Vision Systems Project Report
Project Title: Understanding similarity of Saliency Maps between NR IQA Models and Vision Classification Models

Krishna Srikar Durbha
krishna.durbha@utexas.edu

Abstract

In recent years, convolutional neural networks (CNNs) have been copiously applied in various image and video processing problems giving state-of-the-art results close to human-level performance. They are used for real-time scenarios like quality assessment, object detection, autonomous navigation, satellite imaging, etc. But for the SOTA (state-of-the-art) models to maintain their performance, they need images/videos to follow a distribution on which they are trained. Any example that lies outside this distribution might reduce the model's performance. In this report, we try to understand the similarities or differences in perception of the basic computer vision image classification model and humans. We try to understand the correlation between them by comparing the performance of vision models with image quality rated by no-reference image quality assessment algorithms, which are trained to approximate human judgements.

1. Introduction

Convolutional neural networks (CNNs) are a class of deep neural networks that have become prominent in various image and video processing tasks. They perform incredibly in extracting spatial features of visual data of images and videos. CNN can learn complex features required from images for recognition, detection, segmentation, and retrieval using a hierarchy of trainable filters and feature pooling operations. They have proven to be successful in tasks related to image quality assessment [11], and computer vision [9] give state-of-the-art results close to the performance of humans.

Quality can be considered as the accuracy with which information is contained. In the case of images which are signals, the techniques used to capture, the accuracy of detail, colour, contrast, luminance etc., become various factors in determining the quality of the image.

Image quality assessment algorithms are used in rating the quality of images and are divided into three categories.

- Reference IQA algorithms need a reference and distorted images to estimate quality.
- Reduced IQA algorithms: Need less information about reference image and require the entire distorted image to estimate quality.
- No Reference IQA algorithms: Need no reference image at all. They only need a distorted image to estimate quality.

NR IQA algorithms are trained on ratings of various images collected from humans. Many human studies have been conducted to create massive datasets consisting of subjective image evaluations. Over the years, NR IQA algorithms used NSS (Natural Scene Statistics), which is an important field of vision science [3] as their base to develop various algorithms [7], [5], [6] etc. Recently with the surge of deep learning and its ability to extract features and create better approximations, these NR IQA algorithms started using CNNs in their algorithms [11] etc.

CNNs also played a major in the rise of computer vision algorithms. Every new model or algorithm is developed to achieve state-or-the-art results on major computer vision datasets like ImageNet, Kinetics-700, MS Coco, Open Images etc., on various tasks like classification, segmentation, detection, generation, compression etc. CNN models like [10], [4] etc. reached state-of-the-art on these standard datasets with their performance very close to humans. To understand what these high-accuracy models are learning and to understand these models better, techniques like Vanilla Gradients, Grad-CAM, Grad-CAM++, LIME etc., have been developed to understand the perception of these computer vision models. These are explainable AI techniques or Pixel attribution methods that highlight the relevant pixels for image classification by a neural network.

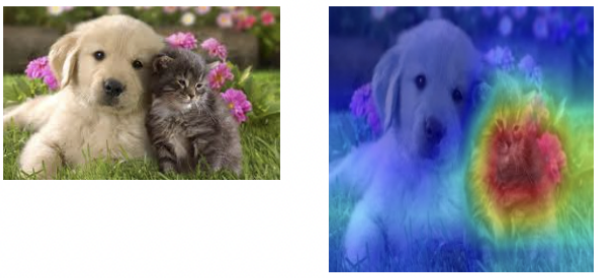

Figure 1. Pixel Attribution Map using Grad-CAM of ResNet50 detecting a 'cat'.

2. Problem Statement

The term perception is very different for humans and computers. Understanding a scene comes naturally to us without any effort, but for computers, there is no such thing as conscious; it is just a bunch of filters leading to a prediction. The distortions in images are usually an aspect of electronic noise produced by an image sensor or the circuitry of a scanner or digital camera. These disturbances occur naturally while capturing or transmitting, or saving an image. These distortions are visible to the human eye. They can also be affected by the brightness, contrast, saturation, settings, zooming, weather conditions, compression etc.

In this report, we will observe if there is a correlation between the performance of the image classification model and quality rating given by an NR IQA algorithm not only in terms of accuracy but also plotting saliency maps and local patch qualities, respectively. As NR-IQA models are trained on human judgements/perception, they help to understand the similarity in the perception of an image as humans rate it and the classification model while they classify it.

As a part of the experiment, we will use images from a dataset rated as good by a no-reference IQA algorithm. As NR-IQA algorithms are the closest approximation of human ratings, they should help us understand how humans would rate images. Then, by splitting the dataset into train and validation sets, we will train an image classification model on the training set and validate its performance on the validation set with different types of natural distortions and various levels. We are training a classification model on high-quality images to understand its perception of distortions wrt humans. We then generate the saliency maps from the image classification model and compare them with the patch-wise quality estimate obtained from the NR-IQA model.

3. Literature

Paper [2] discuss how the quality of the image affects the performance of neural networks. It shows that when a model is trained on uncorrupted images, its performance is affected when tested on images with distortions like Blur, Noise, Contrast, JPEG and JPEG 2000. Also, the performance decreases with the extent of corruption. This follows that a data shift affects the performance of a model trained in a distribution. The paper concludes with the possibility that training models with low-quality images might lead to better performance on low-quality images but also increases the risk of misclassification on high-quality images.

One of the key points to note is that adding noise to images during training is one of the data augmentation techniques. So, training the model on noisy images should help create robust models, but in fact, it is not the case. Paper [1] discuss the correlation between the level of noise that needed to be added to an image to create robust images and the structural similarity index metric. It concludes by saying that to create a robust model against particular noise while maintaining the model's performance, the amount of degradation that should be applied should result around a quality estimate of 0.8 MS-SSIM.

The above literature provides an understanding that there is a correlation between the quality estimate and the performance of a classification vision model. As a reference image might only be available some of the time, we considered no reference image quality estimation techniques to rate the quality of images.

4. Approach

4.1. Dataset and Data Preprocessing

As a part of our experiment, we shall use the mit-indoor-scenes dataset [8]. The dataset consists of 67 indoor scene categories with 15620 images. Among all the images, we get the images which are rated 'Good' or 'Excellent' by our NR IQA model among 'Bad', 'Poor', 'Fair', 'Good' and 'Excellent'. The quality of the dataset is as follows:

We split the rated images into 80-20 as training and validation sets. We train our image classification models on the training set and evaluate them on the validation set. We will observe the model's performance in terms of accuracy and attribution maps on various corruptions like 'brightness', 'contrast', 'jpeg˙compression', and 'saturate' with severity levels {1,2}. We use the term severity to explain how strongly the image is distorted. For further reference check ImageCorruptions

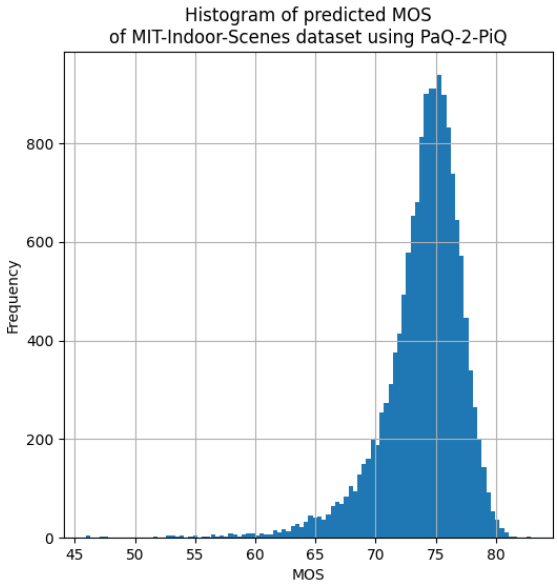

Figure 2. MOS of images in the dataset

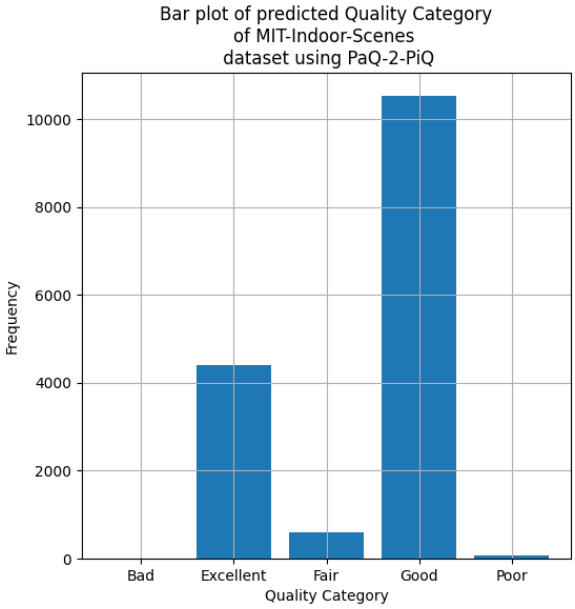

Figure 3. Quality category of images in the dataset

4.2. Models

We use the PaQ-2-PiQ model as a no-reference image quality assessment algorithm for rating images. The PaQ-2-PiQ model has three different variations. In the model, we use the RoIPool layer similar to Fast-RCNN, which allows flexibility to train both patch and picture-sized scales. It has a ResNet18 model as its backbone, followed by RoIPool and two fully connected layers. In this paper, whenever we refer to the NR-IQA PaQ-2-PiQ model, we refer to the model with only the RoIPool layer without feedback. We will use the ResNet18 model with the last two layers modified to maintain consistency with the PaQ-2-PiQ model as a classification model.

5. Results

Consider the image below. We are going to show the variation in attribution maps of both ResNet and Paq-2PiQ for different distortions and different levels.

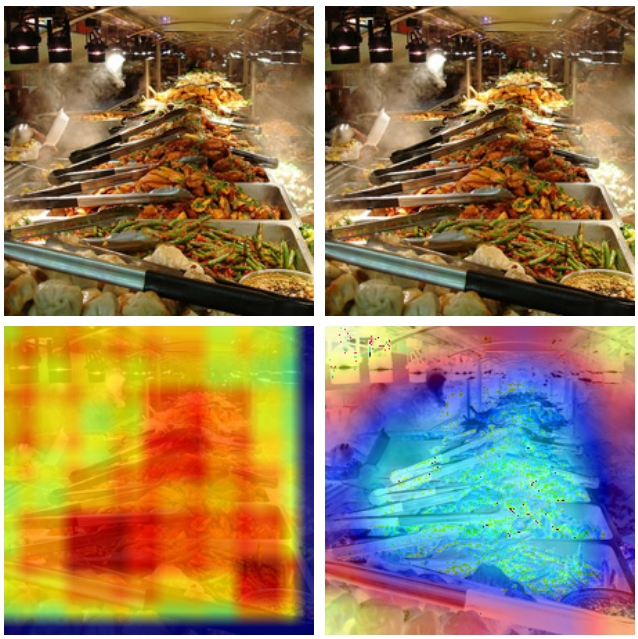

Figure 4. First row: Original image and center cropped version of the original one, respectively. The quality estimations from PaQ-2-PiQ are 76.75, 76.75, respectively. Second row: Attribution maps from PaQ-2-PiQ and ResNet, respectively on the cropped original image.

Distortions are first added to the original image; then the distorted image is center-cropped to shape (224,224). Images are cropped instead of resized to maintain originality as we perform the quality estimation.

In the figures below; In the first row, the left one is cropped original image and the right one is cropped distorted image with a particular similarity; In the second row, the left image is the local patch quality of the cropped distorted image from PaQ-2-PiQ and the right image is attribution map of the cropped distorted image from ResNet. We will also describe the quality ratings of original and distorted images along with severity.

PaQ-2-PiQ for NR-IQA and ResNet for classification have the same backbone model as mentioned. We chose such models to clearly observe the difference in image perception of models with two different goals. The goal of PaQ-2-PiQ is to estimate the image's quality, whereas the goal of ResNet is to classify it. Due to their inherited same backbone structure, the areas they tend to focus on without distortions with be very close. But the benefit is we get to understand the change in their perception of distortions in images.

5.1. Brightness Distortion

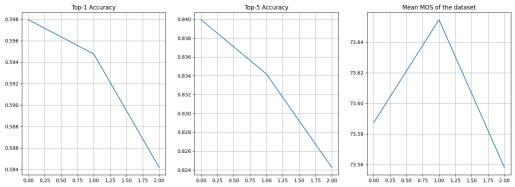

Figure 5. The figure show the change in Top@1, Top@5 Accuracies and Mean MOS on the validation dataset by changing severity of brightness corruption.

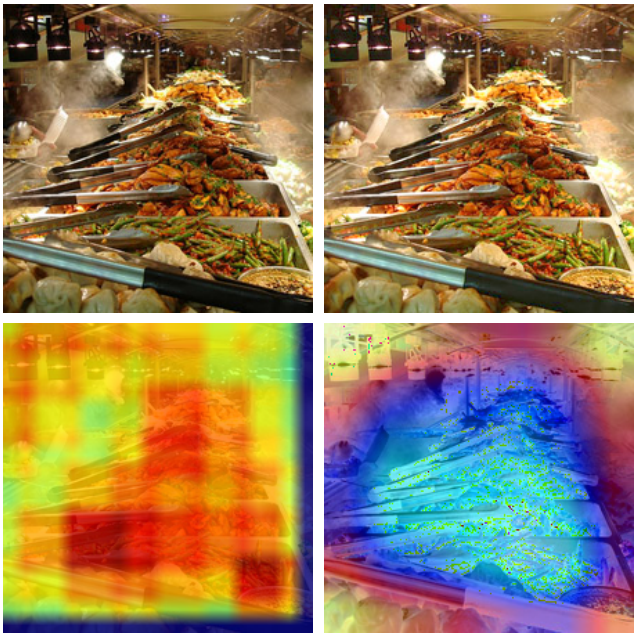

Figure 6. Brightness Severity = 1. The quality estimations from PaQ-2-PiQ are 76.75, 76.97 respectively (left to right).

The variation in accuracy and MOS wrt brightness distortion are minor. This is also reflected in both quality maps and attribution maps. Also, the change in brightness is causing a shift in the quality map, particularly the quality of certain patches. The slight differences can be seen in the quality map's top left and central areas. This difference can also be observed inside the blue region of the attribution map.

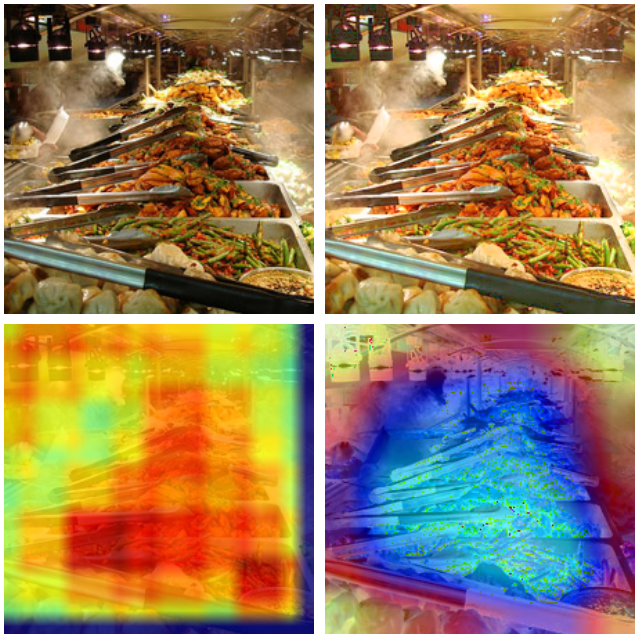

Figure 7. Brightness Severity = 2. The quality estimations from PaQ-2-PiQ are 76.75, 76.66 respectively (left to right).

5.2. Contrast Distortion

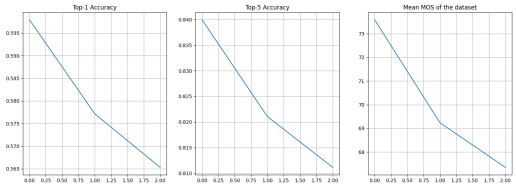

Figure 8. The figure show the change in Top@1, Top@5 Accuracies and Mean MOS on the validation dataset by changing severity of contrast corruption.

The variation in accuracy and MOS wrt change in contrast distortion is very significant. This is also reflected in both quality maps and attribution maps. The contrast distortion has shifted the focus of attribution maps. The area where the ResNet lost its focus is also where the image quality is low.

5.3. JPEG Compression Distortion

The variation in accuracy and MOS wrt change in jpeg compression distortion is very significant but unique. From the attribution maps, it can be understood that the JPEG compression didn't change the area the ResNet focuses on but has changed the importance of those areas. A similar

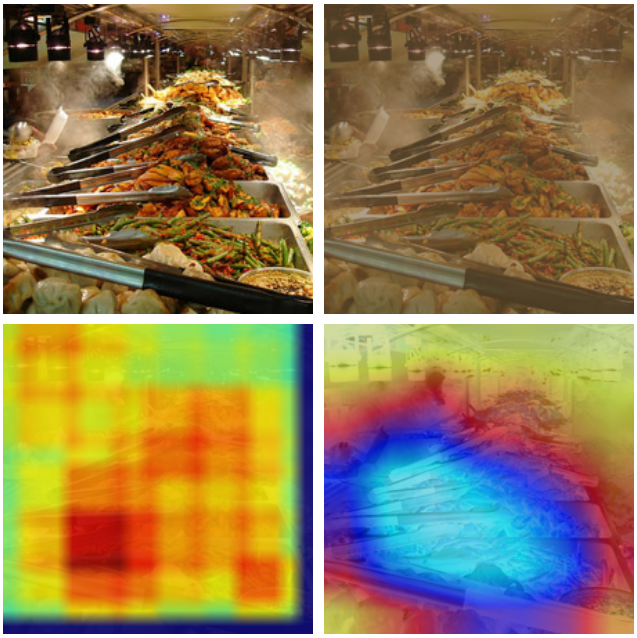

Figure 9. Contrast Severity = 1. The quality estimations from PaQ-2-PiQ are 76.75, 71.19 respectively (left to right).

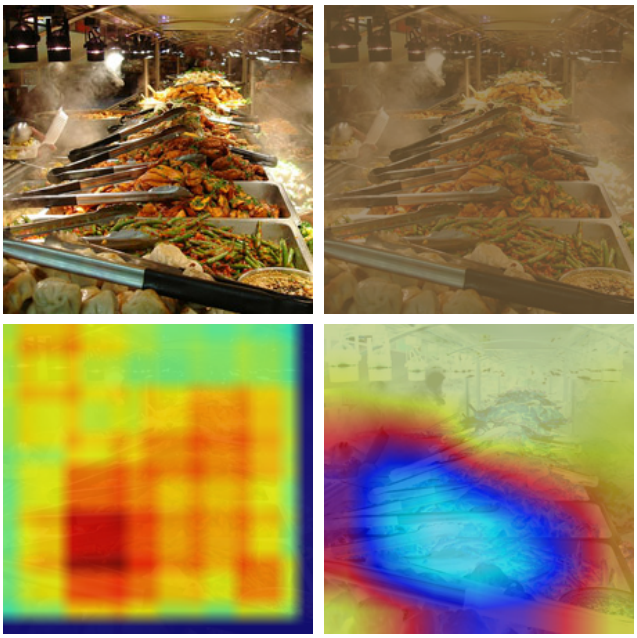

Figure 10. Constrast Severity = 2. The quality estimations from PaQ-2-PiQ are 76.75, 70.31 respectively (left to right).

goes for the case of quality maps; it has affected the quality of individual patches and the quality of the entire image, but the areas considered for overall quality prediction are not changed.

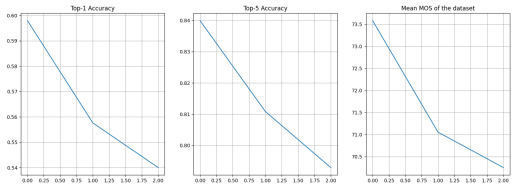

Figure 11. The figure show the change in Top@1, Top@5 Accuracies and Mean MOS on the validation dataset by changing severity of jpeg compression corruption.

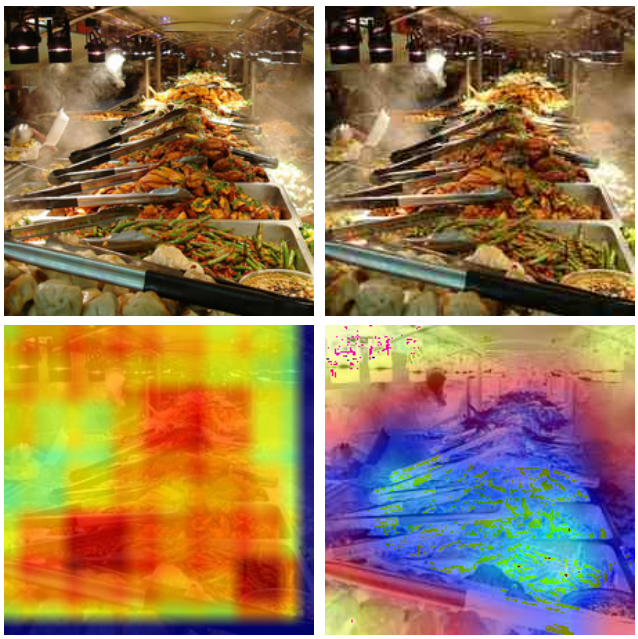

Figure 12. JPEG Compression Severity = 1. The quality estimations from PaQ-2-PiQ are 76.75, 72.19 respectively (left to right).

5.4. Saturate Distortion

The variation in accuracy and MOS wrt change in the saturation level of distortion is significant. The effect of saturation is very peculiar. In this case, saturation reduces the quality of the patches making the quality estimate of the image lower. But in attribution maps, the effect is quality is reflected, but there is less similarity between the change in quality map and attribution map.

6. Conclusion

Noise in an image causes performance degradation in classification models. While rating an image, humans observe various portions to rate the quality. PaQ-2-PiQ is an NR-IQA model that is trained on human ratings of patches and images. To understand the similarities of perception of humans and computer vision models, we have considered two models PaQ-2-PiQ for NR-IQA and ResNet, for classi-

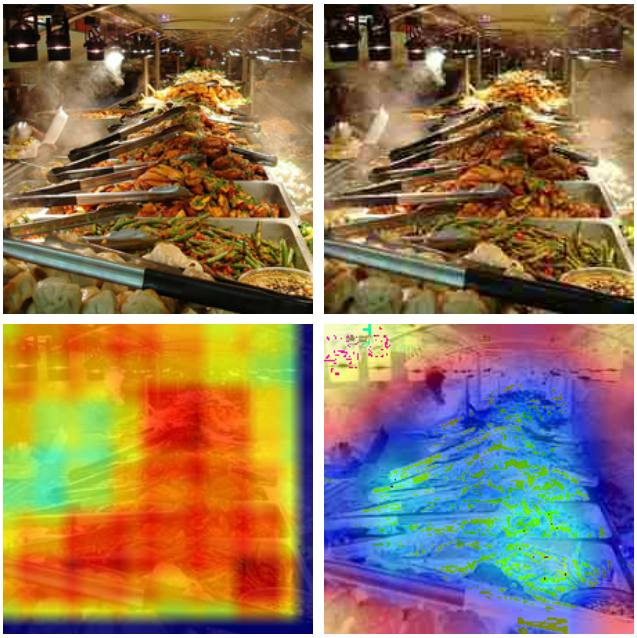

Figure 13. JPEG Compression Severity = 2. The quality estimations from PaQ-2-PiQ are 76.75, 71.58 respectively (left to right).

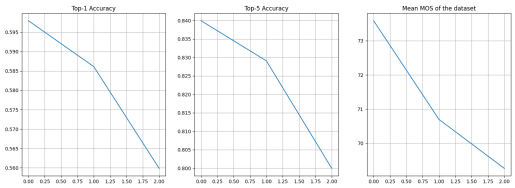

Figure 14. The figure show the change in Top@1, Top@5 Accuracies and Mean MOS on the validation dataset by changing severity of saturate corruption.

fication with a similar backbone structure of ResNet18 but with two different goals. The change in MOS of the image rated using PaQ-2-PiQ seems to be related to the accuracy of the classification model. This can not be explained by the plots having the variation of accuracies and MOS but also the saliency/attribution maps. The quality maps from PaQ-2-PiQ and attribution maps are affected similarly due to distortions. So, one can understand that there is some correlation between how humans perceive images while rating them and how computers see images while classifying them.

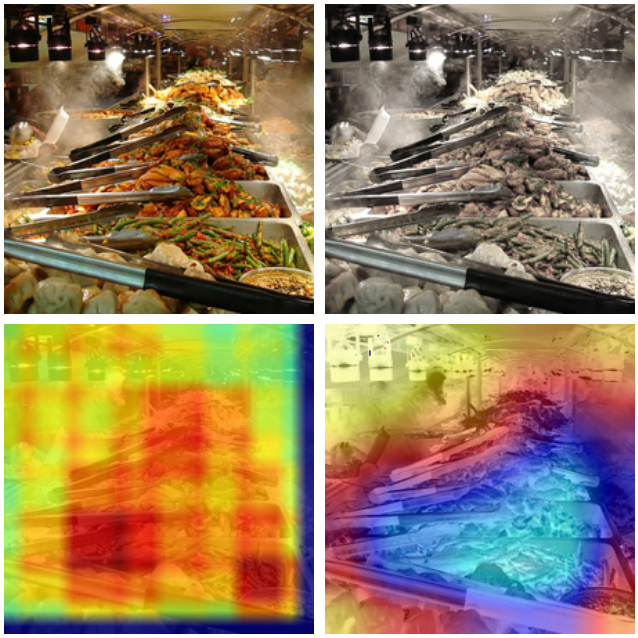

Figure 15. Saturation Severity = 1. The quality estimations from PaQ-2-PiQ are 76.75, 72.73 respectively (left to right).

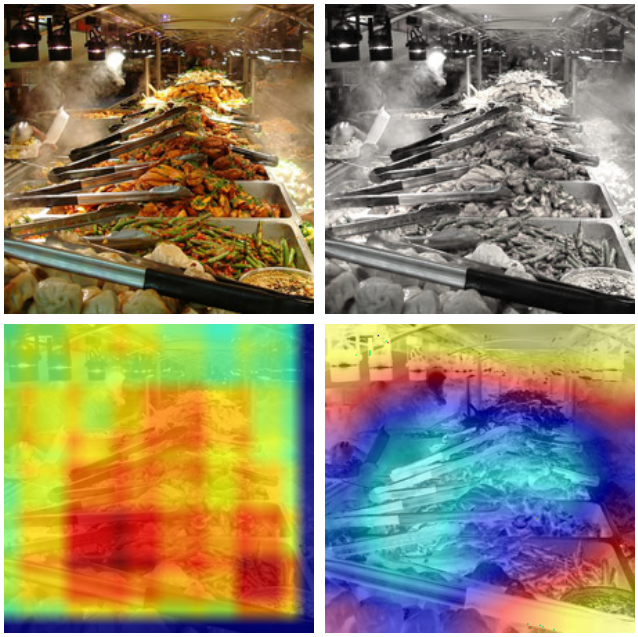

Figure 16. Saturation Severity = 2. The quality estimations from PaQ-2-PiQ are 76.75, 70.6 respectively (left to right).

References

[1] Murtaza Eren Akbiyik. Data augmentation in training {cnn}s: Injecting noise to images, 2020. 2

[2] Samuel F. Dodge and Lina J. Karam. Understanding how image quality affects deep neural networks. In *Eighth International Conference on Quality of Multimedia Experience, QoMEX 2016, Lisbon, Portugal, June 6-8, 2016*, pages 1–6. IEEE, 2016. 2

[3] Wilson S Geisler. Contributions of ideal observer theory to vision research. *Vision research*, 51(7):771–781, 2011. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE*

Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 770–778. IEEE Computer Society, 2016. 1

[5] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.*, 21(12):4695–4708, 2012. 1

[6] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.*, 20(3):209–212, 2013. 1

[7] Anush Krishna Moorthy and Alan Conrad Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364, 2011. 1

[8] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 413–420, 2009. 2

[9] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015. 1

[10] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1–9. IEEE Computer Society, 2015. 1

[11] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan C. Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3572–3582. Computer Vision Foundation / IEEE, 2020. 1