An improved training scheme for deep neural network ultrasound beamforming

Emelina Vienneau
Department of Biomedical Engineering
Vanderbilt University
Nashville, TN, USA
emelina.p.vienneau@vanderbilt.edu

Adam Luchies
Department of Biomedical Engineering
Vanderbilt University
Nashville, TN, USA
adam.c.luchies@vanderbilt.edu

Brett Byram
Department of Biomedical Engineering
Vanderbilt University
Nashville, TN, USA
brett.c.byram@vanderbilt.edu

Abstract— Deep neural networks have been shown to be effective adaptive beamformers for ultrasound imaging. However, when training with traditional $\mathcal{L}_p$ norm loss functions, model selection is difficult because lower loss values are not always associated with higher image quality. This ultimately limits the maximum achievable image quality with this approach and raises concerns about the optimization objective. In an effort to align the optimization objective with the image quality metrics of interest, we implemented a novel ultrasound-specific loss function based on the spatial lag-one coherence and signal-to-noise ratio of the delayed channel data in the short-time Fourier domain. We employed the R-Adam optimizer with lookahead and cyclical learning rate to make the training more robust to initialization and local minima, leading to better model performance and more reliable convergence. With our custom loss function and optimization scheme, we achieved higher contrast-to-noise-ratio, higher speckle signal-to-noise-ratio, and more accurate contrast ratio reconstruction than with previous deep learning and delay-and-sum beamforming approaches.

Keywords—beamforming, image quality, deep neural networks, loss function, optimization

I. INTRODUCTION

Feed-forward neural networks can approximate any continuous function through nonlinear transformations [1], making them a powerful tool for image reconstruction tasks such as ultrasound beamforming that can be cast as such [2], [3]. Deep neural networks (DNNs) have shown great efficacy in adaptive beamforming by successfully suppressing sources of acoustic clutter such as off-axis scattering in challenging imaging scenarios [3]–[6]. However, our previous work in this area has been limited by the difficulty of model selection. When training with traditional $\mathcal{L}_1$ or mean squared error (MSE) loss functions, we have observed a consistent trend that lower loss upon convergence does not necessarily lead to higher image quality metrics such as contrast-to-noise ratio (CNR), signal-to-noise ratio (SNR), and contrast ratio (CR) reconstruction accuracy [5]. This phenomenon motivated the creation of a new ultrasound-specific loss function.

II. METHODS

A. Loss Function

Training for CNR directly poses a significant challenge with our current deep learning architecture since CNR is calculated on the beamformed data and our neural network operates on delayed channel data in the short-time Fourier domain before summation. In order to incorporate the notion of CNR in the loss function in a computationally efficient manner, we added a spatial lag-one coherence (LOC) term to the original loss function. Lag-one coherence captures the contribution of thermal noise and spatially incoherent acoustic noise, providing a local measure of image quality [7]. We also added a noise-to-signal ratio (NSR) term to the loss function. The LOC and NSR terms do not include any information about overall signal amplitude and are based solely on the training data, so in order to retain amplitude information in a supervised learning approach, it is necessary to include a data fidelity term such as smooth $\mathcal{L}_1$ or MSE.

Using the notation of [8], the lag-one ensemble correlation was calculated on the delayed channel data in the short-time Fourier transform domain according to (1), where ξ_1 indicates the set of all lag-one pairs of channels and * denotes the complex conjugate. An axial kernel of one sample was used to improve computational efficiency as suggested in [8].

$$\text{LOC} = \hat{R}_1^{ens} = \frac{\Sigma_{\xi_1} ab^*}{\sqrt{\left(\Sigma_{\xi_1}|a|^2\right)\left(\Sigma_{\xi_1}|b|^2\right)}} \qquad (1)$$

The overall coherence term used in the loss function was calculated as in (2) in order to optimize for the maximum achievable lag-one coherence according to the van Cittert-Zernike theorem [9]. M is the number of elements (65 in this case) and λ_1 is a weighting factor that must be tuned to determine the term's relative contribution to the entire loss function.

$$\lambda_1 \left| \left(1 - \tfrac{1}{M}\right) - \text{LOC} \right| \qquad (2)$$

The NSR term used in the loss function was calculated as in (3), where λ_2 is the weighting factor.

$$\lambda_2 \left(\tfrac{\sigma}{\mu}\right) \qquad (3)$$

The final loss function employed used smooth $\mathcal{L}_1$ instead of MSE due to its superior performance.

$$\text{Loss} = \text{Smooth } \mathcal{L}_1 + \lambda_1 \left| \left(1 - \tfrac{1}{M}\right) - \text{LOC} \right| + \lambda_2 \left(\tfrac{\sigma}{\mu}\right) \qquad (4)$$

B. Optimization

Adding LOC and NSR terms to the original loss function should allow us to maximize the image quality metrics of interest, but in practice it seemed to create a rough loss landscape

full of sub-optimal local minima that was very difficult to optimize using traditional approaches. To address this, we began using the R-Adam optimizer, an adaptive optimizer based on the standard Adam optimizer that includes built-in warmup [10], [11]. Warmup reduces the variance in adaptive momentum calculations at the beginning of training, leading to stability and robustness to bad initializations. In the case of R-Adam, the warmup is itself adaptive and is based on the underlying variance of the data [10].

To further improve the long-term stability of the optimization while still converging relatively quickly, we employed lookahead [12]. Lookahead keeps two sets of weights: one from a slower version of the optimizer and one from a faster version of the optimizer that "looks ahead" by taking k additional training steps for every single training step the slower optimizer takes. The weights of the slower optimizer are updated in the direction of the faster optimizer's weights via linear interpolation. How far in that direction the weights are updated depends on a hyperparameter α (set to 0.55). After updating the slower optimizer's weights, the faster optimizer starts over again at the same place before advancing ahead k steps (set to 5). This optimization scheme has been shown to provide the long-term stability of stochastic gradient descent optimization with the convergence speed of adaptive optimizers such as Adam [12]. We used an open-source PyTorch implementation of lookahead [13].

Finally, we utilized a cyclical learning rate scheduler to add additional robustness throughout the entire training process [14]. We found that its stabilization effects were complimentary to those of R-Adam and lookahead. Reference [15] also reports added benefit of using a scheduled learning rate multiplier with any adaptive optimizer, corroborating our observations. We used a triangular mode, a base learning rate of 0.001, and a maximum learning rate of 0.003.

C. Training, Validation, and Testing Data

For a detailed description of the Field II simulation scheme, please see Ref. [6]. Briefly, the training data consisted of 24 hypoechoic targets each generated with a native contrast of -22, -24, -26, -28, and -30 dB, yielding a total of 120 unique cysts. All samples inside of the cysts were included in training as well as the same number of samples in the background in order to keep balanced classes. Seventy-five percent (43,344 samples) were used for training and 25% (14,448 samples) were used for validation during training. The hypoechoic cyst test set was separate from the training set and included five hypoechoic targets each with a native contrast of -22, -24, -26, -28, and -30 dB. The training data were processed with the short-time Fourier transform for input into the network as in Ref. [6].

D. Model Architecture

The deep neural network was a feed-forward network with 30 hidden layers and 1000 nodes in each layer. ReLU activations functions were used and batch normalization was applied during training before the activation function. Dropout of 0.2 probability was used for the input layer and each successive hidden layer during training. The batch size was 1024 samples. Since the number of active elements in the transducer was 65,

the size of the input and output layers were 130 x 1, where 130 is made up of the real (in-phase) values for each element concatenated with the imaginary (quadrature) values for each element. The minimum number of epochs for training was 300, after which training would cease when the validation loss did not improve after 50 epochs. For testing and model selection, the loss, CNR, SNR inside and outside the cyst, and CR were each averaged across the five targets for each native contrast ratio case. The beamformer that performed the best on all of these measurements would then be selected.

III. RESULTS

Fig. 1 shows results from standard delay-and-sum (DAS) beamforming, the proposed approach ($\lambda_1 = 0.02$, $\lambda_2 = 0.01$), and three sub-optimal DNN approaches. DNN approach #1 used a traditional optimization scheme (Adam optimizer with a fixed learning rate of 0.001) with a traditional loss function (smooth $\mathcal{L}_1$). DNN approach #2 used the proposed optimization scheme (R-Adam with lookahead and cyclical learning rate) with a traditional loss function (smooth $\mathcal{L}_1$). DNN approach #3 used the proposed loss function with a traditional optimization scheme. Careful inspection of Fig. 1 reveals a much improved speckle pattern when using the proposed approach, particularly inside the cyst.

Table 1 shows the image quality metrics for each of the beamforming approaches shown in Fig. 1. The letters *a* and *b* refer to using MSE loss and smooth $\mathcal{L}_1$ loss, respectively. To ensure a fair comparison of each approach, three beamformers with the same hyperparameters were trained for all four DNN approaches and their results were averaged for all 25 hypoechoic cysts in the test set. To indicate the contrast ratio reconstruction accuracy across the -22 to -30 dB range, the mean squared error between the model's contrast ratio for each cyst and the true value was computed. The speckle SNR (SNRs) was computed inside and outside the cyst. Table 1 clearly shows the benefit of employing the proposed optimization scheme and loss function. From the optimization strategy alone, there is an improvement in CNR of about 0.8 dB over the previous DNN approach, just barely surpassing DAS. From the loss function alone, there is a bump in SNRs inside the cyst of about 0.4 dB, showing that the cyst is being reconstructed more accurately. The contrast ratios of these DNN approaches have huge biases, leading to large CR MSE values. However, the full benefit of the loss function is not realized until it is combined with the optimization strategy. CNR, SNRs, and CR accuracy for the proposed approach are higher than the previous DNN approaches and DAS, showing the benefit of our method.

TABLE I.   IMAGE QUALITY METRICS FOR BEAMFORMING APPROACHES

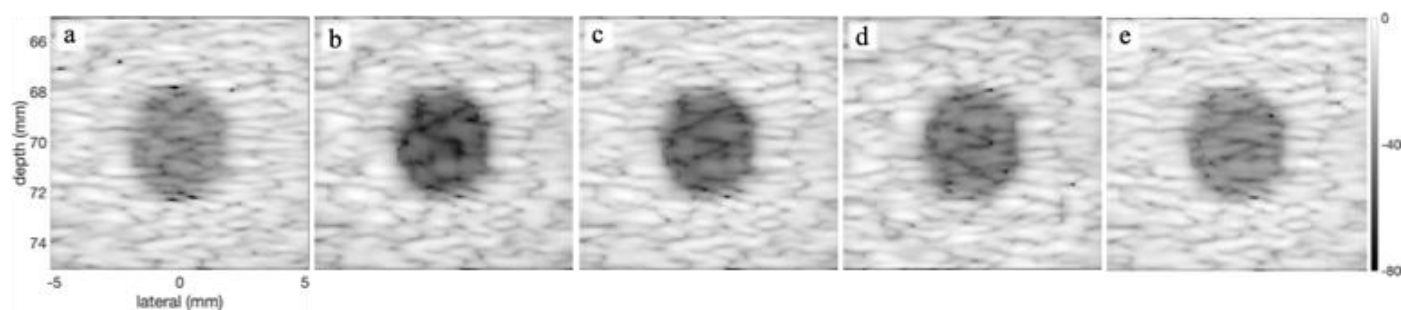| Method	Image Quality Metric (dB)			
	CR MSE	CNR	SNRs outside	SNRs inside
DAS	8.8757	5.3235	1.9923	1.7834
DNN 1a	127.1490	4.5826	1.7236	1.1563
DNN 1b	128.5730	4.6117	1.7282	1.1315
DNN 2a	59.7852	5.3678	1.9073	1.3541
DNN 2b	65.7790	5.4572	1.9166	1.3430
DNN 3	12.5730	4.5085	1.7442	1.7782
Proposed	**3.6053**	**5.6648**	**2.0395**	**2.0747**

Fig. 1. *Qualitative comparison of beamforming approaches on a -26 dB hypoechoic cyst. Note the differences in speckle pattern on the inside of the cyst. a) delay-and-sum beamforming. b) DNN approach #1: smooth $\mathcal{L}_1$ loss with traditional optimization scheme. c) DNN approach #2: smooth $\mathcal{L}_1$ loss with proposed optimization scheme. d) DNN approach #3: proposed loss function with traditional optimization scheme. e) Proposed deep learning approach with custom loss function and advanced optimization scheme. All images are shown on the same dynamic range and same axes. The colorbar is in units of dB.*

IV. CONCLUSION

In summary, we have shown that an ultrasound-specific loss function is useful for maximizing image quality metrics of interest and ultimately improving the maximum achievable image quality. An appropriate optimization scheme that provides sufficient robustness to local minima and bad initializations leads to reliable convergence despite the complexity of the loss landscape.

REFERENCES

[1] K. Hornik and M. Stinchcombe, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, Jan. 1989.

[2] B. Byram, K. Dei, J. Tierney, and D. Dumont, "A model and regularization scheme for ultrasonic beamforming clutter reduction," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 62, no. 11, pp. 1913–1927, 2015.

[3] A. C. Luchies and B. C. Byram, "Deep neural networks for ultrasound beamforming," *IEEE Trans. Med. Imaging*, vol. 37, no. 9, pp. 2010–2021, 2018.

[4] A. C. Luchies and B. C. Byram, "Suppressing off-axis scattering using deep neural networks," *Proc. SPIE Med. Imaging 2018 Ultrason. Imaging Tomogr.*, vol. 10580, 2018.

[5] A. C. Luchies and B. C. Byram, "Training improvements for ultrasound beamforming with deep neural networks," *Phys. Med. Biol.*, vol. 64, no. 4, pp. 1–15, 2019.

[6] A. Luchies and B. C. Byram, "High dynamic range ultrasound beamforming using deep neural networks," in *Medical Imaging 2019: Ultrasonic Imaging and Tomography*, 2019, vol. 10955, p. 19.

[7] W. Long, N. Bottenus, and G. E. Trahey, "Lag-one coherence as a metric for ultrasonic image quality," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 65, no. 10, pp. 1768–1780, 2018.

[8] D. Hyun, A. L. C. Crowley, and J. J. Dahl, "Efficient strategies for estimating the spatial coherence of backscatter," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 64, no. 3, pp. 500–513, 2017.

[9] R. Mallart and M. Fink, "The van Cittert-Zernike theorem in pulse echo measurements," *J. Acoust. Soc. Am.*, vol. 90, no. November, pp. 2718–2727, 1991.

[10] L. Liu *et al.*, "On the variance of the adaptive learning rate and beyond," *arXiv Prepr. arXiv1908.03265*, pp. 1–14, 2019.

[11] D. P. Kingma and J. Lei Ba, "Adam: a method for stochastic optimzation," *Proc. Int. Conf. Learn. Represent.*, pp. 1–15, 2015.

[12] M. R. Zhang, J. Lucas, G. Hinton, and J. Ba, "Lookahead optimizer: k steps forward, 1 step back," *arXiv Prepr. arXiv1907.08610*, pp. 1–16, 2019.

[13] LonePatient, "Lookahead-PyTorch." GitHub Repository, 2019. https://github.com/lonePatient/lookahead_pytorch

[14] L. N. Smith, "Cyclical learning rates for training neural networks," *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, pp. 464–472, 2017.

[15] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *Proc. Int. Conf. Learn. Represent.*, pp. 1–8, 2019.