A. Adversarial variational information bottleneck

In this section, we extend the idea of Adversarial Neural Pruning to Variational Information Bottleneck (VIB). Variational information bottleneck (Dai et al., 2018) uses information theoretic bound to reduce the redundancy between adjacent layers. Let $p(\boldsymbol{h_i}|\boldsymbol{h_{i-1}})$ define the conditional probability and $I(\boldsymbol{h_i};\boldsymbol{h_{i-1}})$ define the mutual information between hidden layer activations $\boldsymbol{h_i}$ and $\boldsymbol{h_{i-1}}$ for every hidden layer in the network. For every hidden layer $\boldsymbol{h_i}$, we would like to minimize the information bottleneck (Tishby et al., 2000) $I(\boldsymbol{h_i};\boldsymbol{h_{i-1}})$ to remove interlayer redundancy, while simultaneously maximizing the mutual information $I(\boldsymbol{h_i};\boldsymbol{y})$ between $\boldsymbol{h_i}$ and the output $\boldsymbol{y}$ to encourage accurate predictions of adversarial examples. The layer-wise energy $\mathcal{L}_i$ can be written as:

$$\mathcal{L}_i = \beta_i I(\boldsymbol{h_i};\boldsymbol{h_{i-1}}) - I(\boldsymbol{h_i};\boldsymbol{y}) \tag{10}$$

The output layer approximates the true distribution $p(\boldsymbol{y}|\boldsymbol{h_L})$ via some tractable alternative $q(\boldsymbol{y}|\boldsymbol{h_L})$. Using variational bounds, we can invoke the upper bound as:

$$\begin{aligned}\mathcal{L}_i = \beta_i \mathbb{E}_{\boldsymbol{h_{i-1}}\sim p(\boldsymbol{h_{i-1}})}[\mathrm{D_{KL}}[p(\boldsymbol{h_i}|\boldsymbol{h_{i-1}})||q(\boldsymbol{h_i})]] - \\ \mathbb{E}_{\{\boldsymbol{x},\boldsymbol{y}\}\sim D, h\sim p(\boldsymbol{h}|\tilde{\boldsymbol{x}})}[\log q(\boldsymbol{y}|\boldsymbol{h_L})] \geq \mathcal{L}_i\end{aligned} \tag{11}$$

$\mathcal{L}_i$ in Equation 11 is composed of two terms, the first is the KL divergence between $p(\boldsymbol{h_i}|\boldsymbol{h_{i-1}})$ and $q(\boldsymbol{h_i})$, which approximates information extracted by $\boldsymbol{h_i}$ from $\boldsymbol{h_{i-1}}$ and the second term represents constancy with respect to the adversarial data distribution. In order to optimize Equation 11, we can define the parametric form for the distributions $p(\boldsymbol{h_i}|\boldsymbol{h_{i-1}})$ and $q(\boldsymbol{h_i})$ as follow:

$$\begin{aligned}p(\boldsymbol{h_i}|\boldsymbol{h_{i-1}}) &= \mathcal{N}(\boldsymbol{h_i}; f_i(\boldsymbol{h_{i-1}}) \odot \mu_i, \mathrm{diag}[f_i(\boldsymbol{h_{i-1}})^2 \odot \sigma_i^2] \\ q(\boldsymbol{h_i}) &= \mathcal{N}(\boldsymbol{h_i}; 0, \mathrm{diag}[\boldsymbol{\xi_i}])\end{aligned} \tag{12}$$

where $\boldsymbol{\xi}_i$ is an unknown vector of variances that can be learned from data. The gaussian assumptions help us to get an interpretable, closed-form approximation for the KL term from Equation 11, which allows us to directly optimize $\boldsymbol{\xi}_i$ out of the model.

$$\begin{aligned}\mathbb{E}_{\boldsymbol{h_{i-1}}\sim p(\boldsymbol{h_{i-1}})}[\mathrm{D_{KL}}[p(\boldsymbol{h_i}|\boldsymbol{h_{i-1}})||q(\boldsymbol{h_i})]] = \\ \sum_j \left[\log\left(1 + \frac{\mu_{i,j}^2}{\sigma_{i,j}^2}\right)\right]\end{aligned} \tag{13}$$

The final variational information bottleneck can thus be obtained using Equation 13:

$$\begin{aligned}\mathcal{L} = \sum_{i=1}^{L} \beta_i \sum_{j=1}^{r_i} \left[\log\left(1 + \frac{\mu_{i,j}^2}{\sigma_{i,j}^2}\right)\right] - \\ \mathbb{E}_{\{\boldsymbol{x},\boldsymbol{y}\}\sim D, \boldsymbol{h}\sim p(\boldsymbol{h}|\tilde{\boldsymbol{x}})}[\log q(\boldsymbol{y}|\boldsymbol{h_L})]\end{aligned} \tag{14}$$

where $\beta \geq 0$ is a coefficient that determines the strength of the bottleneck that can be defined as the degree to which we value compression over robustness.

B. Experiment setup

In this section, we describe our experimental settings for all the experiments. We follow the two-step pruning procedure where we pretrain all the networks using the standard-training procedure followed by network sparsification using various sparsification methods. We train each model with 200 epochs with a fixed batch size of 64. All the results are measured by computing mean and standard deviation across 5 trials upon randomly chosen seeds.

Our pretrained standard Lenet 5-Caffe baseline model reaches over 99.29% accuracy on MNIST and VGG-16 architecture reaches 92.76% and 67.44% on CIFAR-10 and CIFAR-100 dataset respectively after 200 epochs. We use Adam (Kingma & Ba, 2014) with the learning rate for the weights to be 0.1 times smaller than those for the variational parameters as in (Neklyudov et al., 2017; Lee et al., 2018). For Beta-Bernoulli Dropout, we set $\alpha/K = 10^{-4}$ for all the layers and prune the neurons/filters whose expected drop probability are smaller than a fixed threshold 10^{-3} as originally proposed in the paper. For Beta-Bernoulli Dropout, we scaled the KL-term by different values of trade-off parameter β where $\beta \in \{1, 4, 8, 10, 12\}$ for Lenet-5-Caffe and $\beta \in \{1, 2, 4, 6, 8\}$ for VGG-16. For Variational Information Bottleneck (VIBNet), we tested with trade-off parameter β in Equation 14 where $\beta \in \{10, 30, 50, 80, 100\}$ for Lenet-5-Caffe with MNIST and $\beta \in \{10^{-4}, 1, 20, 40, 60\}$ for VGG-16 with CIFAR-10 and CIFAR-100 dataset. For generating black-box adversarial examples, we used an adversarial trained full network for adversarial neural pruning and the standard base network for the standard Bayesian compression method.

C. More experimental results

Due to the length limit of our paper, some results are illustrated here.

C.1. Robustness of adversarial variational information bottleneck

The results for ANP-VS with Variational Information Bottleneck are summarized in Table 4. We can observe that ANP-VS with Variational Information Bottleneck significantly outperforms the base adversarial training for robustness of adversarial examples by achieving an improvement of $\sim 2\%$ in adversarial accuracy. Note that, ANP-VS leads to $\sim 50\%$ and $\sim 25\%$ reduction in vulnerability for CIFAR-10 and CIFAR-100 dataset with memory and computation efficiency. We emphasize that ANP can similarly be ex-

Table 4. Robustness and compression performance for MNIST on Lenet-5-Caffe, CIFAR-10 and CIFAR-100 on VGG-16 architecture under ℓ_∞-PGD attack for ANP-VS with Variational Information Bottleneck. All the values are measured by computing mean and standard deviation across 5 trials upon randomly chosen seeds. The best results over adversarial baselines are highlighted in bold.

			Adversarial accuracy (↑)		Vulnerability (↓)		Computational efficiency		
	Model	Clean accuracy (↑)	White box attack	Black box attack	White box attack	Black box attack	Memory (↓)	xFLOPS (↑)	Sparsity (↑)
MNIST	Standard	99.29±0.02	0.00±0.0	8.02±0.9	0.129±0.001	0.113±0.000	100.0±0.00	1.00±0.00	0.00±0.00
	BP	99.32±0.04	5.66±0.4	15.47±0.3	0.091±0.001	0.078±0.001	4.34±0.34	9.39±0.25	82.46±0.61
	AT	99.14±0.02	88.03±0.7	94.18±0.8	0.045±0.001	0.040±0.000	100.0±0.00	1.00±0.00	0.00±0.00
	AT BNN	99.16±0.05	88.44±0.4	94.87±0.2	0.364±0.023	0.199±0.031	200.0±0.00	0.50±0.00	0.00±0.00
	Pretrained AT	**99.18±0.06**	88.26±0.6	94.49±0.7	0.412±0.035	0.381±0.029	100.0±0.00	1.00±0.00	0.00±0.00
	ADMM	99.01±0.02	88.47±0.4	94.61±0.7	0.041±0.002	0.038±0.001	100.00±0.00	1.00±0.00	80.00±0.00
	TRADES	99.07±0.04	89.67±0.4	95.04±0.6	0.037±0.001	0.033±0.001	100.0±0.00	1.00±0.00	0.00±0.00
	ANP-VS (ours)	98.86±0.02	**90.11±0.9**	**95.14±0.8**	**0.017±0.001**	**0.015±0.001**	**4.87±0.21**	**10.06±0.87**	**78.48±0.42**
CIFAR-10	Standard	92.76±0.1	13.79±0.8	41.65±0.9	0.077±0.001	0.065±0.001	100.0±0.00	1.00±0.00	0.00±0.00
	BP	92.73±0.1	12.28±0.3	76.35±0.8	0.035±0.002	0.032±0.001	12.38±0.12	2.38±0.0.05	76.35±0.23
	AT	87.50±0.5	49.85±0.9	63.70±0.6	0.050±0.002	0.047±0.001	100.0±0.00	1.00±0.00	0.00±0.00
	AT BNN	86.69±0.5	51.87±0.9	64.92±0.9	0.267±0.013	0.238±0.011	200.0±0.00	0.50±0.00	0.00±0.00
	Pretrained AT	87.50±0.4	52.25±0.7	66.10±0.8	0.041±0.002	0.036±0.001	100.0±0.00	1.00±0.00	0.00±0.00
	ADMM	78.15±0.7	47.37±0.6	62.15±0.8	0.034±0.002	0.030±0.002	100.00±0.00	1.00±0.00	75.00±0.00
	TRADES	80.33±0.5	52.08±0.7	64.80±0.5	0.045±0.001	0.042±0.005	100.0±0.00	1.00±0.00	0.00±0.00
	ANP-VS (ours)	**87.56±0.2**	**53.41±0.5**	**68.12±0.7**	**0.025±0.002**	**0.021±0.001**	**12.09±0.26**	**2.43±0.02**	**77.02±0.32**
CIFAR-100	Standard	67.44±0.7	2.81±0.2	14.94±0.8	0.143±0.007	0.119±0.005	100.0±0.00	1.00±0.00	0.00±0.00
	BP	69.09±0.5	2.73±0.3	19.53±0.4	0.084±0.001	0.073±0.001	18.46±0.42	1.95±0.03	63.84±0.62
	AT	57.79±0.8	19.07±0.8	32.47±1.4	0.079±0.003	0.071±0.003	100.0±0.00	1.00±0.00	0.00±0.00
	AT BNN	53.75±0.7	19.40±0.6	30.38±0.2	0.446±0.029	0.385±0.051	200.0±0.00	0.50±0.00	0.00±0.00
	Pretrained AT	57.14±0.9	19.86±0.6	35.42±0.4	0.071±0.001	0.065±0.002	100.0±0.00	1.00±0.00	0.00±0.00
	ADMM	52.52±0.5	19.65±0.5	31.30±0.3	0.060±0.001	0.056±0.001	100.00±0.00	1.00±0.00	65.00±0.00
	TRADES	56.70±0.7	21.21±0.3	32.81±0.6	0.065±0.003	0.060±0.003	100.0±0.00	1.00±0.00	0.00±0.00
	ANP-VS (ours)	**59.77±0.4**	**21.53±0.6**	**36.82±0.7**	**0.048±0.002**	**0.042±0.002**	**16.46±0.34**	**2.06±0.02**	**67.19±0.57**

tended to any existing or future sparsification method to improve performance. Table 5 further shows the number of units for the baselines and our proposed method.

C.2. Features vulnerability

Figure 6 shows the histogram of the feature vulnerability for various datasets. We consistently observe that standard Bayesian pruning zeros out some of the distortions, AT reduces the distortion level of all the features and ANP-VS does both, with the most significant number of features with zero distortion and low distortion level in general which confirms that our proposed method works successfully as a defense against adversarial attacks. All these results overall confirm the effectiveness of our defense.

C.3. Features visualization

One might also be curious about the representation of the robust and vulnerable features in the latent-feature space. We visualize the robust and vulnerable features based on the vulnerability of a feature in the latent-feature space from our paper. Figure 7 shows the visualization of robust and vulnerable features in the latent space for adversarial training. Note that, AT contains features with high vulnerability (vulnerable feature) and features with less vulnerability (robust feature), which aligns with our observation that the latent features have a varying degree of susceptibility to adversarial perturbations to the input. As future work, we plan to explore more effective ways to suppress perturbation at the intermediate latent features of deep networks.

	Model	No of neurons
MNIST	Standard	20 – 50 – 800 – 500
	BP (BBD)	14 – 21 – 150 – 49
	BP (VIB)	12 – 19 – 160 – 37
	AT	20 – 50 – 800 – 500
	ANP-VS (BBD)	7 – 21 – 147 – 46
	ANP-VS (VIB)	10 – 23 – 200 – 53
CIFAR-10	Standard	64 – 64 – 128 – 128 – 256 – 256 – 256 – 512 – 512 – 512 – 512 – 512 – 512 – 512 – 512
	BP (BBD)	57 – 59 – 127 – 101 – 150 – 71 – 31 – 41 – 35 – 10 – 46 – 48 – 16 – 16 – 25
	BP (VIB)	49 – 56 – 106 – 92 – 157 – 74 – 26 – 43 – 32 – 10 – 39 – 40 – 7 – 7 – 13
	AT	64 – 64 – 128 – 128 – 256 – 256 – 256 – 512 – 512 – 512 – 512 – 512 – 512 – 512 – 512
	ANP-VS (BBD)	42 – 57 – 113 – 96 – 147 – 68 – 25 – 37 – 27 – 9 – 39 – 40 – 13 – 13 – 12
	ANP-VS (VIB)	40 – 57 – 104 – 93 – 174 – 96 – 30 – 48 – 39 – 9 – 49 – 57 – 10 – 10 – 12
CIFAR-100	Standard	64 – 64 – 128 – 128 – 256 – 256 – 256 – 512 – 512 – 512 – 512 – 512 – 512 – 512 – 512
	BP (BBD)	62 – 64 – 128 – 123 – 244 – 203 – 84 – 130 – 95 – 18 – 152 – 157 – 32 – 32 – 101
	BP (VIB)	52 – 64 – 119 – 116 – 229 – 179 – 83 – 99 – 71 – 17 – 107 – 110 – 12 – 11 – 49
	AT	64 – 64 – 128 – 128 – 256 – 256 – 256 – 512 – 512 – 512 – 512 – 512 – 512 – 512 – 512
	ANP-VS (BBD)	60 – 64 – 126 – 122 – 235 – 185 – 77 – 128 – 101 – 17 – 165 – 177 – 35 – 35 – 45
	ANP-VS (VIB)	44 – 58 – 110 – 109 – 207 – 155 – 81 – 86 – 66 – 19 – 88 – 86 – 15 – 15 – 36

Table 5. Distribution of neurons for all the layers of Lenet-5 Caffe for MNIST and VGG-16 architecture for CIFAR-10 and CIFAR-100 datasets.

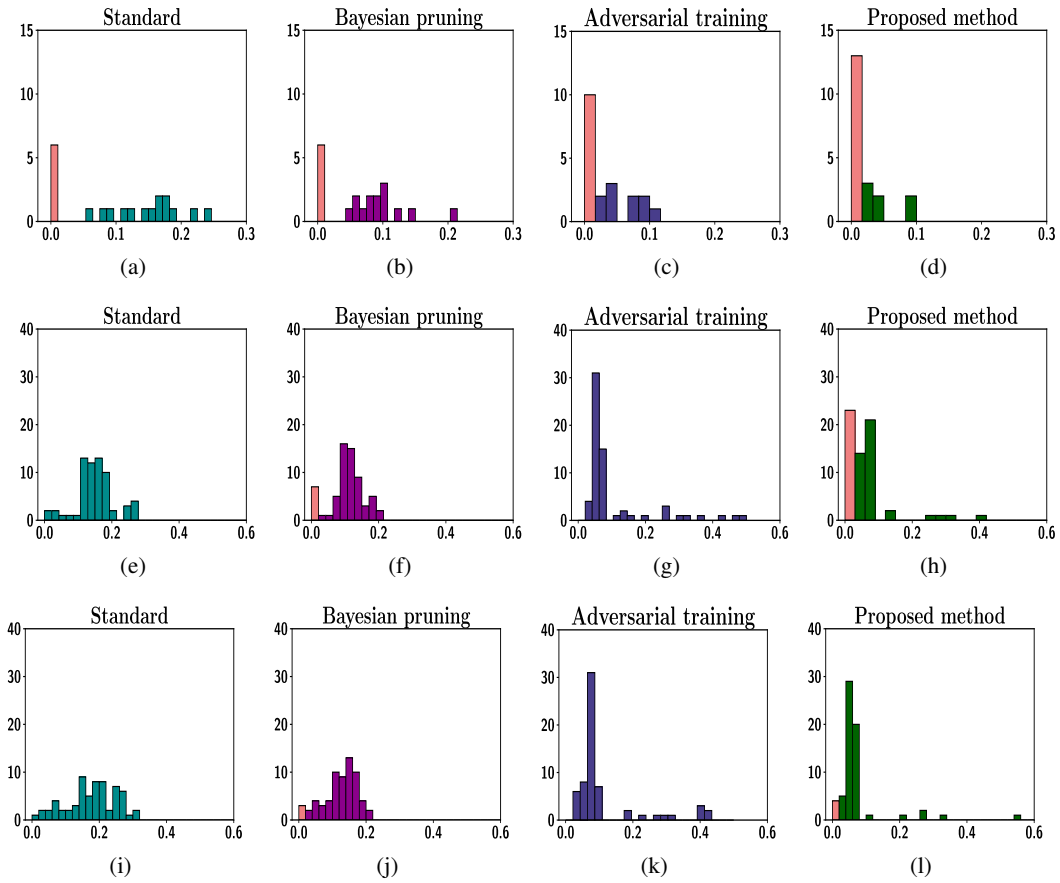
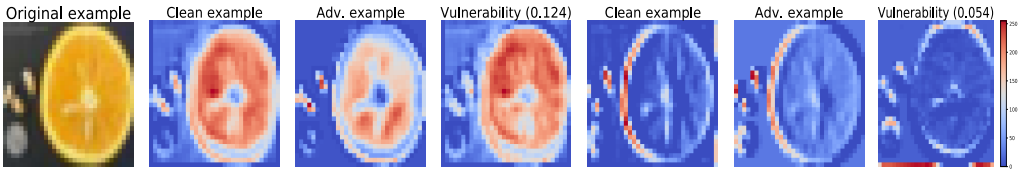


Figure 6. Histogram of vulnerability of the features for the input layer for MNIST in the top row, CIFAR-10 in the middle and CIFAR-100 in the bottom with the number of zeros shown in orange color.



(a) (b) (c) (d) (e) (f) (g)

Figure 7. Visualization of convolutional features of first layer of adversarial trained VGG-16 network with CIFAR-100 dataset. **b) - d)** represents the vulnerable latent-feature with high vulnerability (vulnerable feature) on b) clean example, c) Adversarial example d) Vulnerability (difference between clean and adversarial example) **e) - f)** represents the vulnerable latent-feature with low vulnerability (robust feature) on e) clean example, f) Adversarial example g) Vulnerability (difference between clean and adversarial example)