

# Effects of Degradations on Deep Neural Network Architectures

Prasun Roy\*, Subhankar Ghosh\*, Saumik Bhattacharya\* and Umapada Pal, *Senior member, IEEE*  
 Indian Statistical Institute, 203 B.T. Road, Kolkata, India - 700108

**Abstract**— Recently, image classification methods based on capsules (groups of neurons) and a novel dynamic routing protocol are proposed. The methods show promising performances than the state-of-the-art CNN-based models in some of the existing datasets. However, the behavior of capsule-based models and CNN-based models are largely unknown in presence of noise. So it is important to study the performance of these models under various noises. In this paper, we demonstrate the effect of image degradations on deep neural network architectures for image classification task. We select six widely used CNN architectures to analyse their performances for image classification task on datasets of various distortions. Our work has three main contributions: 1) we observe the effects of degradations on different CNN models; 2) accordingly, we propose a network setup that can enhance the robustness of any CNN architecture for certain degradations, and 3) we propose a new capsule network that achieves high recognition accuracy. To the best of our knowledge, this is the first study on the performance of CapsuleNet (CapsNet) and other state-of-the-art CNN architectures under different types of image degradations. Also, our datasets and source code are available publicly to the researchers.

**Index Terms**—CapsuleNet, convolutional neural networks, image degradations.

## I. INTRODUCTION

VISUAL quality is an important parameter in machine vision problems. Though there are several no-reference image quality measures available in literature [1], visual quality of an input image is a subjective quantity and traditionally we rely on human perception to conclude about it. However, computer vision algorithms work differently from human vision system (HVS), and the concept of image quality for computer vision problems does not always match with human perception. Convolutional neural networks (CNNs) depend on several sets of filter outputs to perform the final classification task. Thus, it is difficult to predict the outcome of a degraded input in an intuitive way and the classification accuracy largely depends on the model architecture and the nature of the degradation. In most of the cases, we train and validate a CNN model with high quality images with minimum noise. However, in practical applications, several different kinds of degradations can be introduced in the input image that can heavily affect the performances of CNN models. These image degradations can be obtained due to poor image sensor, lighting conditions, focus, stabilization, exposure time etc. To overcome such effects, some researchers have suggested to include noisy data in the training itself [2]. Though, this



Fig. 1. Examples of some typical samples of data from our datasets: (a) Synthetic digits dataset; (b) Natural images dataset.

technique produces better results than training only with high quality images, it is practically not possible to train a network with all probable degradation types that may appear in real scenarios.

Starting from multi-class classification tasks to generative models, CNNs are used in numerous computer vision algorithms. State-of-the-art CNN architectures such as ResNet50, Inception v3, DenseNet etc. [3]–[5] have achieved exceptional results for large image classification task in ILSVRC 2010 challenge (ImageNet). Interestingly, it has been shown recently that well-trained complex CNN models might produce wrong results even in the presence of a small amount of carefully selected noise, although such noise does not create any problem in visual recognition [6]. Though the probability of occurrence of such adversarial noises might be low, it is important to know the performances of different CNN architectures under different noise conditions to build more robust systems in future. Thus, in presence of different image degradations, the performances of different deep CNN architectures in classification task consisting of different challenging images are considered in this paper.

### A. Related Works

In most of the recent applications, it is generally assumed that the CNN models accept good quality images. However in many cases, it is not possible to have good quality images in computer vision problems. Thus, several authors have recently proposed different architectures and preprocessing steps to work with low quality images. In [7], the authors used coupled kernel embedding to recognize faces in low resolution images that also suffer from degradations due to high compression.

\*These authors contributed equally in this paper.

\*\*The source code and datasets are available at: <https://goo.gl/EZ43Lx>

Zou and Yuen [8] addressed the same problem of face recognition in low resolution by introducing discriminative constraints to achieve high quality images for recognition. In [9], authors introduced a modified version of well-known MNIST dataset, that includes synthetically generated noisy handwritten images, and using this dataset they proposed a novel framework that learns representations using probabilistic quadtrees and deep belief network. A noisy face database is developed in [10] to act as a benchmark in robust face recognition algorithms. However, in [10], authors did not mention any model that can achieve robust recognition results. Later, Tao et al. [11] proposed joint kernel sparse representation based classification algorithm that performs satisfactorily on the database proposed in [10]. In [12], Ullman et al. tried to show that human vision system (HVS) and CNNs process an image differently by defining minimal recognizable configurations (MIRC) which are the smallest cropped version of an input image for which human being can still recognize a class or action perfectly. It was shown in [12] that even though the cropping action reduces the information content in an image, CNNs are still not comparable to HVS. In [2], the authors rigorously analyzed the effect of image degradation under different noise conditions on the accuracy of CNNs in classification task. Though the authors include degradations like blur, compression, contrast etc., they did not include different common degradations like motion blur, salt and pepper noise etc. in their work. Also, the work in [2] does not include models like ResNet50 and CapsuleNet, that integrate different architectures along with conventional CNN layers to find more complex features from an input image. In this work, we not only consider more number of image degradations, we also compare latest CNN architectures using two completely different types of datasets. The major contributions of our work are as follows.

- Though CNN is widely used in several image classification tasks, the effect of image degradation is not well-explored. The models considered in [2] are simple CNN models unlike ResNet, Inception or capsule network. Moreover, the effect of adversarial attacks on these networks are not discussed. In this paper, we consider state-of-the art CNN models, and consider larger set of degradation types to understand their effects.
- Though the recently proposed capsule architecture has shown promising results for MNIST dataset, the performance of the architecture is not investigated for natural images. The susceptibility of the capsule model under degradations are also not reported in literature. We propose a novel capsule based architecture to show that as we increase depth of a CNN model, the accuracy might improve, but that may significantly reduce the robustness against severe noises. Thus, we propose a novel architecture that can enhance the robustness of any CNN architecture for several degradations maintaining an accuracy-robustness trade off.
- We found that structural similarity index measure (SSIM) [13] can be used coarsely as a metric to measure the effects of degradations on a CNN model.

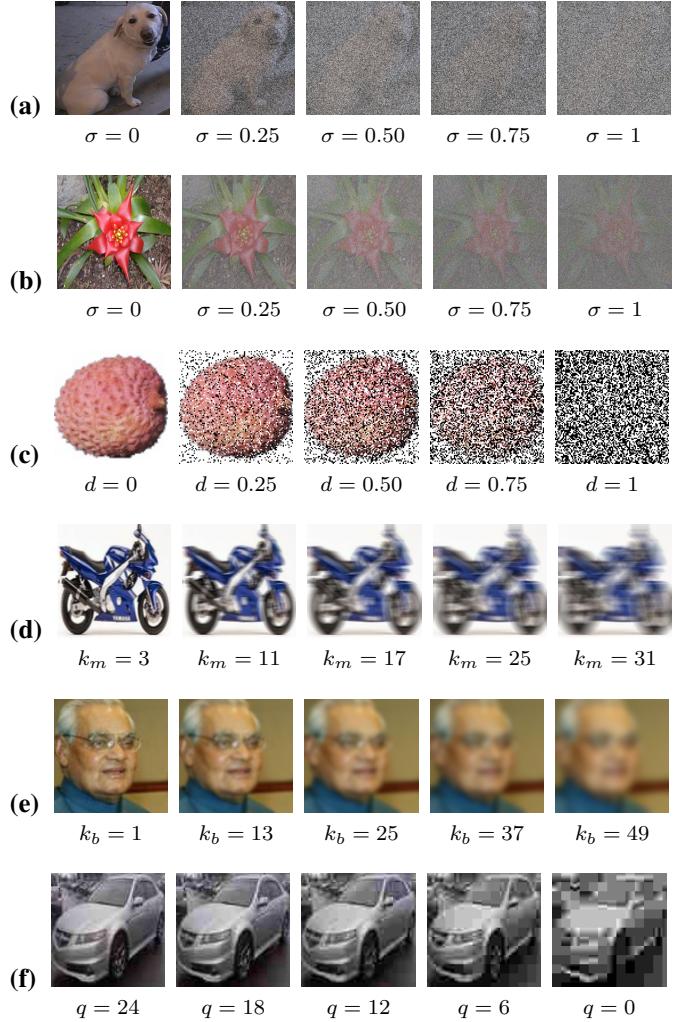


Fig. 2. Examples of images after different image degradations: (a) Gaussian white noise, (b) Colored Gaussian noise, (c) salt and pepper noise, (d) motion blur, (e) Gaussian blur, (f) Degradation due to JPEG compression (JPEG quality).

## II. DATASET AND EXPERIMENTAL SETUP

To understand the effects of different distortions on classification task, we select some state-of-the-art CNN architectures that achieved impressive results in ImageNet challenge. To evaluate the models, we have developed two datasets. One of them is a synthetic digits dataset and another is a natural images dataset. After training a model with a particular dataset, we apply different image degradations, e.g., motion blur, Gaussian blur, additive noise, salt and pepper noise etc., on each image individually and measure the accuracy of individual models. Details of these datasets are as follows.

### A. Dataset

*1) Synthetic Digits Dataset:* Getting the inspiration from MNIST dataset, we build our own synthetic numeral dataset with 16 different English fonts. The numerals are randomly rotated with rotation angle -30 degree to 30 degree. Each digit has a random color and random font size ranging from 30 to 240. Additionally, to increase the difficulty level of the dataset,

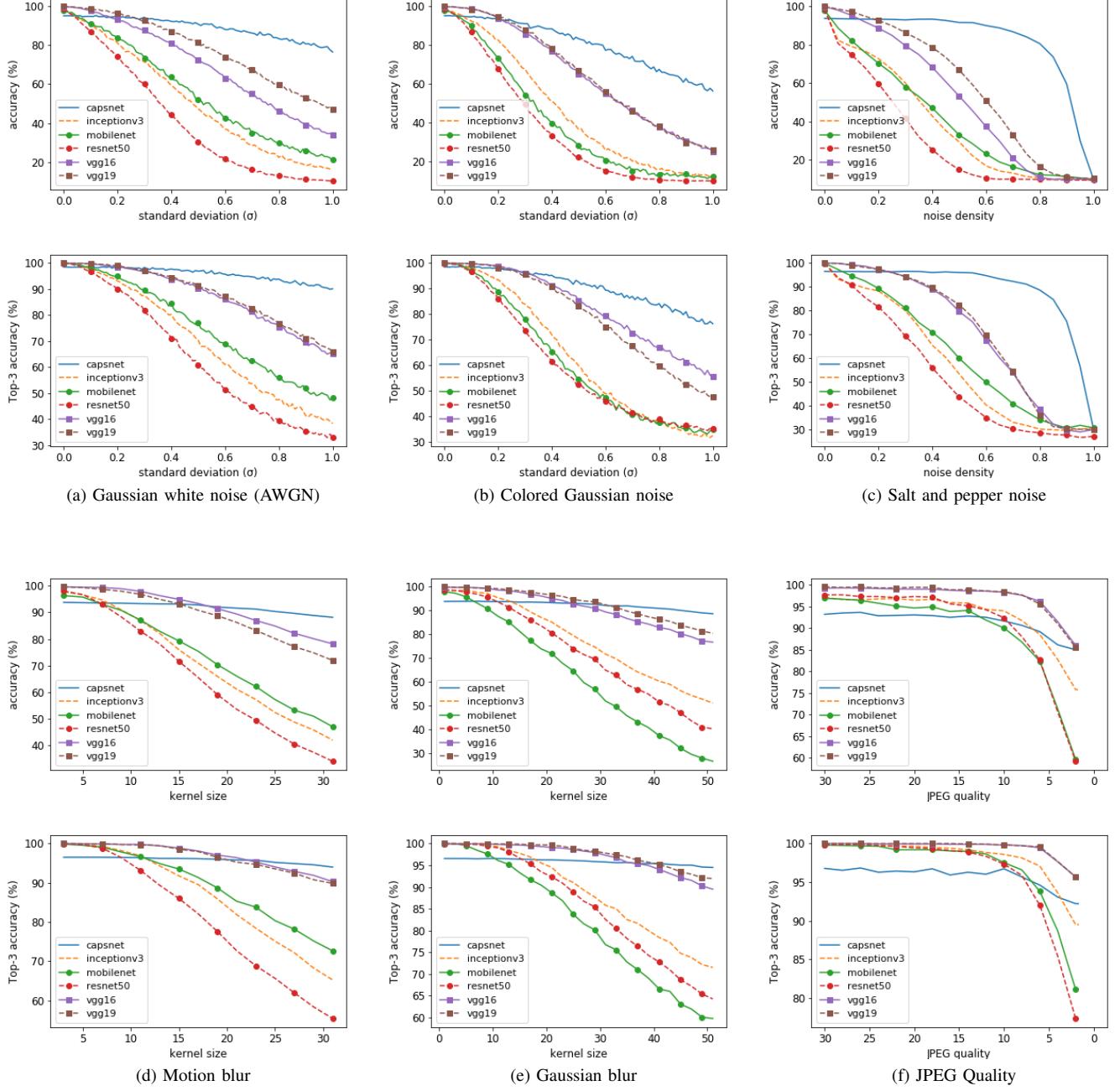


Fig. 3. Comparison of classification accuracies of different CNN architectures under different image degradations on synthetic digits dataset. For each type of degradation, the top figure shows accuracy (top-1 accuracy) vs. respective degradation parameter and the bottom figure shows top-3 accuracy vs. respective degradation parameter.

we arbitrarily select image patches from COCO dataset [14] and place them in the background of the digits. The dataset has 10 different English numeral classes and each class contains 1200 images, and hence the total size of this dataset is 12000. We call this dataset as ‘*ISID*’ dataset. Using 6 fold cross-validation technique, we use 10000 images for training and 2000 images for testing the CNN models.

2) *Natural Images Dataset*: State-of-the-art CNN models trained on the ImageNet dataset can correctly classify 1000 classes even in very complex environmental context. On the

other hand, the behavior of capsule network is not well-explained for natural images like ImageNet or COCO dataset. As it is difficult to train CapsuleNet with large number of classes, we compiled a dataset containing natural images to evaluate the performance of different CNNs including CapsuleNet for complex input images under various image degradations. The dataset contains 8 different classes- airplane, car, cat, dog, flower, fruit, motorbike and person having 727, 968, 885, 702, 843, 1000, 788 and 986 image samples respectively from the 8 classes. We call this dataset as ‘*ISINI*’ dataset.

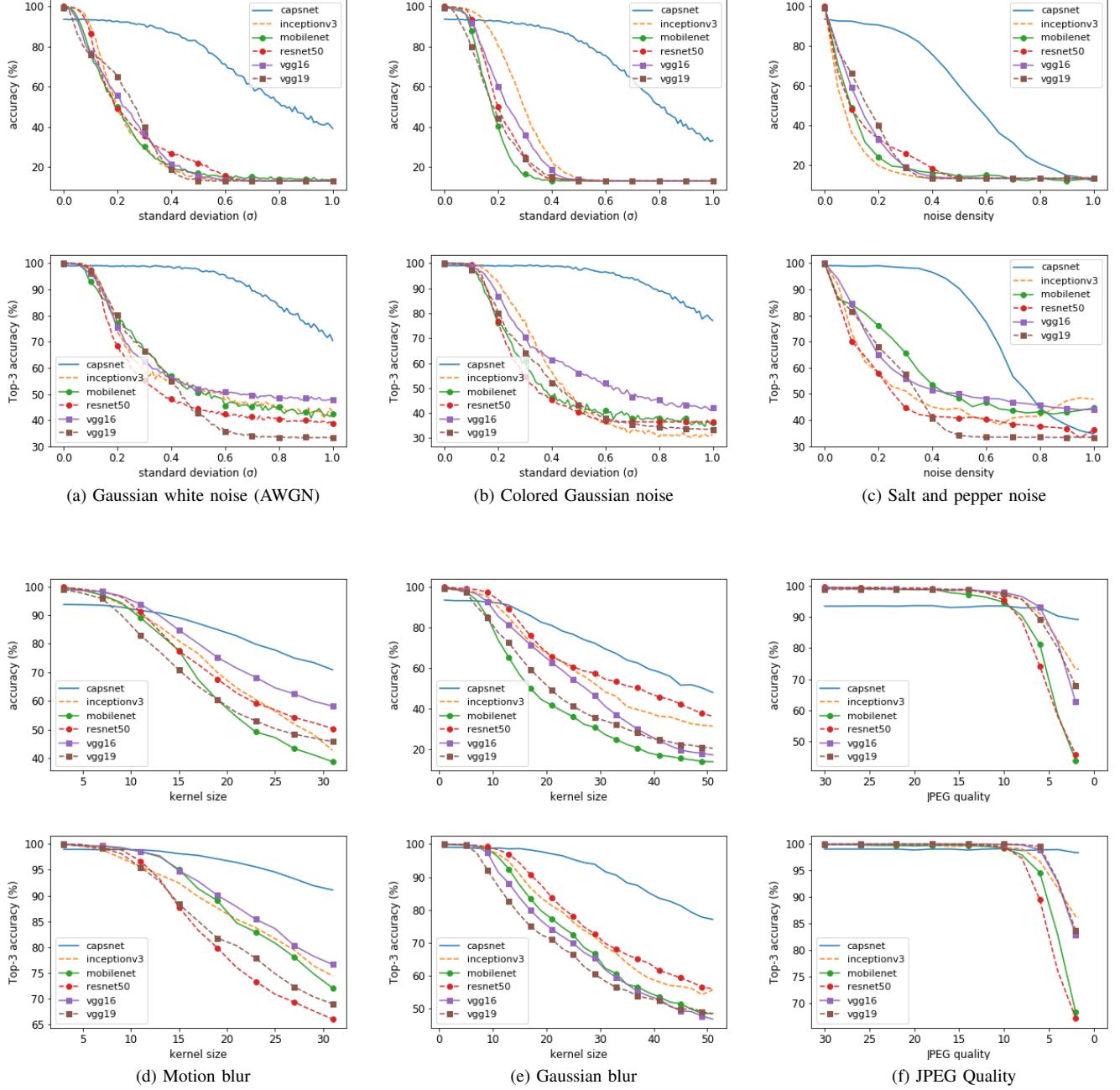


Fig. 4. Comparison of classification accuracies of different CNN architectures under different image degradations on natural images dataset. In each sub-figures, the top figure shows accuracy (top-1 accuracy) vs. respective degradation parameter and the bottom figure shows top-3 accuracy vs. respective degradation parameter.

Of these total 6899 images, using 5 fold cross-validation technique, we use 5724 images for training and 1175 images for testing the CNN models.

Fig. 1 shows some typical examples from synthetic digits dataset and natural images of the proposed datasets.

### B. Deep Neural Networks

In this paper we consider six CNN based architectures- MobileNet, VGG16, VGG19, ResNet50, InceptionV3 along with CapsuleNet to evaluate the respective performances. Though

numerous CNN architectures are available in the literature, the networks that are tested here are the popularly known standard deep CNN architectures and hence they are used here for the comparisons.

The first network that we considered for the experiment is MobileNet which is based on streamlined architecture consisting of depth-wise separable convolutional layers [3]. Considering both depth-wise and point-wise convolutions as separate layers, this model has 28 layers. It maintains a

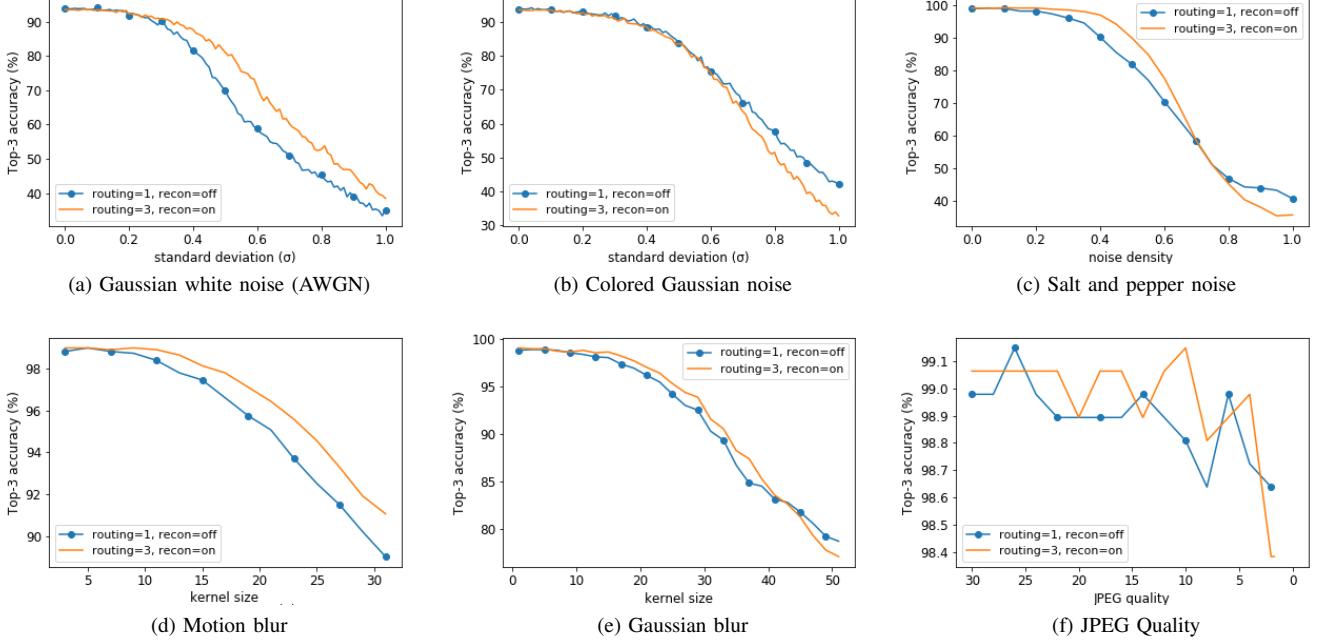


Fig. 5. Top-3 performance comparison of CapsuleNet architectures on different image degradations for two different parameter settings.

TABLE I  
SUMMARY OF THE MODELS USED

Model	#Trainable Parameters	Input Size
MobileNet	55.7M	(224 × 224 × 3)
VGG16	41.5M	(224 × 224 × 3)
VGG19	54.6M	(256 × 256 × 3)
ResNet50	26.7M	(224 × 224 × 3)
Inception v3	157.1M	(299 × 299 × 3)
CapsuleNet	47.6M	(104 × 104 × 3)
V-CapsNet	15.7M	(256 × 256 × 3)

trade-off between latency and accuracy to consume minimal computational resources. This architecture is preferable in mobile and embedded based vision application where the availability of resources is limited.

VGG16 [4] is a deep architecture with 13 convolutional layers with very small convolutional filters and 3 fully connected layers. Including input, output, pooling and activations, there are 41 layers in this model. From the literature, it is noted that it outperformed many well-known CNN like Caffe Reference Model and AlexNet in ImageNet classification task.

VGG19 is even deeper architecture than VGG16 with similar architecture but 16 convolutional layers and 3 fully connected layers. Including input, output, pooling and activations, VGG19 has 47 layers.

As CNN based models go deeper, they become difficult to train for large datasets. It was found that instead of training the filtered outputs, it is easier to train on the residuals of the outputs [5]. Following this concept, ResNet architecture is designed. In this paper, we consider ResNet50 architecture for the comparisons. ResNet50 depends on a different kind of architecture than VGG to achieve better accuracy for

ImageNet classification problem. ResNet50 typically contains fewer filters and has lower complexity than VGG.

Inception architectures are built upon so called ‘inception modules’ that contains different convolutional kernels stacked together along the depth dimension to get multi-scale features [15]. Here, we consider Inception v3 model that also includes the concept of residual training for better accuracy.

CapsuleNet is based on a relatively new neural network concept of ‘dynamic routing between capsules’. A capsule is a collection of neurons whose activity vector represents the instantiation of parameters of an object [16]. CapsuleNet removes the maxpooling layers and relies on dynamic routing protocol between capsules to achieve translational invariance. The CapsuleNet architecture that has been used in this paper is different from the original architecture proposed in [16]. Our model takes input image of dimension 104 × 104. Unlike the model of [16], our model has two generic convolutional layers along with a primary capsule layer and one classification capsule layer. The decoder module contains three fully connected layers and a reshaping operation after the last layer. The loss function and the routing protocol of the architecture are as described in [15].

The number of parameters that are used here for different CNN architectures are summarized in Table I. For the datasets used in this paper except for CapsuleNet, we train the models after initializing with pre-trained ImageNet weights for faster convergence. We also propose a modified capsule architecture based model named V-CapsNet that we will discuss in Sec. III.

### C. Degradation Types

We choose six well-known degradations which are common in any vision based tasks. We consider the following types of

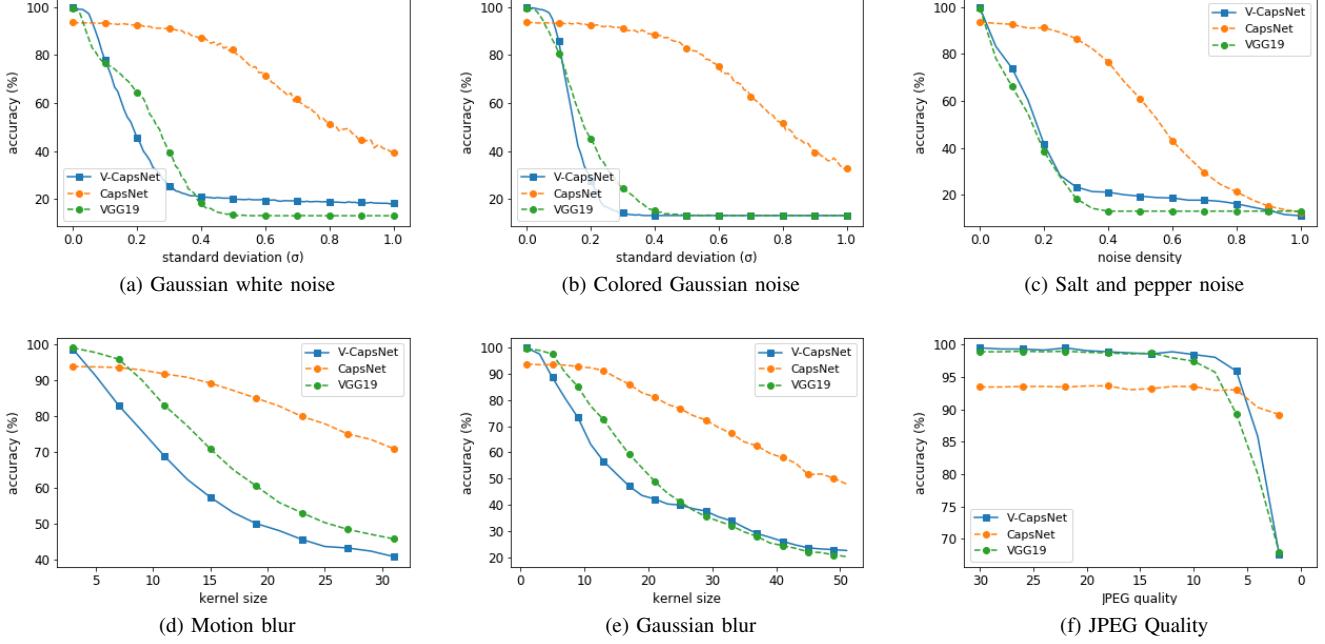


Fig. 6. Performance comparison of V-CapsNet architecture with VGG19 and CapsNet under different image degradation.

degradations for our experiments.

1) *Gaussian noise*: Additive white Gaussian noise (AWGN) can be introduced if an image is transmitted through a channel or it is captured using a low quality sensor [2]. To understand the effect of AWGN, we apply two different types of AWGN noise.

In AWGN, we generate a noise matrix that has same spatial dimension with the image and add the same noise with the three color channels.

In the second type, noises are added independently in three color channels and thus the noise can cause color artifacts. In this paper, we call the second type of AWGN as ‘colored Gaussian noise’. It is important to note that the colored Gaussian noise considered in this paper is different from additive colored Gaussian noise. In both the cases, the AWGN have zero mean with standard deviation  $\sigma > 0$ . We change the variance values in the experiment to see the effect of AWGN noises on different networks.

2) *Salt and pepper noise*: Salt and pepper is a typical impulse noise that can be observed in images due to sparse but intense disturbances. This noise replaces the original pixel values with random black and white pixels. We define a parameter  $d$  that controls the noise density in the image. For example,  $d = 0.1$  indicates that the 10% pixels in an image is degraded with salt and pepper noise. We vary the noise density from 0 to 1 to check its effect on a classification task.

3) *Blur*: One of the most common degradations that can be observed in real scenes is blurring. In this paper, we consider two different types of blurs: motion blur and Gaussian blur. Motion blur typically occurs due to poor stabilization of camera or movement of an object during exposure. Gaussian blur roughly approximate defocus blurring and blurs that may arise in different post-processing operations. In this paper, we

consider only horizontal blur with kernel width  $k_m$  signifying the number of pixels that contributes in the motion blurring. We vary  $k_m$  from 1 to 31 with an interval of 2 and normalize the kernel accordingly to observe the effect of motion blur. To generate zero mean Gaussian blur, we vary the kernel size  $k_b$  from  $3 \times 3$  to  $51 \times 51$  with an interval of 2 to maintain odd kernel size and the standard deviation of the blur is calculated as  $\sigma_b = 0.3 * ((k_b - 1) * 0.5 - 1) + 0.8$ , where  $k_b$  is the size of the square Gaussian kernel.

4) *Degradation due to JPEG compression*: Often after capturing, raw image goes through multiple compression steps for storage or processing. Thus, to understand the effect of compression, we consider JPEG compression as a distortion type in our experiment. We use standard JPEG encoder and vary the JPEG quality level ( $q$ ) from 30 to 0 in our experiments where a higher value in the quality level parameter indicates better visual quality of the compressed image with less compression.

In Fig. 2, we show the effects of different degradations that are used in this work. To generate a degraded image in our system, we first apply the degradation on the input image and then resize the degraded image into input shape of a specific CNN model.

### III. EXPERIMENTAL RESULTS, ANALYSIS AND PROPOSED SOLUTION

To understand the effects of degradations on different CNN architectures, we apply the degradations on the input images and measure the top recognition accuracy and top-3 recognition accuracy of the six models considered in this paper. The output of the networks are the recognition probability of each class of the dataset. We include the top-3 accuracy measure

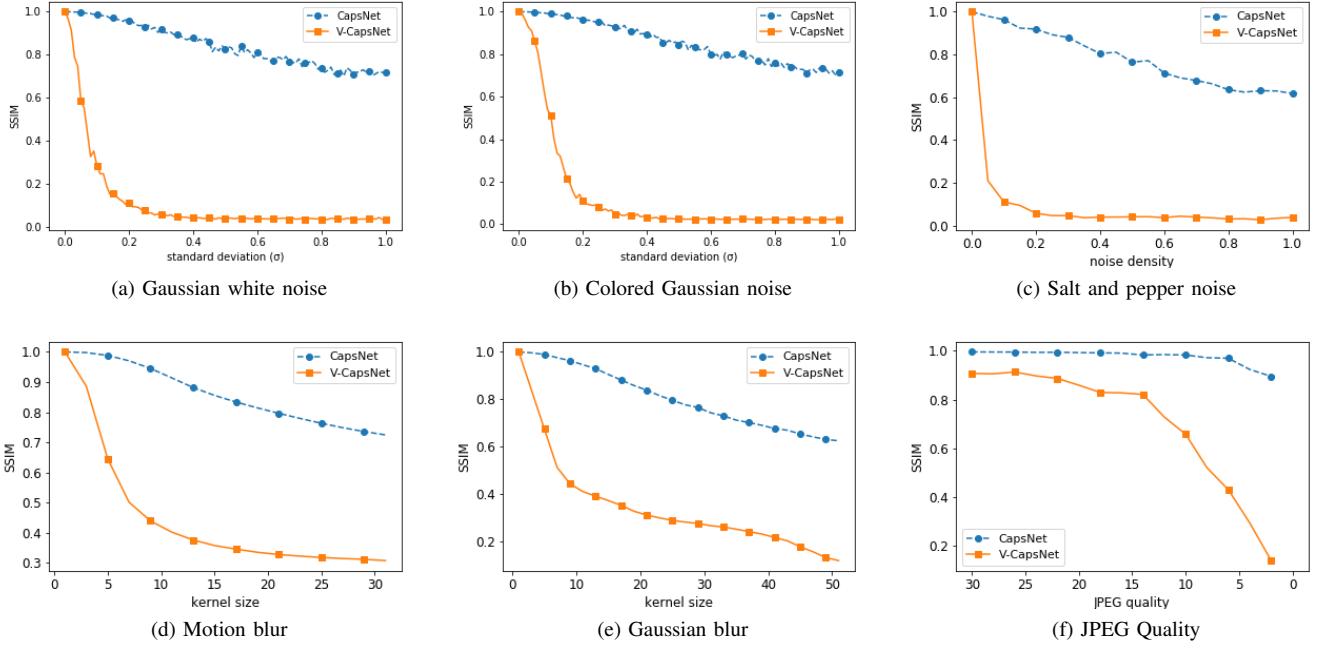


Fig. 7. SSIM comparisons between the last convolution layers of basic CapsuleNet and V-CapsNet under different image degradations.

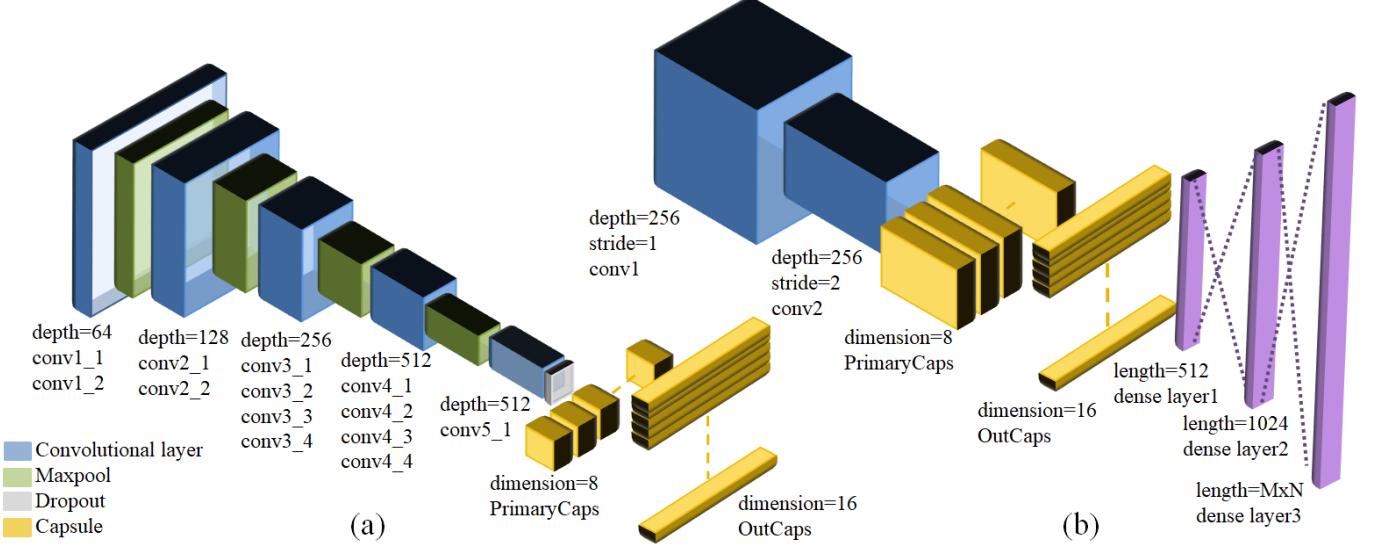


Fig. 8. Two different capsule architectures used in this paper: (a) V-CapsNet; (b) basic CapsuleNet.

as it is popular in many complex recognition tasks. Fig. 3 and Fig. 4 show the accuracies of different models under different degradations on ISISD and ISINI respectively.

#### A. Effects of degradations on CNN architectures:

From the experiments, it can be observed that for both AWGN and colored Gaussian, the recognition accuracies decrease as we increase the variance of the additive noises. VGG architectures show more robustness against the additive noises than the other conventional CNN architectures. The decrement in recognition accuracy for CapsuleNet is notably small for Gaussian noises.

Salt and pepper noise badly affects MobileNet architecture. Though the recognition accuracy using CapsuleNet is slightly lower than the other models in absence of impulse noise, CapsuleNet starts to outperform all the models even when the noise density is 0.2. CapsuleNet retains the robustness upto noise density 0.8. For both the accuracy measures (Top-1 and Top-3), CapsuleNet outperforms all the state-of-the-art models in presence of salt-and-pepper noise.

Both motion blur and Gaussian blur degrade the performance of CNN models. However, VGG architectures are more robust to blurring than ResNet or Inception models for synthetic digits dataset. CapsuleNet is robust to both blurring

degradations and outperforms MobileNet, Inception v3 and ResNet.

As we can see from Figs. 3(f) and 4(f), JPEG noise does not affect the recognition performance of any model till the compression quality is 20 as the structural quality of an input image remains almost unchanged. Beyond that the performance of ResNet and MobileNet degrade sharply. It can be seen that Inception, VGG and CapsuleNet architectures perform significantly well even when the JPEG quality is close to 0.

To understand the robustness of the capsule architecture, we perform the same test on two different CapsuleNet models - (i) with single routing iteration and with only marginal loss in the loss function and (ii) with 3 routing iterations and reconstruction loss combined with marginal loss in the loss function. In both the cases, no significant change in accuracy under different image degradations was observed, though architecture with higher routing and reconstruction loss as regularizer performed better in most of the cases. In Fig. 5, we compare the recognition accuracy for two different hyper parameter settings of the CapsuleNet. It can be observed that except for the additive noises, CapsuleNet with higher routing performs slightly better.

#### B. Effect of depth on capsule architecture:

As the basic capsule architecture cannot achieve substantially high accuracy for real image dataset, we design a novel capsule-based architecture. The proposed architecture takes input shape  $256 \times 256$  and the convolutional layers are same as VGG19 architecture up to its 1st convolutional layer of fifth block. The output of the last convolutional layer goes to primary capsule layer with a dropout of 80% where we have 32 number of 8-dimensional capsules generated using convolution with kernel size 3 and stride 2. The final 16-dimensional capsule layer is same as in [16], but to reduce the number of parameters, we remove the decoder module of basic capsule architecture and minimize only the marginal loss. For the ease of discussion, we call this novel architecture as V-CapsNet (VGG19 + CapsuleNet) in rest of our discussion. Using the architecture, we achieve 99.83% accuracy in the natural image dataset, which is almost 6.2% higher than the accuracy of basic CapsuleNet model discussed in Sec. II-B. However, even after this significant improvement in classification accuracy, we observe that in almost all the cases, V-CapsNet is more susceptible to degradations than the basic CapsuleNet architecture. The performance of V-CapsNet under different image degradations are shown in Fig. 6. It is evident from the figure that though the novel V-CapsNet architecture achieves significantly higher accuracy than the conventional CapsuleNet architecture, V-CapsNet is more sensitive to image degradations than both basic CapsuleNet and VGG19.

While investigating the reason behind the resilience of basic CapsuleNet model over V-CapsNet under degradations, we observe that not only the capsule layers, but also the shallowness of the basic CapsuleNet model shows robustness when degradation is present in the input image. To test our hypothesis, we take the output of last convolution layer of

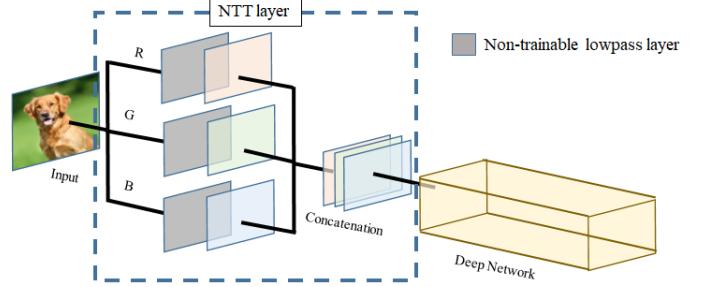


Fig. 9. Nontrainable-trainable (NTT) layer placement with any conventional CNN architecture.

basic CapsuleNet and the output of last convolution layer of V-CapsNet and measures the change in the filter outputs independently when we introduce degradation in input image. To quantify the change in output, we normalize the output and measure mean structural similarity index (SSIM) [13] of the output in presence of noise with respect to the output in absence of noise. It is observed that the structural information of the features extracted by basic CapsuleNet change much slower than the structural information of the features extracted by V-CapsNet. In other words, the features extracted by basic CapsuleNet is more robust to image degradation. Thus, the shallowness of the basic CapsuleNet might be one of the major reasons behind its robustness against different types of degradation. The SSIM comparison of the features extracted by basic CapsuleNet and V-CapsNet is shown in Fig. 7. The two capsule architectures, i.e., basic CapsuleNet and V-CapsNet, that are used in this paper are shown in Fig. 8.

#### C. Proposed architecture for robustness:

As the depth of a network allows to perform complex tasks by incorporating more nonlinearity, the experiment leads to a major question- whether it is possible to increase the robustness of any CNN architecture for certain type of degradation. One popular way to handle degradation is to train a network with both degraded and undegraded image samples. But there are several limitations with this approach. This not only increases the training time to large extent, but it is also not possible to have samples to capture all degradation variations during training phase. For example, it is straight forward to generate synthetic degraded image samples with linear blur kernel, but it is difficult to have samples with all probable nonlinear blur kernels. Thus, it is desirable if the network is inherently robust against different types of probable degradations. To achieve that, we propose a variant of depth-wise filtering approach, where on top of any existing network, we include one depth-wise nontrainable filter layer followed by another depth-wise trainable layer. We set the weights of the nontrainable layers as the weights of a lowpass filter. In Fig. 9 we depict the composite nontrainable-trainable (NTT) layer. Both the trainable and nontrainable layers have depth 1 and linear activation. The size of the filters are same in both trainable and nontrainable layers with stride 1 and padding ‘same’. In the figure, ‘R’, ‘G’ and ‘B’ indicate three color

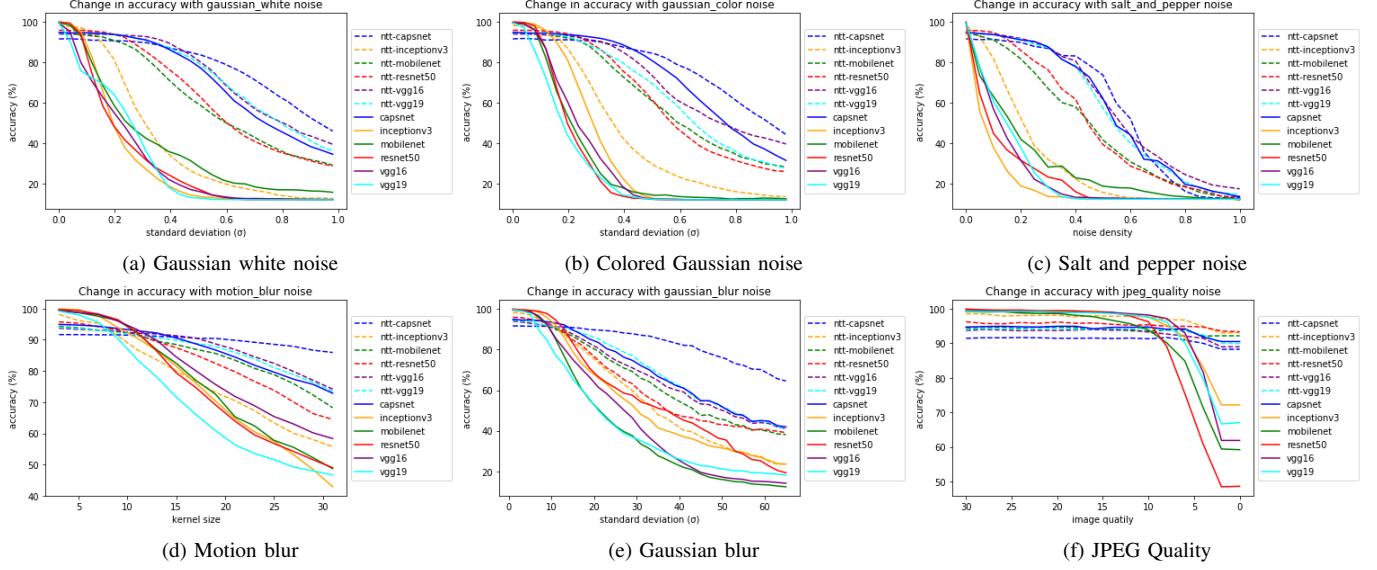


Fig. 10. Performance comparison of different CNN architecture with and without NTT layer.

channels. Though, the arrangement may look like conventional lowpass filtering that we use in case of additive noise, the key difference is that NTT layer is part of the network itself, which means it will remain on top of the conventional network in the training and testing phase like any usual layer, and any input image will pass through this layer. Because of this, during training phase, the undegraded images will also pass through the low pass filtering which is against the conventional approach of preprocessing. The motivation of NTT layer is as follows. If we train a network with only good quality images, and try to classify any degraded image after any post-processing like denoising, deblurring, we get less accuracy because there can be several subtle structural information that will be missing in the processed image. Thus it is necessary to create invariance in the network to the subtle structures that can be present in an image, and that can be easily modified by noises or different post-processing steps. In Fig. 10 we demonstrate the robustness of different CNN architectures with and without NTT layers. It can be observed that with NTT-layer, all the CNN architectures exhibit significant improvement under all the degradation types considered in this work. Though, for a given architecture, we may compromise certain amount of accuracy in absence of noise, we gain significant robustness against degradations. For example, in case of VGG19, we sacrifice 6% accuracy in absence of noise, but in Gaussian color noise and salt and pepper noise, we may gain up to 68% increase in accuracy in presence of noise. Depending on the network depth and construction, the drop in maximum accuracy, and gain in presence of noise will vary, but the behaviour remains same. The trade-off between the maximum accuracy and the robustness can be achieved by varying the complexity of the network and the size of the filter in the NTT layer. We use average filter as the non-trainable lowpass layer. As different network accepts different input sizes, we have different filter sizes in NTT layer. The filter dimensions are mentioned in Table II.

With this idea, we add the NTT layer with the designed V-capsnet. The performance of the modified network is shown in Fig. 11. It can be observed that in most of the cases V-CapsNet with NTT layer has maximum accuracy close to VGG19 and robustness close to CapsuleNet. Interestingly, in all the cases, NTT-CapsuleNet has equal or more robustness than the CapsuleNet architecture.

Though degradation of input image quality is one of the prime reasons for poor performance of CNN models, researchers have found out that a well-calculated imperceptible change may drastically reduce the accuracy of a CNN architecture. This perturbed images, known as adversarial examples, can be generated in various ways [17]. Even simple crafted attack [18] can significantly affect the accuracy of a neural network. To understand the performances of CNN models under adversarial attacks, we apply fast gradient sign method (FGSM) [18] on ISINI dataset. FGSM calculates the gradient of the loss with respect to the intensity at each pixel location and then modifies the actual pixel intensity by an amount  $\epsilon$  in the direction such that the loss increases. We use this method to perform an untargeted adversarial attack, which means we update the intensity such that the image is misclassified to any other class. The amount of  $\epsilon$  decides the visible change in the modified image. We generate the adversarial examples for each trained model and compare their robustness as higher  $\epsilon$  is required to change the accuracy of a more robustly trained model. We found that for ISINI dataset, capsule network and VGG architectures perform poorly under untargeted adversarial attack, whereas Inception v3, ResNet50 and proposed V-CapsNet models are proved to be quite robust. As NTT-layer is integral part of the network architecture, it is important to observe the effect of NTT-layer under the adversarial attack. As shown in Fig. 12(a), presence of NTT layer makes it easier to create adversarial examples for a particular architecture; but as depicted in Fig. 12(b), average PSNR of the generated samples drop significantly in presence

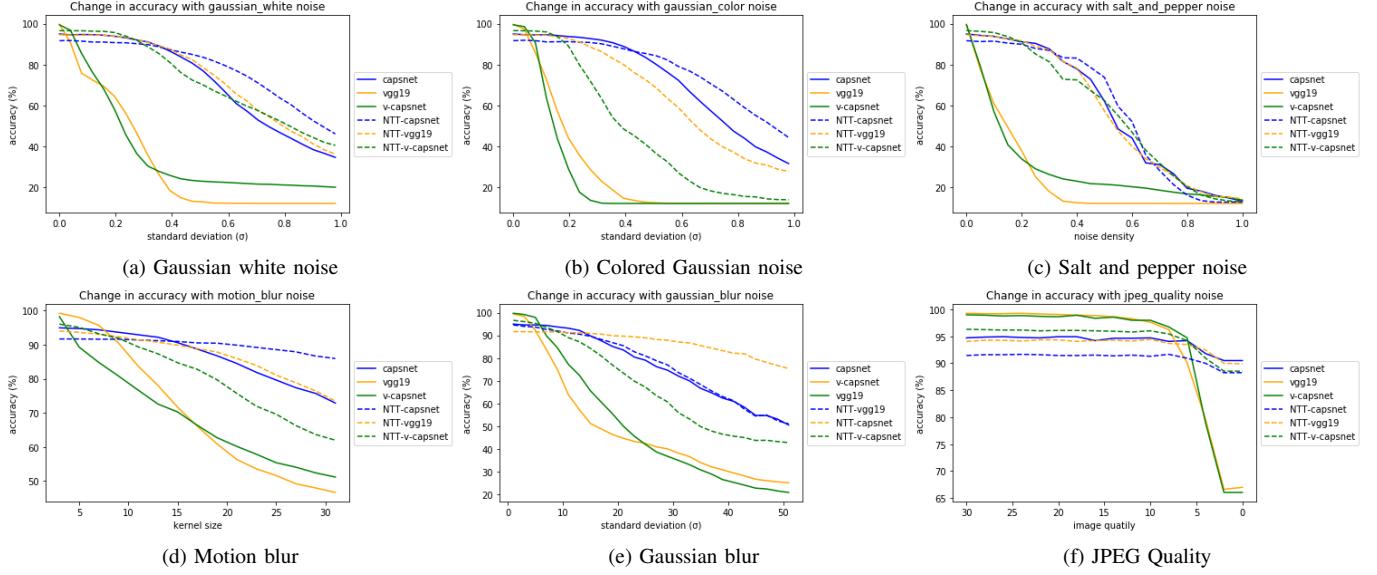
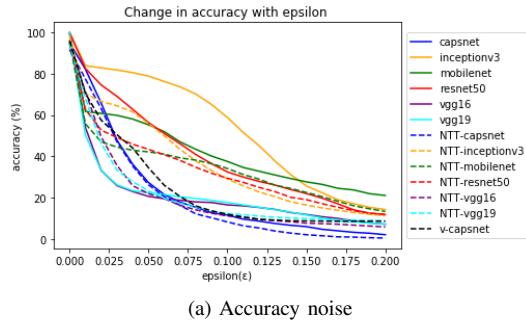
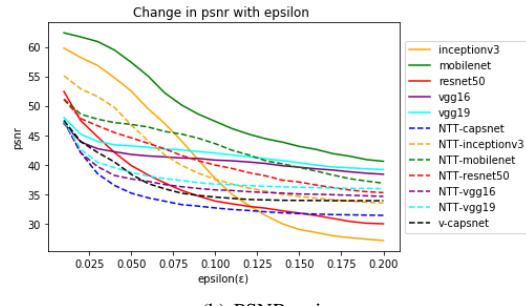


Fig. 11. Performance comparison of modified V-Capsnet with V-CapsNet, VGG19 and CapsuleNet.



(a) Accuracy noise



(b) PSNR noise

Fig. 12. Comparison of classification accuracies of different architectures under FGSM attack using natural images dataset.

of NTT layer in most of the architectures. That means at lower  $\epsilon$ , where the adversarial changes are difficult to detect, presence of NTT layer degrades the quality, and the adversarial noise might become perceptible to the viewers.

As the performances of the architectures under adversarial attack do not have any coherence with the performances under perceptible degradations, like noise, blur, compression etc., it is evident that individual analysis of the models under different degradations is necessary to understand the overall robustness of an architecture.

TABLE II  
 FILTER SIZE USED IN NTT LAYER

Model	#Filter Size	Input Size
MobileNet	(21 × 21)	(224 × 224 × 3)
VGG16	(21 × 21)	(224 × 224 × 3)
VGG19	(23 × 23)	(256 × 256 × 3)
ResNet50	(21 × 21)	(224 × 224 × 3)
Inception v3	(23 × 23)	(299 × 299 × 3)
CapsuleNet	(7 × 7)	(104 × 104 × 3)
V-CapsNet	(23 × 23)	(256 × 256 × 3)

#### IV. CONCLUSION

In this paper we demonstrate the effects of different image degradations on CNN models for image classification task. It is evident that all the CNN architectures are susceptible to image degradations. It is interesting to observe that some shallower models like VGGs that achieve less accuracy in many classification tasks are more resilient to degradations. It is also important to notice that the conventional capsule architecture is remarkably robust against several image degradations, particularly against salt and pepper noise and blurring. We proposed a novel capsule network based architecture that achieves highest accuracy in classification task among the six CNN architectures considered here. We also observe that having small number of convolution layers, basic CapsuleNet architecture is resilient to degradation and simply going deeper in the architecture in conventional way will affect the robustness. Thus, it is important to design CNN models that can achieve high accuracy without increasing the depth of the network. We also observe that it is difficult to comment about the robustness of CNN architectures under adversarial attacks by only seeing their performances under perceptible image degradations. Though, we failed to achieve very high accuracy for recognition task using CapsuleNet where the dataset contains very large number of classes like ImageNet, CapsuleNet has shown promising results in all the cases. It

will be important to come up with new networks in future that can provide robustness against image degradations maintaining high accuracy.

## V. ACKNOWLEDGEMENT

We would like to thank Nvidia® for providing a Titan X GPU to our research group.

## REFERENCES

- [1] A. Mittal, A. K. Moorthy, and A. C. Bovik, “No-reference image quality assessment in the spatial domain,” *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.
- [2] S. Dodge and L. Karam, “Understanding how image quality affects deep neural networks,” in *IEEE Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, 2016, 2016, pp. 1–6.
- [3] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” *arXiv preprint arXiv:1704.04861*, 2017.
- [4] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [6] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “Deepfool: a simple and accurate method to fool deep neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2574–2582.
- [7] C.-X. Ren, D.-Q. Dai, and H. Yan, “Coupled kernel embedding for low-resolution face image recognition,” *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3770–3783, 2012.
- [8] W. W. Zou and P. C. Yuen, “Very low resolution face recognition problem,” *IEEE Transactions on Image Processing*, vol. 21, no. 1, pp. 327–340, 2012.
- [9] S. Basu, M. Karki, S. Ganguly, R. DiBiano, S. Mukhopadhyay, S. Gayaka, R. Kannan, and R. Nemani, “Learning sparse feature representations using probabilistic quadtrees and deep belief nets,” *Neural Processing Letters*, vol. 45, no. 3, pp. 855–867, 2017.
- [10] L. J. Karam and T. Zhu, “Quality labeled faces in the wild (qlfw): a database for studying face recognition in real-world environments,” in *Human Vision and Electronic Imaging XX*, vol. 9394. International Society for Optics and Photonics, 2015, p. 93940B.
- [11] J. Tao, W. Hu, and S. Wen, “Multi-source adaptation joint kernel sparse representation for visual classification,” *Neural Networks*, vol. 76, pp. 135–151, 2016.
- [12] S. Ullman, L. Assif, E. Fetaya, and D. Harari, “Atoms of recognition in human and computer vision,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 10, pp. 2744–2749, 2016.
- [13] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.
- [15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.
- [16] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic routing between capsules,” in *Advances in Neural Information Processing Systems*, 2017, pp. 3859–3869.
- [17] J. Rauber, W. Brendel, and M. Bethge, “ Foolbox: A python toolbox to benchmark the robustness of machine learning models,” *arXiv preprint arXiv:1707.04131*, 2017. [Online]. Available: <http://arxiv.org/abs/1707.04131>
- [18] A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” *arXiv preprint arXiv:1607.02533*, 2016.