ROBUSTNESS OF DEEP CONVOLUTIONAL NEURAL NETWORKS FOR IMAGE DEGRADATIONS

Sanjukta Ghosh[⋆†] *Rohan Shet*[⋆†] *Peter Amon*[†] *Andreas Hutter*[†] *André Kaup*[⋆]

[⋆]Multimedia Communications and Signal Processing,
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany
[†]Sensing and Industrial Imaging, Siemens Corporate Technology, Munich, Germany

ABSTRACT

Deep convolutional neural networks (CNNs) have achieved tremendous success in image recognition tasks. However, the performance of CNNs degrade in situations where the input image is degraded by compression artifacts, blur or noise. In this paper, we analyze some of the common CNNs for degradations in images caused by Gaussian noise, blur as well as compression using JPEG and JPEG 2000 for the full range of quality factors. Moreover, we propose a method to improve the performance of CNNs for image classification in the presence of input images with degradations based on a master-slave architecture. Our method was found to perform well for individual and combined degradations.

Index Terms— Deep convolutional neural networks, robustness, image compression, noise, blur

1. INTRODUCTION

Deep neural networks perform well in various tasks like classification, detection, semantic segmentation, super-resolution to name a few. Especially in image classification [1], excellent levels of performance have been reached on datasets like ImageNet and others. Moreover, the classification network is also one of the key building blocks of detection algorithms like R-CNN [2], Fast R-CNN [3] and Faster R-CNN [4]. However, images acquired in practice are often degraded by sensor noise, compression artifacts or motion blur to name some causes. In order to effectively use deep neural networks in applications, it is required that the model is robust against such degradations. The main contributions of our paper are
1) Analysis of common CNN architectures for image classification performance to degradations in images due to compression artifacts caused by JPEG and JPEG2000, noise and blur.
2) Proposing a master-slave architecture and its alternative for improving the performance.

The research leading to these results has received funding from the German Federal Ministry for Economic Affairs and Energy under the VIRTUOSE-DE project.

3) Testing and analyzing the performance on not just individual degradations but also combinations of degradations.

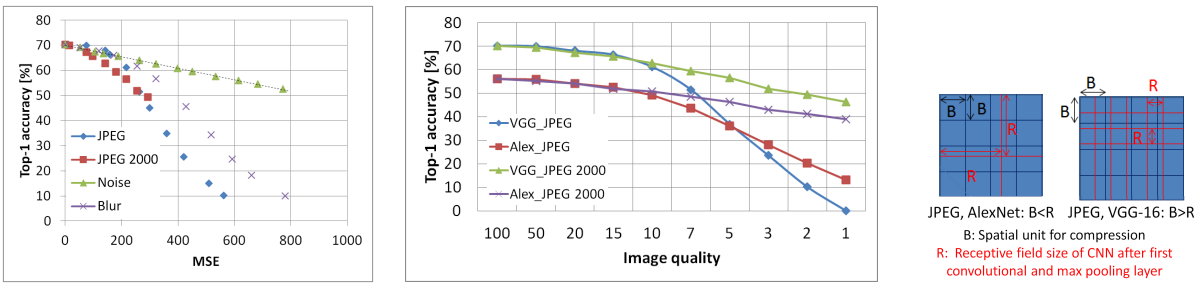
2. RELATED WORK

Dodge at al. [5] analyze various CNN architectures for degradations due to JPEG and JPEG 2000, blur, noise and contrast. However, there is no common parameter for comparing the performance across various degradations. While it provides valuable analysis, no method is proposed for making the CNNs robust to degradations. Data augmentation is a commonly used mechanism to increase the robustness of CNNs [6], [7]. Zheng et al. [8] propose a method for increasing the robustness of deep neural networks by stability training. In this method, Gaussian noise is added to the training samples. The cost function has a factor that tries to minimize the distance between the clean and degraded image features. Wang et al. [9] describe a theoretical framework to analyze the robustness of CNNs against adversarial examples. Gao et al. [10] propose a technique for dealing with adversarial examples by masking out features. A popular related topic is to investigate the robustness of deep neural nets to adversarial examples [11], [12]. In [13], though not explicitly trained for robustness to degradations, using a similarity metric in a Siamese network brings about robustness to degradations like motion blur and poor illumination.

3. ANALYSIS

Images with blur, noise and compression artifacts due to JPEG and JPEG 2000 compression are used to test the performance of CNN based image classification. We model blur by convolution with a Gaussian blur kernel and sensor noise by additive Gaussian noise. Images are compressed using the JPEG and JPEG 2000 compression scheme at various quality factors. The classification performance is measured using accuracy at top-1 score in percentage.

Fig. 1a shows the performance of image classification using the VGG-16 network [14] on the 50,000 images of the validation set of the ImageNet [15] classification challenge,



(a) (b) (c)

Fig. 1: (a)VGG-16 and various degradations. (b)VGG-16 and AlexNet for JPEG and JPEG 2000. (c)Relative size of compression unit and receptive field of CNN.

ILSVRC-12. It is observed that the tested CNNs are more robust to noise than other artifacts. This is because Gaussian noise is independent of the content of the image and the spatial distribution of the noise across the image is uniform. Under mean square error (MSE) as metric, it is observed that compression artifacts have significant performance degradation in comparison to that caused by Gaussian noise. So it is worthwhile to explore the robustness of CNNs for degradations in images due to compression artifacts.

Fig. 1b shows the performance of AlexNet [16] and VGG-16 [14] networks for JPEG and JPEG 2000 compressed images. It is observed that the tested CNNs are robust to compression artifacts up to a point and then the performance starts degrading. As can be observed from the graph, VGG-16 has a higher classification accuracy than AlexNet for almost all the cases except JPEG compressed images for quality factor 5 and below. However, the percentage degradation is less in case of AlexNet than VGG-16 implying that AlexNet is more robust to compression artifacts than VGG-16. Thus, there is a dependency of the robustness of classification on the architecture of the CNN.

The increased robustness of AlexNet can be explained by its higher receptive field size (R) than VGG-16 at the output of the first max pooling layer and the relation to the size of the unit of compression (B). The relative sizes of the compression unit for JPEG and receptive field are shown in Fig. 1c. During training, AlexNet extracts key features over a larger receptive field to develop contextual features. Whereas, VGG-16 extracts feature from a smaller receptive field and learns context with increased depth. JPEG uses 8x8 blocks. At a higher compression ratio, high frequency components within the block are lost which are relevant features for VGG-16 due to its smaller receptive field. However, the overall structures formed by low frequency components are retained across blocks which are relevant features for AlexNet due to its higher receptive field spanning multiple JPEG blocks. So a model with higher receptive field is robust against JPEG. Analogous analyses can be done for other compression schemes and CNN architectures.

4. PROPOSED METHOD

Section 3 showed that the slope of the performance degradation of the tested CNNs is dependent on the type of degradation and the architecture of the CNN. Hence, a neural network whose weights adapt to various degradation types on the basis of incoming image samples is required. We propose a degradation adaptive network robust to varying degradation types and extents based on a master-slave architecture. There also exists the possibility of an alternative.

4.1. Master-Slave Architecture

Fig. 2 shows the master-slave architecture proposed.

Master CNN: The master is a 'Quality Prediction Network'. Here we train a small CNN to predict the quality of the input image. Based on the predicton by the master, a specific branch of the slave CNN is selected. This CNN comprises three convolutional layers followed by three fully connected layers. The number of nodes in the final fully connected layer corresponds to the number of quality bins to be predicted. Each quality bin corresponds to a set of quality factors.

Slave CNN: It comprises a common trunk which is the first eleven layers of the VGG-16 network. This is followed by 3 branches which are trained using the cross-entropy loss function for different compression ranges. Branch 1 comprises of the pre-trained VGG-16 weights and during inference caters to the clean images and compressed images whose classification accuracy does not fall significantly. Branch 2 is trained using images compressed by JPEG with quality factors in the range of 20 to 5. Branch 3 is trained using images compressed by JPEG2000 using a range of quality factors of 10 to 2. The loss function is as follows:

$$L(x, \boldsymbol{\theta}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} t_{ij} \log y_{ij} \qquad (1)$$

where L is the loss which is a function of the parameters, $\boldsymbol{\theta}$, comprising the weights and biases. The number of training

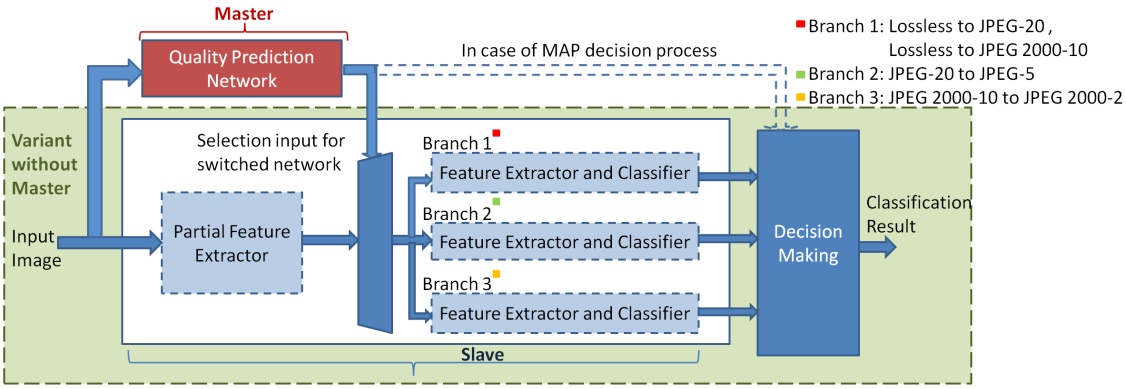


Fig. 2: Proposed method for improving robustness of CNNs for image degradations.

images is denoted by N and the number of classes is denoted by C. The predicted class is denoted by y and the ground truth is denoted by t. For the branches, $y = f(x, \theta)$ where x is the clean image or the image degraded by blur, noise and/or compression using JPEG or JPEG 2000 and θ are the network parameters. The motivation for the common layers of the slave CNN is to make use of transfer learning and avoid training a network from scratch. Moreover, instead of one network for all cases, the motivation for different branches is to learn features distinct from that of clean images or images with no significant degradation. Images with significant degradation are used to tune a separate branch so that a different set of features is learned. The intuition is that when there is a severe degradation in the image, features different from that of the clean image could become more important to take the correct decision.

Decision Making: The decision based on the Maximum a Posteriori (MAP) scheme. With reference to the slave network, the quality parameter is prior information. This prior information can be obtained by decoding the quantization parameter from the encoded bitstream or by using the quality prediction network.

However, due to high computational costs and since the master CNN has high accuracy, instead of the MAP based approach, we use a switched approach as.

$$P(y|x) = p_i(y|x), i = \{k|_{k=1\ to\ K} \max p_k(d|x)\} \quad (2)$$

where k indexes the number of branches of the slave CNN, K is the number of branches of the slave CNN, x is the input image, y is the predicted class of the slave CNN, d is the predicted class of the master CNN which corresponds to a degradation bin. In the architecture described in Fig. 3, K is 3.

4.2. Alternative to Master-Slave Architecture

Since the different branches of the slave have been trained for different quality bins, if an input is passed through all the branches of the slave CNN without using a master CNN, the highest confidence score should be obtained from the branch that was suited best for the specific kind of degraded image. This motivates the removal of the master CNN and processing the input by all branches of the slave CNN in the alternative. The decision is taken by max-pooling as

$$P(y|x) = \max_{k=1\ to\ K} p_k(y|x) \quad (3)$$

where the symbols are as described above.

5. EXPERIMENTS

5.1. Individual and Combined Degradations

The master-slave architecture and its alternative were tested using 50,000 images of the ImageNet validation set. The images were compressed at various quality levels using JPEG and JPEG 2000. Fig. 3 show the graphs for the baseline network, the master-slave system and its alternative based on the max pooled decision for JPEG and JPEG 2000. It is observed that the proposed master-slave architecture shows a definite improvement in the top-1 classification accuracy. Moreover, there is not much degradation in performance between the master-slave architecture and its alternative. This is especially advantageous for practical applications where computations can be reduced by removing the master.

In practical systems, the image is often affected by combinations of degradations. In these experiments, we test for the classification accuracy using 2 combinations of degradations: 1) Noise and compression artifacts and 2) Blur (of radius 4 and standard deviation 1) and compression artifacts. Fig. 4 shows the results for JPEG and JPEG 2000 compressed images. It is observed that the degradation in performance due to combined effects of noise and compression is bounded by the degradation due to noise for high quality factors and compression for lower quality factors. This is demonstrated in the graph by the trend line for the combined degradation starting

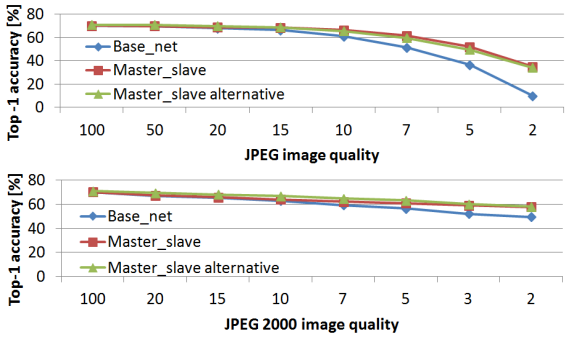


Fig. 3: Top-1 accuracy for degradations due to compression.

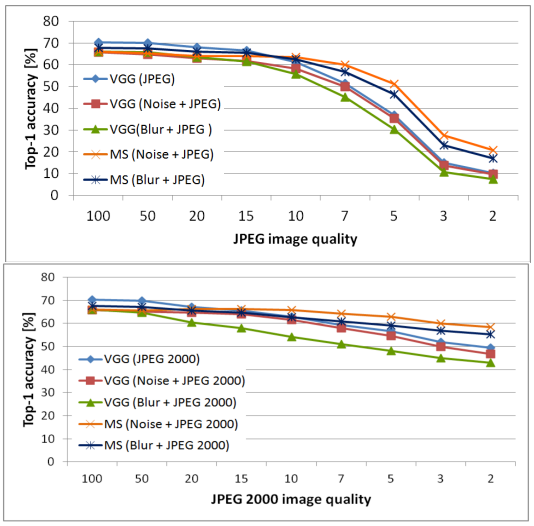


Fig. 4: Top-1 accuracy for combined degradations.

with degradation solely due to noise at high quality factors and converges to degradation solely due to compression at lower quality factors. On the other hand, it is observed that combined degradation due to blur and compression is additive. The degradation curve is similar to compression with a constant offset determined by blur. Moreover, our method using the master-slave architecture improves the classification accuracy for the combined degradations inspite of not being trained explicitly for the combinations.

5.2. Comparison of Different Techniques

The performance of our method using the master-slave architecture (MS) and its alternative (AMS) are compared with techniques for increasing CNN robustness like data augmentation (DA), stability training (ST) [8] and DeepCloak (DC) [10] on 50,000 images of the ImageNet validation set. Table 1 shows the top-1 classification accuracy for clean images and JPEG compressed images at a quality factor 10. As can be

Table 1: Top-1 accuracy (in %) for different methods for increasing robustness of CNNs.

Method	With Method		Without Method		% Change
	Clean (c)	JPEG-10 (d)	Clean (c')	JPEG -10 (d')	$(d'-d)/(c-d)$
DA, fc8	70.2	61.2	70.3	63	19.8
fc67	70.2	61.2	70.5	63.5	25.3
fc678	70.2	61.2	69.3	64.4	35.7
ST[8]	77.8	61.1	77.9	67.9	40
MS	70.2	61.2	70.3	66.7	60.5
AMS	70.2	61.2	70.2	66.0	52.5
DC[10]	58.2	54	54.8	52.2	-

observed, the improvement in performance increases as more fully connected layers (fc6, fc7 and fc8) are tuned for data augmentation using VGG-16 as the base network. Stability training uses a base network of Inception [17] and shows an improvement of 40%. Our method using the master-slave architecture has the highest improvement of 60.5% followed by the alternative with 52.5%. While in data augmentation, the network is left to discover meaningful features from the mixed training dataset of clean and degraded images, stability training explicitly communicates to the network to reduce the distance in the representation of the clean and degraded images via the loss function. However, stability training uses additive Gaussian noise to degrade images for training and increases robustness of the network in a limited neighborhood of the clean image. Whereas, our method is able to learn features for a greater degradation. Since DeepCloak [10] is class dependent, for ImageNet with 1000 classes, it does not scale well. We sampled the 1000 classes for varying degrees of performance with 1% masking parameter and report the average results. While [10] shows improvement at a class level for adversarial examples, we observed no clean improvement for compressed images.

6. CONCLUSION

The performance analysis of common CNN architectures for input images degraded by compression artifacts revealed a dependency on the relative sizes of the compression units for a specific type of compression and the receptive field size of the CNN being used for classification. A master-slave architecture for catering to different kinds and extents of degradation was proposed and shown to be effective for individual and combined degradations. Moreover, an alternative to the proposed method without a master and using max pooled decision was also shown to be effective. The advantage of our proposed method and its alternative is that there is no explicit dependency on a quality parameter. No quality parameter needs to be decoded from the bitstream. This is especially advantageous when (multi-generation) transcoding may have taken place. Moreover, training our proposed method is simple.

7. REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25.*, 2012, pp. 1106–1114.

[2] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *CoRR*, vol. abs/1311.2524, 2013.

[3] Ross Girshick, "Fast r-cnn," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, Washington, DC, USA, 2015, ICCV '15, pp. 1440–1448, IEEE Computer Society.

[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, Cambridge, MA, USA, 2015, NIPS'15, pp. 91–99, MIT Press.

[5] 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*, 2016, pp. 1–6.

[6] Ishan Misra, Abhinav Shrivastava, and Martial Hebert, "Watch and learn: Semi-supervised learning of object detectors from videos," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015, pp. 3593–3602.

[7] Igor Vasiljevic, Ayan Chakrabarti, and Gregory Shakhnarovich, "Examining the impact of blur on recognition by convolutional networks," *CoRR*, vol. abs/1611.05760, 2016.

[8] Stephan Zheng, Yang Song, Thomas Leung, and Ian J. Goodfellow, "Improving the robustness of deep neural networks via stability training," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 4480–4488.

[9] Beilun Wang, Ji Gao, and Yanjun Qi, "A theoretical framework for robustness of (deep) classifiers under adversarial noise," *CoRR*, vol. abs/1612.00334, 2016.

[10] Ji Gao, Beilun Wang, and Yanjun Qi, "Deepmask: Masking DNN models for robustness against adversarial samples," *CoRR*, vol. abs/1702.06763, 2017.

[11] Ling Huang, Anthony D. Joseph, Blaine Nelson, Benjamin I.P. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, New York, NY, USA, 2011, AISec '11, pp. 43–58, ACM.

[12] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," *CoRR*, vol. abs/1412.6572, 2014.

[13] Luca Bertinetto, Jack Valmadre, João F. Henriques, Andrea Vedaldi, and Philip H. S. Torr, "Fully-convolutional siamese networks for object tracking," in *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II*, 2016, pp. 850–865.

[14] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., pp. 1097–1105. Curran Associates, Inc., 2012.

[17] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Computer Vision and Pattern Recognition (CVPR)*, 2015.