Report 06

Yiming Bian

Jul 2nd 2022

Major Professor

Arun Somani(arun@iastate.edu)

Committee Members

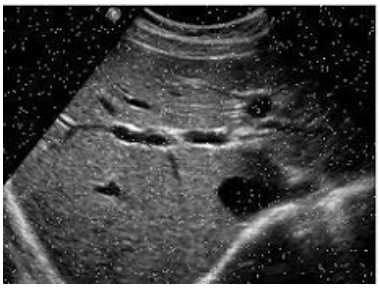
Henry Duwe (duwe@iastate.edu)
Aditya Ramamoorthy (adityar@iastate.edu)
Alexander Stoytchev (alexs@iastate.edu)
Cindy Yu (cindyyu@iastate.edu)

1 Introduction

In the past two weeks, I was working with Dr. Somani revising the paper "Effective Management of Both Sparse and Intense Salt and Pepper Noise Image Classification". We add the models trained on a mixture of images with different noise intensities. We name those mix-noise-trained, or mix-trained models. On the contrary, the previous noise-trained models are specified as single-noise-trained, or single-trained models. As explained before, some single-trained models suffer from very poor classification accuracy when the noise is sparse in the image and median filter is preferred in this scenario. However, the latest results show that mix-trained models have an improved and comparable performance to median filter when the noise is sparse. Therefore, we conclude that mix-training a model is an effective method to improve the classification accuracy on salt-and-pepper noise images. This paper has been submitted to TransAI 2022. The notification of acceptance date is August 15th. The submitted version is attached to this report.

2 Plan for next two weeks

Currently, I am working on three papers: 1) speckle noise image classification 2) survey: flawed image processing 3) survey: neural architecture search on



(a) Ultrasound images of liver[3]

(b) A satellite image. Credit to: SpaceNet

Figure 1: Left: an ultrasound image. Right: an SAR image: Both images inevitably contain speckle noise

object detection(cooperate with Krishna). For two survey papers, I am in the phase of reviewing literature. The speckle noise image classification paper is the focus as of now because we aim to submit it to KDIR 2022. The submission deadline is July 14th.

Speckle noise is the noise that arises due to the effect of environmental conditions on the imaging sensor during image acquisition. Speckle noise is mostly detected in case of medical images, active Radar images and Synthetic Aperture Radar (SAR) images. [1] Two examples are shown in Fig. 1. The most widely-used transforms for denoising purposes are wavelet, contourlet, and shearlet.[2] Many speckle noise suppression methods were invented and proved to be effective. However, previous works mainly focus on fixing the image for human being to recognize. What I plan to do is to improve the classification accuracy of CNN models on, for example, determining if a nodule is benign or malignant.

The big-picture of this project is similar to the previous one: test a CNN model on noise images after applying denoising measures and compare it to the noise-trained models. Current obstacles are 1) specifying several CNN architectures 2) dataset preparation 3) searching for mighty rivals. To prove the noise-training method is universal, CNN selection is not trickiest. Noise-training should work for all CNN models but the improvement is dependent. I also found several public medical image datasets such as JSRT(Japanese Society of Radiological Technology) database, which contains 247 chest radiographs with and without a lung nodule. This dataset is not ideal as it is not large enough but it is a good start point. Mighty rivals are not rare as so many denoising methods have been proposed such as adaptive median filter[3], wavelet-based threshold techniques[4], ResNet-based denoising method[5] etc. I will look for those with top performance and off-the-shelf implementations.

Moreover, Dr. Somani suggested me scheduling the prelim exam after the flawed image processing survey paper is done. The estimated finish date will be

the end of July and it will be submitted to a top IEEE/ACM journal.

References

[1] A. Maity, A. Pattanaik, S. Sagnika, S. Pani, A comparative study on approaches to speckle noise reduction in images, in: 2015 International Conference on Computational Intelligence and Networks, 2015, pp. 148–155. doi:10.1109/CINE.2015.36.

[2] H. R. Shahdoosti, Z. Rahemi, Edge-preserving image denoising using a deep convolutional neural network, Signal Processing 159 (2019) 20–32.

[3] O. Magud, E. Tuba, N. Bacanin, Medical ultrasound image speckle noise reduction by adaptive median filter, Wseas Trans. Biol. Biomed 14 (2017) 38–46.

[4] R. Sivakumar, D. Nedumaran, Comparative study of speckle noise reduction of ultrasound b-scan images in matrix laboratory environment, International Journal of Computer Applications 10 (9) (2010) 46–50.

[5] W. Jifara, F. Jiang, S. Rho, M. Cheng, S. Liu, Medical image denoising using convolutional neural network: a residual learning approach, The Journal of Supercomputing 75 (2) (2019) 704–718.

Effective Management of Both Sparse and Intense Salt and Pepper Noise Image Classification

Yiming Bian and Arun K. Somani
Department of Electrical and Computer Engineering
Iowa State University, Ames, Iowa 50010
{ybian, arun}@iastate.edu

***Abstract*—In the past decade, machine learning based classification of high-quality images has made remarkable progress. However, the classification of low-quality images poses new challenges. This work focuses on the improvement of classification accuracy on images with salt and pepper noise. Compared to widely adopted pre-processing techniques such as median filter, our method retrains the model on noise data carefully and the noise-trained models gain very high resilience. We analyze median filters, single-noise-trained, and mix-noise-trained models. When salt and pepper noise is sparse in an image, the classification accuracy of single-noise-trained models varies on convolutional neural networks (CNN). VGG16 models show close performance compared to median filter when the noise intensity is low but ResNet-18 models perform otherwise. Therefore, median filter remains reliable, effective and preferred in the light noise case. When the noise is intense, median filter is outperformed by the single-noise-trained models with the largest accuracy gap of 11.23% and an average of 6.97%. Nevertheless, our proposed mix-noise-trained model, *mix_404020*, achieves excellent performance in all scenarios as it has comparable, if not better than, classification accuracy to the optimal median filter/single-noise-trained model when the noise is sparse/intense.**

***Index Terms*—salt and pepper noise, noise image classification, noise training, mix training, noise-trained CNN models**

I. INTRODUCTION

Image classification is one of the most substantial and comprehensive problems in computer vision. In the past decade, numerous machine learning based techniques were introduced and remarkable progress and achievements have been made.

In 2009, a large database, which contains over 14 million hand-annotated images, designed for visual object recognition research called ImageNet was presented by Li Fei-Fei et al. [1]. The next year, ImageNet project began an annual contest called the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [2]. This challenge uses a subset of ImageNet that has $1,000$ non-overlapping classes of objects. As it is commonly acknowledged that the average recognition ability of human has a top-5 error rate of 5%, the winner in 2015, *ResNet* [3], beat human level for the first time with an error rate of 3.6%. The state-of-art model, *Florence* [4], achieved an extraordinary error rate of 0.98%. Behind all these intriguing figures, it is reasonable to conclude that the improvement of classification accuracy on high-quality images is merely a challenge anymore.

Expensive lens and camera systems are manufactured to capture top-quality photos. They incur high cost to store, process and transfer. On the other hand, low-quality images are more ubiquitous. Processing low-quality image, however, poses greater challenges. First, the definition of being low-quality is vague as there are numerous reasons that lead to a flawed image such as taking photos using a cheap webcam, hardware failure during capture, storage, transmission and so on. Many flaws can degrade the quality of an image such as low resolution, noise pixels, over and underexposure etc. There is no panacea for processing all kinds of low-quality images so far and it is hardly possible in the future due to the complicate nature of the problem.

A. Related works

Previous researchers have done embryonic works on stressing noise image processing and proposing experimental solutions. For example, authors in [5] show several networks including VGG16, GoogleNet, VGG-CNN-S and Caffe Reference are susceptible to quality distortions, particularly to blur and noise. In [6], authors investigate the relationship between object recognition and image quality and their results show that the selected algorithms can estimate recognition performance under certain conditions.

Many achievements have been made in biometrics, particularly fingerprint and face recognition. In [7], authors propose a two-stage enhancement scheme that handles low-quality fingerprint images. In [8], authors' investigations indicate that both hand-crafted and deep-learning based face detectors are not robust enough for low-quality images. In [9], authors concluded several techniques to improve the performance of low-quality face recognition (LQFR) such as super-resolution processing, deblurring and learning a relationship between different resolution domains. In [10], authors proposed a framework for recognizing objects in very low resolution images through the collaborative learning of two deep neural networks including image enhancement network and object recognition network.

B. Contributions

In this paper, we focus on one specific kind of noise that degrades the image quality called salt and pepper noise. We explore measures to improve the accuracy of salt and pepper noise image classification for several convolutional neural networks. A pre-processing technique called median filter is widely adopted to remove salt and pepper noise in

both greyscale and RGB images. In [11], [12], [13], authors emphasize that median filter is a great general-purposed salt and pepper removal measure. More works, such as [14], [15], [16], focus on optimizing the median filter as determining the median of a set of numbers is an inefficient process. Our method directly retrains the CNN model on salt and pepper noise data of various intensities. We apply this method on iconic ResNet-18 and VGG16 architectures because ResNet-18 is the first network beating human's recognition level in ILSVRC and VGG16 was proved surprisingly resilient to many distortions as shown in [5]. There is a concept in computer science called GIGO, thus garbage in, garbage out. GIGO describes that flawed input generates useless output. It is true in most cases in machine learning, but if the useful information can be extracted in the flawed input, the output could be meaningful. Our results prove that noise-trained models obtain solid resilience to salt and pepper noise data and we propose a mix-noise-trained model, mix_040420, that achieve exceptional performance on all noise test sets.

The rest of the paper is organized as follows. Section II provides preliminary knowledge of salt and pepper noise, median filter, deep convolutional networks, image classification and transfer learning. Section III describes how we construct noise data and develop CNN models. In section IV, we present our experiments, results and detailed analysis. Conclusions and future research directions are discussed in section V.

II. BACKGROUND

A. Salt and pepper noise

We focus on salt and pepper noise that degrades the quality of an image in this paper. An image is stored in the form of an array of pixel values and each value, ranging from 0 to 255, denotes the intensity of the current color channel. Two common image color models are greyscale and RGB. In a greyscale image, it has $H \times W$ pixels, where H and W stand for the height and width of the image. Each pixel value reflects the intensity of greyness as 0 being black and 255 being white. On the other hand, since RGB is an additive color model, an RGB image is the product of stacking three color channels and each channel describes how red, green and blue of every pixel. Taking the red channel as an example, the pixel value ranges from 0 to 255 with 0 being black and 255 being red. Therefore, an RGB image is a 3D array with dimension $H \times W \times C$ where C stands for channel and is fixed to 3.

Salt and pepper noise is a type of impulse noise. When the original pixel value is lost and set to an extreme value: 0 or 255, it becomes a salt and pepper noise pixel. When it happens in a grayscale image, the noise appears as either a white or black pixel, thus how the noise gets its name from. Salt and pepper noise can happen in an RGB image as well and the color of a noise pixel is more complicated. Since there are three channels, a pixel is denoted as a tuple with three values: (r, g, b). When the noise occurs to a pixel, all three channels lose their values and be set to either 0 or 255. It creates a noise pixel with eight possible colors (i.e. white $(255, 255, 255)$, cyan $(0, 255, 255)$, magenta $(255, 0, 255)$, yellow $(255, 255, 0)$, red $(255, 0, 0)$, green $(0, 255, 0)$, blue $(0, 0, 255)$ and black $(0, 0, 0)$). Fig. 1 shows the salt and pepper noise pattern in the aforementioned two cases. In Fig. 2, we add salt and pepper noise with intensity equals to 0.1 and 0.5, denoted as SNP_0.1 and SNP_0.5, to an image. This example shows how sparse and intense salt and pepper noise affects the image quality. For simplicity, salt and pepper is referred to as SNP.

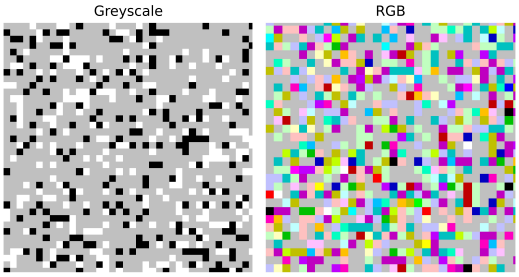


Figure 1: Salt and pepper pixels in greyscale and RGB images

Figure 2: An original image and two salt and pepper noise images of intensity 0.1 and 0.5

B. Median filter

Median filter is one of the most popular noise reduction techniques in digital image processing. For salt and pepper noise, it is particularly effective [17]. To apply a median filter to an image, which is denoted as a 2D array, we need to specify the median filter size, or kernel size, as $k \times k$ where k is an odd number by convention. This kernel traverses the image array from left to right and top to the bottom with a step size of one pixel. It first sweeps horizontally and when the whole row is covered, it moves one pixel down and starts from the leftmost position. In each move, the kernel covers k^2 pixels and output the median value among them. Fig. 3 explains this process using a small proportion of an image. On the left is a 7×7 pixel area of an input image, SNP noise pixels are marked with a red box and the 3×3 kernel locates at its initial position in this area. The output of the current kernel, 129, is the median of the covered pixel values. After the kernel traversing the whole area, we have the output image on the right and all noise pixels are filtered and removed. The shrinkage of output image size puts a spotlight on the boundary issue of median filtering. There are multiple methods to mitigate this problem and maintain the image size such as padding a circumference

using 0s or the pixel values on the boundary. Other schemes may be preferred in particular circumstances.

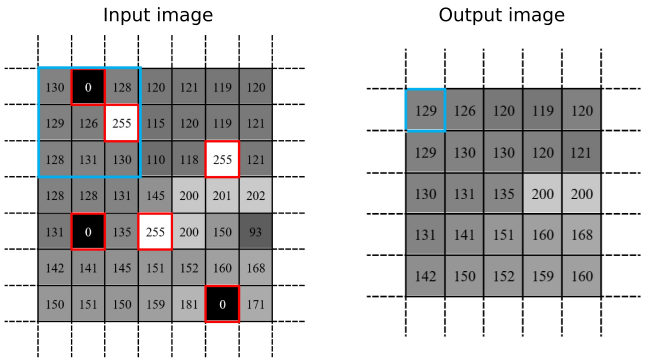


Figure 3: Median filter: an example

Compared to other filters such as mean filer, which replaces the pixel value with the mean of its neighbours' pixel values, median filter shows more robustness in SNP noise removal because unrepresentative pixel values deteriorate the mean value while barely affect the median value. SNP noise pixels have two extremely unrepresentative values as 0 and 255. Moreover, median filter guarantees the output pixels come from the original image and does not create unnatural-color pixels. Fig. 4 shows the restoration ability of median filter of different kernel sizes on images with three levels of SNP noise. When the SNP noise intensity is 0.1, the median filter of kernel size 3 removes nearly all the noise pixels and almost perfectly restores it back to the original looking. However, when the noise intensity increases, noise pixels may not be removed entirely with smaller kernels. Tuning up the kernel size does get rid of all the noise pixels but sacrifices sharpness of the image. In other words, restored image becomes more blurry and vague with larger kernel size.

C. Deep convolutional networks and image classification

A concise introduction to deep neural network has been provided in [5]. The big picture of deep learning, convolutional neural network, image understanding with deep convolutional networks were discussed in [18]. In this section, we introduce the structure of two specific neural network architectures that we experiment on: ResNet-18 and VGG16. Both are popular CNN models that achieved top results for image classification in ILSVRC. There are four kinds of basic blocks, or residual blocks in ResNet-18 [3] and skip connections are adopted to address the vanishing or exploding gradient problem. VGG16 has an architecture of sixteen layers including thirteen convolutional layers and three fully connected layers [19]. The layer-wise details of ResNet-18 and VGG16 are presented in Table I and II where k, s, p and OC stand for kernel size, stride, padding and output channels.

D. Transfer learning

Existing CNN models can be modified and retrained to solve particular problems by transfer learning. Transfer learning

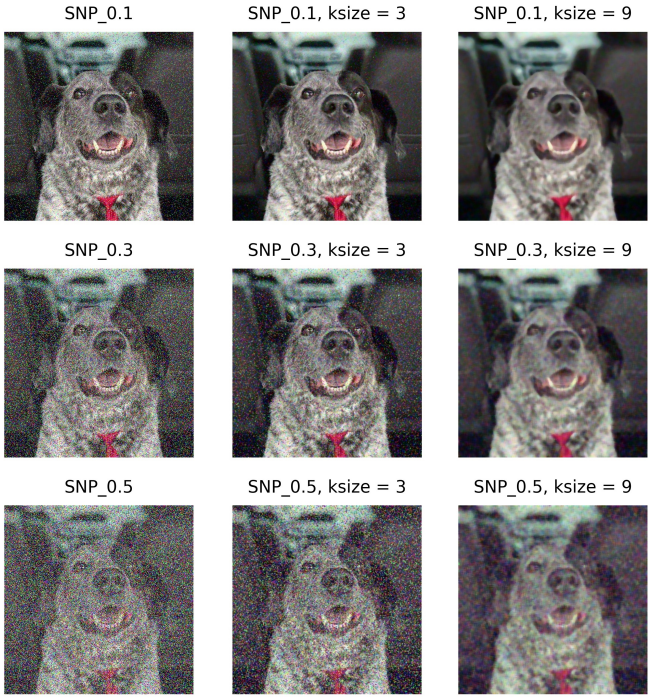


Figure 4: Applying median filter of kernel size 3 and 9 to SNP noise image of intensity 0.1, 0.3 and 0.5

Table I: Layer-wise details of ResNet-18: BB_m_n stands for the n-th appearance of basic block m. Downsampling is performed by BB_2_1, BB_3_1 and BB_4_1 with s=2.

#	Layer		Input size	k	s	p	OC
1	Conv		$224 \times 224 \times 3$	7	2	2	64
-	MaxPool		$112 \times 112 \times 64$	3	2	1	64
2	BB_1_1	Conv	$56 \times 56 \times 64$	3	1	1	64
3		Conv	$56 \times 56 \times 64$	3	1	1	64
4	BB_1_2	Conv	$56 \times 56 \times 64$	3	1	1	64
5		Conv	$56 \times 56 \times 64$	3	1	1	64
6	BB_2_1	Conv	$56 \times 56 \times 64$	3	2	1	128
7		Conv	$28 \times 28 \times 128$	3	1	1	128
8	BB_2_2	Conv	$28 \times 28 \times 128$	3	1	1	128
9		Conv	$28 \times 28 \times 128$	3	1	1	128
10	BB_3_1	Conv	$28 \times 28 \times 128$	3	2	1	256
11		Conv	$14 \times 14 \times 256$	3	1	1	256
12	BB_3_2	Conv	$14 \times 14 \times 256$	3	1	1	256
13		Conv	$14 \times 14 \times 256$	3	1	1	256
14	BB_4_1	Conv	$14 \times 14 \times 256$	3	2	1	512
15		Conv	$7 \times 7 \times 512$	3	1	1	512
16	BB_4_2	Conv	$7 \times 7 \times 512$	3	1	1	512
17		Conv	$7 \times 7 \times 512$	3	1	1	512
-	AvgPool		$7 \times 7 \times 512$	7	7	0	512
18	FC		$1 \times 1 \times 512$	-	-	-	1000

Table II: Layer-wise details of VGG16

#	Layer	Input size	k	s	p	OC
1	Conv	$224 \times 224 \times 3$	3	1	1	64
2	Conv	$224 \times 224 \times 64$	3	1	1	64
-	MaxPool	$224 \times 224 \times 64$	2	2	0	128
3	Conv	$112 \times 112 \times 128$	3	1	1	128
4	Conv	$112 \times 112 \times 128$	3	1	1	128
-	MaxPool	$112 \times 112 \times 128$	2	2	0	256
5	Conv	$56 \times 56 \times 256$	3	1	1	256
6	Conv	$56 \times 56 \times 256$	3	1	1	256
7	Conv	$56 \times 56 \times 256$	3	1	1	256
-	MaxPool	$56 \times 56 \times 256$	2	2	0	512
8	Conv	$28 \times 28 \times 512$	3	1	1	512
9	Conv	$28 \times 28 \times 512$	3	1	1	512
10	Conv	$28 \times 28 \times 512$	3	1	1	512
-	MaxPool	$28 \times 28 \times 512$	2	2	0	512
11	Conv	$14 \times 14 \times 512$	3	1	1	512
12	Conv	$14 \times 14 \times 512$	3	1	1	512
13	Conv	$14 \times 14 \times 512$	3	1	1	512
-	MaxPool	$14 \times 14 \times 512$	2	2	0	512
14	FC	$1 \times 1 \times 25088$	-	-	-	4096
15	FC	$1 \times 1 \times 4096$	-	-	-	4096
16	FC	$1 \times 1 \times 4096$	-	-	-	1000

reuses knowledge from past related tasks to ease the process of learning to perform a new task [20]. A CNN model can be roughly viewed as the combination of a feature extractor and a classifier. When performing image tasks, the feature extractor in CNN transforms pixel values of an image into a suitable feature vector so that the classifier could detect and classify patterns in the input image [18]. Since it requires very careful engineering and considerable domain expertise to design a high-quality feature extractor, one could take advantage of existing pretrained CNN models and apply to the similar tasks instead of randomly initializing a model and training it from scratch. Fig. 5 presents the idea of transfer learning. Based on a pretrained CNN model CNN_x, CNN_y shares the same feature extractor architecture but has its own classifier. To tune the model CNN_y, it requires retraining using the data for the specific task.

Feature extraction and fine-tuning are two types of transfer learning. In feature extraction, CNN_y shares the entire feature extractor in CNN_x, including its architecture and weights, thus Feature Extractor_y = Feature Extractor_x. And the weights of its feature extractor do not change during the retraining process. In fine-tuning, Feature Extractor_y is initialized using the weights in Feature Extractor_x and its weights are updated in the retraining process.

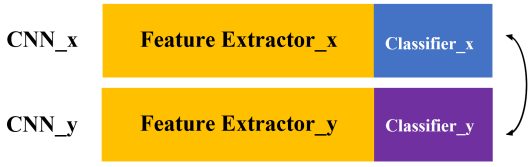


Figure 5: Transfer learning between two CNNs

III. CONSTRUCTING CNN MODELS

To explore the effectiveness of training on noise data and evaluate the noise resilience of noise-trained models, we first generate training, validation and test sets by adding five levels of SNP noise to a subset of ImageNet. Then we retrain multiple CNN models using transfer learning based on pretrained models for both ResNet-18 and VGG16. The classification accuracy of a tested model is defined as one minus its top-1 error. We provide the classification accuracy of each model on various test sets and form our conclusions. In this section, we present the details of constructing noise datasets and developing CNN models.

To prepare datasets, we randomly select five objects, each containing $1,300$ sample images, from ImageNet: goldfinch, hamster, vizsla, upright and pitcher. The proportions of the original sample images for training, validation and test purpose are 80%, 10% and 10%, thus $1,040$ training images, 130 validation images and 130 test images for each object. Next, we add SNP noise of intensity 0.1–0.5 to each original sample image for every purpose and yield fifteen noise sets as shown in Table III. We also mix SNP_0.x_train/val ($x = 1, 2, 3$) with a proportion of $(25\%, 50\%, 25\%)$, $(33\%, 33\%, 33\%)$, $(20\%, 40\%, 40\%)$ and $(40\%, 40\%, 20\%)$ for mix-noise-training.

Table III: Dataset details: In the column named "Dataset Name", the purpose of each set is conveyed by its name. For SNP noise sets, $0.x$ is the noise intensity where x is an integer ranging from 1 to 5. In the column named "Configuration", it shows the number of images and objects in each set. e.g. $1,040 \times 5$ means the set contains 5 objects and each has $1,040$ images, thus a total of $5,200$ images.

Dataset Name	Contents	Configuration
Original_train	Original images	$1,040 \times 5$
Original_val		130×5
Original_test		130×5
SNP_0.x_train	Images with SNP_0.x noise	$1,040 \times 5$
SNP_0.x_val		130×5
SNP_0.x_test		130×5

Table IV shows that based on the pretrained model, ten retrained models are developed: one RtO, five single-noise-trained models RtN_0.x and four mix-noise-trained models mix_abc. Since the size of our training set $(5,200)$ is not comparable to that of the pretrained training set (over 1.2 million), feature extraction is preferred over fine-tuning to avoid potential overfitting issues. Therefore, all retrained models only have a modified classifier but share the same feature extractor with the pretrained model. For both ResNet-18 and VGG16, the output dimension of the last layer is changed from $1,000$ to 5. In the retraining process, only the weights of the classifier are updated.

IV. EXPERIMENTS AND RESULTS

We design experiments to show the classification accuracy improvement brought by transfer learning and the performance

Table IV: Pretrained model and retrained model details: "RtO" stands for **Ret**rianed on **O**riginal images. A noise-trained model is denoted by"RtN", which is short for **Ret**rianed on **N**oise images. "mix_abc" is mix-noise-trained models and the corresponding training set is a mixture of $a\%$ SNP_0.1, $b\%$ SNP_0.2 and $c\%$ SNP_0.3 noise images. All retrained models are developed from the pretrained model using transfer learning. "OD" is the acronym for output dimension, thus the number of output channels in the last fully connected layer.

Model Name	Based on	Retrained on	OD
Pretrained	-	-	$1,000$
RtO	Pretrained	Original_train	5
RtN_0.x	Pretrained	SNP_0.x_train	5
mix_abc	Pretrained	Mix noise images	5

comparisons among multiple noise-trained models. Comprehensive results show the superiority of different models in particular scenarios and we propose a mix-noise-trained model as an universal solution to SNP noise image classification. Below we present the design details, results and analysis of each experiment.

A. Pretrained model and RtO

In this experiment, we test the classification accuracy of the pretrained model and RtO on all six test sets: Original_test and SNP_0.x_test, where x is an integer ranging from 1 to 5. RtO shares the feature extractor with the pretrained model. Differences are that the output dimension of classifier and the classifier of RtO is retrained with Original_train.

The accuracy comparison between two models are shown in Fig. 6. We notice that the classification accuracy reduces as the SNP noise intensity increases, which is an expected outcome as neither model has ever seen a noise image. However, the performance of the pretrained model drops sharply, which indicates its more significant lack of robustness on SNP noise data. Benefiting from transfer learning, RtO reduces the possible output to 5 from $1,000$ and gets about 20% accuracy improvement immediately. Another obstacle for the pretrained model to produce the right classification is the interference by similar objects such as the Vizsla and the Great Dane. As a result, RtO is an obvious winner on all test sets compared to the pretrained model and the performance gap tends to increase with the noise intensity: 19.54%, 37.54%, 54.46%, 59.54%, 58.31% and 50.46%.

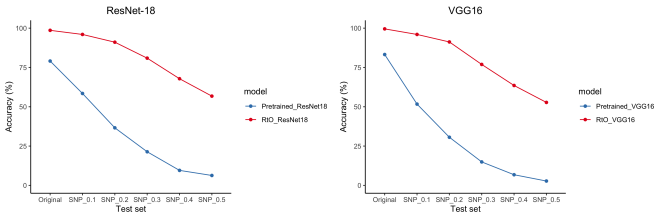


Figure 6: A classification accuracy comparison between the pretrained model and RtO

This experiment shows the great performance leap after applying transfer learning on the pretrained model. Therefore, we take RtO as the baseline model to compare the performance with in the rest experiments.

B. RtO, RtO_MF, RtN_single and RtN_mix

In this experiment, we compare the classification accuracy among RtO and RtN models on multiple test sets. RtO is the model retrained on original images and RtO_MF is the same model but tested on noise images restored by median filter. RtN_single and RtN_mix are two types of RtN models. The former is trained on a single intensity of noise images and the latter is trained on a mix. Hence they are also refer to as single-/mix-noise-trained, or single-/mix-trained models for simplicity. This experiment is divided into four parts and each shows one of the following points:

- RtO vs. RtO_MF shows the accuracy improvement brought by applying median filter on noise images
- RtO vs. RtN_single shows the performance gain of single-trained models over RtO
- RtN_single vs. RtN_mix shows the advantage of mix-trained models over single-trained models
- Top models compare the best performance in all aforementioned models including RtO, RtO_MF, RtN_single and RtN_mix.

Below we present the details of each part of this experiment.

a) RtO vs. RtO_MF: In the first part of the experiment, we compare the classification accuracy of RtO models before and after applying different median filters to the noise test sets. As mentioned in section II, there is a trade-off between the unfiltered noise pixels and the sharpness of the restored image. Thus, the larger the kernel size is, more noise pixels will be removed but the output image will be more blurry. Previous work [5] proved that blur in an image will create, if not more, similar difficulties for CNNs to process as noise does. Since all the images are from ImageNet, the average resolution is roughly 480×410 and if the kernel size of the applied median filter is larger than 11, the restored image loses majority of sharpness when the noise intensity is high. Therefore, kernel sizes are carefully selected as 3, 5, 7 and 9.

We take the RtO model, which is retrained on a set of original images, and test its accuracy on five SNP noise sets before and after applying the median filter of each selected kernel size. We record the accuracy changes with the noise intensity in all five scenarios as shown in Fig. 7. When the SNP noise intensity is 0.1, median filter makes little accuracy difference to RtO. Generally, median filter improves the classification accuracy by 9.54%–14.04% for ResNet-18 RtO and 11.35%–15.5% for VGG16 RtO when obvious noise pixels (intensity ≥ 0.3) are observed in an image. Nevertheless, ResNet-18 RtO is more sensitive to the kernel size change while VGG16 RtO shows more immunity to it.

b) RtO vs. RtN_single: In the second part, we compare the classification accuracy among RtN_single models and the RtO model on all original and noise test sets. As its has been observed that RtO performs poorly on noise test sets, we are

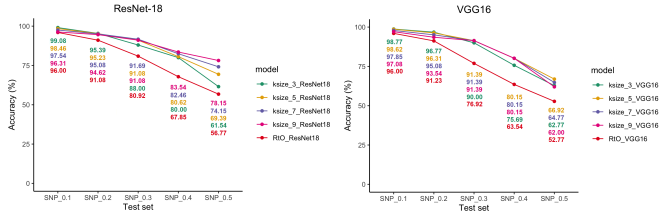


Figure 7: Comparisons of RtO accuracies on noise data with and without applying median filters

interested in how RtN_single performs on 1) the original test set 2) the noise set of the model's training intensity 3) noise sets other than the model's training intensity. Below we answer all of these questions in the perspective of CNN.

The left line chart in Fig. 8 shows the accuracy of ResNet-18 RtO and RtN_single models. All RtN_single models perform poorly on original images, especially RtN_0.5, its accuracy (27.85%) is merely better than the expected value of a random guess (20%). Only RtN_0.1 outperforms RtO on set SNP_0.1, SNP_0.2 and the rest RtN_single models show disastrous performance on these two test sets. However, as the noise level increases, more RtN_single models perform better RtO and they have a decent classification accuracy averaging 89.26% on noise images of their training intensities. When the noise intensity reaches 0.5, RtN_single models achieve an accuracy ranging from 72.92% to 85.39% compared to RtO's 56.77%. To conclude, all ResNet-18 RtN_single models have a better classification accuracy as the noise intensity increases and finally all of them outperform the RtO model. They have a very decent performance on the noise images of their own training intensities, an approximately satisfying performance when the test noise intensity is slightly deviated but if the test image is much more or less noisy than the data the model has ever seen, the result is unpredictable, usually disappointing.

On the contrary, VGG16 RtN_single models have quite different behavior as shown in the lower chart in Fig. 8. Every RtN_single model except RtN_0.5 either has a comparable performance or crushes RtO model in each test case. They all have a performance line that drops steadily and not much fluctuation occurs especially RtN_0.5 model. Therefore, a claim can be made that RtO can be entirely replaced by RtN_0.3 as the latter model triumphs in all scenarios. Unfortunately, this claim cannot be made easily in the ResNet-18 case.

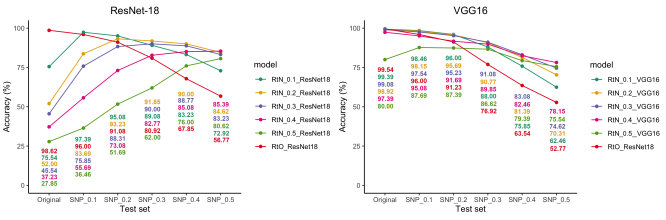


Figure 8: Accuracy comparisons among RtO and RtN_single models

When the intensity of SNP noise is below 0.3, RtO models produce a solid accuracy as the sparse noise pixels do not degrade the feature extraction process to a large extent. The overall performance of VGG16 RtN_single models is surprisingly good because noise-trained models have never seen a noise-free image. However, the performance of ResNet-18 RtN_single models in this scenario is completely unacceptable. When the intensity increases to 0.3 and above, noise pixels start to dominate the image. The performance of RtO drops quickly while noise-trained models begin to show its superiority in noise image processing for both CNNs as expected. This experiment shows that single noise training is not suitable for all CNNs such as ResNet-18 because of its unpredictable performance on light noise data. On the contrary, VGG16 noise-trained models have a more stable and decent performance. Single noise training deserves a try when the noise is intense but median filter is obviously preferred when the noise is sparse. Therefore, we conclude single noise training as a flawed and unreliable method.

c) RtN_single vs. RtN_mix: In part two, we show the classification accuracy improvement brought by training the model on a single level of noise data. The results in Fig. 8 indicate that single noise training, or single training, is not always beneficial to all CNN architectures and trade-offs need to be made in most cases. In this part, we further explore noise training by introducing mix noise training, or mix training. Proportions of images of different noise levels are carefully selected for the model to retrain on. Our proposed mix-trained models show exceptional competence in all testing scenarios.

Among all the single-noise-trained, or single-trained, models in both CNNs, RtN_0.1, RtN_0.2 and RtN_0.3 are three models that have a better overall performance. We conclude that training on images with a noise intensity of 0.1, 0.2 and 0.3 help generalization for retrained models. Not as many features can be extracted and learned when the training image has a higher noise intensity. They confuse the model and hurt its classification ability in the end. Therefore, we decide to train the model on a mixture of light noise images. Mix proportions of SNP_0.1, SNP_0.2 and SNP_0.3 images evaluated are $(25\%, 50\%, 25\%)$, $(33\%, 33\%, 33\%)$, $(20\%, 40\%, 40\%)$ and $(40\%, 40\%, 20\%)$.

We provide the classification accuracy comparisons between each mix-trained model and top three single-trained models in Fig. 9. An obvious improvement on sparse noise image classification is observed for all mix-trained models and the accuracy on intense noise images remain comparable or slightly better. Model mix_404020 is a clear winner among them. Compared to ResNet-18, single-trained VGG16 models have a more stable performance and its RtN_mix_404020 model achieves an equally good accuracy on all test sets.

d) Top models: We have discussed the advantage of applying median filters and noise training. In this part, we analyze median filter and single-training in terms of their worst and best accuracies on each noise set compared to RtO. The selective classification accuracy figures of ResNet-18 and VGG16 models are presented in Table V and VI. We

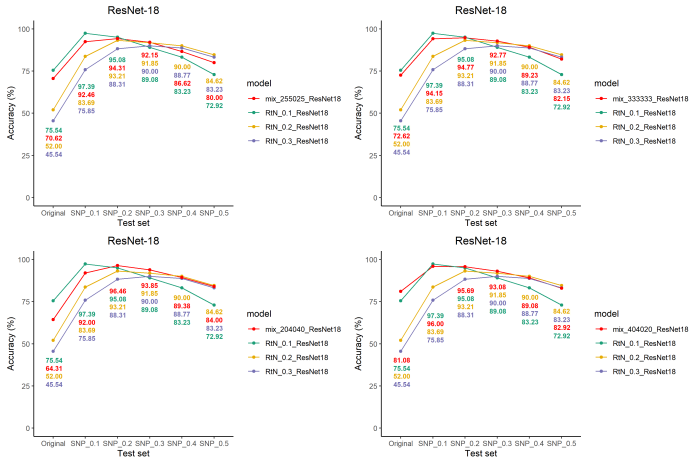


Figure 9: Performance comparisons among four mix-trained models and top three single-trained models for ResNet-18: six digits in the name of a mix-trained model stands for the proportion of SNP_0.1, SNP_0.2 and SNP_0.3 images in the training set. e.g. "mix_255025_ResNet18" is interpreted as a ResNet-18 model retrained on a set that contains 25% SNP_0.1, 50% SNP_0.2 and 25% SNP_0.3 noise images.

omitted "_train" in the test set name for simplicity. Median filter has four kernel size options: 3, 5, 7 and 9. There are five RtN models and each was retrained on a single intensity of SNP noise data as mentioned in Table IV. Table VII shows the worst and the best median filter and RtN_single model selections. For median filter, the best kernel size option tends to increase with higher noise intensity. A mixture of unfiltered noise and blur generated by the filter is usually the worst case, which can happen when the kernel size is small and the noise intensity is high. For single-trained models, the worst one is often extremely-trained, either RtN_0.1 or RtN_0.5. The best accuracy is always achieved by the model that was trained on the noise data with a slightly milder intensity.

Table V: Classification accuracy of ResNet-18 RtO and RtN_single models, including median filters and RtN_single models of the worst and best performance

Test set	RtO	RtO_MF		RtN_single	
		worst	best	worst	best
Original	98.62%	NA	NA	27.85%	75.54%
SNP_0.1	96.00%	96.31%	99.08%	36.46%	97.39%
SNP_0.2	91.08%	94.62%	95.39%	51.69%	95.08%
SNP_0.3	80.92%	88.00%	91.69%	62.00%	91.85%
SNP_0.4	67.85%	80.00%	83.54%	76.00%	90.00%
SNP_0.5	56.77%	61.54%	78.15%	72.92%	85.39%

In Fig. 10, we plot the accuracies of six models: RtO, RtO_MF_ksize_k, RtO_MF_optimal, RtN_0.x, RtN_optimal and mix_404020. RtO_MF_optimal and RtN_optimal show the ceiling performance of median filter and single-trained models on each test set as provided in Table V and VI. As the best accuracy on each noise level is usually achieved by models with different options and the noise intensity of an

Table VI: Classification accuracy of VGG16 RtO and RtN_single models, including median filters and RtN_single models of the worst and best performance

Test set	RtO	RtO_MF		RtN_single	
		worst	best	worst	best
Original	99.54%	NA	NA	80.00%	99.39%
SNP_0.1	96.00%	97.08%	98.77%	87.69%	98.46%
SNP_0.2	91.23%	93.54%	96.77%	87.39%	96.00%
SNP_0.3	76.92%	90.00%	91.39%	86.62%	91.08%
SNP_0.4	63.54%	75.69%	80.15%	75.49%	83.08%
SNP_0.5	52.77%	62.00%	66.92%	62.46%	78.15%

Table VII: The best and the worst kernel size and RtN model selections: "SNP" is omitted for noise sets except SNP_0.1 due to the table width limitation. "w" and "b" stand for worst and best respectively. For MF (median filter), the entry is the kernel size. For RtN(_single), the entry is the noise intensity of its training set. Multiple values in an entry indicate a tie.

Model			SNP_0.1	_0.2	_0.3	_0.4	_0.5
ResNet-18	MF	w	9	9	3	3	3
		b	3	3	7	9	9
	RtN	w	0.5	0.5	0.5	0.5	0.1
		b	0.1	0.1	0.2	0.2	0.4
VGG16	MF	w	9	9	3	3	9
		b	3	3	5, 7, 9	5, 7, 9	5
	RtN	w	0.5	0.5	0.5	0.1	0.1
		b	0.1	0.1	0.3	0.3	0.4

image is unknown when feeding in, these best accuracies only can be achieved at the same time theoretically. Optimal models are included here to illustrate the gap between the most optimal case and our best practical model, mix_404020. The single-trained model and median filter that have the best overall performance are also included. They are $k = 9, x = 2$ for ResNet-18 and $k = 5, x = 3$ for VGG16. As the median filter is preferred when noise intensity is low (≤ 0.2) and noise-training performs better when intensity is high (≥ 0.3). We compare mix_404020 to both the best and optimal median filters and single-trained models in particular scenarios.

For ResNet-18, when the noise intensity is 0.1, mix_404020 is 3.08% and 0.31% shy to the optimal and the best RtO_MF model. However, it surpasses both by 0.3% and 1.07% when the noise is still sparse (0.2). As the noise getting intense, mix_404020 achieves either the best or close to the best ($\leq -2.47\%$) compared to all the rest opponents including the theoretical optimal ones.

For VGG16, mix_404020 tops or gets an accuracy very close to the first place ($\leq -1.23\%$) when the noise intensity is no greater than 0.4. RtN_optimal, achieved by RtN_0.4, is 4.92% better than mix_404020 when the intensity is 0.5. Since the extremely noisy images are not the most common case and the excellent performance of mix_404020 in other noise scenarios, it is still considered as a very successful model.

V. CONCLUSION

Our experiments substantiate that noise training, especially mix-noise training, improves the performance on SNP noise image classification. The noise-trained models including

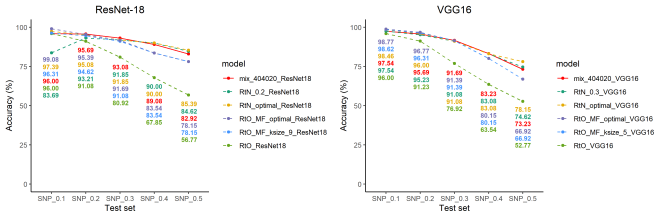


Figure 10: Performance of RtO, the best and optimal RtO_MF and single-trained models, and the best mix-trained model

single-trained and mix-trained, show more resilience to SNP noise as the intensity increases. Single-trained models of different CNNs vary in the performance on images with sparse noise. Due to this lack of reliability, median filter is preferred to process light noise images. Nevertheless, our proposed mix-trained model, mix_404020, shows exceptional ability to deal with both sparse and intense noise images. Its training set contains 40% SNP_0.1, 40% SNP_0.2 and 20% SNP_0.3 noise images. This mix-trained model is an universal solution to improve SNP noise image classification, regardless of the noise intensity.

For future research directions, we will extend this work by adding more objects, increasing the output dimension up to $1,000$ and creating a complete noise-proof CNN. We will also explore how other CNNs benefit from noise training by bringing in classic and state-of-art architectures such as AlexNet [21], SqueezeNet [22], EfficientNet [23], Mobilenetv2 [24], NFNet-F4 [25] etc. Moreover, we will investigate other training set constructions by introducing more intense noise images and tuning the proportions. We also plan to test if other noise types can be generalized and learned through training such as Gaussian noise and speckle noise. We are interested in designing a noise-trained model that works for all kinds of noise. Taking a step further, other image flaws such as partial evidence, low resolution, over and underexposure can be explored for their generalization potentials.

Acknowledgment

THE RESEARCH REPORTED IN THIS PAPER IS PARTIALLY SUPPORTED BY THE PHILIP AND VIRGINIA SPROUL PROFESSORSHIP AT IOWA STATE UNIVERSITY.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International journal of computer vision 115 (3) (2015) 211–252.

[3] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[4] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li, et al., Florence: A new foundation model for computer vision, arXiv preprint arXiv:2111.11432 (2021).

[5] S. Dodge, L. Karam, Understanding how image quality affects deep neural networks, in: 2016 eighth international conference on quality of multimedia experience (QoMEX), IEEE, 2016, pp. 1–6.

[6] D. Temel, J. Lee, G. AlRegib, Cure-or: Challenging unreal and real environments for object recognition, in: 2018 17th IEEE international conference on machine learning and applications (ICMLA), IEEE, 2018, pp. 137–144.

[7] J. Yang, N. Xiong, A. V. Vasilakos, Two-stage enhancement scheme for low-quality fingerprint images by learning from the images, IEEE transactions on human-machine systems 43 (2) (2012) 235–248.

[8] Y. Zhou, D. Liu, T. Huang, Survey of face detection on low-quality images, in: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), IEEE, 2018, pp. 769–773.

[9] P. Li, L. Prieto, D. Mery, P. Flynn, Face recognition in low quality images: a survey, arXiv preprint arXiv:1805.11519 (2018).

[10] J. Seo, H. Park, Object recognition in very low resolution images using deep collaborative learning, IEEE Access 7 (2019) 134071–134082.

[11] K. K. V. Toh, H. Ibrahim, M. N. Mahyuddin, Salt-and-pepper noise detection and reduction using fuzzy switching median filter, IEEE Transactions on Consumer Electronics 54 (4) (2008) 1956–1961.

[12] E. J. Leavline, D. A. A. G. Singh, Salt and pepper noise detection and removal in gray scale images: An experimental analysis, International Journal of Signal Processing, Image Processing and Pattern Recognition 6 (5) (2013) 343–352.

[13] B. Wang, Q. Xiang, Fast median filter image processing algorithm and its fpga implementation, Front Signal Process 4 (4) (2020) 88–94.

[14] B. Weiss, Fast median and bilateral filtering, in: ACM SIGGRAPH 2006 Papers, 2006, pp. 519–526.

[15] M.-H. Hsieh, F.-C. Cheng, M.-C. Shie, S.-J. Ruan, Fast and efficient median filter for removing 1–99% levels of salt-and-pepper noise in images, Engineering Applications of Artificial Intelligence 26 (4) (2013) 1333–1338.

[16] B. R. Jana, H. Thotakura, A. Baliyan, M. Sankararao, R. G. Deshmukh, S. R. Karanam, Pixel density based trimmed median filter for removal of noise from surface image, Applied Nanoscience (2021) 1–12.

[17] G. R. Arce, Nonlinear signal processing: a statistical approach, John Wiley & Sons, 2005.

[18] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, nature 521 (7553) (2015) 436–444.

[19] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[20] L. Yang, S. Hanneke, J. Carbonell, A theory of transfer learning with applications to active learning, Machine learning 90 (2) (2013) 161–189.

[21] A. Krizhevsky, One weird trick for parallelizing convolutional neural networks, arXiv preprint arXiv:1404.5997 (2014).

[22] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, K. Keutzer, Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size, arXiv preprint arXiv:1602.07360 (2016).

[23] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR, 2019, pp. 6105–6114.

[24] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.

[25] A. Brock, S. De, S. L. Smith, K. Simonyan, High-performance large-scale image recognition without normalization, in: International Conference on Machine Learning, PMLR, 2021, pp. 1059–1071.