IMAGE QUALITY ASSESSMENT BASED LABEL SMOOTHING IN DEEP NEURAL NETWORK LEARNING

Zhuo Chen[1] *Weisi Lin*[2] *Shiqi Wang*[3] *Long Xu*[4] *Leida Li*[5]

[1] ROSE Lab, Interdisciplinary Graduate School, Nanyang Technological University, Singapore
[2] School of Computer Science and Engineering, Nanyang Technological University, Singapore
[3] Department of Computer Science,City University of Hong Kong, Hong Kong, China
[4] NAOC, Chinese Academy of Sciences, China [5] China University of Mining and Technology, China

ABSTRACT

For many computer vision problems, deep neural networks are trained and validated based on the assumption that the input images are pristine (i.e., artifact-free). However, digital images are subject to a wide range of distortions in real application scenarios, while the practical issues regarding image quality in high level visual information understanding have been largely ignored. In this paper, in view of the fact that most widely deployed deep learning models are susceptible to various image distortions, distorted images are involved for data augmentation in the deep neural network training process to learn a reliable model for practical applications. In particular, an image quality assessment based label smoothing method, which aims at regularizing the label distribution of training images, is further proposed to tune the objective functions in learning the neural network. Experimental results show that the proposed method is effective in dealing with both low and high quality images in the typical image classification task.

Index Terms— Deep learning, image quality assesment

1. INTRODUCTION

Recently, deep neural networks (DNNs) have demonstrated state-of-the-art performance in various computer vision tasks since the AlexNet model [1] achieved 9% better classification accuracy than the previous hand-crafted methods in ILSVRC 2012 [2]. In contrast to the handcrafted features such as Scale-Invariant Feature Transform (SIFT) [3], deep learning based approaches are able to learn representative features directly from the vast amounts of data, which makes it feasible to achieve outstanding performance with the explosion of big data. However, such property gives rise to the fact that the capability of deep models heavily relies on the training samples. In particular, most DNN models were trained and tested based on the assumption that the input image samples are pristine without any distortions injected. As such, they can achieve promising performance on high quality samples, but the performance will be seriously degraded when encountering with low quality images. Fig. 1 provides some examples in CIFAR-10 dataset [4] and it is shown that DNN model fails in predicting the correct classes when the input images are distorted. A recent work [5] evaluated several classical deep models for image classification by injecting different types of distortion into the test images. The results show that all the evaluated neural networks are susceptible to typical distortions such as blur and noise. For example, more than 20% Top 1 and Top 5 accuracy drop can be observed when the images are distorted by Gaussian blur.

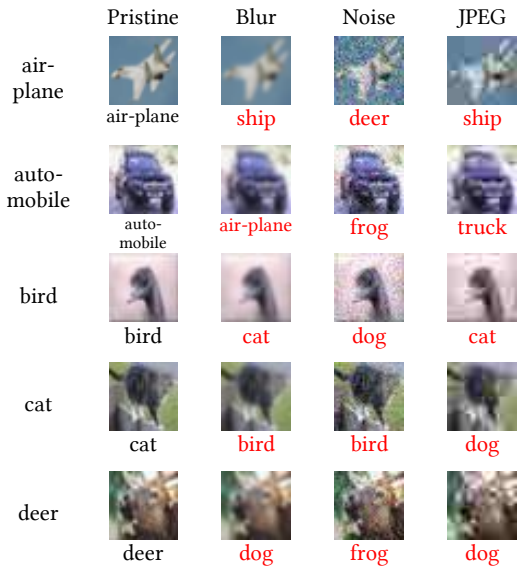

Fig. 1. Illustration of classification results in CIFAR-10. The first column is the pristine sample image of CIFAR-10, whose ground-truth label is on the left and the predicted result is on the bottom of the image. The second, third and fourth columns are the images corrupted by blur, noise and JPEG compression respectively.

In real application scenarios, distortions will be introduced in image acquisition, compression, processing, transmission and reproduction. Evaluating the visual quality of

these distorted images becomes meaningful. In the literature, there are numerous approaches proposed to assess the degradation of visual quality [6]. Popular image quality assessment (IQA) algorithms such as SSIM [7], FSIM [8], GSIM [9], VSNR [10], PCQI [11], etc., focus on the perception of quality degradation from the perspective of viewing experience. Due to the fact that the distortions can also bring difficulties in image understanding, it becomes more and more important to further investigate the applications of these IQA algorithms in the context of computer vision, as computer vision systems aim to automatically achieve the high-level understanding tasks that the human visual system can perform.

This inspires us to incorporate the quality measure in the DNN learning process to deal with the visual understanding with low quality images. In particular, we first train the deep neural network by augmenting data with mixture of pristine and distorted data. Then an IQA-based label smoothing technique is proposed to enhance the performance of deep models by fine-tuning the network with the IQA measure. In this manner, the robustness of DNN models with distorted input data can be significantly improved. Experimental results show that the proposed scheme can significantly improve the classification performance of both high and low quality images.

2. DATA AUGMENTATION WITH DISTORTED IMAGES FOR DEEP LEARNING

It is generally acknowledged that the capability of the deep model largely depends on the training data. As such, to deal with the low quality test images, the straightforward solution is augmenting the training data with low quality versions and mixing them together with the original high quality data. In view of this, in this work we involve the pristine as well as the distorted images as the input data to learn the CNN model. The distorted versions are generated by injecting different levels of distortion into the pristine training images manually. As one of the first attempts using distorted images as the training data, here we are particularly interested in three types of commonly encountered distortions: blur, noise and JPEG compression.

All these three types of distortions may introduce disturbance in representing the semantic information of images by damaging the image content. According to the study in [5], not only the human visual system, but also computer vision algorithms cannot efficiently recognize and understand the visual information with these three kinds of distortions. Therefore, it is meaningful to start from these artifacts, which pose a unique set of challenges to computer vision tasks.

However, there are also side effects when both pristine and distorted images are used as the training data. In particular, although the model can provide promising prediction performance on the distorted images in principle, it may also significantly degrade the testing performance when we feed the high quality images as the input. Thus, in this work we further seek a good balance between the high quality and low quality training images with IQA to learn a more robust model.

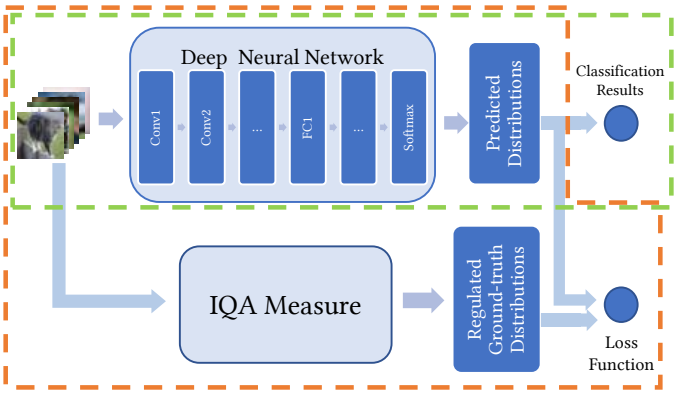

Fig. 2. The deep learning framework with IQA-based label smoothing. The modules in the orange dashed box represent the training process while the modules in the green dashed box stand for the inference process.

3. IQA-BASED LABEL SMOOTHING

As mentioned early, although the corrupted images are involved in the training process to deal with the scenario of low quality images as the input, such a method may not permanently solve the problem, and there are several challenging issues:

1) Since both high and low quality images are used as the training data, the learned model is lack of the generalization capability and exhibit strong bias to the low quality images. As such, though we can improve the accuracy of low quality images, the prediction performance of the high quality images will be degraded.
2) The human visual perception has been largely ignored in the learning process. It is generally hypothesized that the human visual system evolves through learning from the natural images that possess certain statistical properties. As such, low quality images which belong to unnatural images should play a less importance role compared with pristine images since low-quality images are more difficult to understand. A reasonable way to manipulate this is to make the expected probabilities corresponding to the ground-truth labels of the low-quality images lower. However, the commonly used ground-truth distribution does not follow this trend.

To avoid these drawbacks and incorporate the brain-like perception in the deep learning framework, an IQA-based label smoothing method (IQA-LS) is proposed, as shown in Figure 2. In particular, given the label $k \in \{1 \cdots K\}$ and a single input x with ground-truth label y, instead of one-hot encoded label distribution $q(k|x) \in \{0, 1\}$, we reformulate

the label distribution as follows

$$q'(k|x) = \begin{cases} T(s(x)) & k = y \\ (1 - T(s(x)))/(K-1) & k \neq y \end{cases} \quad (1)$$

where $s(\cdot)$ denotes the score of IQA measure and $T(\cdot)$ transforms the IQA score to the range of $(0, 1]$. This implies that the distribution of the label k is obtained based on the IQA score of the input image x when $k = y$, while the uniform distribution is employed for the rest labels. Therefore, the confidence value is directly determined by the image quality, and better quality implies higher confidence in the network learning. This is in line with the human perception when understanding the image content, as low quality images may be perceived with higher uncertainties from the perspective of free-energy theory [12].

In this work, we adopt the SSIM [7] as the IQA measure and identity function (i.e. $f(x) = x$) for $T(\cdot)$, due to its good trade-off between the accuracy and computational complexity. In particular, it is computed by comparing the original and distorted images based on the degradation of the structural information. IQA measure is only employed in the training procedure where we can get access to both the distorted and its corresponding pristine images, as shown in Figure 2. It is also worth mentioning that other IQA algorithms including reduced-reference (RR)[13, 14, 15] and no-reference (NR)[16, 17, 18, 19] methods are also compatible with our proposed IQA-based label smoothing framework.

4. EXPERIMENTAL RESULTS

4.1. Dataset and learning architecture

In this paper, the proposed scheme is evaluated on CIFAR-10 dataset which is a labelled subset of 80 million tiny images [20]. The dataset contains 50,000 training samples and 10,000 testing samples in 10 different classes for performing the classification tasks. In order to compare the models trained with different strategies, except for specific image distortion, data enhancement strategies (e.g. image contrast, brightness and saturation adjustment) are prohibited in the training process. Here, as discussed in Section 2, we only apply Gaussian blur, Gaussian noise and JPEG compression on the pristine training images.

The learning architecture is designed following Alex Krizhevsky's work [1] with a few modifications on Tensorflow [21]. Specific descriptions regarding the proposed learning architecture is illustrated in Table 1.

4.2. Parameter setting

The proposed deep architectures are trained with stochastic gradient descent method on a NVIDIA GeForce 980Ti GPU with batch size 100 for 2,000 epochs. All our experiments use the initial learning rate of 0.1 which decays for every 350 epochs with an exponential rate of 0.1. In addition, L2Loss weight decay multiplied by 0.004 is added to the two full-connected layers.

Table 1. Descriptions of the learning architecture. The input and output sizes are specified as $rows \times cols \times channels$, and the kernel is characterized in terms of $rows \times cols, stride$.

layer	size-in	size-out	kernel
$conv1$	$32 \times 32 \times 3$	$32 \times 32 \times 64$	$5 \times 5, 1$
$pool1$	$32 \times 32 \times 64$	$16 \times 16 \times 64$	$3 \times 3, 2$
$lrn\&ReLU$	$16 \times 16 \times 64$	$16 \times 16 \times 64$	/
$conv2$	$16 \times 16 \times 64$	$16 \times 16 \times 64$	$5 \times 5, 1$
$lrn\&ReLU$	$16 \times 16 \times 64$	$16 \times 16 \times 64$	/
$pool2$	$16 \times 16 \times 64$	$8 \times 8 \times 64$	$3 \times 3, 2$
$fc1\&ReLU$	4096	384	/
$fc1\&ReLU$	384	192	/
$softmax$	192	10	/

Regarding the training data, pristine images and the mixture data are used. The pristine data are totally from the CIFAR-10 training dataset, and the mixture data are the combination of pristine and distorted images with different distortion levels in a fixed ratio. In each training epoch, 60% of the pristine training samples are maintained, the rest samples are corrupted by $level$ $1/2/3$ distortions in the ratio of 15% / 15% / 10%. Specifically, for the blur distortion, we use the Gaussian kernels with $\sigma = 0.7, 1.0, 1.2$ for $levels$ from 1 to 3 respectively. With respect to the noise artifacts, white Gaussian noise with variance values $v = 0.005, 0.01, 0.02$ are employed. Regarding to JPEG compression, we compress the images with the JPEG quality factors of $12, 8, 4$.

In summary, five different types of training set are utilized in the experiments: one pristine set and four mixture data sets. For convenience, the four mixture data sets are denoted as MIX_{blur}, MIX_{noise}, MIX_{JPEG} and MIX_{all3}, where MIX_{blur}, MIX_{noise} and MIX_{JPEG} represent the training sets of mixture data with pristine and distorted images degraded by one certain type of distortion, while MIX_{all3} is the combination of pristine and all the three types of distorted samples.

With respect to the testing data, a pristine set and nine distorted sets (with three different types and each one has three levels) are generated for evaluating each learned model. Example images with the three types of artifact are also illustrated in Fig. 1.

4.3. Performance comparisons

Nine different training approaches are implemented to evaluate the performance of the proposed scheme. All these nine approaches share the same architecture but different training strategies. These training strategies are with different combinations of data augmentation strategies and label distributions. Here, we will detail these training strategies and analyse their performance in terms of the top 1 classification accuracy, as illustrated in Table 2.

Table 2. Performance comparisons of the models with different training strategies.

Strategy	Regularization Method	Training Set	Pristine	Blur			Noise			JPEG		
				Level 1	Level 2	Level 3	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
1	Original	$Pristine$	0.794	0.676	0.524	0.436	0.677	0.566	0.392	0.600	0.529	0.391
2	Original	MIX_{blur}	0.781	0.777	0.766	**0.751**	0.735	0.681	0.585	0.669	0.621	0.469
3	IQA-LS	MIX_{blur}	**0.798**	**0.782**	**0.766**	0.749	0.749	0.698	0.607	0.681	0.624	0.465
4	Original	MIX_{noise}	0.794	0.716	0.606	0.526	0.779	0.773	**0.749**	0.693	0.640	0.496
5	IQA-LS	MIX_{noise}	**0.807**	0.736	0.612	0.525	**0.792**	**0.776**	0.743	0.699	0.642	0.503
6	Original	MIX_{JPEG}	0.753	0.728	0.711	0.656	0.720	0.662	0.616	0.710	0.687	**0.607**
7	IQA-LS	MIX_{JPEG}	**0.791**	0.751	0.704	0.622	0.742	0.667	0.607	**0.722**	**0.693**	0.590
8	Original	MIX_{all3}	0.767	0.753	0.737	0.721	0.761	0.744	0.725	0.721	0.686	**0.599**
9	IQA-LS	MIX_{all3}	**0.790**	**0.773**	**0.753**	**0.738**	**0.771**	**0.757**	**0.731**	**0.726**	**0.692**	0.585

4.3.1. Training on pristine dataset

Strategy 1 in Table 2 aims to train deep models with the pristine CIFAR-10 training samples without the augmented data. Moreover, the label distribution of each training image is the classical 0-1 distribution. Such training strategy follows the widely adopted benchmark models such as AlexNet and VGG. Therefore, we consider this approach as the baseline.

From Table 2, we can see that the model trained with this strategy is sensitive to all the three involved distortions as the classification performance decreases dramatically when the distortion level increases. This phenomenon can be explained by the fact that the distortions can heavily remove the texture and edge information in an image, which is important to the DNN models learned with pristine images since such DNN models may always attempt to look for specific textures and edges for the classification task.

4.3.2. Training on mixture dataset

Strategies 2,4,6,8 in Table 2 train DNN models with the mixture data of pristine and distorted images while the label distribution maintains the typical 0-1 distribution. Such kind of training approach is a straightforward solution to make the network better adapt to the distorted images.

As shown in Table 2, the results exhibit that training strategy with low quality samples improves the performance on the corresponding distorted images. For instance, the classification accuracy of Strategy 2 only decays about 1% when the distortion level rises from pristine to level 1. Such decreasing speed is an order of magnitude slower than that of the baseline strategy. However, it is noticed that the performance of these strategies on high-quality pristine images cannot approach as high as the baseline method. As discussed in Section 3, this is due to the fact that 0-1 label distribution teaches the model to be equally confident about the classification results of both high and low quality images.

4.3.3. Training with IQA-based label smoothing

Strategies 3,5,7,9 in Table 2 target at training the model based on the mixture data as well. Moreover, in contrast to the previous strategies, the label distribution is regularized by the proposed IQA-based label smoothing method.

The performance of models trained with IQA-LS is credibly better than the original ones on relative high quality images (e.g.,pristine, distortion levels 1 and 2). Regarding to the performance on strongly distorted images (e.g., distortion level 3), although the models trained with IQA-LS are slightly weaker than those without IQA-LS, the performance drop is marginal and acceptable. Therefore, it is concluded that, comparing to the straightforward way that trains the deep models on mixture data, our proposed IQA-LS technique is not only effective in maintaining the high classification performance for distorted samples, but also promising in improving the accuracy on high quality test data. Moreover, it is observed that when training on mixture of samples with multiple types of artifacts rather than a certain type the superiority of IQA-LS is more apparent. This can be explained by the reason that the regularized label distribution penalises the false inference based on the quality levels, which provides the DNN models with stronger generalizing ability.

5. CONCLUSION

We have proposed a quality assessment based label smoothing approach for deep neural network learning. The novelty of the proposed approach lies in that the distorted images are included in the training process in learning the reliable neutral network model, and IQA is adopted in regularizing the label distribution of training samples to obtain a more robust representation. The performance of the proposed scheme is evaluated based on image classification and it is shown that the proposed scheme achieves high prediction accuracy across different distortion types and levels.

Acknowledgement

This work was partially supported by Singapore Ministry of Education Tier-2 Fund MOE2016-T2-2-057(S), NSFC 61572461,1179305,11433006,61771473,61379143, the Six Talent Peaks High-level Talents in Jiangsu Province (XYDXX-063) and the Qing Lan Project. The research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the Infocomm Media Development Authority, Singapore.

6. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[3] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[4] Alex Krizhevsky and Geoffrey Hinton, "Learning multiple layers of features from tiny images," 2009.

[5] Samuel Dodge and Lina Karam, "Understanding how image quality affects deep neural networks," in *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*. IEEE, 2016, pp. 1–6.

[6] Weisi Lin and C-C Jay Kuo, "Perceptual visual quality metrics: A survey," *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297–312, 2011.

[7] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

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

[9] Anmin Liu, Weisi Lin, and Manish Narwaria, "Image quality assessment based on gradient similarity," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1500–1512, 2012.

[10] Damon M Chandler and Sheila S Hemami, "Vsnr: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE transactions on image processing*, vol. 16, no. 9, pp. 2284–2298, 2007.

[11] Shiqi Wang, Kede Ma, Hojatollah Yeganeh, Zhou Wang, and Weisi Lin, "A patch-structure representation method for quality assessment of contrast changed images," *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2387–2390, 2015.

[12] Guangtao Zhai, Xiaolin Wu, Xiaokang Yang, Weisi Lin, and Wenjun Zhang, "A psychovisual quality metric in free-energy principle," *IEEE Transactions on Image Processing*, vol. 21, no. 1, pp. 41–52, 2012.

[13] Lin Ma, Xu Wang, Qiong Liu, and King Ngi Ngan, "Reorganized dct-based image representation for reduced reference stereoscopic image quality assessment," *Neurocomputing*, vol. 215, pp. 21–31, 2016.

[14] Xu Wang, Qiong Liu, Ran Wang, and Zhuo Chen, "Natural image statistics based 3d reduced reference image quality assessment in contourlet domain," *Neurocomputing*, vol. 151, pp. 683–691, 2015.

[15] Xu Wang, Lin Ma, Sam Kwong, and Yu Zhou, "Quaternion representation based visual saliency for stereoscopic image quality assessment," *Signal Processing*, vol. 145, pp. 202–213, 2018.

[16] Qiuping Jiang, Feng Shao, Weisi Lin, Ke Gu, Gangyi Jiang, and Huifang Sun, "Optimizing multi-stage discriminative dictionaries for blind image quality assessment," *IEEE Transactions on Multimedia*, 2017.

[17] Qiuping Jiang, Feng Shao, Gangyi Jiang, Mei Yu, and Zongju Peng, "Supervised dictionary learning for blind image quality assessment using quality-constraint sparse coding," *Journal of Visual Communication and Image Representation*, vol. 33, pp. 123–133, 2015.

[18] Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang, "Using free energy principle for blind image quality assessment," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 50–63, 2015.

[19] Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang, "Hybrid no-reference quality metric for singly and multiply distorted images," *IEEE Transactions on Broadcasting*, vol. 60, no. 3, pp. 555–567, 2014.

[20] Antonio Torralba, Rob Fergus, and William T Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 11, pp. 1958–1970, 2008.

[21] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.