	0.95	0.953	0.952	0.959	0.929	0.948
Proposed	0.977	0.975	0.968	0.972	0.98	0.978

Table 8. Comparison with the existing methods

The good performance of the proposed approach can be explained by the use of a representative feature set. Perceptual quality of digital images is based on various factors such as color, contrast, noise, sharpness, artifacts and some factors which are not related to quality such as content, viewing angle and composition. Handcrafted features are therefore focused to some specific aspects of quality such as artifacts generating due to compression or some specific image processing. Moreover, these handcrafted features can model some specific classes of blur or noise but they cannot be generic to be used for all sort of impairments appearing in digital images. Deep features on the other hand are learned automatically on the basis of quality score (MOS). Therefore, deep features seem to be better candidates for image quality assessment.

Extraction of deep features which are quality aware is a tough task as the features can be quality aware only when the training algorithm is provided with a sufficient size of training data having different content. The size of training data is a very important factor when using a deep learning algorithm as these algorithms have a large number of parameters which are required to be trained, and over-fitting can easily occur if the training data is not sufficient. The limitation of the training data is slightly overcome by using augmentation which increased the effective dataset size, but a larger database will definitely be of help. The experimentation with different architectures with and without fine-tuning have highlighted the factors affecting the extraction of quality aware features, and we therefore selected NASNet-Large pre-trained model and obtain deep features after its fine tuning.

Most of the approaches highlighted in the related work used support vector machines for quality prediction which is a convenient and easy way. It provides

a reasonably good performance, but it is not an optimized algorithm in our observation. We have analyzed the MOS and observed that it nearly follows a Gaussian distribution and therefore can be modeled with Gaussian process regression. The optimization of hyperparameter for GPR resulted in final model which has high performance and good generalization. The resulting model therefore outperformed most of the existing approaches. The further improvement can be brought by training the CNN architecture with a larger and representative dataset and using the ensemble learning methods, and these two will be explored in our further work.

4.3.1 Statistical Significance Test

The Pearson and Spearman's correlation is provided in the Table 9 for comparison of the proposed scheme with the existing schemes. It is, however, worth mentioning that the absolute comparison of correlation coefficients can be sometime misleading and therefore statistical significance tests are performed to check if the propose scheme is statistically superior to the existing approaches. We have used one-sided t-test for hypothesis testing, whereas the null hypothesis is stated as the mean correlation of the row algorithm is greater than the mean correlation of the column algorithm. The hypothesis testing is performed with 95 % confidence interval. A value of '1' indicates that the row algorithm is statistically superior to the column algorithm whereas a value of '−1' indicates that the row algorithm is statistically not superior to the column algorithm. The value of '0' indicates an indistinguishable scenario of the row and column algorithm.

	[19]a	[19]b	[9]	[17]	[18]	[24]	[10]	[20]	[16]	[11]	[26]	[27]	Proposed
[19]a	0	1	1	1	1	1	1	1	1	1	1	1	1
[19]b	1	0	1	1	1	1	1	1	1	1	1	1	1
[9]	1	1	0	1	1	1	1	1	1	1	1	1	1
[17]	1	1	1	0	1	1	1	1	1	1	1	1	1
[18]	1	1	1	1	0	1	1	1	1	1	1	1	1
[24]	1	1	1	1	1	0	1	1	1	1	1	1	1
[10]	1	1	1	1	1	1	0	1	1	1	1	1	1
[20]	1	1	1	1	1	1	1	0	1	1	1	1	1
[16]	1	1	1	1	1	1	1	1	0	1	1	1	1
[11]	1	1	1	1	1	1	1	1	1	0	1	1	1
[26]	1	1	1	1	1	1	1	1	1	1	0	1	1
[27]	1	1	1	1	1	1	1	1	1	1	1	0	1
Proposed	1	1	1	1	1	1	1	1	1	1	1	1	0

Table 9. One-sided T-Test

4.4 Ablation Study

The ablation studies have been widely used in the area of neuroscience to tackle the complexities of these systems. Similarly the ablation experiments are being

used in the area of artificial neural networks owing to their increasing complexity. These experiments involve removal of a certain part of the neural network architecture to check their effect on the overall performance of the artificial neural network. These studies investigate the efficacy of the key components of the model and the experiments are done using TID2013 database. Table 1 and Table 2 provide the performance of the pre-trained CNN architectures by selecting an intermediate layer for feature extraction and discarding the layers following this layer. It was noted that without fine-tuning, the complete CNN architecture is not important for image quality assessment as the later layers have learned the features specific to object recognition task. However, the fine-tuning will make the later layers to learn the complex representations for image quality assessment and the last layer of the network performed better for image quality assessment. Table 10 highlights the performance of the selected architecture by keeping the complete architecture, removing last 41 layers and removing last 82 layers. It can be noted that the highest performance is obtained by keeping the complete architecture. Moreover, the complete architecture compacts the size of feature set, making it easy to train a regression algorithm.

#	Ablation Experiment	FeatureSet	RMSE	PLCC	SROCC	KROCC
1	Keeping complete architecture	1 000	0.4432	0.9357	0.9327	0.7768
2	By removing last 41	487 872	0.5117	0.9129	0.9006	0.732
3	By removing last 82	487 872	0.5242	0.9041	0.8879	0.7139

Table 10. Ablation experiment on NASNet-Large using TID2013 database

4.5 Computational Complexity

The experiments are performed on the Intel® Xeon® Processor E5-2687W with 512 GB SSD, 32 GB RAM and RTX 2070 GPU. The training of NASNet-Large for fine-tuning on image quality database is performed for 30 epochs for a batch size of 16 and it took almost 120 hours for training. The training of the NASNet-Large is a one-time job and feature extraction can be performed for each image in order to access the quality. The training and hyperparameter optimization of GPR took 23 minutes. The total training time is therefore 120.5 hours. Whereas in the testing phase, deep feature extraction takes 1.8 seconds per image and score prediction takes less than 120 milliseconds making it a total of less than 2 seconds per image. The testing is reported based on single core CPU only.

5 CONCLUSION

The paper presents a comprehensive insight to the use of deep features for the task of image quality assessment. As HVS is a naïvely understood subject and NSS does not perform consistently better for image quality assessment, the use of CNN

can help to overcome this limitation. Shallow CNN cannot learn the quality aware features and becomes a poor candidate, whereas the deep CNN requires a large number of training images which is not possible due to the subjective nature of image quality. Owing to the visual recognition performance of CNN, we have experimented with 13 popular pre-trained CNN models for feature extraction and eight of these were used to perform fine-tuning on a combination of five image quality assessment databases. Deep activation of NASNet-Large provided best quality estimate and a Gaussian process regression based model is trained using these features. Averaging of quality score over multiple image crops is used as the input image has a larger size then the input of NASNet architecture. The proposed methodology provided good results which are comparable with the state of the art in no-reference image quality assessment. An extensive analysis is performed to demonstrate the robustness and generalization of the proposed model.

5.1 Future Work

1. Experimental testing revealed that GPR is a good algorithm for assessment of image quality. However, ensemble learning approaches should be explored to further increase the performance.
2. The training dataset size can be improved to obtain better features, the dataset size can be increased by using weakly supervised approaches.
3. Moreover, a self-collected dataset with subjective evaluation from local users and having distortions introduced during the process of image acquisition will be used for generalization testing.
4. Combination of extracted features from a different pre-trained model may provide better performance. As the deep feature has a large size, a dimensionality reduction technique may be employed before training the regression algorithm.

REFERENCES

[1] Mittal, A.—Moorthy, A. K.—Bovik, A. C.: No-Reference Image Quality Assessment in the Spatial Domain. IEEE Transactions on Image Processing, Vol. 21, 2012, No. 12, pp. 4695–4708, doi: 10.1109/TIP.2012.2214050.

[2] Liu, L.—Liu, B.—Huang, H.—Bovik, A. C.: No-Reference Image Quality Assessment Based on Spatial and Spectral Entropies. Signal Processing: Image Communication, Vol. 29, 2014, No. 8, pp. 856–863, doi: 10.1016/j.image.2014.06.006.

[3] Sazzad, Z. M. P.—Kawayoke, Y.—Horita, Y.: No Reference Image Quality Assessment for JPEG2000 Based on Spatial Features. Signal Processing: Image Communication, Vol. 23, 2008, No. 4, pp. 257–268, doi: 10.1016/j.image.2008.03.005.

[4] Saad, M. A.—Bovik, A. C.—Charrier, C.: Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain. IEEE Transactions on Image Processing, Vol. 21, 2012, No. 8, pp. 3339–3352, doi: 10.1109/TIP.2012.2191563.

[5] Ma, L.—Li, S.—Ngan, K. N.: Reduced-Reference Image Quality Assessment in Reorganized DCT Domain. Signal Processing: Image Communication, Vol. 28, 2013, No. 8, pp. 884–902, doi: 10.1016/j.image.2012.08.001.

[6] Liu, L.—Dong, H.—Huang, H.—Bovik, A. C.: No-Reference Image Quality Assessment in Curvelet Domain. Signal Processing: Image Communication, Vol. 29, 2014, No. 4, pp. 494–505, doi: 10.1016/j.image.2014.02.004.

[7] He, L.—Tao, D.—Li, X.—Gao, X.: Sparse Representation for Blind Image Quality Assessment. 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 1146–1153, doi: 10.1109/CVPR.2012.6247795.

[8] Chang, H.-W.—Yang, H.—Gan, Y.—Wang, M.-H.: Sparse Feature Fidelity for Perceptual Image Quality Assessment. IEEE Transactions on Image Processing, Vol. 22, 2013, No. 10, pp. 4007–4018, doi: 10.1109/TIP.2013.2266579.

[9] Gao, F.—Yu, J.—Zhu, S.—Huang, Q.—Tian, Q.: Blind Image Quality Prediction by Exploiting Multi-Level Deep Representations. Pattern Recognition, Vol. 81, 2018, pp. 432–442, doi: 10.1016/j.patcog.2018.04.016.

[10] Guan, J.—Yi, S.—Zeng, X.—Cham, W.-K.—Wang, X.: Visual Importance and Distortion Guided Deep Image Quality Assessment Framework. IEEE Transactions on Multimedia, Vol. 19, 2017, No. 11, pp. 2505–2520, doi: 10.1109/TMM.2017.2703148.

[11] Hou, W.—Gao, X.—Tao, D.—Li, X.: Blind Image Quality Assessment via Deep Learning. IEEE Transactions on Neural Networks and Learning Systems, Vol. 26, 2015, No. 6, pp. 1275–1286, doi: 10.1109/TNNLS.2014.2336852.

[12] Zhou, B.—Lapedriza, A.—Xiao, J.—Torralba, A.—Oliva, A.: Learning Deep Features for Scene Recognition Using Places Database. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., Weinberger, K. Q. (Eds.): Advances in Neural Information Processing Systems 27 (NIPS 2014), 2014, 9 pp.

[13] Babenko, A.—Lempitsky, V.: Aggregating Local Deep Features for Image Retrieval. Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1269–1277, doi: 10.1109/ICCV.2015.150.

[14] Zhou, B.—Khosla, A.—Lapedriza, A.—Oliva, A.—Torralba, A.: Learning Deep Features for Discriminative Localization. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2921–2929, doi: 10.1109/CVPR.2016.319.

[15] Zhang, R.—Isola, P.—Efros, A. A.—Shechtman, E.—Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595, doi: 10.1109/CVPR.2018.00068.

[16] Bare, B.—Li, K.—Yan, B.: An Accurate Deep Convolutional Neural Networks Model for No-Reference Image Quality Assessment. 2017 IEEE International Conference on Multimedia and Expo (ICME), 2017, pp. 1356–1361, doi: 10.1109/ICME.2017.8019508.

[17] Bianco, S.—Celona, L.—Napoletano, P.—Schettini, R.: On the Use of Deep Learning for Blind Image Quality Assessment. Signal, Image and Video Processing, Vol. 12, 2018, No. 2, pp. 355–362, doi: 10.1007/s11760-017-1166-8.

[18] Fan, C.—Zhang, Y.—Feng, L.—Jiang, Q.: No Reference Image Quality Assessment Based on Multi-Expert Convolutional Neural Networks. IEEE Access, Vol. 6, 2018, pp. 8934–8943, doi: 10.1109/ACCESS.2018.2802498.

[19] Bosse, S.—Maniri, D.—Müller, K.-R.—Wiegand, T.—Samek, W.: Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment. IEEE Transactions on Image Processing, Vol. 27, 2018, No. 1, pp. 206–219, doi: 10.1109/TIP.2017.2760518.

[20] Bosse, S.—Maniri, D.—Wiegand, T.—Samek, W.: A Deep Neural Network for Image Quality Assessment. 2016 IEEE International Conference on Image Processing (ICIP), 2016, pp. 3773–3777, doi: 10.1109/ICIP.2016.7533065.

[21] Brandao, T.—Queluz, M. P.: No-Reference Image Quality Assessment Based on DCT Domain Statistics. Signal Processing, Vol. 88, 2008, No. 4, pp. 822–833, doi: 10.1016/j.sigpro.2007.09.017.

[22] Ponomarenko, N.—Jin, L.—Ieremiev, O.—Lukin, V.—Egiazarian, K.—Astola, J.—Vozel, B.—Chehdi, K.—Carli, M.—Battisti, F.—Kuo, C.-C. J.: Image Database TID2013: Peculiarities, Results and Perspectives. Signal Processing: Image Communication, Vol. 30, 2015, pp. 57–77, doi: 10.1016/j.image.2014.10.009.

[23] Rasmussen, C. E.—Williams, C. K. I.: Gaussian Processes for Machine Learning. The MIT Press, 2006.

[24] Liu, L.—Hua, L.—Zhao, Q.—Huang, H.—Bovik, A. C.: Blind Image Quality Assessment by Relative Gradient Statistics and Adaboosting Neural Network. Signal Processing: Image Communication, Vol. 40, 2016, pp. 1–15, doi: 10.1016/j.image.2015.10.005.

[25] Ma, K.—Liu, W.—Liu, T.—Wang, Z.—Tao, D.: dipIQ: Blind Image Quality Assessment by Learning-to-Rank Discriminable Image Pairs. IEEE Transactions on Image Processing, Vol. 26, 2017, No. 8, pp. 3951–3964, doi: 10.1109/TIP.2017.2708503.

[26] Ma, K.—Liu, W.—Liu, T.—Wang, Z.—Tao, D.: dipIQ: Blind Image Quality Assessment by Learning-to-Rank Discriminable Image Pairs. IEEE Transactions on Image Processing, Vol. 26, 2017, No. 8, pp. 3951–3964, doi: 10.1109/TIP.2017.2708503.

[27] Xu, J.—Ye, P.—Du, H.—Liu, Y.—Doermann, D.: Blind Image Quality Assessment Based on High Order Statistics Aggregation. IEEE Transactions on Image Processing, Vol. 25, 2016, No. 9, pp. 4444–4457, doi: 10.1109/TIP.2016.2585880.

Nisar Ahmed is Ph.D. student in the Department of Computer Engineering, University of Engineering and Technology, Lahore, Pakistan. His area of interest includes machine learning and digital image processing.

Hafiz Muhammad Shahzad Asif obtained his Ph.D. degree in informatics from the University of Edinburgh, UK in 2012. He is working as Chairman and Associate Professor at the Department of Computer Science, University of Engineering and Technology, Lahore, Pakistan.