Analytic Expressions for Probabilistic Moments of PL-DNN with Gaussian Input

Adel Bibi*, Modar Alfadly*, Bernard Ghanem

King Abdullah University of Science and Technology (KAUST)

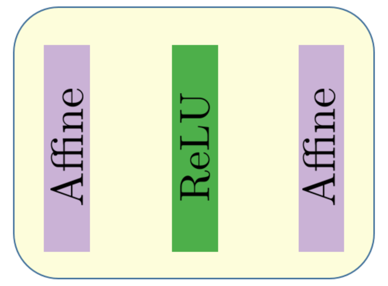
Access the source code

2018 CVPR SALT LAKE CITY • JUNE 18-22

Motivation

The uncouth reaction of deep neural networks (DNNs) to noisy input has spawned research on developing more adversarial input attacks and defenses. Ideally, we want to study the output probability density function of any given DNN for any given input distribution. Instead, we will derive exact analytic expressions for the first and second moments, i.e., $\mathbb{E}[\mathbf{g}^{k\in\{1,2\}}(\mathbf{n})]$ (mean and variance), of a small piecewise linear (PL) neural network (Affine-ReLU-Affine) subject to general Gaussian input.

$\mathbf{x}\sim\mathcal{N}(\mu,\Sigma)$ →



→ $\mathbf{g}(\mathbf{x})$ $\begin{Bmatrix}\mathbb{E}[\mathbf{g}(\mathbf{x})]\\ \text{var}[\mathbf{g}(\mathbf{x})]\end{Bmatrix}$

$\mathbf{g}(\mathbf{x}) = \mathbf{B}\max(\mathbf{A}\mathbf{x}+\mathbf{c}_1, \mathbf{0}_p) + \mathbf{c}_2$

where $\mathbf{A}\in\mathbb{R}^{p\times n}, \mathbf{B}\in\mathbb{R}^{d\times p}, \mathbf{c}_1\in\mathbb{R}^p, \mathbf{c}_2\in\mathbb{R}^d$

Useful Remarks and Lemmas

Remark: Probabilistic Properties of the Affine Function

Let $f(\mathbf{x}) = \mathbf{A}\mathbf{x}+\mathbf{b}$ *and* $\mathbf{x}$ *be a random variable with* $\mathbb{E}[\mathbf{x}]=\mu$ *and* $Cov[\mathbf{x}]=\Sigma$*, then:*

$\mathbb{E}[f(\mathbf{x})] = \mathbf{A}\mu+\mathbf{b}$, $Cov[f(\mathbf{x})] = \mathbf{A}\Sigma\mathbf{A}^\top$ *and if* $\mathbf{x}$ *is Gaussian, then* $f(\mathbf{x})$ *is Gaussian.*

Where $\Sigma = Corr[\mathbf{x}] - \mu\mu^\top$, $Corr[\mathbf{x}] = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$*, and* $var[\mathbf{x}] = \mathbb{E}[\mathbf{x}^2] - \mathbb{E}[\mathbf{x}]^2 = diag(\Sigma)$.

Lemma 1: The PDF of the Squared ReLU Function

Let $x\sim\mathcal{N}(0,\sigma^2)$ *and* $q^2(x) = \max^2(x,0):\mathbb{R}\to\mathbb{R}$*, then the PDF of* $q^2(x)$ *is* $f_{q^2}(x)$ *where:*

$$f_{q^2}(x) = \frac{1}{2}\delta(x) + \frac{1}{2\sqrt{x}}f_x(\sqrt{x})u(\sqrt{x}) \text{ and } \mathbb{E}[q^2(x)] = \frac{\sigma^2}{2}$$

Lemma 2: Extension of Price's Theorem [1]

Let $\mathbf{x}\in\mathbb{R}^n\sim\mathcal{N}(\mu,\Sigma)$*, where* $\sigma_{ij}=\Sigma(i,j)\ \forall i\neq j$*, for any even* p*, and under mild assumptions on the nonlinear map* $\Psi:\mathbb{R}^n\to\mathbb{R}$*, we have* $\frac{\partial^{\frac{p}{2}}\mathbb{E}[\Psi(\mathbf{x})]}{\prod_{\forall i_{odd}}\partial\sigma_{ii+1}} = \mathbb{E}[\frac{\partial^p\Psi(\mathbf{x})}{\partial x_1\ldots\partial x_p}]$.

Lemma 3: The First Moment of Bivariate Gaussian ReLUs Product

Let $\mathbf{x}\sim\mathcal{N}(\mathbf{0}_2,\Sigma)$ *be a bivariate Gaussian and* $T(x_1,x_2) = \max(x_1,0)\max(x_2,0)$*, then:*

$$\mathbb{E}[T(x_1,x_2)] = \frac{1}{2\pi}\left(\sigma_{12}\sin^{-1}\left(\frac{\sigma_{12}}{\sigma_1\sigma_2}\right) + \sigma_1\sigma_2\sqrt{1-\frac{\sigma_{12}^2}{\sigma_1^2\sigma_2^2}}\right) + \frac{\sigma_{12}}{4}$$

where $\sigma_{ij} = \Sigma(i,j)\ \forall i\neq j$ *and* $\sigma_i^2 = \Sigma(i,i)$.

Gaussian Network Moments (GNM)

Main Theorems: Gaussian Moments of the ReLU Function

Let $\mathbf{x}\sim\mathcal{N}(\mu,\Sigma)$, $q(\mathbf{x}) = \max(\mathbf{x},\mathbf{0})$*, then:*

$$\mathbb{E}[q(\mathbf{x})] = \frac{1}{2}\mu\odot\left(\mathbf{1}+erf(\sqrt{\mathbf{u}})\right) + \frac{1}{\sqrt{2\pi}}\sqrt{\mathbf{v}}\odot exp\left(-\mathbf{u}\right) \quad (Theorem\ 1)$$

$$Corr\left[q(\mathbf{x})\right]|_{\mu=\mathbf{0}} = \frac{1}{2\pi}\left(\Sigma\odot\sin^{-1}(\mathbf{V}) + \mathbf{S}\odot\sqrt{1-\mathbf{V}^2}\right) + \frac{1}{4}\Sigma \quad (Theorem\ 2)$$

where $\mathbf{v} = diag(\Sigma)$, $\mathbf{u} = \mu^2\oslash 2\mathbf{v}$, $\mathbf{S} = \sqrt{\mathbf{v}\mathbf{v}^\top}$, $\mathbf{V} = \Sigma\oslash\mathbf{S}$*, and* $erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$.
Additionally, $\odot$ *and* $\oslash$ *are element wise product and division, respectively.*

Corollary: Gaussian Network Moments of Affine-ReLU-Affine

Let $\mathbf{x}\sim\mathcal{N}(\mu,\Sigma)$ *and for any network of the form* $\mathbf{g}(\mathbf{x}) = \mathbf{B}\max\left(\mathbf{A}\mathbf{x}+\mathbf{c}_1,\mathbf{0}\right)+\mathbf{c}_2$*:*

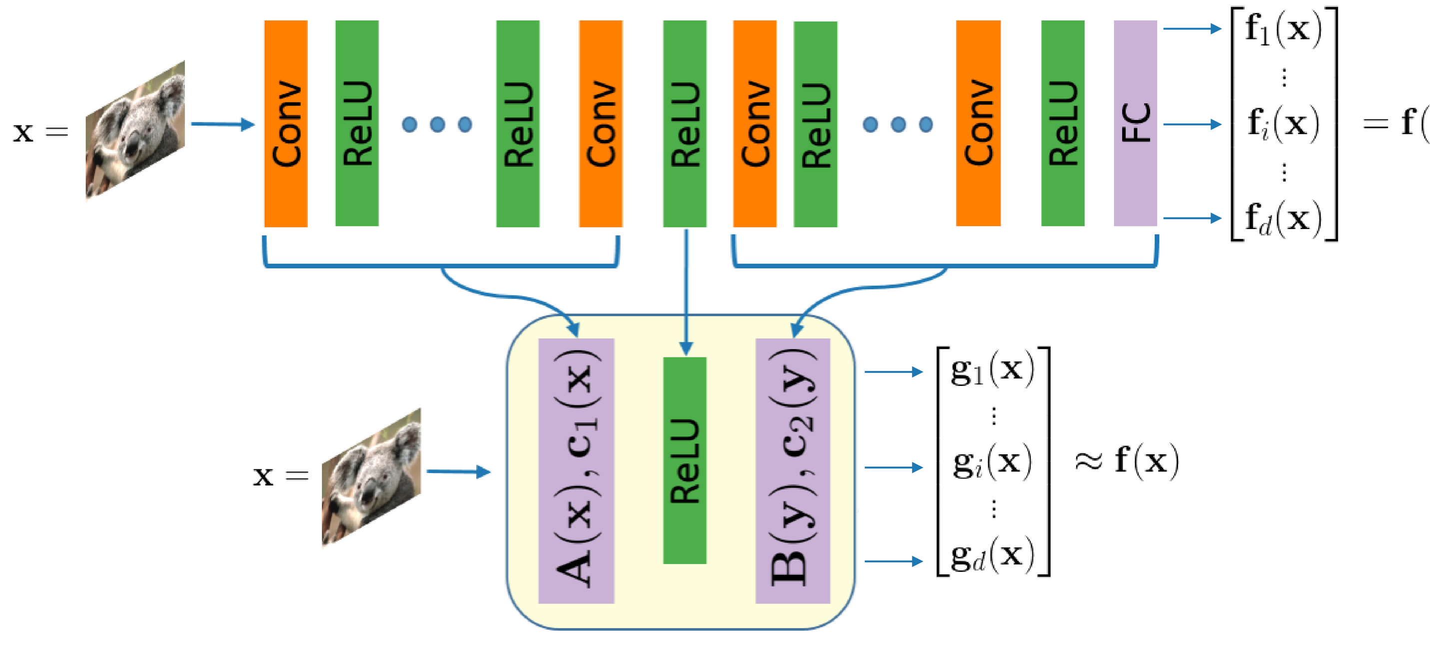
$$\mathbb{E}[\mathbf{g}(\mathbf{x})] = \mathbf{B}\mathbb{E}[q(\mathbf{y})] + \mathbf{c}_2$$

$$Cov[\mathbf{g}(\mathbf{x})]|_{\mathbf{c}_1=-\mathbf{A}\mu} = \mathbf{B}\left(Corr[q(\mathbf{y})] - \mathbb{E}[q(\mathbf{y})]\mathbb{E}[q(\mathbf{y})]^\top\right)\mathbf{B}^\top$$

where $\mathbf{y}\sim\mathcal{N}(\mathbf{A}\mu+\mathbf{c}_1, \mathbf{A}\Sigma\mathbf{A}^\top)$ *and* $q(\mathbf{y}) = \max(\mathbf{y},\mathbf{0})$.

Working with deep models and large datasets

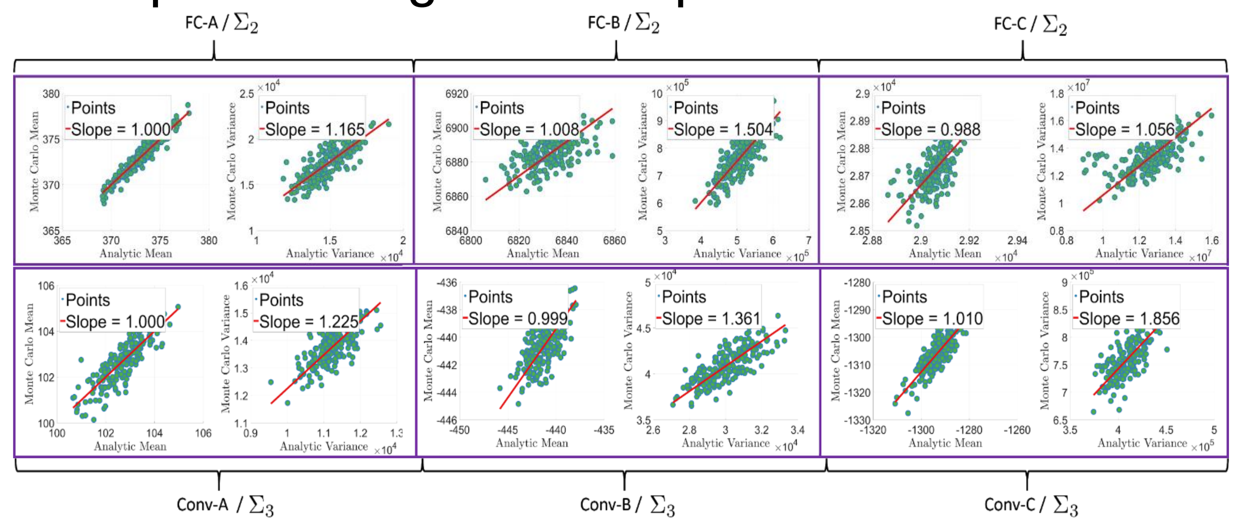
We experimentally show that these expressions are tight under simple two-stage linearization of deeper PL-DNNs, especially popular architectures in the literature (e.g. LeNet and AlexNet), where each of the affine layers will be the first-order Taylor approximation of the layers surrounding a ReLU.



For large datasets, we will consider k-mean clustering and use the cluster centers as linearization points in the two-stage linearization method. In our experiments, the expressions were tight even with small number of clusters.

Tightness Experiments

Fitting a line on a plot of Monte-Carlo estimations of the mean and variance vs. our expressions gives a slope that is close to one on different networks.



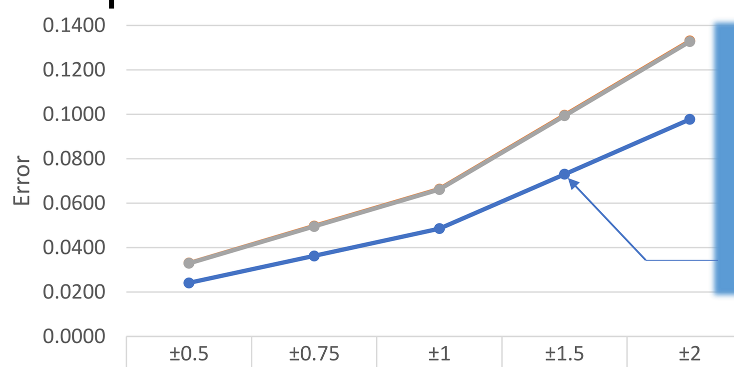
Logits	$\mathbb{E}[\mathbb{E}_{ratio}]$	$\sigma(\mathbb{E}_{ratio})$	$\mathbb{E}[\text{var}_{ratio}]$	$\sigma(\text{var}_{ratio})$
$\mathbf{g}_0(\mathbf{x})$	1.007	0	0.5431	0.021
$\mathbf{g}_1(\mathbf{x})$	1.043	0.001	0.6192	0.021
$\mathbf{g}_2(\mathbf{x})$	0.916	0.001	0.3681	0.014
$\mathbf{g}_3(\mathbf{x})$	1.061	0.001	0.517	0.019
$\mathbf{g}_4(\mathbf{x})$	0.581	0.006	0.4878	0.016
$\mathbf{g}_5(\mathbf{x})$	0.699	0.004	0.4267	0.015
$\mathbf{g}_6(\mathbf{x})$	0.999	0	0.4795	0.017
$\mathbf{g}_7(\mathbf{x})$	0.990	0	0.446	0.014
$\mathbf{g}_8(\mathbf{x})$	1.001	0	0.4476	0.014
$\mathbf{g}_9(\mathbf{x})$	1.164	0.002	0.5931	0.020

Top-k scores ordering in AlexNet is accurately predicted by the GNMs.

k	1	2	3	4	5
Top-k accuracy	98.30%	95.31%	91.16%	86.09%	80.51%

K-mean clustering, helps to reduce the computation cost on MNIST.

Logits	k=5000		k=500		k=250	
	$\mathbb{E}[\mathbb{E}_{ratio}]$	$\mathbb{E}[\text{var}_{ratio}]$	$\mathbb{E}[\mathbb{E}_{ratio}]$	$\mathbb{E}[\text{var}_{ratio}]$	$\mathbb{E}[\mathbb{E}_{ratio}]$	$\mathbb{E}[\text{var}_{ratio}]$
$\mathbf{g}_0(\mathbf{x})$	0.9998	0.2725	0.752	0.2851	1.1406	0.2849
$\mathbf{g}_1(\mathbf{x})$	0.9983	0.3782	1.0235	0.3668	1.0587	0.3668
$\mathbf{g}_2(\mathbf{x})$	0.992	0.2298	1.6994	0.2366	1.8291	0.2499
$\mathbf{g}_3(\mathbf{x})$	0.9959	0.2604	1.2021	0.2673	1.3191	0.2582
$\mathbf{g}_4(\mathbf{x})$	0.9932	0.2761	1.2705	0.3012	1.5868	0.2945
$\mathbf{g}_5(\mathbf{x})$	0.9936	0.2441	0.9686	0.2523	1.5272	0.2429
$\mathbf{g}_6(\mathbf{x})$	0.991	0.3592	1.1909	0.3739	1.257	0.3539
$\mathbf{g}_7(\mathbf{x})$	0.9844	0.213	2.5221	0.2156	1.3419	0.2169
$\mathbf{g}_8(\mathbf{x})$	0.9906	0.2554	1.1877	0.25	1.9278	0.2541
$\mathbf{g}_9(\mathbf{x})$	0.9917	0.2685	1.3469	0.2796	1.9592	0.2832

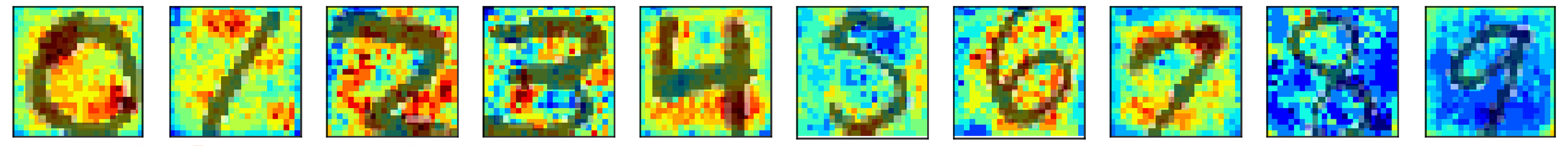


	±0.5	±0.75	±1	±1.5	±2
layer 1	0.0241	0.0362	0.0485	0.0730	0.0977
layer 2	0.0330	0.0497	0.0663	0.0996	0.1330
layer 3	0.0329	0.0495	0.0661	0.0993	0.1327

Early layers increase the accuracy at memory cost

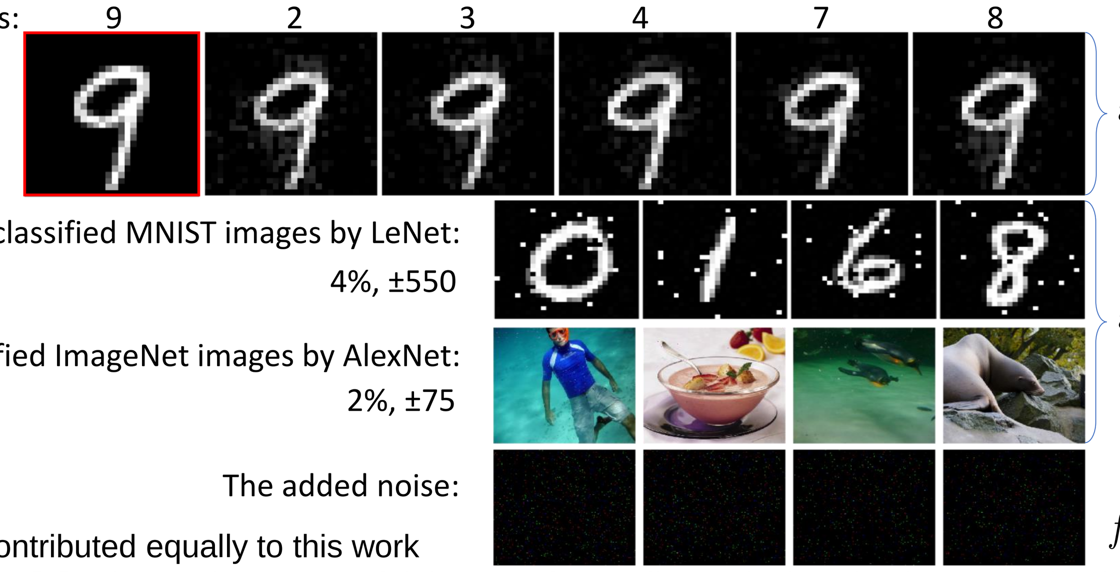
Sensitivity Analysis

Localized spatial noise identifies the pixels that are more sensitive to noise.



Adversarial Attacks

GNMs enable targeted and non-targeted adversarial attacks.



$$\begin{array}{ll}\arg\min_{\mu,\sigma} & \left(\max_{i\neq j}\left(\mathcal{E}_i^{\mathbf{M}}(\mu,\sigma^2)\right) - \mathcal{E}_j^{\mathbf{M}}(\mu,\sigma^2)\right)\\ s.t. & 0<\sigma^2\leq 2, \quad -\beta\mathbf{1}_n\leq\mu\leq\beta\mathbf{1}_n\end{array}$$

$$\begin{array}{ll}\arg\min_{\mu^\alpha,\sigma} & \left(\mathcal{E}_i^{\mathbf{M}}(\mu^\alpha,\sigma^2) - \max_{j\neq i}\left(\mathcal{E}_j^{\mathbf{M}}(\mu^\alpha,\sigma^2)\right)\right)\\ s.t. & 0<\sigma^2\leq 2, \quad -\beta\mathbf{1}_{\alpha n}\leq\mu^\alpha\leq\beta\mathbf{1}_{\alpha n}\end{array}$$

where $\mathcal{E}_i^{\mathbf{M}} \equiv \mathbb{E}[\mathbf{g}_i\left(\mathbf{M}+\mathbf{x}_{(\mu,\sigma^2\mathbf{I}_n)}\right)$
for any image $\mathbf{M}$ *of predicted class label* i.

* Authors contributed equally to this work

References: [1] Price, Robert. "A useful theorem for nonlinear devices having Gaussian inputs." IEEE Trans. Inf. Theory (1958): 69-72.

Acknowledgment: This research was supported by competitive funding from King Abdullah University of Science and Technology.