Image Quality for Deep Learning

Pooryaa Cheraaqee

September 2, 2022

1 Introduction

Digitized images are a prevalent medium in the Internet communications and a large portion of our information are conveyed by images and visual signals [1]. It is apparent that the unpleasant quality of such media can prevent us from obtaining the information that we desire or may cause disturbing viewing experiences. The same thing is possible to happen for the machines. To be clear, artificial intelligence (AI) algorithms use images as training data and if the quality of the data that is fed to such machine learning algorithms is poor, the algorithm may deviate from its normal performance.

Convolutinal neural networks (CNNs) are at the heart of a large number of computer vision applications, such as license plate recognition, face recognition, image classification, and object detection [2]. These networks operate by training on a set of images (dataset) and applying the gained knowledge in their task. Since CNNs are vitial for automating many tedious tasks, it is important to assure that the data which they are training on, is of adequate quality. One way to acheive this, is comming up with an objective measure of image quality that represents how well the CNN will perform if it is fed with this image for training or inference. Having such method for quantifying the quality of images for machine perception, we can automatically optimize deep learning datasets or discard missinformative image samples.

Image quality has been investigated by researchers and practitioners since the early days of photograpy [3]. Human has usually been considered as the ultimate viewer of images and hence the most reliable measures of quality were provided by *subjective* assessments [4]. During a subjective image quality assessment (IQA), a statistically sufficient number of observers (subjects) view an image and give a score to its quality (either as a number in the $[0, 10]$ interval or with any other psychometric scales [5]). Their mean opinion score (MOS[1]) provides a quantification of image quality. Due to the inconvinience of subjective experiments for on-line and large-scale applications, computational models are desired that can predict human judgements on image quality. The results of subjective experiments are used as a ground-truth for validating and training objective models [6].

As mentioned, the IQA research has been mostly concerned with human satisfaction of images, while it has been shown that metrifying image quality for other objectives is quite different [7]. In particular, to see how image quality can affect the performance deep neural networks, we need a model, f. f inputs an image, I, and produces a **bounded** value, such as in $[0, 1]$. That is, $f(I) \in [0, 1]$, and $f(I)$ correlates with the performance of CNN in learning from I or making inferences about I. It is also desired that the model can point out the flaws or distortions that have possibly made I inappropriate for using in a deep learning pipeline.

Apart from physically measuring image attributes or quality dimensions and devising explicit features, the problem can be tackled in an end-to-end paradigm. After presenting a background, a potential solution will be formulated in the section 3.

[1]MOS is in a direct proportion with the perceived quality and *difference*-MOS (DMOS) is in inverse proportion with the perceived quality. To make it convinient, we can state that MOS measures quality and DMOS measures *distotion* or artefact's severity.

2 Background

The most common researches for quality assessment (QA) of digitized images has recently belonged to the category of *computational metrics* as classified by Fry et. al. [8]. These methods are validated with subjective experiments [9]. These subjective experiments are published as datasets which are essentially sets of $(image, MOS/DMOS)$ ordered pairs. An objective method is applied to the images and its output is expected to correlate with the MOS/DMOS of images in the dataset. (The correlation is expressed with common correlation indices, such as Pearson linear correlation index (PLCC) and Spearman rank order correlation index (SROCC).)

A classification of IQA methods based on the availability of a pristine signal is good to be made here. In some applications, such as image compression, there are two images available: the initial raw image and the compressed image that may suffer from blocking or ringing artefacts. Suppose the lossy compression algorithm embeds a quality metric to assess the compressed image and in the case of a dire quality, automatically regulates its loss parameters. This quality metric has the pristine image in addition to the test image at its disposal and can use it as a reference. This is an example of *full-reference* (FR) IQA. In communcation settings, where the pristine signals are bulky to transmit, certain features can be extracted and sent over to aid the assessment of a distorted image (*reduced-reference* (RR) IQA). *No-reference* (NR) IQA is the case that the quality of an image is to be assessed without any information of its pristine version. (If such a reference even exists, like what happens in a DSLR camera.) It is believed that NR IQA is more difficult than the other cases [10].

Early FR computational metrics were based on error visibility. If $ref(x, y)$ and $dst(x, y)$ are the luminance components of the reference and the distorted images and they are both of dimensions $M \times N$, the mean squared error (MSE) for them is defined as:

$$MSE(ref, dst) = \frac{1}{M \times N} \sum_{x=1}^{M} \sum_{y=1}^{N} \left(ref(x, y) - dst(x, y)\right)^2 \tag{1}$$

Peak signal to noise ratio (PSNR) is also based on MSE. It has been shown that these methods do not necessarily correlate with human perception of quality [11].

Wang et. al. [12] showed that structures can be a reasonable criterion for image quality and proposed a method to metrify structural similarity (SSIM) in 2004 [13]. With the success of SSIM, other methods tried to express structures with other features, such as image edges [14] and gradients [15]. Natural scene statistics [16] and information fidelity [17] are other criteria for image quality. End-to-end machine learning is also exploited for the task of objective IQA [18, 19, 20].

Document image quality assessment (DIQA) is an example of measuring quality for an objective other than human perception [7]. The ground-truth for DIQA methods is not human opinion, but the accuracy of optical character recognition (OCR) methods. The DIQA metric needs to predict how accurate an OCR method will perform on the given image. The same approach can be taken for evaluating an image's suitability for deep learning tasks. A similar problem is video coding for machine [21], where the goal is to optimize the coding with the assumption that an algorithm is going consume the uncompressed data for some recognition task.

3 Proposed Method

Convolutional neural networks can be considered as a function that maps an image to some output. This output can be a number, multiple numbers, another image, a label, a combination of numbers and labels, etc. No matter what is the output, the network needs a ground-truth to guide it in tuning its weights. If x is the input image, then $y = CNN(x)$ is the output of the network and $\hat{y}$ is the correct answer for x. Some measure of error (*loss function*) tells the network how bad y is; according to $\hat{y}$: $\text{error}_x = loss(y, \hat{y})$. The network then tries to minimize the error with each labeled sample that it is fed with.

We need a model, f, that given x, predicts network's error for x. That is, ideally:

$$f(x) = \text{error}_x, x \in D \tag{2}$$

3.1 End-to-End Approach

Given a labeled dataset, D, for a CNN task, the network can be applied to each sample and an error can be computed accordingly. Each image, x, in D can be distorted using synthesized artefacts, such as Gaussian blur, Gaussian noise, and JPEG compression as in image quality datasets [22]. Each artefact, a, can be applied at various levels of severity, l. It is hoped that the network performance, $error$, has a meaningful relation with l.

The ground-truth for x in D is represented as $\hat{y}_x$. It is clear that the distorted version of x with artifcat a at level l, $x_{a,l}$, will have the same ground-truth for the task of D. To compute the error of the CNN for each $x_{a,l}$, the same $\hat{y}_x$ can be used: $\text{error}_{x_{a,l}} = loss(CNN(x_{a,l}), \hat{y}_x), a \in A, l \in L$, where A is the set of all distortion types employed and L is the set of all levels of severity.

By computing $\text{error}_{x_{a,l}}$ for all combinations of xs, as, and ls, we will have a set of ordered pairs like $(x_{a,l}, error_{x_{a,l}})$. This set, $D\prime$, can serve as a dataset for training a second CNN, called $CNN_{metrifier}$. $CNN_{metrifier}$ receives an image and tries to learn what error the CNN will make on this image. A sucessful $CNN_{metrifier}$ can serve as the model f in (2).

3.2 Bottom-up Approach

Characteristics of human visual system (HVS) have been discovered, modeled, and employed in predicting our judgements of image quality [9, 8]. CNNs can be analogous to HVS and trying to interpret their inner-mechanism may reveal some traits that can be explicitly used to measure image dimensions and their effect on network performance. A probable phenomenon to witness is that the effect of distortions may be more apparent in the early layers of a network in comparison to the ending layers that are more semantic. The network architectural characteristics, such as drop-out layers, can also have alleviating effects in propagating the distortions. Studies such as [23, 24, 25, 26, 27] may bring insights to this matter.

It is worth to note that the approach proposed in section 3.1 was task agnostic. However, it is probably different to predict the performance of an image classifier on an image than an object detector. A more detailed formulation of the proposed method may be found in the Jupyter notebooks at https://github.com/cheraaqee/iqa_for_cnn.

4 Timing& Plan

Figure 1 provides an overview of the tasks considered for the research.

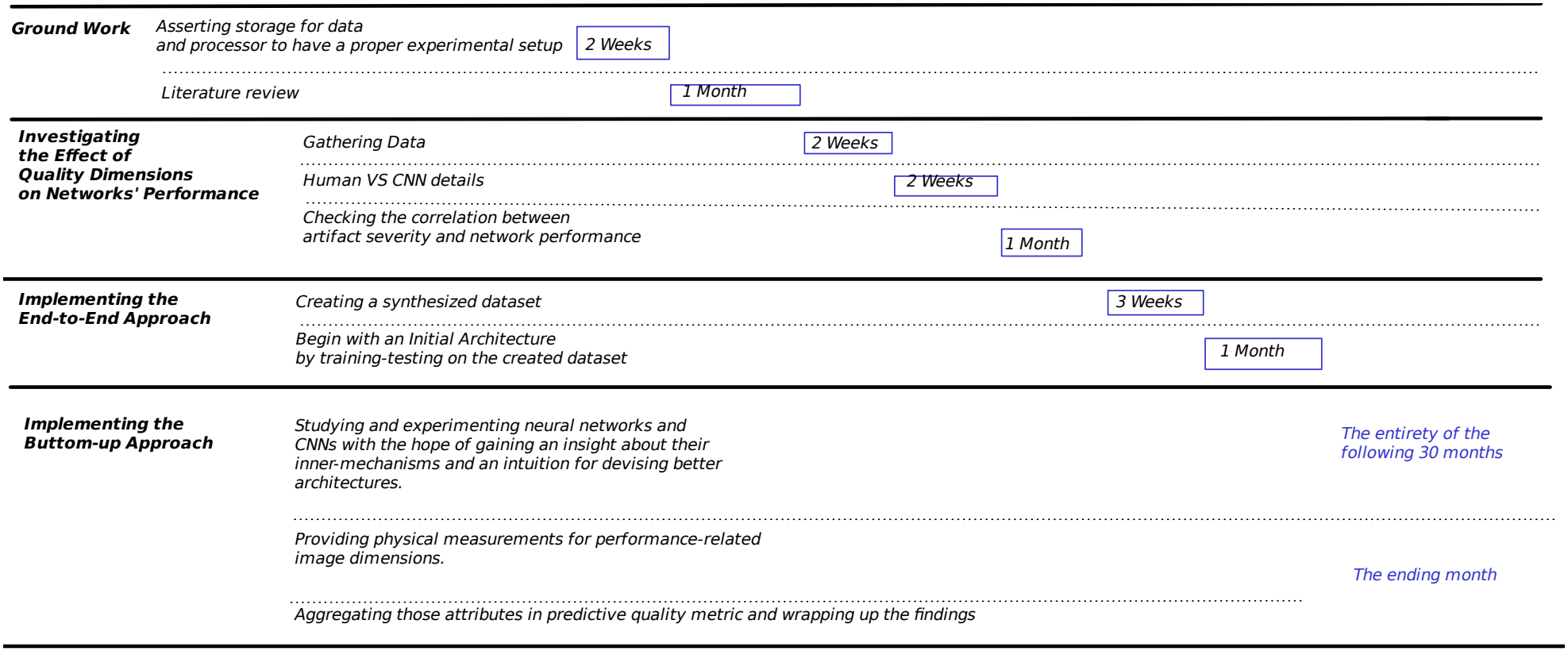

Figure 1: Gantt chart for the proposed method

References

[1] CISCO, "Cisco annual internet report." `https://www.cisco.com/c/en/us/solutions/executive-perspectives/annual-internet-report/index.html`, 2022. Accessed: 2022-July-15.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[3] E. Allen and S. Triantaphillidou, *The manual of photography*. CRC Press, 2012.

[4] D. M. Chandler, "Seven challenges in image quality assessment: past, present, and future research," *ISRN Signal Processing*, vol. 2013, 2013.

[5] S. S. Stevens, "On the theory of scales of measurement," *Science*, vol. 103, no. 2684, pp. 677–680, 1946.

[6] K. Seshadrinathan and A. C. Bovik, "Automatic prediction of perceptual quality of multimedia signals—a survey," *Multimedia Tools and Applications*, vol. 51, no. 1, pp. 163–186, 2011.

[7] P. Ye and D. Doermann, "Document image quality assessment: A brief survey," in *2013 12th International Conference on Document Analysis and Recognition*, pp. 723–727, IEEE, 2013.

[8] E. W. Fry, S. Triantaphillidou, R. E. Jacobson, J. R. Jarvis, and R. B. Jenkin, "Bridging the gap between imaging performance and image quality measures," *Electronic Imaging*, vol. 2018, no. 12, pp. 231–1, 2018.

[9] G. Zhai and X. Min, "Perceptual image quality assessment: a survey," *Science China Information Sciences*, vol. 63, no. 11, pp. 1–52, 2020.

[10] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1733–1740, 2014.

[11] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE signal processing magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[12] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE signal processing letters*, vol. 9, no. 3, pp. 81–84, 2002.

[13] Z. 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.

[14] G.-H. Chen, C.-L. Yang, L.-M. Po, and S.-L. Xie, "Edge-based structural similarity for image quality assessment," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 2, pp. II–II, IEEE, 2006.

[15] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 668–695, 2014.

[16] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

[17] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on image processing*, vol. 15, no. 2, pp. 430–444, 2006.

[18] X. Yang, F. Li, and H. Liu, "A survey of dnn methods for blind image quality assessment," *IEEE Access*, vol. 7, pp. 123788–123806, 2019.

[19] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *2012 IEEE conference on computer vision and pattern recognition*, pp. 1098–1105, IEEE, 2012.

[20] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal processing magazine*, vol. 34, no. 6, pp. 130–141, 2017.

[21] L. Duan, J. Liu, W. Yang, T. Huang, and W. Gao, "Video coding for machines: A paradigm of collaborative compression and intelligent analytics," *IEEE Transactions on Image Processing*, vol. 29, pp. 8680–8695, 2020.

[22] H. Sheikh, "Live image quality assessment database release 2," *http://live. ece. utexas. edu/research/quality*, 2005.

[23] M. Kumar, N. Houlsby, N. Kalchbrenner, and E. D. Cubuk, "On the surprising tradeoff between imagenet accuracy and perceptual similarity," *arXiv preprint arXiv:2203.04946*, 2022.

[24] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6228–6237, 2018.

[25] A. Berardino, V. Laparra, J. Ballé, and E. Simoncelli, "Eigen-distortions of hierarchical representations," *Advances in neural information processing systems*, vol. 30, 2017.

[26] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

[27] D. Amir and Y. Weiss, "Understanding and simplifying perceptual distances," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12226–12235, 2021.