Capabilities Statement

PI: Hao Wang, Assistant Professor, Department of Computer Science, Rutgers University.
Email: hw488@cs.rutgers.edu. Homepage: `http://www.wanghao.in`

1 Technical Expertise

Probabilistic and Interpretable Deep Learning. PI Wang's Ph.D. thesis introduced one of the first Bayesian deep learning frameworks [36, 33, 27, 31] that unify DL, which learns representations efficiently from high-dimensional signals, and probabilistic graphical models (PGM) [24, 14], which handle complex conditional dependencies and uncertainty in data. Since 2014, the framework has impacted the design of models in various domains and spawned hundreds of follow-up works [37]. In particular, it has led to the first adaptive defense method against adversarial attack [21], the first causal explainer for graph neural networks [16] (Best Paper Finalist at CVPR'22), and the first deep hybrid recommender system [33, 34] (most cited paper at ACM SIGKDD 2015 and deployed in Amazon Personalize [1]). Besides, our framework has been integrated into a interpretable health monitoring system to improve patients' medication adherence (published in Nature Medicine) [49] and detect Parkinson's disease (Ten Notable Advances in Nature Medicine 2022) [43].

Causal and Adversarial Robustness. PI Wang and his colleagues have been developing new causal inference algorithms [35, 22, 23, 16, 38] to improve DL models' robustness to spurious, confounding, and incomplete data. For example, we have explored how causal inference improves visual recognition models' robustness to spurious features [22], boosts their generalization performance across environments [23], and produces causally robust explanations (against spurious features) for graph neural network predictions [16] (Best Paper Finalist at CVPR'22). We have developed methods to improve DL models' adversarial robustness using regularized projected gradient descent (PGD) [21] and to improve DL robustness to distribution shift using domain adaptation [19, 40, 28, 41], invariant risk minimization [17], and causal transportability [23, 22].

Deep Learning Applications. PI Wang also brings a unique combination of expertise in machine learning, data mining, computer vision, and healthcare. Besides foundational methods on probabilistic neural networks [31, 30, 28], he has created algorithms for healthcare [49, 45, 46], network analysis [16, 32, 15], forecasting [25, 26], image generation [29, 18, 8], and speech recognition [12, 11]. Notably his work, Convolutional LSTM (ConvLSTM) [25], has become one of today's *de facto* algorithms for spatio-temporal forecasting (deployed in Microsoft's MSN Weather).

Correcting Algorithmic Bias against Under-Exposed Items. We have extensive expertise in investigating theoretical guarantees of correcting algorithmic bias with minimal accuracy sacrifice by adjusting DL training objectives. For example, we developed a set of unbiased, low-variance estimators and the associated new loss functions to correct exposure bias (a type of algorithmic bias) when training DL models for link recommendation (e.g., recommending articles to cite for a given abstract) [7]. Due to exposure bias, items enjoying more exposure to users (e.g., articles from high-profile research groups) contain more positive labels, leading the DL model to recommend these over-exposed items more frequently and creating a feedback loop; this is unfair for under-exposed items (e.g., articles from smaller research groups). Our preliminary work [7] successfully corrects such bias and improves fairness for under-exposed items.

Correcting Data Bias against Minority Groups. We have expertise in correcting data bias by adjusting DL representations. For example, we developed a set of deep imbalanced learning methods [39, 44, 42] to correct such bias in single-domain [44], multi-domain [42], and uncertainty-aware [39] settings. Due to data bias, DL models tend to have higher accuracy in majority groups where labeled data is abundant; this is unfair for minority groups, and such unfairness exacerbates in multi-domain settings. Our preliminary work [44, 42] successfully corrects such data bias in DL models and improves fairness for minority groups.

Large Language Models. Large language models (LLMs) are typically built on the Transformer architecture [2, 3]. PI Wang has extensive expertise in such an architecture and developed the first Trasnformer models for interpretable painting generation [18], cross-domain forecasting [13], and earth system forecasting (including precipitation and weather forecast) [6] as well as the first prompt tuning for test-time domain adaptation and transfer learning [5]. In terms of LLMs, PI Wang and his colleagues are among the first to study the bias introduced by LLM embeddings [7]. They also developed the first LLM-based zero-shot recommender system [4], the first LLM that directly runs as a recommender system [48], and the first LLM that enables privacy-aware generation on the cloud [47].

2 Past Performance

PI Wang has an active grant from NSF IIS-2127918 "Enabling Interpretable AI via Bayesian Deep Learning" (10/2021-09/2024; $499,926) as PI. *Intellectual Merits:* The project establishes a general "interpreter" framework for modern deep learning with theoretical guarantees via Bayesian deep learning [35, 40, 9, 10, 19, 41, 16, 20]. *Broader Impacts:* Throughout the project, the PI has been disseminating results, promote the awareness of interpretability in AI, and enhance the understanding of relevant techniques by presentations in conferences/seminars/classes. The PI also organized related workshops at machine learning conferences (e.g., NeurIPS, ICML, and ICLR).

3 Facilities and Equipment

Laboratory. The Department of Computer Science of Rutgers University provides lab room for the research group of the PI and his graduate students.

Computing and Testing Facility. Rutgers University provides computational environments that will support this work. The PI and his group can use local servers and clusters, and machines with multi-core architectures, including an iLab computing cluster with 1TB memory, 80 cores, and 8 1080Ti GPUs, the Aurora CPU/GPU High Performance Computing Cluster, local GPU servers with A100 and A5000 GPUs (to support LLM training), and Dell desktop computers. The clusters will provide storage and management of large datasets (up to 50TBs of data), and computational resources for the proposed research and student training via courses. Another computing cluster with 64-core CPU processors, 4 GPU processors, 128GB memory, 4TB storage space, and extended 50TB virtual mount storage space, will be used solely for the proposed research.

Data and Information Security. The accessibility, integrity and confidentiality is important to the appropriate storage and maintenance of data. Rutgers Department of Computer Science has a professional IT support team to ensure the privacy and integrity of all data. All servers and systems are consciously scanned and protected from potential vulnerabilities. Systems are patched immediately, and virus protection is updated weekly or more frequently when necessary. Data encryption mechanisms are available for users if desired. Access control lists are maintained, and the IT staff monitor all systems daily.

4 Relevance

Our team, led by PI Wang, possesses a robust blend of expertise that aligns closely with the goals and aspirations of the BENGAL initiative. Below is how our proficiencies cater to the U.S. Government's interest in the responsible and efficient deployment of LLMs:

1. **Understanding and Mitigating LLM Threats.** PI Wang's extensive grounding in probabilistic and interpretable deep learning, particularly Bayesian frameworks that bridge DL and probabilistic graphical models, forms a sturdy foundation to explore, understand, and quantify LLM threat modes. His significant contributions, like the adaptive defense method against adversarial attacks and the causal explainer for graph neural networks, attest to our team's capability to devise innovative techniques for threat detection and mitigation.
2. **Probing and Detecting Vulnerabilities.** With our history of addressing adversarial robustness, notably via regularized projected gradient descent, our team is primed to develop technologies that efficiently probe LLMs. The goal is to identify and rectify biases, vulnerabilities, and potential hazards, ensuring model resilience even when exposed to unforeseen challenges.
3. **Addressing Biases in LLMs.** PI Wang's pioneering efforts in correcting algorithmic biases, especially in under-exposed items and among minority groups, are directly pertinent to BENGAL's objective of avoiding unwarranted biases in LLM applications. This expertise will be instrumental in ensuring fairness and eliminating toxic outputs from LLMs.
4. **Extensive LLM Expertise.** PI Wang's profound knowledge of LLMs, especially their foundation on the Transformer architecture, equips us to work seamlessly with the U.S. Government's interest in diverse LLM applications. Not only has our team studied the biases introduced by LLM embeddings, but we have also ventured into LLM-based recommender systems and domain adaptation techniques, making us adept at addressing the multifaceted challenges presented by LLMs.
5. **Diverse Applications.** Our capabilities extend beyond theoretical expertise. PI Wang's significant contributions in various domains, including healthcare, network analysis, forecasting, and speech recognition, underscore our ability to apply our deep learning expertise to real-world scenarios. This, combined with our focus on correcting biases, positions us to contribute to the BENGAL initiative's aim of making LLMs safer for a wide variety of applications.

In conclusion, our team's technical prowess resonates deeply with the BENGAL initiative's goals. We are poised to make significant strides in ensuring the safe and optimal utilization of LLMs for the U.S. Government and the broader Intelligence Community. Our commitment to innovation, fairness, and the responsible deployment of technology aligns seamlessly with the objectives of this opportunity.

References

[1] Amazon Personalize. `https://aws.amazon.com/personalize/`. Accessed: 2023-07-16.

[2] T. B. Brown, B. Mann, M. S. Nick Ryder, J. Kaplan, A. N. Prafulla Dhariwal, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[4] H. Ding, Y. Ma, A. Deoras, Y. Wang, and H. Wang. Zero-shot recommender systems. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022.

[5] Y. Gao, X. Shi, Y. Zhu, H. Wang, Z. Tang, X. Zhou, M. Li, and D. N. Metaxas. Visual prompt tuning for test-time domain adaptation. *arXiv preprint arXiv:2210.04831*, 2022.

[6] Z. Gao, X. Shi, H. Wang, Y. Zhu, Y. Wang, M. Li, and D.-Y. Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. In *NeurIPS*, 2022.

[7] S. Gupta, H. Wang, Z. Lipton, and Y. Wang. Correcting exposure bias for link recommendation. In *ICML*, 2021.

[8] H. He, H. Wang, G. Lee, and Y. Tian. Probgan: Towards probabilistic GAN with theoretical guarantees. In *ICLR*, 2019.

[9] H. Huang, X. Gu, H. Wang, C. Xiao, H. Liu, and Y. Wang. Extrapolative continuous-time bayesian neural network for fast training-free test-time adaptation. In *NeurIPS*, 2022.

[10] H. Huang, H. Liu, H. Wang, C. Xiao, and Y. Wang. Strode: Stochastic boundary ordinary differential equation. In *ICML*, 2021.

[11] H. Huang, H. Wang, and B. Mak. Recurrent poisson process unit for speech recognition. In *AAAI*, volume 33, pages 6538–6545, 2019.

[12] H. Huang, F. Xue, H. Wang, and Y. Wang. Deep graph random process for relational-thinking-based speech recognition. In *ICML*, 2020.

[13] X. Jin, Y. Park, D. C. Maddix, H. Wang, and Y. Wang. Domain adaptation for time series forecasting via attention sharing. In *ICML*, 2022.

[14] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In *Learning in graphical models*, pages 105–161. Springer, 1998.

[15] M. Kampffmeyer, Y. Chen, X. Liang, H. Wang, Y. Zhang, and E. P. Xing. Rethinking knowledge graph propagation for zero-shot learning. In *CVPR*, pages 11487–11496, 2019.

[16] W. Lin, H. Lan, H. Wang, and B. Li. Orphicx: A causality-inspired latent variable model for interpreting graph neural networks. In *CVPR*, 2022.

[17] Y. Lin, H. Dong, H. Wang, and T. Zhang. Bayesian invariant risk minimization. 2022.

[18] S. Liu, T. Lin, D. He, F. Li, R. Deng, X. Li, E. Ding, and H. Wang. Paint transformer: Feed forward neural painting with stroke prediction. In *ICCV*, 2021.

[19] T. Liu, Z. Xu, H. He, G. Hao, G.-H. Lee, and H. Wang. Taxonomy-structured domain adaptation. In *ICML*, 2023.

[20] W. Liu, Y. Cheng, H. Wang, J. Tang, Y. Liu, R. Zhao, W. Li, Y. Zheng, and X. Liang. " my nose is running."" are you also coughing?": Building a medical diagnosis agent with interpretable inquiry logics. 2022.

[21] C. Mao, M. Chiquier, H. Wang, J. Yang, and C. Vondrick. Adversarial attacks are reversible with natural supervision. 2021.

[22] C. Mao, A. Gupta, A. Cha, H. Wang, J. Yang, and C. Vondrick. Generative interventions for causal learning. In *CVPR*, 2021.

[23] C. Mao, K. Xia, J. Wang, H. Wang, J. Yang, E. Bareinboim, and C. Vondrick. Causal transportability for visual recognition. In *CVPR*, 2022.

[24] J. Pearl. Bayesian networks. 2011.

[25] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *NIPS*, pages 802–810, 2015.

[26] X. Shi, Z. Gao, L. Lausen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo. Deep learning for precipitation nowcasting: A benchmark and a new model. In *NIPS*, pages 5617–5627, 2017.

[27] H. Wang. *Bayesian Deep Learning for Integrated Intelligence: Bridging the Gap between Perception and Inference*. PhD thesis, Hong Kong University of Science and Technology, 2017.

[28] H. Wang, H. He, and D. Katabi. Continuously indexed domain adaptation. In *ICML*, 2020.

[29] H. Wang, X. Liang, H. Zhang, D.-Y. Yeung, and E. P. Xing. ZM-Net: Real-time zero-shot image manipulation network. *arXiv preprint arXiv:1703.07255*, 2017.

[30] H. Wang, C. Mao, H. He, M. Zhao, T. S. Jaakkola, and D. Katabi. Bidirectional inference networks: A class of deep bayesian networks for health profiling. In *AAAI*, pages 766–773, 2019.

[31] H. Wang, X. Shi, and D. Yeung. Natural-parameter networks: A class of probabilistic neural networks. In *NIPS*, pages 118–126, 2016.

[32] H. Wang, X. Shi, and D.-Y. Yeung. Relational deep learning: A deep latent variable model for link prediction. In *AAAI*, pages 2688–2694, 2017.

[33] H. Wang, N. Wang, and D. Yeung. Collaborative deep learning for recommender systems. In *KDD*, pages 1235–1244, 2015.

[34] H. Wang, S. Xingjian, and D.-Y. Yeung. Collaborative recurrent autoencoder: Recommend while learning to fill in the blanks. In *NIPS*, pages 415–423, 2016.

[35] H. Wang and J. Yan. Self-interpretable time series prediction with counterfactual explanations. In *ICML*, 2023.

[36] H. Wang and D.-Y. Yeung. Towards bayesian deep learning: A framework and some existing methods. *TDKE*, 28(12):3395–3408, 2016.

[37] H. Wang and D.-Y. Yeung. A survey on bayesian deep learning. *CSUR*, 53(5):1–37, 2020.

[38] Y. Wang, V. Menkovski, H. Wang, X. Du, and M. Pechenizkiy. Causal discovery from incomplete data: A deep learning approach. 2020.

[39] Z. Wang and H. Wang. Variational imbalanced regression: Fair uncertainty quantification via probabilistic smoothing. In *NeurIPS*, 2023.

[40] Z. Xu, G. Hao, H. He, and H. Wang. Domain indexing variational bayes: Interpretable domain index for domain adaptation. In *ICLR*, 2023.

[41] Z. Xu, G.-H. Lee, Y. Wang, and H. Wang. Graph-relational domain adaptation. *ICLR*, 2022.

[42] Y. Yang, H. Wang, and D. Katabi. On multi-domain long-tailed recognition, generalization and beyond. In *ECCV*, 2022.

[43] Y. Yang, Y. Yuan, G. Zhang, H. Wang, Y.-C. Chen, Y. Liu, C. Tarolli, D. Crepeau, J. Bukartyk, M. Junna, A. Videnovic, T. Ellis, M. Lipford, R. Dorsey, and D. Katabi. Artificial intelligence-enabled detection and assessment of parkinsonâs disease using nocturnal breathing signals. *Nature medicine*, 1(1):1–1, 2022.

[44] Y. Yang, K. Zha, Y. Chen, H. Wang, and D. Katabi. Delving into deep imbalanced regression. In *ICML*, 2021.

[45] S. Yue, H. He, H. Wang, H. Rahul, and D. Katabi. Extracting multi-person respiration from entangled rf signals. *IMWUT*, 2(2):1–22, 2018.

[46] S. Yue, Y. Yang, H. Wang, H. Rahul, and D. Katabi. Bodycompass: Monitoring sleep posture with wireless signals. *IMWUT*, 4(2):1–25, 2020.

[47] M. Zhang, T. He, T. Wang, F. Mireshghallah, B. Chen, H. Wang, and Y. Tsvetkov. Latticegen: A cooperative framework which hides generated text in a lattice for privacy-aware generation on cloud. *arXiv preprint arXiv:2309.17157*, 2023.

[48] Y. Zhang, H. Ding, Z. Shui, Y. Ma, J. Zou, A. Deoras, and H. Wang. Language models as recommender systems: Evaluations and limitations. In *NeurIPS ICBINB Workshop*, 2021.

[49] M. Zhao, K. Hoti, H. Wang, A. Raghu, and D. Katabi. Assessment of medication self-administration using artificial intelligence. *Nature Medicine*, 2021.