A Stability Analysis of Neural Networks and Its Application to Tsunami Early Warning

Donsub Rim*, Sanah Suri*, Sanghyun Hong**, Kookjin Lee†, and Randall J. LeVeque‡

*Washington University in St. Louis **Oregon State University

†Arizona State University ‡University of Washington

Feedforward Neural Networks

- *Feedforward neural network* Neural network f with L layers

$$f(\mathbf{x}) = A_L \circ \sigma \odot A_{L-1} \circ \ldots A_2 \circ \sigma \odot A_1(\mathbf{x})$$

- The parameters of the NN f are the entries of $\mathbf{W}_\ell$ and $\mathbf{b}_\ell$

$$\mathbf{W}_\ell = \begin{bmatrix} w_{11,\ell} & w_{12,\ell} & \cdots & w_{1n_{\ell-1}} \\ w_{21,\ell} & w_{22,\ell} & \cdots & w_{2n_{\ell-1}} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n_\ell 1,\ell} & w_{n_\ell 1,\ell} & \cdots & w_{1n_{\ell-1}} \end{bmatrix} \in \mathbb{R}^{n_\ell \times n_{\ell-1}} \quad \mathbf{b}_\ell = \begin{bmatrix} b_{1,\ell} \\ b_{2,\ell} \\ \vdots \\ b_{n_\ell,\ell} \end{bmatrix} \in \mathbb{R}^{n_\ell}$$

$\mathbf{W}_\ell$ is called the *weight matrix*, $\mathbf{b}_\ell$ the *bias*

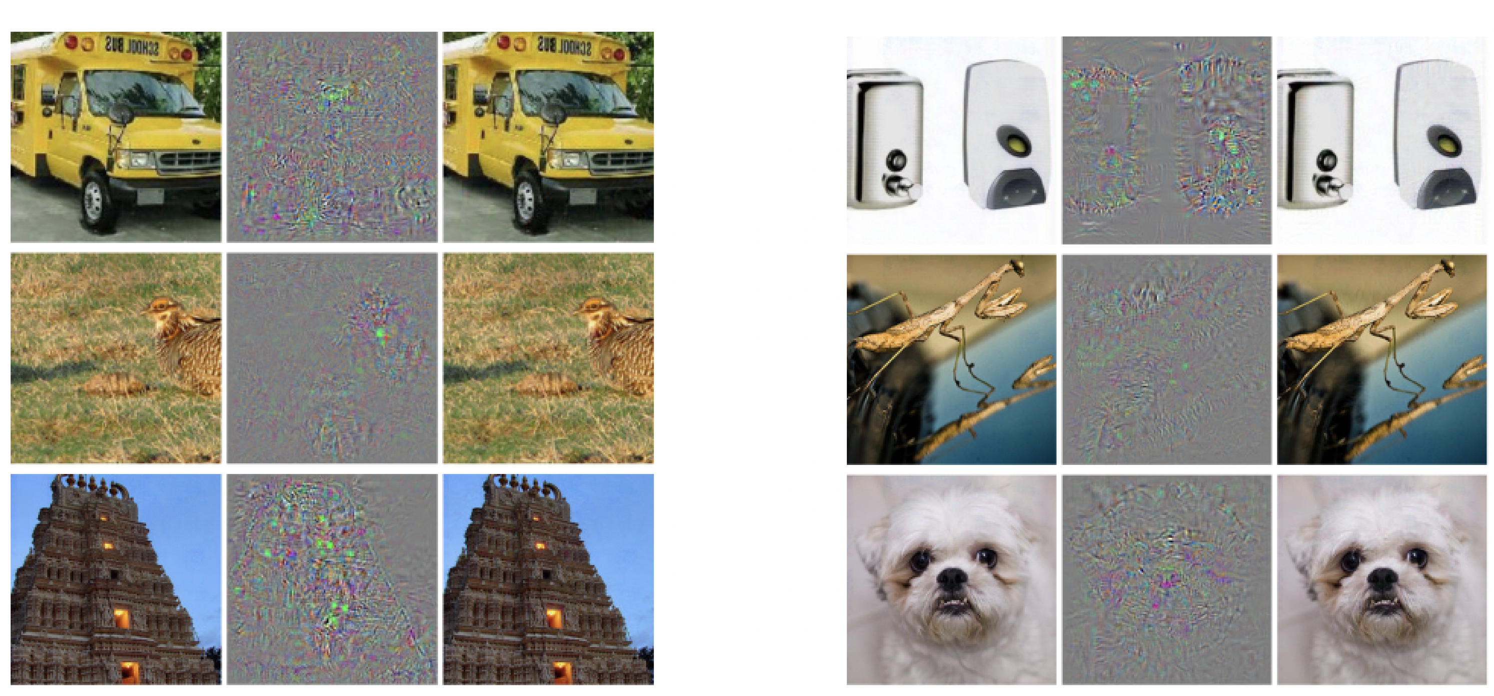
- Affine maps are defined as

$$A_\ell(\mathbf{z}) = \mathbf{W}_\ell \mathbf{z} + \mathbf{b}_\ell$$

- σ is the ReLU activation $\sigma(x) = \max\{x, 0\}$
- *Examples* Convolutional NNs (CNNs), Residual NNs (ResNets)

Need for Stability Analysis: Adversarial Examples

- NNs are not robust with respect to input noise
- *Intriguing property of NNs* Fix the input $\mathbf{x}_0$ then bad perturbations $\mathbf{x}_0 + \delta\mathbf{x}$ that yield very different output can be found
- *Image classification task* A NN classifier that accurately predicts the class of the image $\mathbf{x}_0$ missclasifies a perturbed image $\mathbf{x}_0 + \delta\mathbf{x}$ even when the size of the perturbation $||\delta\mathbf{x}||$ is negligible



(left columns) original image $\mathbf{x}_0$

(middle columns) perturbation $\delta\mathbf{x}$

(right columns) perturbed image $\mathbf{x}_0 + \delta\mathbf{x}$

Perturbed images in the right columns are predicted as *Ostrichs*

- These examples are called *adversarial examples* and can be found through optimization, e.g. the projected gradient descent (PGD)

Low Rank Householder Expansion (LRHE)

- *Low-Rank Householder Expansion* feedforward NNs are written

$$f(\mathbf{x}) = F(\mathbf{x})\,\mathbf{x} = [F_0 + F_\sigma(\mathbf{x})]\,\mathbf{x}$$

The input-dependent matrix F_σ has rank at most $L-1$

$$F_\sigma(\mathbf{x}) = \sum_{\ell=1}^{r} d_\ell(\mathbf{x})\zeta_\ell(\mathbf{x})\xi_\ell(\mathbf{x})^\top \quad r \le L-1$$

- The row and column spaces Φ and Ψ can be computed based on the trained weights

$$\Phi = \text{span}\{\zeta_\ell\} \qquad \Psi = \text{span}\{\xi_\ell\}$$

- low-rank since number of layers L is much smaller than the input dimension (# of data points or pixels)

Householder Reflectors Approximation of ReLU

- Householder reflectors are symmetric, orthogonal matrices

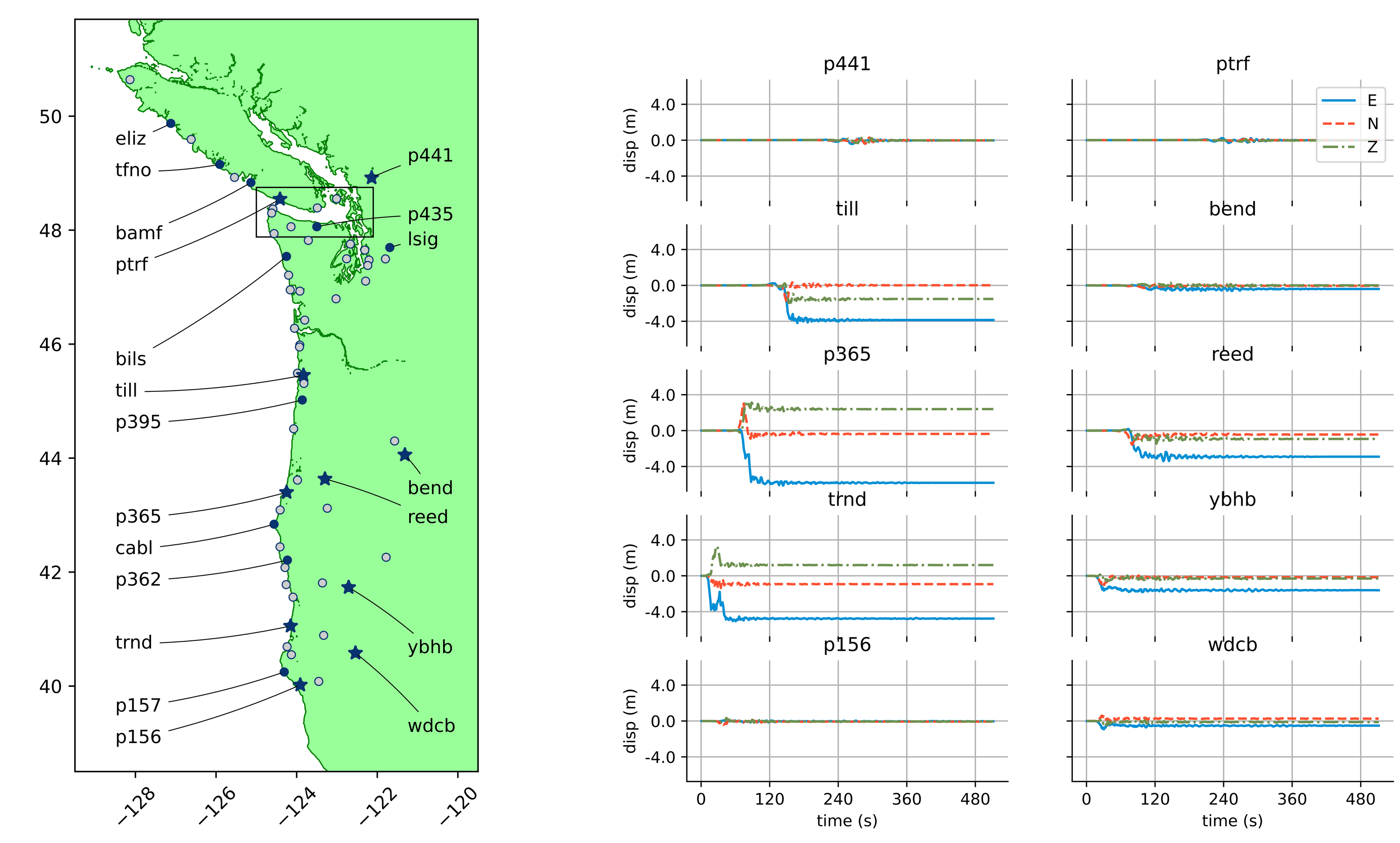
$$\sigma(\mathbf{z}) = \mathbf{H}_\mathbf{z}\mathbf{z} = (\mathbf{I} - 2\mathbf{v}_\mathbf{z}\mathbf{v}_\mathbf{z}^T)\mathbf{z} \qquad ||\mathbf{v}_\mathbf{z}||_2 = 1$$

Rank-1 perturbation of the identity matrix

Tsunami Prediction (Cascadia Subduction Zone)

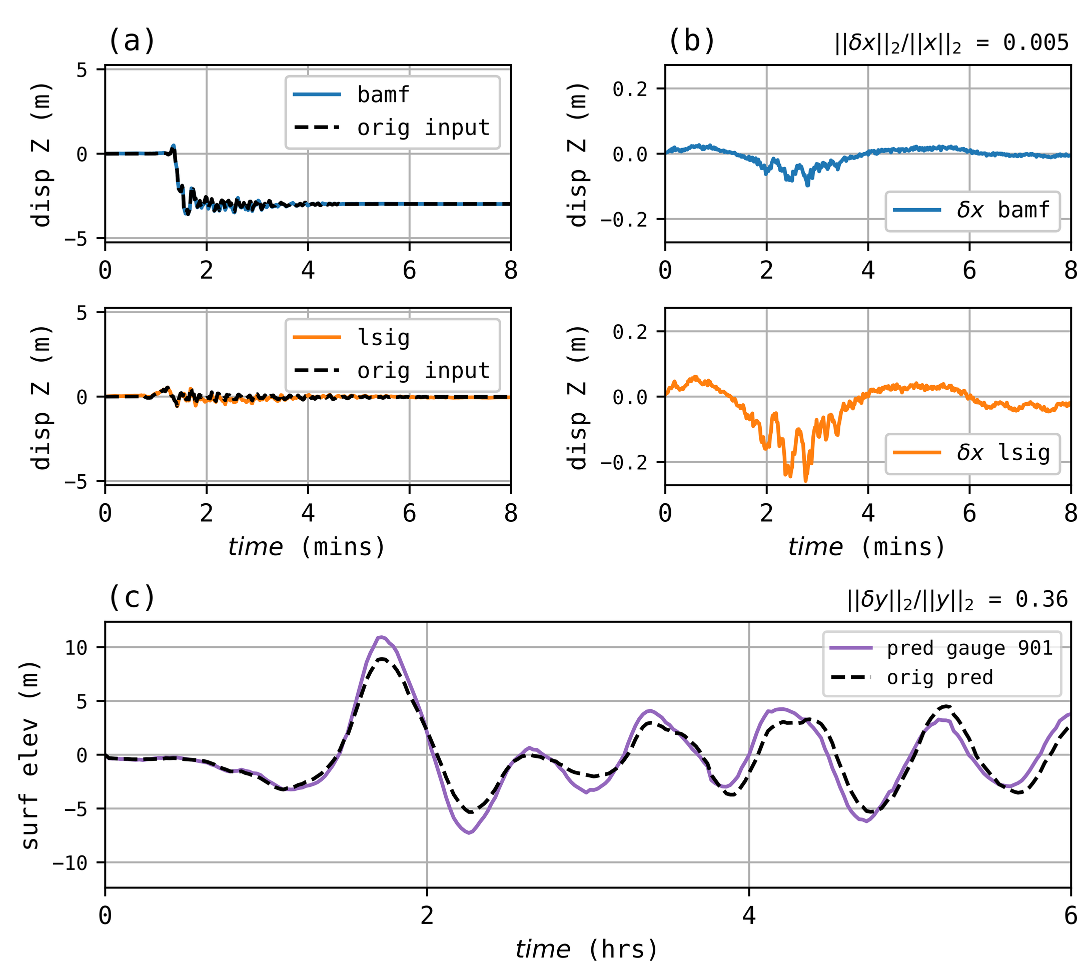
Train NNs to predict tsunami waveforms (6 hrs) at geographical locations using geodetic measurements from GNSS stations (< 8 mins)

GNSS stations　　Geodetic signals (input to NNs)



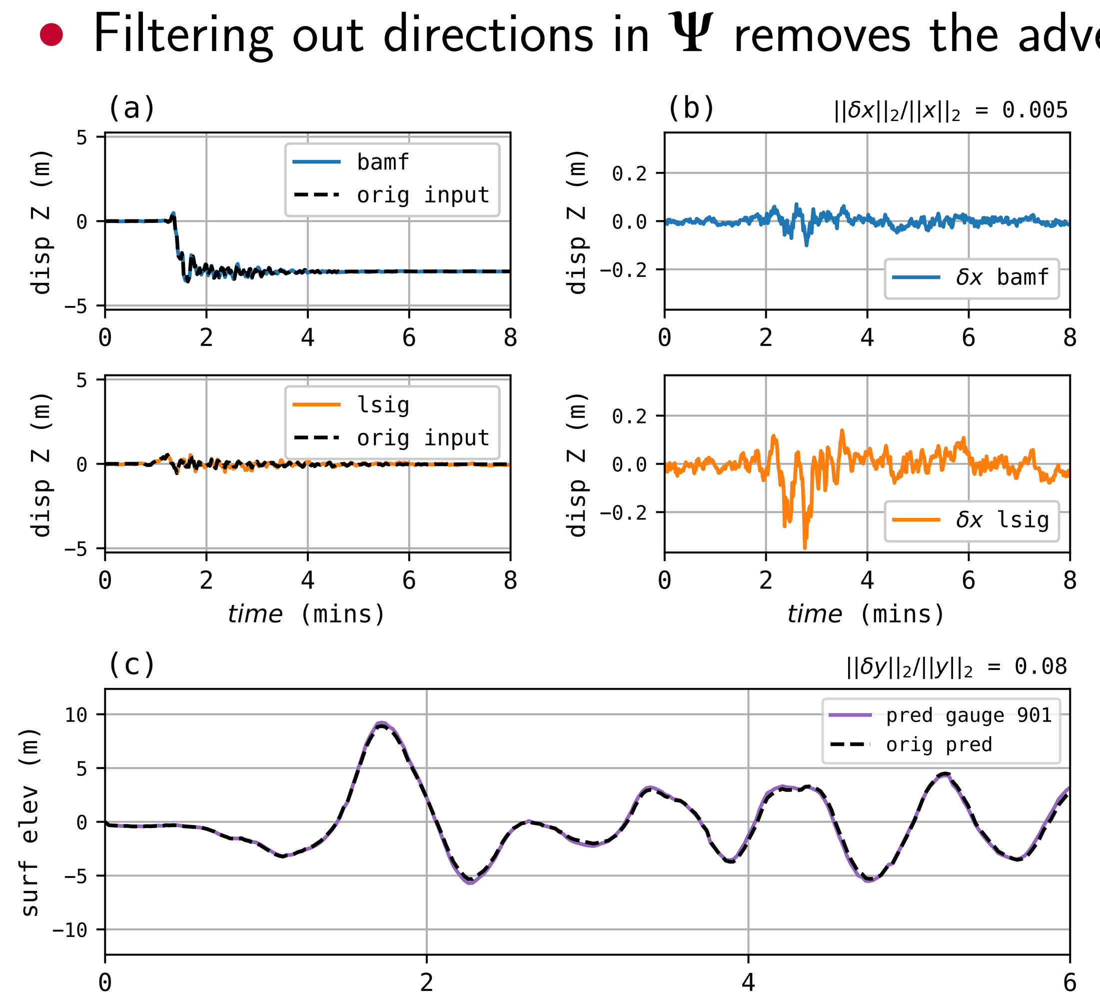
Adversarial Examples for NN Tsunami Model

- An adversarial example found by PGD



(a) The perturbed input $\mathbf{x}_0 + \delta\mathbf{x}$ at two selected stations
(b) The perturbation $\delta\mathbf{x}$
(c) The resulting perturbed output $f(\mathbf{x}_0 + \delta\mathbf{x})$ at gauge 901. An imperceptible 0.5% change in the input causes a large 36% change in the output.

- Filtering out directions in $\boldsymbol{\Psi}$ removes the adversarial effect



(a) The perturbed input $\mathbf{x}_0 + (\delta\mathbf{x})_{\text{filter}}$ at two selected stations
(b) The filtered perturbation $(\delta\mathbf{x})_{\text{filter}}$
(c) The resulting perturbed output $f(\mathbf{x}_0 + (\delta\mathbf{x})_{\text{filter}})$. The amount of output perturbation is at 8%, closer to that of the input perturbation.

References

[1] D. Rim, R. Baraldi, C.M. Liu, R. J. LeVeque, K. Terada
Tsunami Early Warning from Global Navigation Satellite System Data using Convolutional Neural Networks
Geophys. Res. Lett. **49** e2022GL099511 (2022)

[2] D. Rim, S. Suri, S. Hong, K. Lee, R. J. LeVeque
A Stability Analysis of Neural Networks and Its Application to Tsunami Early Warning
J. Geophys. Res.: Machine Learning and Computation 1 e2024JH000223 (2024)

[3] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus
Intriguing properties of neural networks
International conference on learning representations (2014)

Contact Donsub Rim (rim@wustl.edu | dsrim.github.io)