SIMULTANEOUS ESTIMATION OF IMAGE QUALITY AND DISTORTION VIA MULTI-TASK CONVOLUTIONAL NEURAL NETWORKS

Le Kang[1], *Peng Ye*[2], *Yi Li*[3], *David Doermann*[1]

[1]University of Maryland, College Park; [2]SONY US Research Center; [3]NICTA and ANU

[1]{lekang,doermann}@umiacs.umd.edu; [2]pengye@umiacs.umd.edu; [3]yi.li@nicta.com.au

ABSTRACT

In this work we describe a compact multi-task Convolutional Neural Network (CNN) for simultaneously estimating image quality and identifying distortions. CNNs are natural choices for multi-task problems because learned convolutional features may be shared by different high level tasks. However, we empirically argue that simply appending additional tasks based on the state of the art structure (e.g., [1]) does not lead to optimal solutions. We design a compact structure with nearly 90% fewer parameters compared to [1], and demonstrate its learning power.

Index Terms— Image quality assessment, image distortion classification, no-reference, CNN

1. INTRODUCTION

This work focuses on the problem of No-Reference (NR) Image Quality Assessment (IQA), where the objective is to evaluate the quality of digital images without access to reference images and without prior knowledge on the types of distortions. NR-IQA is the most challenging category of objective IQA tasks and its performance is usually worse than the performance of Full-Reference (FR) IQA systems, where non-distorted reference images are available for quantifying image distortions. For many practical tasks, however, there do not exist perfect versions of the distorted images, so NR-IQA is the only solution.

In most standard image quality assessment datasets, distorted images are grouped into classes according to the types of distortion such as Gaussian noise, blur, and compression methods, and the images in each group are degraded with different levels of the same distortion. While many image quality assessment methods have been proposed, it is interesting to notice that most methods estimate quality scores without fully utilizing the distortion type information. Some methods can only work well under certain types of distortion, while other methods ignore distortion types in training when estimating the quality for a collection of images with different types of distortion.

Identifying the distortion type is an important part for NR-IQA. A quality score cannot fully characterize the distortion present in an image. It will be a much better description if both the distortion type and quality score are determined. Take Fig. 1 for example. The images in the same column have very similar quality scores, but have different types of distortions (Gaussian blur and JPEG compression respectively).

Fig. 1. Images in the same column have similar image quality scores (DMOS, range $[0, 100]$), but they have different distortion types. Higher score denotes worse quality.

We use a multi-task Convolutional Neural Network (CNN) to address this problem. The two tasks, quality estimation and distortion identification, are learned simultaneously with one CNN. For multi-task learning the literature typically specifies a main and several extra tasks. In this work, however, we are optimizing both tasks, therefore we use the term primary and secondary tasks. Since quality estimation are more widely addressed, we treat it as the primary task, and distortion identification as the secondary task.

We show through experiments that the proposed multi-task CNN achieves the similar or better performance on image quality estimation and distortion identification compared to the state of the art.

2. RELATED WORK

Previous research efforts on NR-IQA fall in two categories. The first category of work focuses on designing better hand-crafted features, among which Natural Scene Statistics (NSS) based features are the most successful. Methods in this category includes DIIVINE [2], BLIINDS-II [3], and BRISQUE [4]. It is worth noting that DIIVINE and BRISQUE use a two-stage framework. They first perform distortion identification, i.e. classify the images into one of the distortions, and then perform the distortion-specific quality estimation. However, experiments in [4] show that such a two-stage approach is not superior to the distortion-blind approach.

The second category of work focuses on feature learning, which attempts to learn discriminant visual features for the IQA task automatically without using hand-crafted features. Ye et al. introduced CORNIA [5], an unsupervised framework for NR-IQA and extended CORNIA to a supervised setting by employing a shallow architecture and an empirical optimization procedure [6].

Motivated by the learning framework in [6], where features are learned directly from the normalized raw image patches. Kang et al [1] proposed a CNN based approach that integrates feature learning and regression, and achieved state of the art performance.

Multi-task learning using neural networks has been studied for decades. Abu-Mostafa [7] suggested one multi-task setting called catalytic hints, where one task is the internal features that are used by other tasks. Caruana [8] studied on multi-task learning with neural networks on a number of problems. He demonstrated that learning multiple correlated tasks at the same time may significantly improve the performance of the main task.

3. OUR APPROACH

In this section we first describe the state-of-the-art CNN for NR-IQA and its naive extension to multi-task CNN. Then, we discuss the issues in the network and present our multi-task network by increasing filter layers and adjusting hidden node numbers. While it is unlikely that all tasks can achieve the best performance at the same time [8], our goal is to generate balanced outputs that improve both tasks.

3.1. Overall process

We normalize the image similar to [4] and divide the image into patches (typical size 32×32 pixels) as in [1]. We then make predictions on each patch using a multi-task CNN, and aggregate patch predictions to obtain the result for the image. Under the assumption that the distortion is homogeneous across the entire image, the image quality score is simply the average of patches' quality scores, while the image distortion is decided by a majority voting of the patches, i.e. the most frequently occurring distortion on patches determines the distortion of the image.

Dividing an image into patches provides a large quantity of training samples for the CNN. Accurate predictions on patches lead to superior performance on the image level, especially for distortion identification as we will show in experiments section.

3.2. CNN: from single to multi-task architectures

In this section we briefly introduce the CNN architecture in [1], which we refer to as IQA-CNN in this paper. Then we naively extend it to IQA-CNN+, a multi-task variant by directly adding a minor task in the output layer, as a baseline.

The structure of the IQA-CNN for quality score estimation has one convolutional layer, one pooling layer, two fully connected layers and one output layer as described in [1].

We extend this structure for the multi-task by adding a classification layer and refer to it as IQA-CNN+. The secondary task for distortion identification shares the same structure as the one for quality score estimation except that the output layer is a multi-class logistic regression.

Unfortunately such a simple extension is not ideal for the multi-task scenario. The reason is three fold. First, the IQA-CNN+ has a shallow convolutional structure (one layer), which makes the filter learning less efficient compared to deeper structures, since no pattern is shared among filters. Second, given such a large number of neurons, there are too many parameters to be learned in the network. The larger model size makes it less practical or even impossible in some application scenarios. Third, the arrangement of the fully connected layers may not facilitate multiple tasks when it is specially tuned for one task.

3.3. IQA-CNN++: a compact multi-task network

In the multi-task setting, the goal is to estimate the quality and identify the distortion in one network. Therefore, the network needs to learn features that are shared by both tasks, and the importance of two tasks need to be balanced. We also prefer a compact model that is significantly smaller in size but has at least comparable performance.

As a result, we propose the following two modifications: 1) increasing the number of convolutional layers while reducing the receptive field of the filters, and 2) modifying the fully connected layers to have a "fan-out" shape with significantly fewer neurons. We refer to our new model as IQA-CNN++.

The structure of IQA-CNN++ is shown in Fig. 2. There are two convolutional layers, each with a pooling, two fully connected layers and one output layer. The first convolutional layer contains 8 kernels each of size 3×3, followed by a 2×2 pooling. The second convolutional layer contains 32 kernels, each of size $3 \times 3 \times 8$. The 32 feature maps obtained by the second convolutional layer are pooled to 32 max and 32 min

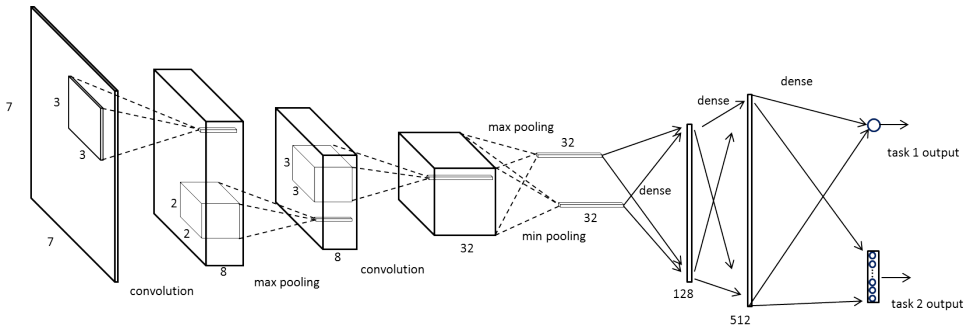

Fig. 2. The architecture of the IQA-CNN++ for simultaneously image quality estimation and distortion identification.

values, i.e. each feature map is pooled to one max and one min value, which form 64 inputs for the next layer. Again, no nonlinear neurons are used in the convolutional layers. There are 128 and 512 Rectified Linear Units (ReLUs) in the two fully connected layers respectively. Both the linear regression layer and logistic regression layer exist in the last part of the multi-task network, and they both take as inputs the second fully connected layer's outputs, i.e. the two tasks share all internal structures.

We employ stochastic gradient descent (SGD) and back-propagation to approximately minimize each task's loss. The loss of the primary task is the l_1 norm of the prediction error, as defined in [1], and the loss of secondary tasks is the negative log likelihood as used in most classification setting. The network parameters are updated using the weighted sum of the gradients from all tasks. Let w_i and p^i denote the ith network parameter and its learning rate respectively. Let D_m^i denote the gradient for w_i from task m, and α_m denote the relative weight for task m. The following updating rule applies at each iteration:

$$w^i \leftarrow w^i - p^i \sum_{m=1}^{2} \alpha_m D_m^i \tag{1}$$

Through some simple computation we can find that IQA-CNN (and IQA-CNN+) has approximately 7.2×10^5 learnable parameters (weights of neurons). By comparison our IQA-CNN++ consists of roughly 7.7×10^4 learnable parameters, which reduces the model size by 90%. Despite a significant reduction in size, the IQA-CNN++ still shows excellent performance as we will see in later experiments.

4. EXPERIMENT

We present experiments on three standard datasets and measure the performance with commonly used metrics. We compare the multi-task CNNs against other models, and compare the performances among CNNs of different architectures. Cross dataset experiments are also conducted to demonstrate the generalization power of the approach.

4.1. Experimental Protocol

Datasets: The following three datasets are used in our experiments.

(1) LIVE [9]: A total of 779 distorted images with five different distortions – JPEG2000 compression (JPEG2K), JPEG compression (JPEG), White Gaussian (WN), Gaussian blur (BLUR) and Fast Fading (FF) at 7 or 8 degradation levels derived from 29 reference images. Differential Mean Opinion Score (DMOS) is provided for each image, roughly in the range $[0, 100]$. Higher DMOS indicates lower quality.

(2) TID2008 [10]: 1700 distorted images with 17 different distortions derived from 25 reference images at 4 degradation levels. Each image is associated with a Mean Opinion Score (MOS) in the range $[0, 9]$. In our experiment we only use 13 global distortions and leave out the other 4 distortions that are either very heterogeneous or related to intensity/contrast change.

(3) CSIQ [11]: 30 original images distorted using six different distortions at four to five levels each, resulting in a total of 866 distorted images.

Evaluation: We follow the same protocol as in [1] to use Linear Correlation Coefficient (LCC) and Spearman Rank Order Correlation Coefficient (SROCC) to evaluate the performance of the quality estimation.

4.2. Evaluation on LIVE

On the LIVE dataset, we train and test on all five distortions together. In Table 1 we compare the performance of multi-task CNNs with previous methods including DIIVINE, BLIINDS-II, BRISQUE, CORNIA and IQA-CNN. A few full-reference methods, including PSNR, SSIM [12] and FSIM [13], are also shown for reference.

Quality score estimation: From Table 1 one can see that in the quality estimation task multi-task CNNs (IQA-CNN+ and IQA-CNN++) outperformed the non-CNN based methods. Both multi-task CNNs achieved similar performance compared to IQA-CNN.

Distortion identification: For the distortion identification task, both multi-task CNNs achieved much higher accuracy.

	LCC	SROCC	Class. Acc.
PSNR	0.856	0.866	-
SSIM	0.906	0.913	-
FSIM	0.960	0.964	-
DIIVINE	0.917	0.916	-
BLIINDS-II	0.930	0.931	0.838
BRISQUE	0.942	0.940	0.886
CORNIA	0.935	0.942	0.875
IQA-CNN	0.953	0.956	-
IQA-CNN+	0.953	0.953	0.921
IQA-CNN++	0.950	0.950	0.951

Table 1. Performance of quality estimation and distortion identification on LIVE. The top three (NR-IQA) performers are highlighted by red, blue, and green.

	LCC	SROCC	Class. Acc.
PSNR	0.652	0.669	-
SSIM	0.857	0.878	-
FSIM	0.913	0.926	-
CORNIA	0.837	0.813	0.862
IQA-CNN	0.873	0.862	
IQA-CNN+	0.870	0.861	0.889
IQA-CNN++	0.880	0.870	0.929

Table 2. Performance of quality estimation and distortion identification on 13 distortions of TID2008.

Compared with the state of the art, the gains are approximately 5% and 8% respectively. IQA-CNN++ achieves the best performance (0.951) among all competitors.

The appealing performance of the multi-task CNNs in distortion identification partly comes from the voting on patches. In fact, at the patch level we observe a classification accuracy around 0.88, therefore the correct prediction is very likely to collect the most votes and stand out as the final prediction at the image level. We can see that being able to predict on small image patches with a reasonable accuracy contributes to the superior performance on distortion identification.

4.3. Evaluation on TID2008

On the TID2008 dataset, we train and test on 13 distortions, similar to the experiments on LIVE described above. In Table 2 we compare the performance of IQA-CNN+, IQA-CNN++ and other methods with available results. From Table 2 we see that both IQA-CNN+ and IQA-CNN++ outperform the previous methods. For the quality estimation task, IQA-CNN++ achieves an LCC and SROCC of 0.880 and 0.870, showing advantage over the IQA-CNN+ as well as other methods. On the distortion identification side, IQA-CNN++ achieves the highest accuracy of 0.929 while other methods are below 0.9.

4.4. Cross Dataset Test

To observe how well our methods generalize, we conduct cross dataset tests. Training and validation are performed on LIVE, and then the obtained model is tested on TID2008 and CSIQ without parameter adaptation. Both the quality estimation and the distortion identification tasks are tested. It is worth noting that the three datasets cover different types of distortions and thus we only use the 4 distortions that are common to all the three datasets, namely JPEG2K, JPEG, WN, and BLUR.

	LCC	SROCC	Class. Acc.
PSNR	0.776	0.901	-
SSIM	0.817	0.903	-
FSIM	0.952	0.954	-
CORNIA	0.890	0.880	0.920
IQA-CNN	0.903	0.920	-
IQA-CNN+	0.893	0.912	0.890
IQA-CNN++	0.895	0.906	0.933

Table 3. Performance of quality estimation and distortion identification on 4 common distortions of TID2008, using models trained on LIVE.

	LCC	SROCC	Class. Acc.
FSIM	0.961	0.962	-
CORNIA	0.914	0.899	0.768
IQA-CNN	0.913	0.923	-
IQA-CNN+	0.910	0.918	0.730
IQA-CNN++	0.928	0.936	0.783

Table 4. Performance of quality estimation and distortion identification on 4 common distortions of CSIQ, using models trained on LIVE.

Since the model trained on LIVE produces a quality score in the same range as DMOS, which is different from MOS in TID2008, we follow the tradition in [5] by applying a nonlinear mapping on the predicted quality scores on TID2008. No mapping is applied for CSIQ since it also uses DMOS as the ground truth quality measure.

Table 3 and Table 4 show the results of the cross dataset tests on TID2008 and CSIQ respectively. For the image quality estimation task, the IQA-CNN++ trained on LIVE achieves an LCC/SORCC of 0.895/0.906 on TID2008, and 0.928/0.936 on CSIQ, outperforming other methods. IQA-CNN++ also wins the distortion identification task for both datasets.

5. CONCLUSIONS

We proposed a multi-task Convolutional Neural Network (CNN) to simultaneously estimate image quality and identify distortion type in a no-reference setting. Our multi-task CNN is compact in size and it shows excellent performance on standard datasets.

6. REFERENCES

[1] Le Kang, Peng Ye, Yi Li, and David Doermann, "Convolutional neural networks for no-reference image quality assessment," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014, pp. 1733–1740.

[2] Anush Krishna Moorthy and Alan Conrad Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *Image Processing, IEEE Transactions on*, vol. 20, no. 12, pp. 3350–3364, 2011.

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

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

[5] Peng Ye, Jayant Kumar, Le Kang, and David Doermann, "Unsupervised Feature Learning Framework for No-reference Image Quality Assessment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1098–1105.

[6] Peng Ye, Jayant Kumar, Le Kang, and David Doermann, "Real-time no-reference image quality assessment based on filter learning," in *Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 987–994.

[7] Yaser S Abu-Mostafa, "Learning from hints in neural networks," *Journal of complexity*, vol. 6, no. 2, pp. 192–198, 1990.

[8] Rich Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[9] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "LIVE image quality assessment database release 2," Online, http://live.ece.utexas.edu/research/quality.

[10] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "TID2008 - a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radio Electronics*, vol. 10, pp. 30–45, 2009.

[11] Eric C. Larson and Damon M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, pp. 011006–011006–21, 2010.

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

[13] Lin Zhang, D. Zhang, Xuanqin Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.