SAND2020-13226C

Confronting Domain Shift in Trained Neural Networks

Carianne Martinez
Sandia National Laboratories
Albuquerque, NM
Arizona State University
Tempe, AZ
cmarti5@sandia.gov

David A. Najera-Flores
Sandia National Laboratories
Albuquerque, NM
danajer@sandia.gov

Adam R. Brink
Sandia National Laboratories
Albuquerque, NM
arbrink@sandia.gov

D. Dane Quinn
University of Akron
quinn@uakron.edu

Eleni Chatzi
ETH Zürich
Zürich, Switzerland
chatzi@ibk.baug.ethz.ch

Stephanie Forrest
Arizona State University
Tempe, AZ
steph@asu.edu

Abstract

Neural networks (NNs) are known as universal function approximators and are excellent interpolators of nonlinear functions between observed data points. However, when the target domain for deployment shifts from the training domain and NNs must extrapolate, the results are notoriously poor. Prior work [1] has shown that NN uncertainty estimates can be used to correct binary predictions in shifted domains without retraining the model. We hypothesize that this approach can be extended to correct real-valued time series predictions. As an exemplar, we consider two mechanical systems with nonlinear dynamics. The first system consists of a spring-mass system where the stiffness changes abruptly, and the second is a real experimental system with a frictional joint that is an open challenge for structural dynamicists to model efficiently. Our experiments will test whether 1) NN uncertainty estimates can identify when the input domain has shifted from the training domain and 2) whether the information used to calculate uncertainty estimates can be used to correct the NN's time series predictions. Success of the proposed technique would unleash the potential of previously underutilized latent features already present in trained NNs and enable the deployment of these models in structural health monitoring systems that directly impact public safety.

1 Introduction

NNs have seen great success in accurately modeling nonlinear functions by learning directly from observed data. Techniques such as Transformers [2] and Long Short Term Memory (LSTM) [3] models have been applied to sequential data and have demonstrated impressive capabilities in the field of natural language processing (NLP) [4], excelling at tasks such as language translation [2] and answering text based questions [5]. These models have been extended to scientific domains where physical laws govern the dynamics of a system [6], [7]; however, while the performance of a NN may be acceptable when the target domain is closely aligned with the training domain, its performance may degrade when the target domain deviates significantly from the training set. This limitation prevents them from use in high consequence environments such as those monitored by structural health monitoring (SHM) systems, where system failure directly implies that the dominant physics of the system shifts, and indications of this failure must be identified and mitigated to ensure public safety.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Techniques to improve deep learning (DL) model performance on targets that have shifted from the training domain have been proposed in the literature and will be discussed in Section 2. These methods often augment the training data set to more closely match the target deployment domain. They require expensive retraining of models and are not feasible when rapid approximations of system dynamics are necessary. **Our approach removes the need for additional data or training by leveraging information that already exists in the weights of the trained model, realized in the form of uncertainty estimation.** The exemplars set forth herein require efficient approximations of future system states and are critical for understanding the risks associated with deploying systems for industries like aviation [6]. Prior work [1] introduced a technique to avoid the need for retraining DL models while extending their applicability to shifted target domains. Results from this work indicated that when the most uncertain predictions were flipped, segmentations were significantly improved. An example of a result from this technique is shown in Figure 1 where a NN trained on a particular image domain is extended for use in a shifted domain with improved predictive capability.

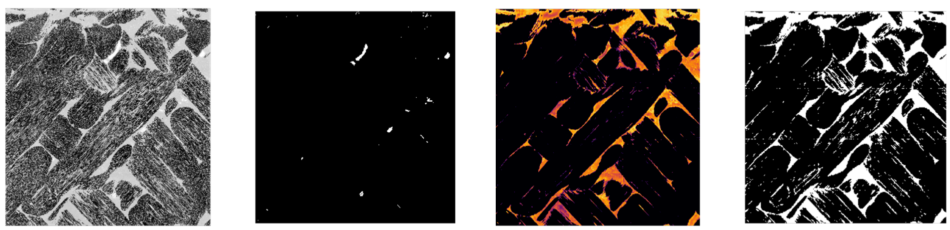

Figure 1: Results from [1] showing that uncertainty maps can be used directly to improved trained NN predictions. (a) Slices of CT scan to be segmented. (b) Predicted binary label for the CT slice from the trained NN without UQ correction. (c) Uncertainty map (brighter colors indicate higher uncertainty). (d) Resulting binary labels after UQ informed improvement.

We hypothesize that this technique can be extended from binary classification to time-dependent regression, where patterns in the sequential input to the DL model can be used to 1) identify that domain shift is occurring and 2) improve the DL model's prediction without retraining. The anticipated contributions of this work are:

- A practical method for applying DL models to time series in shifted domains
- New publicly available datasets from the structural dynamics field of well-defined physical systems
- Open source code implementation that allows replication and extension of our experimental results

2 Related work

The overfitting of DL models to a specific training domain is a known weakness of NNs, and current research efforts seek to overcome this shortfall. Here we review work on domain shift, DL uncertainty quantification, and the structural dynamics involved in our training domain.

2.1 DL model domain shift mitigation and uncertainty quantification

The problem of domain shift from a training domain to a target domain is an open and active area of DL research. Existing works focus on data augmentation, retraining models to better generalize, and training additional models. [8] adds a CORAL loss function that works to effectively transform the features in the network itself to be relevant to a shifted domain. This approach requires unlabeled examples of the shifted domain to learn transformations in the feature space that will reduce the CORAL loss. Li, et al. [9] uses a generative model to [10] augment the data necessary to perform well in a shifted deployment domain. CyCADA [11] also employs a generative model to align the shifted domain with the training domain using both pixel-level and feature-level transformations. In [12], a likelihood ratio is introduced to overcome background statistics that are shown to drive overconfidence in generative model predictions. This method requires training of an additional background-specific model. Domain adaptation techniques [13],[14] can also mitigate shifts in data by training separate models to preprocess the shifted inputs to more closely match the training domain. All of these approaches require additional resources, but in contrast, our proposed method actively uses uncertainty estimates to correct DL model predictions without retraining. Surveys of modern techniques for anomaly detection [15], [16] are also relevant as these approaches could be applied to detecting domain shift.

Several methods have been proposed to quantify uncertainty in DL model predictions. These include ensemble methods [17], Bayesian NNs [18], and dropout networks [19]. We implement dropout networks in this work to quantify the uncertainty in DL model predictions due to their ease of implementation and their effectiveness with only a single model to be trained.

2.2 Structural dynamics modeling and structural health monitoring (SHM)

We obtain our exemplar datasets from the field of structural dynamics where applications such as reduced order modeling of complex systems and SHM require real-time detection of anomalous system behavior. In addition to a mechanical example where the system stiffness shifts dramatically, we will utilize experimental data from a jointed structural system. Frictional joints are well-studied, [20] but current reduced order models (ROMs) cannot practically capture the full extent of the nonlinear physics. The proposed corrective mechanism would advance modeling capabilities.

SHM is defined as a four-level hierarchy [21] [22] aiming to detect, localize, quantify, and finally predict damage on the basis of data extracted from operating engineered systems. In doing so, a large body of recent literature explores utilization of ML and DL methods for damage prognosis. Many existing works focus on outlier classification for damage detection [23]. Generative modeling approaches attempt to reproduce joint probabilistic distributions from monitoring data in order to recognize distinct condition regimes [24]. For achieving the higher steps in the SHM hierarchy, physics-informed learning incorporates domain knowledge into the learning process [25]. In this work, we treat this problem as adaptation to shifted domains.

3 Methodology

When a NN is trained to mimic time series data, it learns a mapping from patterns observed in previous timesteps to the next data point in the time series. When time series deviates from the expected patterns, the NN could fail to make accurate predictions. If successful, our method will extend the applicability of trained NNs to mitigate domain shift by 1) recognizing that the input domain has shifted and 2) using uncertainty quantification to drive the predictions toward a corrective path.

Our method assumes that a NN with dropout layers used to quantify the uncertainty in its predictions is trained to approximate a real-valued function $f(x, t)$. Input to the model is a sequence of values of f over a series of previous timesteps along with the value of x at time t, and output is the value of f over a sequence of subsequent timesteps. When the model's uncertainty exceeds a threshold value, instead of returning the model's nominal prediction for f at time t, our method updates the prediction to incorporate information from the calculated uncertainty to improve accuracy. Using the dropout technique set forth in [19], we infer several predictions for f at time t with different subsets of neuron outputs dropped from the calculation, resulting in a distribution of predicted output values at each time step. Rather than leaving the uncertainty estimation as a simple indication of the model's confidence at time t, our method actively uses statistical properties of the distribution to serve as a corrective factor for the prediction of f at time t. We will explore two corrective methods in this work: 1) We replace the nominal prediction with the mean of the prediction distribution and 2) We add the standard deviation of the prediction distribution to the nominal prediction in the direction of the distribution skew.

4 Experimental protocol

We will use two structural dynamics datasets to test our hypotheses and report results from two DL models. For both datasets, our intent is to answer the following questions:

- RQ1: Does the uncertainty value correctly detect a significant change in the model's accuracy?
- RQ2: Does the corrective factor informed by the uncertainty improve the accuracy of the prediction?

4.1 Datasets

We will first investigate our method's performance on a toy problem consisting of data drawn from simulations of a mass-spring system with one mass element and a fixed stiffness with varying initial conditions, and loaded under a known force. We will also generate simulated data where the stiffness of the spring abruptly changes. The data will consist of a time series of the force on the mass as well as the displacement of the mass, and initial system conditions.

A more challenging dataset will be derived from experimental measurements of a frictional jointed structure subject to a known force. A schematic of the system is shown in Figure 2. This dataset will include the initial conditions, the load on the structure, and accelerometer and displacement measurements from various positions in the structure. We will also develop a reduced order model (ROM) of the system that will predict the displacement over time of structural mass elements. The ability of ROMs for jointed structures to match experimental data is known to degrade as the structural loading on the joint increases and the nonlinear dynamics induced by the joint become more significant.

Each dataset will consist of approximately 100,000 timesteps per example, and the simulations will use on the order of 100 different initial conditions.

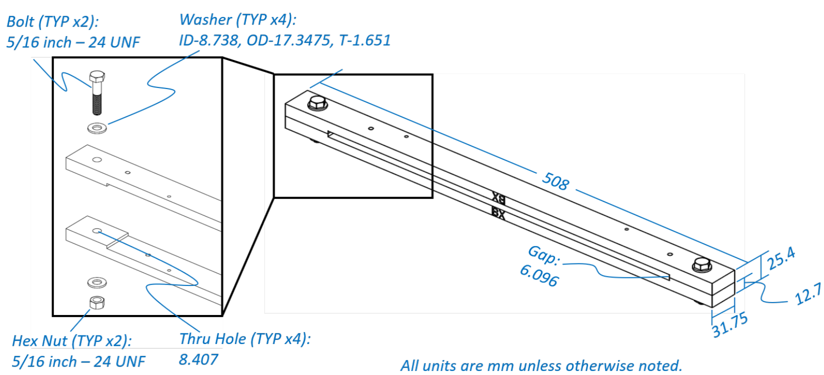

Figure 2: Schematic of jointed structural system from [26] used to obtain displacement dataset.

4.2 Model training

We will implement both a WaveNet with a stack of dilations of size [1,2,4,8] and a receptive field of length 128 as in [27] and a Transformer with the base model architecture as presented in [2], each of which have seen success in predicting sequential data. For WaveNet, we will apply dropout to all convolutional layers. For Transformer, we will apply dropout only to the decoder portion of the network, since we have observed that dropout in encoding layers removes input information necessary for useful encodings. Each model will be evaluated on both datasets.

For the mass-spring system, our DL models will be given the system's initial conditions, the force on the mass elements at each timestep as well as the displacement of the mass elements over a series of previous timesteps, and will be used to predict the displacements at the next timestep. After training on several examples from this system, we will introduce an input series to the trained model that simulates an abrupt change to the spring's stiffness and apply our corrective factor to improve the predictions of the mass displacement.

For the jointed structure, our DL models will be trained to learn the system dynamics solely from the ROM data and will learn to predict the displacement of each discretized mass element modeled by the ROM. We will then apply our trained DL model on the experimental structure data, where the output with the corrective factor will be used to predict the next timestep of the displacements in the real structure. The key idea here is that our ROM will be unable to capture all of the physics necessary to predict the true system dynamics, and that our DL model will identify that the real inputs have shifted from the training domain, and compensate for the missing physics.

One of the primary challenges of employing neural network for predictions in the time domain is the accumulation of error that arises from recursion. To mitigate this challenge, we will enforce physical constraints through the loss function. Terms that require conservation of energy and momentum will encourage the network to learn not only the target output, but its derivatives and the relationship between them. A byproduct of this constraint is that the problem is bounded to produce high-quality predictions in the physical domain in which it was trained. When presented with data from outside its domain, the prediction uncertainty will increase as the physical constraints are harder to enforce.

4.3 Evaluation and significance

We are interested in the impact of our method on the accuracy of sequential predictions, and first must establish baseline behavior of each DL model. We will use the Adam optimizer [28] with learning rate 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and ϵ = 1e-8 (the default Keras [29] settings) with dropout rate of 0.1 for both training and inference to calculate uncertainty. We will exhaustively evaluate each baseline DL model over input sequences of 32, 64, and 128 time steps and output sequence lengths of 1,2,3, and 4. We have proposed these specific hyperparameter settings for concreteness, but we intend to explore other settings such as the dropout rate and the uncertainty threshold value as appropriate to establish baseline performance of the models. We will use the most accurate WaveNet and Transformer model to evaluate the efficacy of our corrective method by calculating the mean squared error with respect to the ground truth sequences with and without our method's corrective factor. From the output of the two competing models, we will estimate the distributions of residuals to quantify the statistical significance of our model improvements using a dependent two-sample t-test or the Wilcoxon-Mann-Whitney U-test as applicable.

When employing our method, we will make 48 predictions for each timestep and use the distribution of predictions to correct the prediction if the standard deviation of the distribution of predictions exceeds 10% of the value of the nominal prediction for the next time step. We will explore two corrective factors: 1) the mean of the predictions and 2) the addition of the standard deviation in the direction of the skew of the prediction distribution.

While DL remains a powerful tool for modeling complex systems, its inability to overcome domain shift severely limits its successful deployment. If we achieve a positive result from these experiments, we will unlock the potential of repurposing unused latent features for improved DL generalization.

Acknowledgments and Disclosure of Funding

We would like to thank Jessica Jones, Justin Jacobs, Benjamin Edwards, Drew Levin, Kevin Potter, and Demitri Maestas for their insightful comments and suggestions.

Supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.

References

[1] C. Martinez, K. M. Potter, M. D. Smith, E. A. Donahue, L. Collins, J. P. Korbin, and S. A. Roberts, "Segmentation certainty through uncertainty: Uncertainty-refined binary volumetric segmentation under multifactor domain shift," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[4] D. W. Otter, J. R. Medina, and J. K. Kalita, “A survey of the usages of deep learning for natural language processing,” *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[6] D. A. Najera-Flores and A. R. Brink, “Efficient random vibration analysis of nonlinear systems with long short-term memory networks for uncertainty quantification,” in *Proceedings of ISMA 2018 International Conference on Noise and Vibration Engineering and USD2018 International Conference on Uncertainty in Structural Dynamics*, 2018.

[7] T. Simpson, N. Dervilis, and E. Chatzi, “On the use of nonlinear normal modes for nonlinear reduced order modelling,” *arXiv preprint arXiv:2007.00466*, 2020.

[8] B. Sun and K. Saenko, “Deep coral: Correlation alignment for deep domain adaptation,” in *European Conference on Computer Vision*, pp. 443–450, Springer, 2016.

[9] X. Li, W. Zhang, and Q. Ding, “Cross-domain fault diagnosis of rolling element bearings using deep generative neural networks,” *IEEE Transactions on Industrial Electronics*, vol. 66, no. 7, pp. 5525–5534, 2018.

[10] Z.-H. Zhou and M. Li, “Tri-training: Exploiting unlabeled data using three classifiers,” *IEEE Transactions on Knowledge & Data Engineering*, no. 11, pp. 1529–1541, 2005.

[11] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, “Cycada: Cycle-consistent adversarial domain adaptation,” *arXiv preprint arXiv:1711.03213*, 2017.

[12] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. Depristo, J. Dillon, and B. Lakshminarayanan, “Likelihood ratios for out-of-distribution detection,” in *Advances in Neural Information Processing Systems*, pp. 14707–14718, 2019.

[13] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

[14] O. Sener, H. O. Song, A. Saxena, and S. Savarese, “Learning transferrable representations for unsupervised domain adaptation,” in *Advances in Neural Information Processing Systems*, pp. 2110–2118, 2016.

[15] R. Wang, K. Nie, T. Wang, Y. Yang, and B. Long, “Deep learning for anomaly detection,” in *Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 894–896, 2020.

[16] M. Braei and S. Wagner, “Anomaly detection in univariate time-series: A survey on the state-of-the-art,” *arXiv preprint arXiv:2004.00433*, 2020.

[17] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” in *Proceedings of the 30th Conference on Advances in Neural Information Processing Systems*, 2017.

[18] R. M. Neal, *Bayesian learning for neural networks*, vol. 118. Springer Science & Business Media, 2012.

[19] Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” in *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

[20] J. Bickford, *Introduction to the Design and Behavior of Bolted Joints*, vol. 4th Edition. CRC Press, 1926.

[21] A. Rytter, *Vibrational based inspection of civil engineering structures*. PhD thesis, Dept. of Building Technology and Structural Engineering, Aalborg University, 1993.

[22] C. R. Farrar and N. A. Lieven, “Damage prognosis: the future of structural health monitoring,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 365, no. 1851, pp. 623–632, 2007.

[23] L. Bull, T. Rogers, C. Wickramarachchi, E. Cross, K. Worden, and N. Dervilis, “Probabilistic active learning: An online framework for structural health monitoring,” *Mechanical Systems and Signal Processing*, vol. 134, p. 106294, 2019.

[24] C. Mylonas, I. Abdallah, and E. Chatzi, “Deep unsupervised learning for condition monitoring and prediction of high dimensional data with application on windfarm scada data,” in *Model Validation and Uncertainty Quantification, Volume 3*, pp. 189–196, Springer, 2020.

[25] F.-G. Yuan, S. A. Zargar, Q. Chen, and S. Wang, “Machine learning for structural health monitoring: challenges and opportunities,” in *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2020*, vol. 11379, p. 1137903, International Society for Optics and Photonics, 2020.

[26] A. R. Brink, R. J. Kuether, M. D. Fronk, B. L. Witt, and B. L. Nation, “Contact stress and linearized modal predictions of as-built preloaded assembly,” *Journal of Vibration and Acoustics*, vol. 142, no. 5, 2020.

[27] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, 2016.

[28] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

[29] F. Chollet *et al.*, "Keras." `https://keras.io`, 2015.