Poster #9320

PremiUm-CNN: Propagating Uncertainty Towards Robust Convolutional Neural Networks

Dimah Dera*, Nidhal Buaynaya†, Ghulam Rasool†, Roman Shterenberg‡, Hassan Fathallah-Shaykh§

*†Department of Electrical and Computer Engineering, ‡Department of Mathematics, §Department of Neurology

*University of Texas Rio Grande Valley, †Rowan University, ‡ §University of Alabama at Birmingham, USA

22-27 MAY 2022

ICASSP 2022

SINGAPORE

INTRODUCTION

Deep neural networks (DNNs) have surpassed human-level accuracy in various learning tasks. However, DNNs cannot express their uncertainty in the output decisions. This limits the deployment of DNNs in mission-critical domains. Bayesian inference provides a principled approach to reason about model's uncertainty by estimating the posterior distribution of the unknown parameters. This paper establishes the theoretical and algorithmic foundations of uncertainty or belief propagation by developing new deep learning models named PremiUm-CNNs. We introduce a tensor normal distribution as a prior over convolutional kernels and estimate the variational posterior by maximizing the evidence lower bound (ELBO). We start by deriving the first-order mean-covariance propagation framework. Later, we develop a framework based on the unscented transformation (correct at least up to the second-order) that propagates sigma points of the variational distribution through layers of a CNN. The propagated covariance of the predictive distribution captures uncertainty in the output decision.

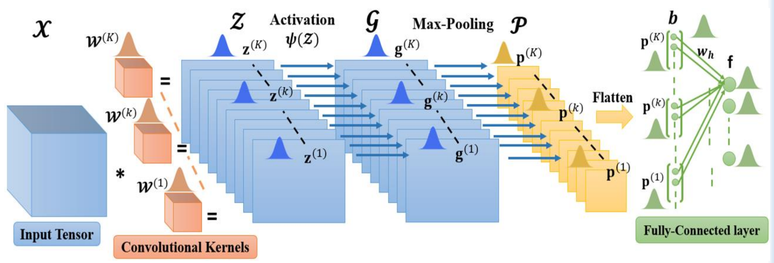

Figure 1: A schematic layout of the proposed PremiUm-CNN. We show the propagation of the variational density through a convolutional layer, activation function, max-pooling, fully-connected layer and a soft-max function. The convolutional kernels, extracted features, the output of activation functions, logits, and the soft-max function output are all random variables.

CONTRIBUTIONS

- We employ the first-order Taylor series approximation (termed Extended Variational Density Propagation, i.e., *exVDP*) for estimating the first two moments of the variational distribution after non-linear activation functions in DNNs.
- We develop the Unscented Variational Density Propagation, i.e., *unVDP* model for approximating the posterior distribution using the unscented transformation (UT). The UT propagates sigma points through the network's layers and results in a posterior approximation that can tackle non-Gaussian distributions and is accurate at least up to the second-order [1].
- We establish superior robustness by analyzing the models' performance (compared to the state-of-the-art DNNs' performance) under noisy conditions and adversarial attacks.

MATERIALS AND METHODS

Variational Density Propagation

- We introduce a prior distribution over the network weights $\boldsymbol{\Omega} \sim \boldsymbol{p}(\boldsymbol{\Omega})$ and estimate the posterior distribution of the weights given the data, D, using variational inference (VI).
- VI method approximates the true posterior $\boldsymbol{p}(\boldsymbol{\Omega}|\boldsymbol{D})$ with a simpler parametrized variational distribution $\boldsymbol{q_\phi}(\boldsymbol{\Omega})$. The optimal parameters of the variational posterior ϕ^* are estimated by minimizing the Kullback-Leibler (KL) divergence between the approximate and the true posterior [2].
- The optimization objective is given by the evidence lower bound (ELBO),

$$\mathcal{L}(\phi; \boldsymbol{y}|\boldsymbol{\mathcal{X}}) = E_{q_\phi(\boldsymbol{\Omega})}\{\log p(\boldsymbol{y}|\boldsymbol{\mathcal{X}}, \boldsymbol{\Omega})\} - \mathrm{KL}\big[q_\phi(\boldsymbol{\Omega})\,||p(\boldsymbol{\Omega})\big]$$

- We build the mathematical foundation of the variational density propagation by deriving the propagation of the mean and covariance of the variational distribution $\boldsymbol{q_\phi}(\boldsymbol{\Omega})$ through a convolutional layer, activation function, maxpooling, fully-connected layer, soft-max function, batch normalization and a skip connection mapping.
- We approximate the mean and covariance after a non-linear activation function $\boldsymbol{\psi}$ using the first-order Taylor series approximation [3]. The model is named Extended Variational Density Propagation, i.e., *exVDP*.

$$\boldsymbol{\mu}_{\mathbf{g}^{(k_c)}} \approx \psi(\boldsymbol{\mu}_{\mathbf{z}^{(k_c)}}),$$
$$\boldsymbol{\Sigma}_{\mathbf{g}^{(k_c)}} \approx \boldsymbol{\Sigma}_{\mathbf{z}^{(k_c)}} \odot \left(\nabla\psi(\boldsymbol{\mu}_{\mathbf{z}^{(k_c)}})\, \nabla\psi(\boldsymbol{\mu}_{\mathbf{z}^{(k_c)}})^T\right),$$

where ∇ is the gradient with respect to $\boldsymbol{z}^{(k_c)}$ and $\odot$ is the Hadamard product.

- The unscented transformation (UT) approximates the mean and covariance after a non-linear transformation with one or two orders of magnitude better than the first-order approximation in the *exVDP* model.
- The UT assures that the estimated mean and covariance are correct, at least up to the second-order [4].
- In the UT framework, the probability density function (pdf) is specified using a set of carefully chosen samples, called sigma points. The model is named Unscented Variational Density Propagation, i.e., *unVDP* .

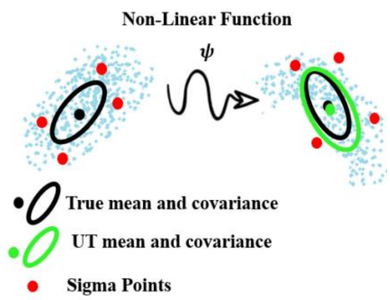

Figure 2: A schematic description of the unscented transformation (UT). We approximate the mean and covariance of a 2D Gaussian after a non-linear function $\boldsymbol{\psi}$ using 4 sigma points.

RESULTS AND DISCUSSION

Table 1: CIFAR-10 Test accuracy using deep CNN (11 layers) and ResNet (18 layers) architectures at varying levels of white and black FGSM and PGD adversarial attacks.

Noise Type and levels		*unVDP*	*unVDP*$_B$	*exVDP*	*exVDP*$_B$	Dropout-CNN	VDP-ResNET	VDP-ResNET$_B$	ResNet
No Noise		91.7%	—	91.8%	—	91.7%	90.0%	—	91.1%
FGSM	Low	91.2%	91.5%	91.2%	91.4%	96.8%	89.2%	89.4%	82.4%
	Medium	83.4%	91.4%	83.8%	91.2%	65.9%	83.6%	86.2%	56.1%
	High	71.1%	88.8%	70.6%	89.1%	56.9%	79.1%	79.0%	47.5%
PGD	Low	91.1%	91.4%	91.2%	91.4%	85.5%	88.6%	89.8%	75.1%
	Medium	82.8%	91.1%	82.2%	91.0%	54.7%	78.4%	85.5%	23.3%
	High	70.5%	88.6%	69.7%	88.5%	42.9%	69.8%	79.3%	13.8%

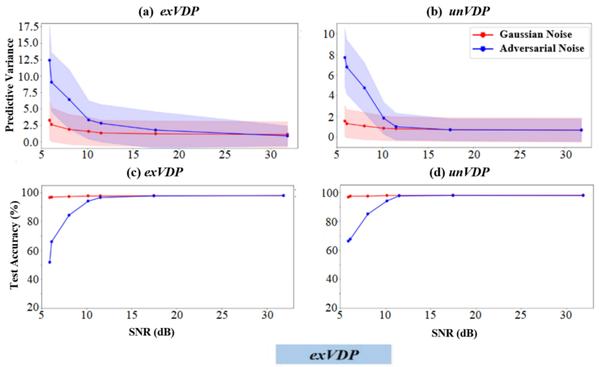

Figure 3: (a) and (b) Average variance values vs. signal-to-noise ratio (SNR) for varying levels of Gaussian and adversarial noise for *exVDP* and *unVDP*, respectively. The variance values are averaged over all 10,000 test examples of the MNIST dataset and the lightly filled areas represent the standard deviation. (c) and (d) Average test accuracy vs. SNR for *exVDP* and *unVDP*, respectively.

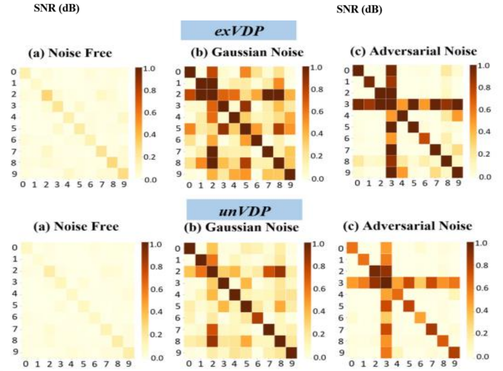

Figure 4: The heat-maps representing the average of the covariance matrices at the output of *exVDP* and *unVDP* models for all 10,000 MNIST test examples. (a) noise-free, (b) Gaussian noise, and (c) adversarial noise. The adversarial examples were generated to fool the models into predicting and image as digit "3".

BIBLIOGRAPHY

[1] D. Simon, Optimal State Estimation: Kalman, H. Infinity, and Non-Linear Approaches. Hoboken, NJ, USA: Wiley-Interscience, 2006.

[2] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," J. Amer. Stat. Assoc., vol. 112, no. 518, pp. 859–877, 2017.

[3] A. Papoulis and S. U. Pillai, Probability, Random Variables, and Stochastic Processes, 4th ed. New York, NY, USA: McGraw-Hill, 2002.

[4] E. A. Wan and R. Van Der Merwe, "The unscented Kalman filter for nonlinear estimation," in Proc. IEEE Adaptive Syst. Signal Process., Commun. Control Symp., 2000, pp. 153–158.

Acknowledgments

This work was supported by the National Science Foundation Awards ECCS-1903466, CCF-1527822, OAC-2008690 and CRII: RI #2153413. The work of Dimah Dera was supported by the ACM SIGHPC/Intel Computational, and Data Science Fellowship Award The authors would like to thank U.K. EPSRC support through EP/T013265/1 Project NSF -EPSRC: ShiRAS towards safe and reliable autonomy in sensor driven systems.