Group-Wise Dynamic Dropout Based on Latent Semantic Variations

Zhiwei Ke,[1,2*] **Zhiwei Wen,**[1*] **Weicheng Xie,**[1,4†] **Yi Wang,**[2†] **Linlin Shen**[1,3,4]
[1]Computer Vision Institute, Shenzhen University, Shenzhen, China
[2]Dongguan University of Technology, Dongguan, China
[3]Shenzhen Institute of Artificial Intelligence and Robotics for Society, PR China
[4]Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, PR China
{kezhiwei2018, wenzhiwei2018}@email.szu.edu.cn, wangyi@dgut.edu.cn, {wcxie, llshen}@szu.edu.cn

Abstract

Dropout regularization has been widely used in various deep neural networks to combat overfitting. It works by training a network to be more robust on information-degraded data points for better generalization. Conventional dropout and variants are often applied to individual hidden units in a layer to break up co-adaptations of feature detectors. In this paper, we propose an adaptive dropout to reduce the co-adaptations in a group-wise manner by coarse semantic information to improve feature discriminability. In particular, we showed that adjusting the dropout probability based on local feature densities can not only improve the classification performance significantly but also enhance the network robustness against adversarial examples in some cases. The proposed approach was evaluated in comparison with the baseline and several state-of-the-art adaptive dropouts over four public datasets of Fashion-MNIST, CIFAR-10, CIFAR-100 and SVHN.

Introduction

Dropout is a stochastic regularization technique commonly used in deep neural networks (DNN) (Hinton et al. 2012). Conventionally, it works in fully-connected (FC) layers by randomly "dropping" out the activation of a neuron with a certain probability p for each training case. The process has the effect of model averaging by simulating a large number of networks with different network structures, which, in turn, making node activations in the network more robust to the inputs.

Inspired by the original dropout, other stochastic model averaging methods were proposed to simulate dynamic sparsity within the network. For example, *spatial dropout* (Tompson et al. 2015) removes the feature map activations in a convolution layer to account for strong spatial correlation of nearby pixels in natural images. In (Wan et al. 2013), *drop connection* transfers the FC layer into a sparsely connected layer in which the connections specified in weight matrices are chosen at random during the training. Previous studies showed that these dropout-inspired regularization often outperformed the original dropout on several visual recognition datasets. We use both the original dropout and spatial dropout as baselines in our experiments.

In conventional dropouts, every hidden unit is treated the same and independently with a constant dropout probability p. Following *Binomial* distribution, the expected number of dropped units in a layer of n units is $n \cdot p$ despite of different layers and samples. This causes deficiency of dropouts (Wang, Zhou, and Bilmes 2019). To improve performance, dropout variants were explored in mainly two aspects (Keshari, Singh, and Vatsa 2019): 1) sampling dropout masks from different distributions other than *Bernoulli*, and 2) adapting the dropout probability. In particular, it was shown that the generalization ability can be improved by dropping nodes *selectively* based on some prior knowledge of the network. For instance, (Keshari, Singh, and Vatsa 2019) learns a strength parameter by stochastic gradient descent (SGD) of the network for guiding dropout regularization of each node. (Wang, Zhou, and Bilmes 2019) adapts the dropout probability by normalizing it at each layer and every training batch such that the effective dropping rate on those activated units is kept the same during the training.

Most of existing dropout regularization methods remove individual activations in each unit independently with a fixed or adaptive probability. We note that for object recognition, visual structures in the input image activate the corresponding regions in the convolution feature maps (He et al. 2015). In other words, feature maps with similar activation patterns tend to have close semantic implications. Intuitively, these feature maps should be stochastically dropped to reduce co-adaptations. However, they also encode information about the *intra-class variation* of latent semantic features (Kim et al. 2017). This motivates us to propose a group-wise dropout that can adapt to the latent semantic variations while simulating dynamic sparseness in the network to improve the object recognition performance.

In this paper, we make the following contributions:

- We propose to represent latent semantic variations with densities of linearly uncorrelated features from the convolution layer. This is done by applying PCA projections to vectorized feature maps and then bin gridding them

[*]Both Z. Ke and Z. Wen contribute equally to this work.

[†]Corresponding authors: W. Xie (OrcID: 0000-0001-8946-7472) and Y. Wang (OrcID: 0000-0002-8448-8570)

in the projected space. In this way, we are able to *group* feature maps with close semantic implications. We show that, after deconvolution, the groups with higher densities demonstrate larger visual variations in the local layouts.

- We propose to assign a self-adaptive dropout probability to *all* entities within the grouped features, i.e. the dropout probability is monotonically decreasing with the estimated density distribution. This is to retain sufficient sampling of interacted layouts for learning more robust features that can account for relatively large intra-class variations.
- We showed with extensive experimental results over four public benchmarks that the proposed group-wise dynamic scheme is able to significantly improve both the object recognition accuracy and the network robustness against white-box adversarial examples bounded by L_2 and L_∞ norms, respectively.

Related Work

In the literature, dropout variants were proposed by studying the convolution feature maps. In *spatial dropout* (Tompson et al. 2015), a random subset of activations in the feature maps are dropped independently to reduce spatial correlations. (Poernomo and Kang 2018) proposed *crossmap dropout* to simultaneously drop or retain elements at the same coordinate on different feature maps. (Zhang, Yang, and Feng 2018) proposed *region dropout* by considering several salient regions with fixed size and relative positions for training. However, these salient regions are not always available for general object recognition problems.

For adaptive dropouts, (Wang and Manning 2013) showed that dropout has a Gaussian approximation and (Kingma, Salimans, and Welling 2015) proposed a *variational dropout* by connecting the global uncertainty with the dropout rates to optimize a generalized Gaussian dropout. (Wager, Wang, and Liang 2013) analyzed the dropout training as a form of adaptive regularization with an approximation of second-order derivative. (Ba and Frey 2013) updated the element-wise probability for the mask matrix generation according to activation output of an overlaid binary belief network. (Zhuo, Zhu, and Zhang 2015) extended the overlaid model so that adaptive dropout rates can be learned for different neurons or group of neurons. However, the side-model based approaches introduce significant computation and memory overhead. (Keshari, Singh, and Vatsa 2019) proposed the *guided dropout* to drop network nodes with high strength to encourage low strength nodes. (Wang, Zhou, and Bilmes 2019) proposed to *Jumpout* samples the dropout probability from a monotone decreasing distribution (e.g., the right half of a Gaussian) such that each linear piece of the network can learn better for data points from nearby than more distant regions to improve generalization of DNNs with ReLU activations. These methods drop the unit activations independently, whereas there are feature visualization studies have shown that the interactive information between feature nodes can be useful for improving the performance of object recognition (Kim et al. 2017; Du et al. 2018).

To the best of our knowledge, most of the existing dropout regularizations were proposed to improve the generalization performance on natural image examples. However, DNNs are also known to be vulnerable to *adversarial examples* that are carefully crafted to cause intended misclassifications with high probability (Goodfellow, Shlens, and Szegedy 2014). The security threat is known as *evasion attacks* at test time (Biggio et al. 2013; Yuan et al. 2019). An increasing number of work is proposed for hardening DNNs to make them more robust against adversarial examples, e.g., by adversarial training (Tramèr et al. 2018) or gradient obfuscation (Akhtar, Liu, and Mian 2018). In particular, *stochastic gradients* are found useful to prevent the attacker from getting the critical information of loss gradients from the target model in generating adversarial examples (Akhtar, Liu, and Mian 2018). For instance, (Feinman et al. 2017) proposed to turn on dropout randomization as well during the test-time for adversarial detection by analyzing uncertainty of the network. (Dhillon et al. 2018) proposed Stochastic Activation Pruning (SAP) to enhance the adversarial robustness of DNNs. During the forward pass, SAP stochastically prunes a subset of the activations in each layer with preference of retaining activations with larger magnitudes (i.e., strengths of the nodes). The surviving activations are then scaled up to normalize the dynamic range of the inputs to the subsequent layer.

Table 1 summarizes some key differences between the proposed approach and some representative dropouts that will be compared in our evaluations. In particular, spatial dropout is applied to individual feature maps such that adjacent pixels in the dropped-out feature map are either all dropped-out or all active. In contrast, the proposed dropout dynamically groups the FC features or convolutional feature maps to assign a self-adaptive dropout probability. While guided dropout screens the neurons for dropout and SAP adjusts the dropout probabilities according to the magnitudes of FC features, the proposed dropout negatively correlates the dropout probability with the *feature density* distribution in linear uncorrelated subspaces after PCA projection. In the following sections, we shall show that the proposed strategy of dynamic dropout can help to improve both classification accuracy and adversarial robustness of DNNs.

Proposed Approach

In the original dropout, the same dropout probability is shared among features, i.e., FC neurons or convolution feature maps, in a network layer. Recent studies showed that DNN-learned features are not acting individually by themselves but rather having interactions in such a way that together they contribute to training a discriminative network (Du et al. 2018). This is consistent with the interpretability study in (Bau et al. 2017), which indicates that there exists a special basis that aligns explanatory factors with individual units in a hidden layer. In other words, interpretability of unit interaction is *not* equivalent to random linear combinations of units. We are inspired by these insights of deep visual representations to propose a group-wise dynamic dropout as follows.

Table 1: Comparison of the proposed and the relevant dropout variants. Dropout variants of Original (Hinton et al. 2012), Biased (Poernomo and Kang 2018),Crossmap (Poernomo and Kang 2018), Spatial (Tompson et al. 2015), Guided (Keshari, Singh, and Vatsa 2019) and Stochastic Activation Pruning (SAP) (Dhillon et al. 2018) are used for the comparison. 'Adaptivity=Yes' means the dropout probability is *not* fixed. 'Dynamicity=Yes' means the dropout is employed during training.

Variants	Dropout strategies	Adaptivity	Dynamicity	Hyper-parameter	Time complexity
Original	Random	No	Yes	Dropout probability	-
Biased	Activation magnitudes	Yes	Yes	Two dropout probabilities	Two-group division
Crossmap	Across feature map	No	Yes	Dropout probability	Close to Original
Spatial	Entire feature map	No	Yes	Dropout probability	Close to Original
Guided	Active region	Yes	Yes	Multi-stage dropout probabilities	Bins computation
SAP	Neuron magnitudes	Yes	No	Layers for dropout	Distribution sampling
Proposed	**Feature density**	**Yes**	**Yes**	**Negative correlation function**	**PCA projection**

Feature Density Estimation

We consider that some visual structures in the input image are inherently more difficult to recognize than others due to intra-class variations. Accordingly, we introduce the concept of *feature density* by analysing the number of linearly uncorrelated deep features gathered over equally spaced grids in low-dimension feature space.

Algorithm 1 outlines the main steps of our feature density estimation. In particular, the PCA projection not only performs dimension reduction of feature maps but also helps to decorrelate deep features for density estimation. Taking FDD2D for example, the first two principal components are used to construct the 2D projection features as illustrated in Figure 1 (a). The PCA is conducted on the entire 512 feature maps of the last convolution layer for each of the training samples. Let ρ_i be the number of FC neurons or the feature maps located in the i-th grid, i.e. $GRID_i, i = 1, \cdots, N^k$, where N denotes the number of the equally divided grids in each dimension, k is the dimension of PCA projection space. To obtain the feature sets $\{GRID_i\}$, the grid that the j-th feature $f^{(j)}$ located is obtained as $FLOOR((f_r^{(j)} - min(f_r))/d_r) + 1$ corresponding in the r-th dimension ($1 \leq r \leq k$), where $d_r = (max(f_r) - min(f_r))/N$ is the space between grids and f_r is the vector of the r-th column of f when $k \geq 2$.

Figure 2 illustrates the semantic implication of deep features using a toy model of 5-layer CNN with 512 neurons in the last FC layer. We group the deep features over equally spaced grids into three densitiy levels, namely `low`, `medium` and `high`, and invert them back to the input space by deconvolution. Figure 2(b) shows three typical inverted representations from each group. The corresponding region of interest (ROI) of each inverted representation is marked with a heat map and overlaid on the original image.

It can be seen that the group of `higher` density features contains more dispersed semantic information in terms of ROI, whereas those grouped in `low` density are more concentrated in the local layout. This can be seen more clearly in Figure 3 by averaging the heat maps in Figure 2. To quantify the dispersion, we calculate the average entropy of each grouped heatmap which is 2.045 and 3.616 for the `low` and `high` density group in Figure 3.

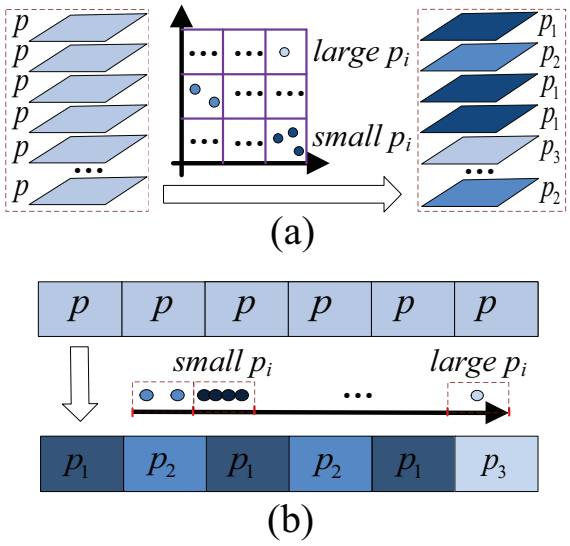


Figure 1: Feature Density-based Dropout (FDD) for (a) convolution feature maps, and (b) FC neurons. The dropout probability p_i is adapted to feature density estimation in the i-th grid, where a smaller p_i is assigned to the feature group with higher density in the PCA-reduced space.

Algorithm 1 Feature Density Estimation

while Not Termination Condition **do**
  if Dropout for Feature Map **then**
    Vectorize the feature maps in the last convolution layer X with the size of $(n_{channel}, n_{width}, n_{height})$ to the size of $(n_{channel}, n_{width} * n_{height})$;
    Set f as the PCA projection of vectorized X, where the j-th feature $f^{(j)} \in k$-D (k=2 or 3);
  else if Dropout for FC **then**
    Set f as the FC features x with the size of $(n_{FC}, 1)$, where $f^{(j)} \in k$-D (k=1);
  end if
  Count the number of features ($f^{(j)}$) located over equally-spaced grids for density estimation;
end while

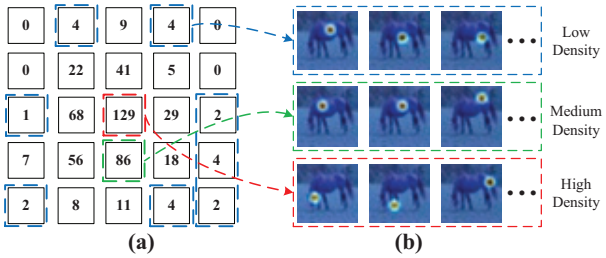


Figure 2: Inverting representations of grouped features: (a) Density estimation over grids; and (b) Heat maps after deconvolution showing regions of interest by the feature maps. Inverted examples are displayed with top three responses of activations in the color-marked groups with low, medium, and high densities, respectively.

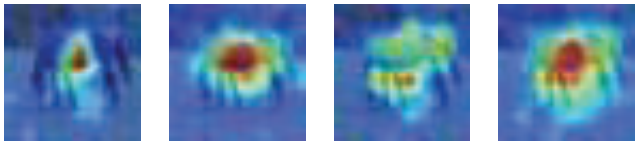


Figure 3: Latent semantic feature variations by averaging heat maps inverted from the typical feature maps of (a) low density, (b) medium density, (c) high density, and (d) all.

Feature Density-based Dropout (FDD)

Both visual and quantitative results in Figure 2 show that feature density is negatively correlated to latent semantic variations. Intuitively, it requires more sampling of "facets" to learn robust representations for `high`-density features with more dispersed semantic information. This motivates us to proposed a feature density-based dropout (FDD) scheme by increasing the activation outputs, i.e. hidden FC neurons or convolution feature maps.

We use the approximate feature densities ρ_i estimated from Algorithm 1 to update the dropout probabilities for the features of FC layer and feature maps in a self adative way as follows

$$p_i = \kappa \frac{\rho_i}{max(\rho_i)} + \alpha. \tag{1}$$

where p_i is the dropout probability for the i-th grid, i.e. $GRID_i$. The negative correlation, i.e. the feature density and dropout probability are negatively linear mapped with $\kappa = -0.8, \alpha = 0.9$ to generate the dropout probabilities in Figure 1.

Feature map dropout. Based on the updated dropout probabilities $\{p_i\}$, the proposed dropout is performed on the j-th FC feature or the j-th convolutional feature map in $GRID_i (1 \le i \le N^k)$ as follows

$$X'^{(j)} \leftarrow m * X^{(j)}, \tag{2}$$

where $m \sim Bernoulli(p_i)$ is generated with probability p_i and multiplied with the j-th feature map of X, denoted by $X^{(j)}$.

FC feature dropout. The proposed dropout is performed on the j-th FC feature in $GRID_i (1 \le i \le N^k)$ as

$$x'^{(j)} \leftarrow m \cdot x^{(j)}, \tag{3}$$

where $x'^{(j)}$ denotes the j-th feature value of x'. In this way, the proposed dropout dynamically updates multiple dropout probabilities, i.e. $\{p_1, \cdots, p_{N^k}\}$, based on feature density. After the dropout, the cross entropy softmax function is applied to the retaining FC features or feature maps, i.e., the updated x' or X' in (2) and (3), respectively. The discrimination loss function is thus updated with

$$\mathscr{L} = -ln \frac{e^{W_y^T \eta + b_y}}{\sum_j e^{W_j^T \eta + b_j}}. \tag{4}$$

where y is the ground truth label of the sample, W_j is the j-th column of the weight matrix W; η is x' or $AvgPool(X')$, where $AvgPool$ is the average pooling operation. The derivatives of the loss $\mathscr{L}$ w.r.t. the feature η and weights W, i.e. $\partial\mathscr{L}/\partial\eta, \partial\mathscr{L}/\partial W$ are automatically calculated with network back propagation.

Computation complexity. For feature map projection, PCA is employed to reduce feature dimension based on the vectorization matrix with $n_{channel}$-rows and $n_{width} \cdot n_{height}$ columns of the feature map tensor X; PCA finally yields the set of features projected in k-D space, i.e. f used in Algorithm 1. The time complexity of PCA is $O(n_{channel} \cdot D) + O(D^3)$, where D is the dimension product of each feature map. The time complexity of the last convolution operator is at least $O(D \cdot n_{channel} \cdot n_{ConvChannel})$, where $n_{ConvChannel} > n_{channel}$ is the number of feature maps at the last-but-one convolution layer. In practice, the runtime cost of PCA over GPU for all the training samples is only 0.74 seconds in each epoch, which is negligible to the average time taken for training the network. Comparing to SAP (Dhillon et al. 2018) with requires an additional distribution sampling, the proposed dropout also needs significant less run-time cost. The computing time is competitive to the guided dropout based on our local runs.

Experimental Results

We run our experiments with ResNet-18 (He et al. 2016) of 512 neurons in the last FC layer on a 4-kernel Nvidia TITAN GPU Card. The learning rate is updated with cosine annealing and a SGD optimizer is used. All our experiments include batch normalization when training ResNet18. The batch size and learning rate are 64 and 0.01, respectively. As described in Algorithm 1, we perform the proposed dropout regularization of FDD-1D on FC features, FDD-2D and FDD-3D on convolution feature maps, respectively. The projected features are partitioned into 25, $5^2 = 25$ and $3^3 = 27$ equally-spaced grids in the 1D, 2D and 3D projection space, respectively.

We evaluated the proposed dropouts over four public benchmarks. The Fashion-MNIST (FM) (Xiao, Rasul, and Vollgraf 2017) is a dataset of Zalando's article images consisting of 60k training samples and 10k testing samples. Each example is a 28x28 grayscale image, associated with

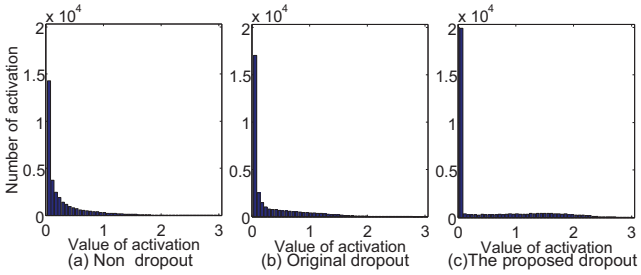


Figure 4: Histogram of FC activations after the training with (a) non-dropout, (b) random dropout, and (c) the proposed group-wise dynamic dropout on Fashion-MNIST.

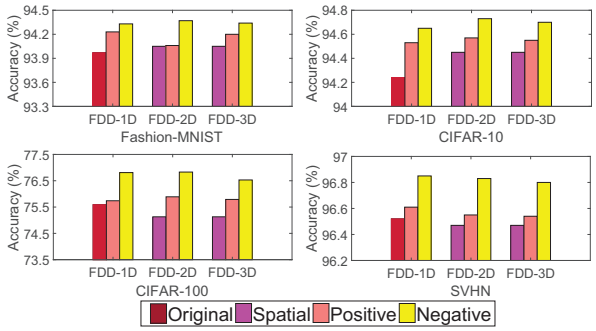


Figure 5: Recognition accuracy by installing positive or negative correlations with feature density in generating the FDD probability. The proposed dropouts with negative correlations consistently outperform the others over all four datasets.

a label from 10 classes. The CIFAR-10 (C10) (Krizhevsky and Hinton 2009) dataset contains 60k color images belonging to 10 classes (32x32 resolution). The experiments utilize 50k training samples and 10k testing samples. CIFAR-100 (C100) (Krizhevsky and Hinton 2009) is a database with 100 classes and also has 50k training samples and 10k testing samples. The Street View House Numbers (SVHN) (Netzer et al. 2011) dataset contains 73,257 training samples and 26,032 testing samples.

Dropout Analysis

Activation sparseness. An important function of dropouts is to simulate dynamic sparseness in the network to reduce co-adaptations of feature encoding for better generalization. Figure 4 plots histograms of FC feature activations before and after applying dropouts for the training. Comparing with the original dropout, the proposed dropout is able to significantly increase the number of de-activations or close-to-zero activations while dampening all other activation values. This suggests that our method works better than random dropout in generating sparseness.

Positive or negative correlations. Setting positive correlations in generating the FDD probability with feature density helps to reduce more co-adaptations of feature encoding. However, it is also at a cost of breaking up meaningful inter-

Table 2: Comparing average recognition rates (%) of adaptive dropout variants and the proposed dropout on the four datasets. Results of the dropout variants are acquired with our local runs, followed by their standard variances (%). The best performances are marked in **bold**.

Methods		FM	C10	C100	SVHN
Dropout variants	Non-dropout	94.05 0.07	94.41 0.04	75.3 0.22	96.51 0.03
	Original	93.97 0.02	94.24 0.2	75.59 0.1	96.52 0.09
	Biased	94 0.1	94.45 0.07	75.61 0.13	96.53 0.06
	Spatial	94.05 0.08	94.45 0.09	75.13 0.02	96.47 0.13
	Crossmap	94 0.05	94.4 0.06	75.45 0.14	96.35 0.07
Proposed	FDD-1D	94.33† 0.04	94.65† 0.08	76.81† 0.07	**96.85**† 0.03
	FDD-2D	**94.37**† 0.05	**94.73**† 0.02	**76.83**† 0.11	96.83† 0.03
	FDD-3D	94.34† 0.04	94.7† 0.06	76.53† 0.07	96.8† 0.01

† denotes that the proposed dropout significantly outperforms the original dropout under the significance level of 0.05 with Student's t-Test (De Winter 2013).

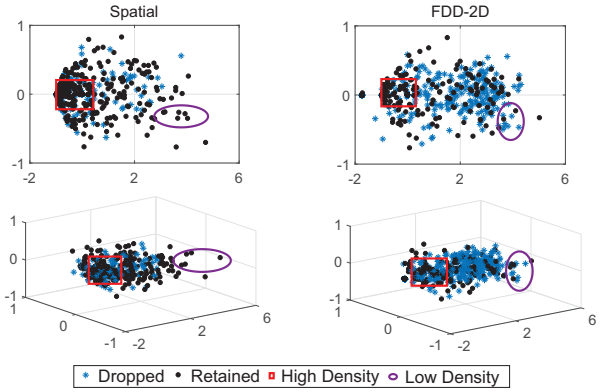


Figure 6: Visualization of 2D- and 3D- PCA projections of the 512 feature maps after applying spatial dropout (Tompson et al. 2015) and the proposed FDD-2D regularization to ResNet-18 for CIFAR-100. FDD-2D is able to drop more features in the low-density regions over all projection space, adapting to the underlying semantic variations while still simulating abundant sparseness for improving performance.

actions between nodes and causing insufficient sampling of "facets" to account for diversity. To verify the hypothesis, we conducted experiments by setting $\kappa = 0.8, \alpha = 0.1$ in Eq. (1) for installing both positive and negative correlations in the FDDs. The results are shown in Figure 5. It can be seen that FDDs with positive correlations do not have significant improvement over baseline dropouts, whereas FDDs with negative correlations consistently outperform the others. Thus, the results support our view of deep feature density.

Comparison of Dropout Variants

To analyze the differences of the proposed dropout, i.e. FDD-2D and the baseline of spatial dropout, the PCA projection of dropout feature maps with the spatial and the proposed dropouts are presented in Figure 6. Figure 6 shows that the numbers of the dropped features with the proposed dropout are more uniform than the spatial dropout (Tompson et al. 2015) in both the high and low density regions. More precisely, the feature maps projected in the ovals, i.e. the low density regions which are representative for relatively concentrated features, are apt to be removed as redundancy to improve the generalization ability. While the feature maps projected in the rectangles, i.e. the high density regions which are representative for variant features, are apt to be retained to enhance the representative ability for features with large semantic variance.

To compare the overall performance of the proposed dropout with the relevant dropout variants, the average performances and their standard variances with different dropout variants on the four datasets are shown in Table 2, where multiple independent experiments are conducted to obtain each average performance. Notations of 'FDD-1D', 'FDD-2D' and 'FDD-3D' are the abbreviations for the proposed feature density-based dropouts with 1D feature neurons and 2D/3D-PCA projections of vectorized feature maps, respectively. Dropout variants of 'Original', 'Biased', 'Spatial' and 'Crossmap' are abbreviated in Table 1.

Regarding to the dropout of the FC features, the performances of 'Original' and 'FDD-1D' in Table 2 show the proposed dropout significantly outperforms the original dropout on the four datsets under the significance level of 0.05 with Student's t-Test (De Winter 2013), where a large improvement of 1.7% is achieved on the CIFAR-100 dataset. Regarding of the dropout of the feature maps, the obtained performances of 'Spatial', 'FDD-2D' and 'FDD-3D' show that the information of negative correlation between dropout probability and feature density is useful to dynamically adapt the dropout probabilities for different datasets. Meanwhile, the proposed FDD-2D balances the performances on the four datasets.

Generating Adversarial Examples

Recent studies show that high-performance classifiers could be still fooled by adversarial examples. The adversarial robustness of a model is evaluated with the recognition performance against each generated adversarial example $\hat{I}$ of its normal counterpart I. In this work, three commonly used attack methods are employed for the robustness evaluation:

Fast Gradient Sign Method (FGSM) (Goodfellow, Shlens, and Szegedy 2014) introduced an efficient one-step attack. This method uses the gradient of the training cost function to determine the direction of the perturbation. The adversarial examples can be generated by $\hat{I} = I + \varepsilon \cdot sign(\nabla_I \mathscr{L}(I,y))$, where ε is the perturbation intensity and $\mathscr{L}(I,y)$ is the training loss function with input I and label y.

Basic Iterative Method (BIM) (Kurakin, Goodfellow, and Bengio 2017) introduced an extension of FGSM, which applied multi-step perturbations and clipped the perturbed values with constrained bounding.

Projected Gradient Descent (PGD) (Madry et al. 2018) proposed an improved adversary method, which is the multi-step variant of FGSM. A small random perturbation is added to the original input data, whose iterative generation is further projected into the valid space for the next step iteration to improve the success attack rate.

We generated adversarial examples using the above methods assuming the scenario of *white-box* attacks (Yuan et al. 2019). That is, the adversary knows everything related to the target network, including model architectures, hyperparameters, activation functions, model weights, etc. In this case, adversarial examples are generated to attack each "white-box" model in our experiments, respectively.

Evaluations of Adversarial Robustness

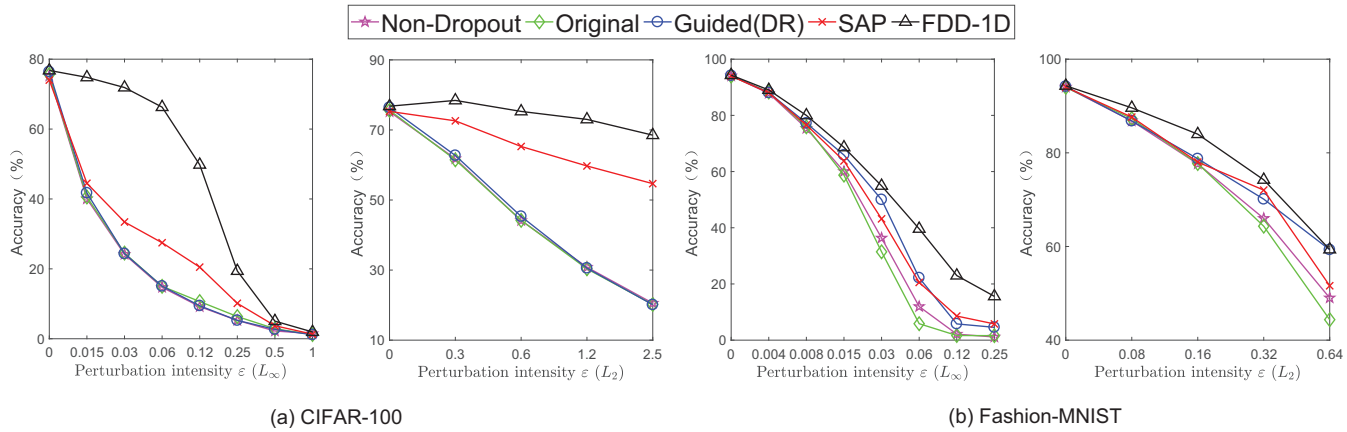
We evaluated the proposed adaptive dropout on improving the DNN robustness against adversarial examples generated by the "white-box" settings as described in the previous section. The evaluation experiments are performed on modified ResNet-18 with different dropout regularizations over Fashion-MNIST and CIFAR-100. Only the input image examples that can be correctly recognized in the original datasets are used to generate adversarial examples. In particular, we normalized the image intensity from Fashion-MNIST with 0-1 normalization and that from CIFAR-100 with z-score normalization for image processing. Taking FGSM for example, in L_∞ norm, the normalized value of perturbation intensity $\varepsilon = 0.03, 0.06, 0.12$ corresponds to changing $8, 16, 32$ pixels for the Fashion-MNIST images while $\varepsilon = 0.015, 0.03, 0.06$ corresponds to changing $1, 2, 4$ pixels for the CIFAR-100 images, respectively. The iteration step in BIM and PGD is set to 10 with a step size of $\varepsilon/10$.

Table 3 reports the results by comparing the proposed approach with non-dropout and the baselines under adversarial perturbations bounded with typical perturbation intensity value ε in L_∞ norm. It can be seen that recognition accuracies of all testing methods drop even with small perturbations. The original random dropout is particularly sensitive to adversarial perturbations, while the proposed FDD modifications are able to achieve significantly improvement over the baseline approaches in all cases. Take our FDD-1D on the CIFAR-100 dataset as an example, it greatly outperforms the original dropout with improvements of 51.28%, 53.61% and 50.71% against the adversarial attacks of FGSM, BIM and PGD with normalized perturbation intensities of $\varepsilon = 0.06$, $\varepsilon = 0.03$ and $\varepsilon = 0.03$, respectively. In general, the gain is 10-50% by the proposed scheme comparing with batch normalization only and other dropout variants. The significant better performance achieved by the proposed dropout illustrate the robustness of feature density for pixel-wise adversarial attacks.

Figure 7 plots adversarial robustness of the network by employing different dropout regularization schemes under an increasing perturbation intensity value of ε in L_∞ and L_2 norms for the two datasets, respectively. In particular, Figure 7 shows that the network performance by employ-

Table 3: Recognition accuracy (%) under normalized perturbation intensity, denoted by ε, in L_∞-norm against adversarial examples generated by the FGSM, BIM and PGD attacks on the Fashion-MNIST and CIFAR-100 images, respectively. FDD-1D and FDD-2D/3D are the proposed dropout with 1D feature neurons and 2D/3D-PCA projections of vectorized feature maps. The top three performances are marked in **bold**.

Datasets	Attacks	Intensity	Non-dropout	Original	**FDD-1D**	Spatial	**FDD-2D**	**FDD-3D**
Fashion-MNIST	FGSM	$\varepsilon = 0.03$	36.35	31.43	**54.92**	38.1	**47.39**	**50.87**
		$\varepsilon = 0.06$	11.95	5.85	**39.62**	9.93	**28.38**	**35.96**
		$\varepsilon = 0.12$	2.13	1.65	**22.98**	2.5	**14.22**	**22.41**
	BIM	$\varepsilon = 0.02$	34.13	30.02	**46.35**	39.08	**40.63**	**43.92**
		$\varepsilon = 0.03$	7.56	2.11	**25.47**	9.08	**16.7**	**20.77**
	PGD	$\varepsilon = 0.02$	34.13	30.02	**46.35**	39.08	**40.63**	**43.92**
		$\varepsilon = 0.03$	8.98	2.54	**27.05**	11.15	**17.34**	**21.39**
CIFAR-100	FGSM	$\varepsilon = 0.015$	40.02	40.56	**74.8**	40.34	**70.9**	**65.65**
		$\varepsilon = 0.03$	24.16	24.51	**71.89**	23.76	**67.1**	**60.04**
		$\varepsilon = 0.06$	14.72	15	**66.28**	14.15	**61.98**	**53.63**
	BIM	$\varepsilon = 0.015$	23.47	24.29	**65.63**	24.49	**59.32**	**50.09**
		$\varepsilon = 0.03$	4.32	5.89	**59.5**	4.72	**49.45**	**36.64**
	PGD	$\varepsilon = 0.015$	33.95	34.09	**67.18**	35.28	**61.68**	**53.99**
		$\varepsilon = 0.03$	7.53	9.25	**59.96**	8.49	**51.01**	**39.44**



(a) CIFAR-100 (b) Fashion-MNIST

Figure 7: Recognition accuracy (%) comparing the network with non-dropout and four dropout variants including the original baseline against the FGSM attack with an increasing the value of normalized perturbation intensity ε in L_∞ and L_2 norm over CIFAR-100 and Fashion-MNIST, respectively. The robustness of all the methods is tested with our local runs.

ing our FDD-1D regularization is significantly more salient to small FGSM-based adversarial perturbations, up till $\varepsilon = 0.06$ in L_∞ norm and $\varepsilon = 2.5$ in L_2 norm, that were applied to CIFAR-100 images. Whereas for all other methods, the recognition accuracy declines quickly by increasing the perturbation intensity in generating adversarial examples. On the Fashion-MNIST, the proposed dropout scheme still outperforms the other comparing variants by improving the network robustness against adversarial examples with larger perturbation intensities.

Conclusion and Future Work

In this paper, we propose an adaptive dropout variant that can simultaneously improve the network performance in terms of the accuracy of objection recognition as well as the robustness against adversarial examples generated by adding small perturbations to the original image intensities. The proposed regularization scheme works by adapting the dropout probability to latent semantic variations of deep representations while simulating dynamic sparseness within the network. This is motivated by the observations and insights that DNN-learned features are not acting individually by themselves but rather having interactions in specific combinations that together contribute to the discriminative function. Accordingly, we propose to encourage diversity of feature representations for the inherently more "difficult" local structures. This is done by negatively correlating the dropout probability with feature density estimated in linearly uncorrelated subspaces of the deep features. Our empirical analysis and experimental results support our hypothesis regarding the dropout. The proposed method outperforms the baseline and other state-of-the-art dropout variants on all four public benchmark datasets. Our future work includes studies of improving the network robustness using dropout-inspired

regularization schemes under more general adversarial perturbation attacks such as C&W (Carlini and Wagner 2017).

Acknowledgement

The work was supported in part by Natural Science Foundation of China under grants no. 61602315, 61672357, 61876038 and U1713214, the Science and Technology Project of Guangdong Province under grant no. 2018A050501014, the Tencent "Rhinoceros Birds"-Scientific Research Foundation for Young Teachers of Shenzhen University, the School Startup Fund of Shenzhen University under grants no. 2018063, and Dongguan University of Technology under Project KCYKYQD2017003.

References

Akhtar, N.; Liu, J.; and Mian, A. 2018. Defense against universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3389–3398.

Ba, J., and Frey, B. 2013. Adaptive dropout for training deep neural networks. In *Advances in Neural Information Processing Systems*, 3084–3092.

Bau, D.; Zhou, B.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6541–6549.

Biggio, B.; Corona, I.; Maiorca, D.; Nelson, B.; Šrndić, N.; Laskov, P.; Giacinto, G.; and Roli, F. 2013. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, 387–402. Springer.

Carlini, N., and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, 39–57. IEEE.

De Winter, J. C. 2013. Using the student's t-test with extremely small sample sizes. *Practical Assessment, Research & Evaluation* 18(10).

Dhillon, G. S.; Azizzadenesheli, K.; Lipton, Z. C.; Bernstein, J.; Kossaifi, J.; Khanna, A.; and Anandkumar, A. 2018. Stochastic activation pruning for robust adversarial defense. In *International Conference on Learning Representations*.

Du, Y.; Yuan, C.; Li, B.; Zhao, L.; Li, Y.; and Hu, W. 2018. Interaction-aware spatio-temporal pyramid attention networks for action classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 373–389.

Feinman, R.; Curtin, R. R.; Shintre, S.; and Gardner, A. B. 2017. Detecting adversarial samples from artifacts. In *International Conference on Machine Learning*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence* 37(9):1904–1916.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hinton, G. E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. R. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*.

Keshari, R.; Singh, R.; and Vatsa, M. 2019. Guided dropout. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 4065–4072.

Kim, S.; Min, D.; Ham, B.; Jeon, S.; Lin, S.; and Sohn, K. 2017. Fcss: Fully convolutional self-similarity for dense semantic correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6560–6569.

Kingma, D. P.; Salimans, T.; and Welling, M. 2015. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems*, 2575–2583.

Krizhevsky, A., and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical report, Citeseer.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2017. Adversarial examples in the physical world. In *International Conference on Learning Representations*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*.

Poernomo, A., and Kang, D.-K. 2018. Biased dropout and crossmap dropout: learning towards effective dropout regularization in convolutional neural network. *Neural Networks* 104:60–67.

Tompson, J.; Goroshin, R.; Jain, A.; LeCun, Y.; and Bregler, C. 2015. Efficient object localization using convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 648–656.

Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; and McDaniel, P. 2018. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*.

Wager, S.; Wang, S.; and Liang, P. S. 2013. Dropout training as adaptive regularization. In *Advances in neural information processing systems*, 351–359.

Wan, L.; Zeiler, M.; Zhang, S.; Le Cun, Y.; and Fergus, R. 2013. Regularization of neural networks using dropconnect. In *International conference on machine learning*, 1058–1066.

Wang, S., and Manning, C. 2013. Fast dropout training. In *international conference on machine learning*, 118–126.

Wang, S.; Zhou, T.; and Bilmes, J. 2019. Jumpout: Improved dropout for deep neural networks with relus. In *International Conference on Machine Learning*, 6668–6676.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*.

Yuan, X.; He, P.; Zhu, Q.; and Li, X. 2019. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*.

Zhang, X.; Yang, Y.; and Feng, J. 2018. Ml-locnet: Improving object localization with multi-view learning network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 240–255.

Zhuo, J.; Zhu, J.; and Zhang, B. 2015. Adaptive dropout rates for learning with corrupted features. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.