Binary and Random Inputs to Rapidly Identify Overfitting of Deep Neural Networks Trained to Output Ultrasound Images

Jiaxin Zhang*, Alycen Wiacek*, and Muyinatu A. Lediju Bell*†‡
*Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD
†Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD
‡Department of Computer Science, Johns Hopkins University, Baltimore, MD

2022 IEEE International Ultrasonics Symposium (IUS) | 978-1-6654-6657-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/IUS54386.2022.9957314

***Abstract*—We developed a novel method to detect overfitting of deep neural networks trained to create ultrasound images. This method only requires the network architecture and trained weights, and does not require loss function monitoring during an otherwise time-consuming training process. Specifically, two binary images and an image of Gaussian random noise were used as inputs to three neural networks submitted to the Challenge on Ultrasound Beamforming with Deep Learning (CUBDL). Comparing the network-created images to the ground truth immediately revealed an overfit to the data used to train one of the three networks, indicating the promise of our method to detect overfitting without requiring lengthy network retraining or the collection of additional test data. This approach holds promise for regulatory oversight of DNNs intended to be deployed on patient data.**

***Index Terms*—deep learning, imaging, beamforming, overfit**

I. Introduction

Deep neural networks (DNNs) have recently emerged as a promising approach to high-quality ultrasound image formation. Conventional ultrasound beamforming methods, such as the delay-and-sum (DAS) algorithm, are typically hand-tailored with array geometry and medium properties [1]. Improved image quality can be achieved by employing the DAS beamformer with multiple plane-wave transmissions instead of a single plane-wave transmission. Compared to the traditional DAS algorithms, ultrasound beamforming with deep learning is advantageous because networks can be trained to directly output high-quality ultrasound images from raw ultrasound channel data [2]. Despite this advantage, one potential problem is overfitting, in which networks perform well on training data, yet fail to generalize across different unseen datasets.

Common methods such as early stopping, k-fold cross-validation, or inference are widely adopted as effective approaches to prevent or detect overfitting [3], [4]. In early stopping, training and validation errors are monitored, and validation errors are measured to represent generalization errors (i.e., the errors associated with predicting outcome values for previously unseen data). In addition, early stopping criteria are implemented to decide when to stop a training process and achieve minimum generalization loss. Common criteria include validation losses, quotient of validation losses, or progress exceeding a particular threshold [5].

In k-fold cross-validation [6], a dataset is split into k groups and enumerates the fitting and evaluation process based on k-1 training sets and 1 test set, k times. The final model skill score shows the generalization of the network quantitatively.

With an inference approach to detect overfitting, additional test data are input to further evaluate DNN performance [7], [8]. This additional ultrasound RF channel data can be obtained through experiments or simulations or from publicly available datasets, such as CUBDL [9]–[11] or PICMUS [12] datasets.

Major limitations of the early stopping, cross-validation, and inference methods are that they require training data, retraining of the network, or the curation of new test data. However, when presented with a new DNN without access to training code, training data, and unseen test data, implementation of these methods are not possible. In addition, considering that the training process typically requires thousands of training examples, it is not always feasible for a user to train a new DNN to perform the same task as that learned with an existing DNN. Ideally, training code and data would not be required to build confidence that an existing publicly available DNN will perform well on new data related to the trained task. This consideration is additionally important with respect to regulatory procedures and trustworthiness of DNNs deployed on patient data.

In this paper, we propose a new method to more rapidly identify the overfitting of DNNs trained to output ultrasound images when compared to conventional approaches. Our method does not require any training code, training data, or test examples. Thus, it is effective when given a DNN and its input data structure. In this case, the user can employ our method to determine if the network is overfitting well before testing on previously unseen ultrasound data.

II. Methods

A. Artificial RF channel data

Robust networks generalize across different datasets while overfitted models perform well only on training data [13]. To test networks on unseen data, we created three types of artificial RF channel data, grouped into two categories: (1) binary samples including zeros and ones and (2) random samples. The proposed artificial channel data were inputs to

a Pytorch DAS plane-wave beamforming algorithm [10], [11] and to three DNN models submitted by Rothlübbers *et al.* [14], Goudarzi *et al.* [15], and Wang *et al.* [16], respectively, to the Challenge on Ultrasound Beamforming with Deep Learning (CUBDL) [10], [11]. For brevity, these three DNNs will be referred to as Network A, Network B, and Network C, respectively. The output single 0° plane wave images from the PyTorch DAS beamformer served as ground truths.

With zeros as the input, the output envelope image contained zeros at each pixel location, resulting in invalid values after normalization. To obtain a valid output, a value close to zero (i.e., 1×10^{-20}) was used instead. In addition, one RF channel data point at the center of the input was set to 1 to achieve a normalized image that was representative of the input and distinguishable from the second binary input. This second binary input was a matrix of ones surrounding a center pixel value of 1×10^{-20} to address the same normalization challenges described above.

A matrix of random samples drawn from a Gaussian distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ was created to be the third type of artificial RF channel data. To maintain the same range as the zeros and ones input data described above, the random input values were normalized to the range $[0, 1]$.

B. Evaluation metrics

For each output, the mean of the envelope detected image was calculated. With the binary inputs, the mean pixel values of the ground truth and DNN outputs are expected to be approximately zero or one. Similarly, with the random input, the DNN-generated images are expected to produce mean pixel values close to that of the ground truth output.

Image-to-image comparisons for identical input data were evaluated based on L1 and L2 losses:

$$l_1 = \frac{1}{N} \sum_{n=1}^{N} |x_n - y_n| \tag{1}$$

$$l_2 = \sqrt{\frac{1}{N} \sum_{n=1}^{N} |x_n - y_n|^2} \tag{2}$$

where x and y denote the Pytorch DAS and the DNN output image being compared after envelope detection, and N is the total number of pixels.

III. Results and Discussion

A. Baseline evaluation

Fig. 1 shows the output log-compressed B-mode images created with the publicly available Plane-wave Imaging Challenge for Medical Ultrasound (PICMUS) data [12], which was also used to train Network C. The network-produced output images were similar to their respective ground truths, confirming that the networks were loaded properly prior to the overfitting evaluation. In particular, Network C performed well on the dataset used for training of this DNN and generated cleaner images than the ground truth. Without further analysis, it remains a question as to whether this is a true improvement or simply a reflection of overfitting.

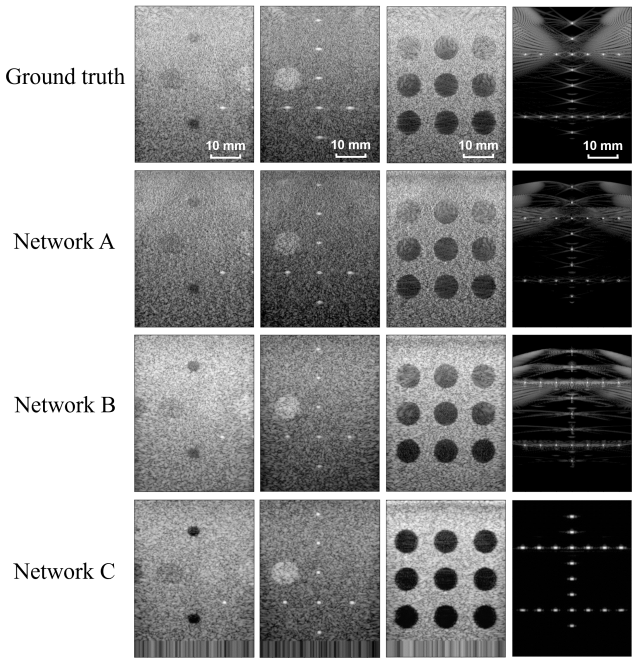

Fig. 1. Baseline evaluation on PICMUS data [12] with images displayed at 60 dB dynamic range.

B. Zeros input

The top row of Fig. 2 shows images created with the zeros input. Networks A and B produced images that look similar to the ground truth. In particular, the point spread function (PSF) of the singular center pixel with a value of 1 seems

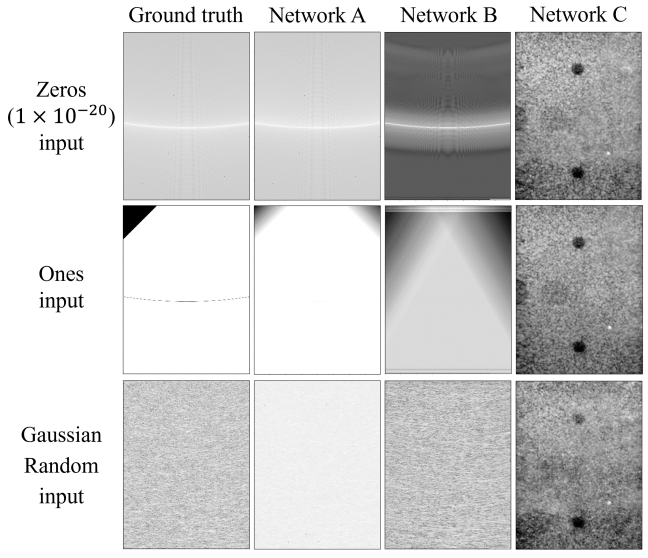

Fig. 2. Network-produced images with artificial RF channel data inputs, including zeros (top), ones (middle), and Gaussian noise (bottom).

TABLE I
MEAN OF ENVELOPE-DETECTED IMAGES AND NUMBER OF TRAINABLE PARAMETERS

	Zeros	Ones	Gaussian Noise	Number of Learned Parameters
Ground truth	0.0052	0.9998	0.2326	0
Network A	0.0068	0.9953	0.0858	3,059
Network B	0.0034	0.8186	0.2366	2,226,146
Network C	0.0871	0.0619	0.0966	54,408,833

TABLE II
L1 LOSS AND L2 LOSS BETWEEN THE GROUND TRUTH AND NETWORK-PRODUCED IMAGES

		Zeros	Ones	Gaussian Noise
L1 loss	Ground truth vs. Network A	0.0018	0.0045	0.1653
	Ground truth vs. Network B	0.0018	0.1812	0.1403
	Ground truth vs. Network C	0.0718	0.9394	0.1849
L2 loss	Ground truth vs. Network A	3.20×10^{-5}	0.0005	0.0415
	Ground truth vs. Network B	3.16×10^{-5}	0.0040	0.0316
	Ground truth vs. Network C	0.0087	0.8846	0.0492

to be represented. However, Network C did not replicate the ground truth PSF and instead created an image that is similar to its training data (see top left of Fig. 1).

Table I reports the mean of envelope detected image output and Table II reports L1 and L2 losses between ground truth and network-produced images. The mean value of pixels in images created with the zeros input are generally similar to their respective ground truths, with the exception of Network C, which shows the greatest deviation from the ground truth. In addition, Network C produced an image that has the largest L1 and L2 losses among the three networks. These qualitative and quantitative results indicate that Network C is overfitting to the training data.

C. Ones input

The middle row of Fig. 2 shows the output images with the ones input. Networks A and B generated images that look like the ground truth. Similar triangular patterns are represented at the top corners. However, Network C created an image similar to one of its training data (see Fig. 1) without reproducing the ground truth.

With the ones input, the mean values of envelope-detected images generated by Networks A and B are similar to that of the ground truth, which is close to one, as shown in Table I. The output image of Network C has a mean value that shows the greatest deviation from the ground truth. Table II shows that Network C produced an image that has the largest L1 and L2 losses among the three neural networks, indicating the worst match between the output image of Network C and the ground truth. The above qualitative and quantitative analyses reveal the overfitting problem of Network C.

D. Gaussian random input

The bottom row of Fig. 2 shows the output B-mode images with the Gaussian random input. Networks A and B produced images that have similar noise samples as the ground truth while Network C still created an image that looks like its associated training data (see Fig. 1).

The mean of the envelope-detected image produced by Network B is close to that of the ground truth while Networks A and C both generated images with greater deviations of mean values from the ground truth, as shown in Table I. The last column of Table II reports similar L1 and L2 losses among the three networks. The above results using the Gaussian random input show that the quantitative measurements are not suitable for identifying overfitting, and qualitative comparisons are more useful in this case.

E. Number of learned parameters

There are various reasons for overfitting of DNNs. Networks with more complexity have the greater potential to overfit [17]. Network complexity is determined by the number of learned parameters (i.e., the number of layers and the number of neurons in each layer) in each network. The same approach employed to obtain the number of learned parameters for Networks A and B [11] was applied to Network C, with all values reported in the last column of Table I. Network C has 1-4 orders of magnitude larger number of trainable parameters compared to that of Networks A and B, which is one of the possible reasons for the overfitting observed with Networks C. Nonetheless, the results in Sections III-B through III-D show that the overfitting of DNNs trained to output ultrasound images can be rapidly identified by the three artificial channel data introduced herein.

IV. CONCLUSION

This work is the first to introduce three types of artificial channel data that are input to DNNs trained to output ultrasound images with the goal of rapidly identifying DNN overfitting. With the binary image inputs, DNN overfitting can be rapidly identified by the qualitative observations, the greatest difference in mean pixel values, and the largest L1 and L2 losses between the network-produced images and the ground truths. With the random image input, overfitting can be rapidly identified by the qualitative observations between the network output and the ground truth. The proposed approach does not require a time-consuming retraining process using the

training code and training data or the collection of additional test data. Instead, images produced by existing DNNs were evaluated after inputting the proposed artificial channel data to provide more rapid identification of network overfitting when compared to traditional overfitting detection approaches. Results demonstrate that the proposed method is promising to be used as a general evaluation approach to overfitting detection with DNNs trained to output ultrasound images, and possibly other types of medical images. In addition, this approach has the potential to provide a new layer of oversight for regulatory bodies tasked with approving the deployment of DNNs on patient data.

REFERENCES

[1] R. J. Van Sloun, R. Cohen, and Y. C. Eldar, "Deep learning in ultrasound imaging," *Proceedings of the IEEE*, vol. 108, no. 1, pp. 11–29, 2019.

[2] A. A. Nair, K. N. Washington, T. D. Tran, A. Reiter, and M. A. L. Bell, "Deep learning to obtain simultaneous image and segmentation outputs from a single input of raw ultrasound channel data," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 12, pp. 2493–2509, 2020.

[3] L. Prechelt, "Automatic early stopping using cross validation: quantifying the criteria," *Neural Networks*, vol. 11, no. 4, pp. 761–767, 1998.

[4] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282, IEEE, 2018.

[5] L. Prechelt, "Early stopping-but when?," in *Neural Networks: Tricks of the trade*, pp. 55–69, Springer, 1998.

[6] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation.," *Encyclopedia of Database Systems*, vol. 5, pp. 532–538, 2009.

[7] C.-J. Wu, D. Brooks, K. Chen, D. Chen, S. Choudhury, M. Dukhan, K. Hazelwood, E. Isaac, Y. Jia, B. Jia, *et al.*, "Machine learning at facebook: Understanding inference at the edge," in *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pp. 331–344, IEEE, 2019.

[8] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, IEEE, 2017.

[9] M. A. L. Bell, J. Huang, A. Wiacek, P. Gong, S. Chen, A. Ramalli, P. Tortoli, B. Luijten, M. Mischi, O. M. H. Rindal, V. Perrot, H. Liebgott, X. Zhang, J. Luo, E. Oluyemi, and E. Ambinder, "Challenge on Ultrasound Beamforming with Deep Learning (CUBDL) Datasets."

[10] M. A. L. Bell, J. Huang, D. Hyun, Y. C. Eldar, R. Van Sloun, and M. Mischi, "Challenge on ultrasound beamforming with deep learning (CUBDL)," in *2020 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–5, IEEE, 2020.

[11] D. Hyun, A. Wiacek, S. Goudarzi, S. Rothlübbers, A. Asif, K. Eickel, Y. C. Eldar, J. Huang, M. Mischi, H. Rivaz, D. Sinden, R. J. Van Sloun, H. Strohm, and M. A. L. Bell, "Deep learning for ultrasound image formation: CUBDL evaluation framework and open datasets," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 68, no. 12, pp. 3466–3483, 2021.

[12] H. Liebgott, A. Rodriguez-Molares, F. Cervenansky, J. A. Jensen, and O. Bernard, "Plane-wave imaging challenge in medical ultrasound," in *2016 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4, IEEE, 2016.

[13] R. Webster, J. Rabin, L. Simon, and F. Jurie, "Detecting overfitting of deep generative networks via latent recovery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11273–11282, 2019.

[14] S. Rothlübbers, H. Strohm, K. Eickel, J. Jenne, V. Kuhlen, D. Sinden, and M. Günther, "Improving image quality of single plane wave ultrasound via deep learning based channel compounding," in *2020 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4, IEEE, 2020.

[15] S. Goudarzi, A. Asif, and H. Rivaz, "Ultrasound beamforming using mobilenetv2," in *2020 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4, IEEE, 2020.

[16] Y. Wang, K. Kempski, J. U. Kang, and M. A. L. Bell, "A conditional adversarial network for single plane wave beamforming," in *2020 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4, IEEE, 2020.

[17] X. Ying, "An overview of overfitting and its solutions," in *Journal of Physics: Conference Series*, vol. 1168, p. 022022, IOP Publishing, 2019.