Deep Learning: Integrating Domain Knowledge and Interpreting the Network Decisions

Final Report for N00014-20-1-2382
Duration: 09/01/2020 - 08/31/2023

Drs. Anil K Jain and Jiayu Zhou
Computer Science and Engineering, Michigan State University

November 12, 2023

1 Major Goals

The major goal of this project is to develop a principled approach to integrate domain knowledge in the lifecycle of deep learning and effectively reduce the model complexity and thereby training robust and accurate deep models using the limited amount of training data available. The proposed approach includes three major tasks:

- Integrate data knowledge from auxiliary data sources to revise the formulation of deep learning, in the form of knowledge-defined structural regularization or constraints on the parametric space;
- Integrate model knowledge, where we exploit the decision surfaces from simpler models on the same task to guide the learning of the deep model, which effectively reduces the model complexity;
- Integrate optimizer knowledge, which seeks to improve the optimization procedure of the training of deep models. By identifying similar learning tasks and observing their gradient trajectories, the optimizer itself can be trained to provide faster convergence and also avoid poor local optimal solutions;
- A byproduct of integrating domain knowledge will be to impart interpretability or explainability to the network decision making, a much desired capability which is currently lacking.

2 Accomplishments Under Goals

2.1 Year 1

The initial year was dedicated to two primary areas within deep learning. Firstly, we explored the concept of robustness in interpretable networks, positing that robustness not only enhances security but also aids in interpretability. Secondly, we worked on the collaborative framework of federated learning, a decentralized approach that amalgamates local users' model parameters on a global server without direct data access. This method leveraged the diverse knowledge from distributed participants, enhancing the learning process.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
11-19-2023	Final Report	09/01/2020 - 08/31/2023

4. TITLE AND SUBTITLE
Deep Learning: Integrating Domain Knowledge and Interpreting the Network Decisions

5a. CONTRACT NUMBER

5b. GRANT NUMBER
N00014-20-1-2382

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Anil K. Jain
Jiayu Zhou

5d. PROJECT NUMBER
PR No. 1000011867

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
MICHIGAN STATE UNIVERSITY
426 AUDITORIUM RD RM 2
EAST LANSING MI 48824-2600
UNITED STATES OF AMERICA

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
875 N. Randolph Street, Suite 1425
Arlington, VA 22203-1995

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The major goal of this project is to develop a principled approach to integrate domain knowledge in the lifecycle of deep learning and effectively reduce the model complexity and thereby training robust and accurate deep models using the limited amount of training data available. The proposed approach includes three major tasks: a) Integrate data knowledge from auxiliary data sources to revise the formulation of deep learning, in the form of knowledge-defined structural regularization or constraints on the parametric space; b) Integrate model knowledge, where we exploit the decision surfaces from simpler models on the same task to guide the learning of the deep model, which effectively reduces the model complexity; c) Integrate optimizer knowledge, which seeks to improve the optimization procedure of the training of deep models. By identifying similar learning tasks and observing their gradient trajectories, the optimizer itself can be trained to provide faster convergence and also avoid poor local optimal solutions.

15. SUBJECT TERMS
Deep learning, model complexity, knowledge integration

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Anil Jain
					19b. TELEPHONE NUMBER *(Include area code)* 517-355-9282

We tackled the challenge of unsupervised anomaly detection through the application of deep neural networks (DNNs). Our focus was on enhancing the conventional methodology, which predominantly relies on autoencoders for learning feature representations of normal data observations. The standard practice involves utilizing the reconstruction error of these autoencoders as an outlier score for anomaly detection. However, a critical limitation arises from the over-parameterization of DNNs, leading to small reconstruction errors for anomalies and consequently diminishing the efficacy of these techniques. To address this issue, we developed an innovative framework based on collaborative autoencoders. This framework is designed to simultaneously discern normal observations and learn their feature representations from the data effectively. We conducted a comprehensive analysis of the framework's theoretical underpinnings and rigorously evaluated its performance through empirical studies. Our results demonstrate a significant enhancement in performance over other DNN-based anomaly detection methods. Additionally, the framework exhibited a remarkable resilience to missing data, surpassing the capabilities of other standard baseline methods in handling incomplete datasets.

We addressed the complexities involved in training deep neural networks under conditions of corrupted supervision, a scenario that poses significant risks to the model's generalization capabilities. We developed an advanced, efficient algorithm characterized by its robustness, which operates effectively without necessitating assumptions about the nature of the data corruption. This algorithm is versatile, offering a comprehensive framework applicable to both classification and regression problems. Distinct from many existing methods that assess data quality on an individual basis, such as evaluating each data point's loss value and filtering them based on these criteria, our approach adopts a novel strategy. It concentrates on moderating the cumulative effect of data points on the model's average gradient. This methodology ensures that even if a corrupted data point is not identified and excluded by our algorithm, its influence on the overall loss remains minimal, especially when compared to conventional filtering techniques based on loss values. We validated the effectiveness and resilience of our algorithm through extensive experimental assessments using various benchmark datasets. These experiments conclusively demonstrated our algorithm's superior performance and robustness in diverse scenarios of data corruption.

In the realm of federated learning, the heterogeneity of user data poses substantial challenges, often leading to the development of global models that suffer from slow convergence rates. A contemporary solution to this issue lies in Knowledge Distillation, which has emerged as a method to refine the server model. This refinement is achieved by integrating aggregated knowledge from diverse users, rather than solely relying on the direct aggregation of their model parameters. However, this approach is contingent upon the availability of a proxy dataset, which limits its practicality in scenarios where such a dataset is absent. Additionally, there is a missed opportunity in fully harnessing the ensemble knowledge to guide the learning of local models, potentially diminishing the quality of the aggregated global model. To overcome these limitations, we developed a novel data-free knowledge distillation approach tailored for heterogeneous FL environments. Our method involves the server learning a lightweight generator that synthesizes user information in a data-free manner. This synthesized knowledge is then disseminated to users, serving as an inductive bias to regulate local training. Theoretical insights underpinning our approach are complemented by empirical studies, which collectively demonstrate that our method not only enhances the generalization performance of FL but also achieves this with fewer communication rounds when compared to leading-edge techniques. This advancement represents a significant stride in addressing the complexities inherent in heterogeneous federated learning systems.

Addressing the challenges arising from user heterogeneity in federated learning, particularly the risk of developing models that are biased against minority groups, is a critical concern. While adversarial learning has been a popular technique in centralized learning environments to mitigate such biases, its implementation in a federated framework encounters substantial obstacles. We developed a pioneering approach, Federated Adversarial DEbiasing (FADE), to tackle this approach. A key advantage of FADE is its ability to perform debiasing without requiring sensitive group information from users. This aspect is crucial in maintaining user privacy and autonomy, as it allows users to opt out of the adversarial component of the process if privacy concerns or computational constraints arise. We provide theoretical evidence demonstrating that FADE, under ideal conditions, can achieve the same level of global optimality as its centralized counterpart. Recognizing that practical scenarios may not always align with ideal conditions, we also investigate potential convergence issues within FADE. To address these, we introduce a straightforward yet effective method, enhancing the framework's robustness.

In our pioneering effort, we have successfully adapted federated learning to the domain of face recognition, a field traditionally reliant on large, centrally aggregated datasets of facial images. This adaptation, FedFace, is particularly significant in light of increasing concerns around data privacy and stringent legal constraints on the access and sharing of facial datasets. FedFace provides a novel federated learning framework specifically designed for the collaborative development of face recognition models, prioritizing privacy at every stage. The framework harnesses face images stored across multiple client devices, such as mobile phones, to train an accurate and generalizable face recognition model. A key feature of FedFace is its strict adherence to privacy principles: facial images remain exclusively on the client device and are not shared with either the central host or other clients. Furthermore, each client device contains images of only a single identity - the device owner - ensuring a one-identity-per-client protocol. Our comprehensive experiments validate the efficacy of FedFace. We have rigorously tested the framework on standard face verification benchmarks, including LFW, IJB-A, and IJB-C. The results from these tests demonstrate a marked improvement in the verification performance of pre-trained face recognition systems.

2.2 Year 2

Our second year was marked by continued advancements in robust deep learning methods. This included developing techniques for density estimation and adaptation in the face of corrupted data sources. Concurrently, we focused on privacy-preserving collaborative learning methods. These methods were designed to be tolerant of data heterogeneity and communication challenges, and they provided flexible adaptation by deploying models of varying sizes to meet diverse requirements.

In our research, we have focused on refining the process of unsupervised anomaly detection through enhanced density estimation methods. Density estimation, a critical tool in this domain, identifies anomalies by learning the data's density function, where data points with exceptionally low densities are flagged as anomalies. However, the incorporation of anomalies within training data can significantly skew the density estimation, particularly when utilizing sophisticated methods like those based on deep neural networks. To address this challenge, we introduce RobustRealNVP, a novel deep density estimation framework tailored to augment the robustness of flow-based density estimation techniques. This enhancement is crucial for their effective application in unsupervised anomaly detection scenarios. RobustRealNVP stands out from existing flow-based models in two key aspects: *Selective Data Discarding*: Unlike traditional models, RobustRealNVP actively discards data points that exhibit low estimated densities during the optimization process. This ap-

proach is designed to prevent such points from adversely impacting the overall density estimation. *Lipschitz Regularization*: We implement Lipschitz regularization within the flow-based model to ensure the smoothness of the estimated density function. This regularization plays a vital role in maintaining the integrity and reliability of the density estimation. Our comprehensive evaluation of RobustRealNVP, encompassing both theoretical analysis and empirical testing, underscores its resilience against anomalies in the training data. The framework demonstrates not only robustness but also competitive performance when benchmarked against the current leading methods in unsupervised anomaly detection. This positions RobustRealNVP as a significant advancement in the field, offering a more reliable and effective tool for identifying anomalies in complex datasets.

We studied the realm of Unsupervised Domain Adaptation (UDA), a burgeoning field in machine learning focused on transferring knowledge from labeled source domains to unlabeled target domains. The typical UDA frameworks, however, often rely on clean, well-distributed training data from the source domain, making them vulnerable to corruption, whether inherent or through adversarial attacks. To combat these vulnerabilities, we have developed a robust and effective framework tailored to address the challenges of UDA when dealing with corrupted source domain data. Our approach is distinguished by two primary strategies: *Knowledge Ensemble from Multiple Models*: We leverage multiple domain-invariant models that are trained on randomly partitioned segments of the training data. This ensemble technique helps to mitigate the impact of corrupted data by diversifying the knowledge base and reducing reliance on any single data partition. *Refinement via Mutual Information Maximization*: We introduce an advanced refinement process for each learned model, utilizing mutual information maximization. This method effectively aligns the models with the target domain by adaptively capturing and utilizing the predictive information relevant to the target domain with high confidence. The effectiveness of our framework is substantiated through extensive empirical studies. These studies demonstrate our approach's resilience against various types of poisoned data attacks, while simultaneously achieving high asymptotic performance in the target domain. This dual capability – robustness against data corruption and efficacy in domain adaptation – positions our framework as a significant contribution to the field of UDA, especially in scenarios involving compromised source domain data.

We addressed the increasingly vital topic of maintaining privacy in machine learning applications, particularly those involving sensitive data. A key focus is on Private Gradient Descent (PGD), a prevalent framework in private learning, which applies the principles of Differential Privacy by introducing noise to gradients. Recent studies have highlighted the potential benefits of dynamic privacy schedules, which utilize decreasing noise magnitudes, in improving the loss outcomes at the final iteration of PGD. However, a thorough theoretical understanding of these dynamic schedules, and their relationship with optimization algorithms, has been limited. To bridge this knowledge gap, our research offers a comprehensive analysis of the influence of noise in dynamic privacy schedules. We make several pivotal contributions: *Optimization of Dynamic Noise Schedules*: We introduce a dynamic noise schedule that minimizes the utility upper bound of PGD. This schedule is tailored to optimize the balance between privacy and model performance. *Collective Impact of Noise on Utility*: Our study examines how the cumulative effect of noise introduced at each optimization step influences the overall utility of the final model. This analysis is crucial in understanding the long-term implications of noise on model performance. *Influence of Momentum*: We delve into how the dynamics of noise influence change when momentum is incorporated into the optimization process. This aspect is particularly significant as momentum is a common element in optimization algorithms. *Empirical Connections for Non-Convex Losses*: Through empirical evaluation, we es-

tablish the relevance of our findings for general non-convex loss functions. We demonstrate that the impact of dynamic noise is significantly affected by the curvature of the loss function. Our work sheds light on the intricate balance between privacy protection and optimization in machine learning, providing valuable insights for the development of more effective and privacy-preserving learning algorithms.

We also addresses the critical challenges hindering the widespread adoption of Federated Learning in *edge computing*, specifically the heterogeneity of edge network topologies and the uncertainty of wireless transmissions. These factors contribute to extended convergence times and elevated communication costs in FL applications. To mitigate these issues, we have developed a novel FL scheme that comprehensively tackles both challenges. Key aspects of our proposed FL scheme include: *Self-Distilled Neural Networks*: We enable edge devices to learn specialized neural networks that are designed for self-distillation. These networks are uniquely capable of being pruned to various sizes without compromising the integrity of the learned knowledge. This feature allows the networks to adapt to the learning domain in a nested and progressive fashion. *Adaptability to System Heterogeneity*: Our approach addresses the issue of system heterogeneity by providing edge devices with varying model architectures. This adaptability ensures that devices with different computational capabilities can participate effectively in the learning process. *Resilience to Connection Uncertainty*: In the face of unstable network connections, a common occurrence in edge computing, our scheme allows for partial transmission of model parameters. This ensures that the knowledge encapsulated in the transmitted parameters is not lost, thus making the communication process more efficient and resilient. The efficacy of our approach has been thoroughly validated through extensive empirical studies. These studies demonstrate that, in scenarios characterized by system heterogeneity and network instability, our FL scheme exhibits remarkable resilience and superior communication efficiency compared to existing state-of-the-art methods.

2.3 Year 3

In the final year, our research pivoted towards deep learning applications in dynamically changing environments. We successfully developed methods for the rapid adaptation of deep learning models post-deployment. Additionally, we developed technical solutions for distilling knowledge from deep models and integrating external knowledge sources. This year's work underscored the importance of adaptability and knowledge integration in the evolving landscape of deep learning.

Our research explores Continual Test-time Adaptation (CTA), an emerging technique essential for maintaining model accuracy in dynamic environments where data distributions continually change. Despite the effectiveness of state-of-the-art CTA methods in enhancing out-of-distribution model accuracy through efficient online test-time gradient descents, they are hindered by a significant drawback: excessive memory consumption, often multiple times higher than that required for inference. This high memory requirement, particularly when updating only a fraction of the parameters, severely limits the application of advanced CTA techniques on memory-constrained devices. To address this limitation, we introduced a novel solution, named Memory-Efficient CTA (MECTA), designed to substantially enhance the memory efficiency of gradient-based CTA processes. Our approach is grounded in an in-depth analysis of memory overhead, which we found primarily originates from the intermediate caching necessary for back-propagation, and scales with batch size, channel number, and layer count. MECTA implements several strategic modifications to counteract these memory overheads: *Reduced Batch Sizes*: By decreasing the batch sizes, we significantly lower the memory requirements without compromising the model's predictive stabil-

ity and accuracy. *Adaptive Normalization Layer*: We incorporate an adaptive normalization layer to ensure consistent and accurate predictions, even with reduced batch sizes. *Heuristic Back-Propagation Caching*: To further conserve memory, we selectively halt back-propagation caching, employing a heuristic approach to determine when this caching is unnecessary. *Network Pruning*: MECTA prunes the network during the optimization phase to cut down both computation and memory overheads. Subsequently, we recover the pruned parameters to prevent the loss of learned information (model forgetting). MECTA's design allows it to be easily integrated into existing state-of-the-art CTA algorithms with minimal impact on computation and memory resources.

In our research, we address a significant challenge in the application of deep neural networks: their vulnerability to out-of-distribution (OoD) samples, which are often misclassified with high confidence. While recent advancements have made strides in OoD detection within centralized training environments, the domain of federated learning has largely been unexplored in this context. This oversight is particularly critical given that many security-sensitive applications, such as autonomous driving and voice recognition authorization, increasingly rely on FL due to its data privacy advantages. A primary obstacle in adapting state-of-the-art OoD detection methods to FL is their dependency on substantial amounts of real OoD samples for training. In real-world scenarios, acquiring such large-scale OoD training data is often impractical or excessively resource-intensive, especially for local devices with limited capabilities. Conversely, FL is inherently challenged by data heterogeneity, with each client gathering non-identically and independently distributed (non-iid) data. Our proposal leverages this heterogeneity, transforming what is typically seen as a drawback into an asset for OoD detection within FL. We introduce the Federated Out-of-Distribution Synthesizer (FOSTER), a novel approach that utilizes a class-conditional generator to create virtual external-class OoD samples. These synthesized samples serve as a viable alternative to real OoD data. FOSTER is uniquely designed to respect the essential requirements of FL, including maintaining data confidentiality and ensuring communication efficiency. By exploiting the inherent data heterogeneity in FL, where non-iid data from other clients can mimic external-class OoD samples, FOSTER provides an innovative solution for effective OoD detection in federated learning environments.

Data-free knowledge distillation (KD) is a technique that facilitates the transfer of knowledge from a large, pre-trained model (the teacher) to a smaller, more efficient model (the student) without requiring access to the original training data used to train the teacher. A critical aspect that has been largely overlooked in data-free KD is the security of the synthetic or out-of-distribution (OOD) data employed in the process. We addressed the security risks associated with data-free KD, particularly in the context of untrusted pre-trained models. We introduce Anti-Backdoor Data-Free KD (ABD), a novel defensive approach specifically designed to safeguard against the inadvertent transfer of potential backdoors from teacher to student models during the distillation process. Our approach, ABD, stands as the first of its kind—a plug-in method that can be seamlessly integrated into existing data-free KD processes to mitigate the risk of backdoor transfer. We conducted extensive empirical evaluations to test the efficacy of ABD in minimizing the transmission of backdoor knowledge, ensuring it does so while maintaining performance levels comparable to traditional KD methods. The findings from our study serve as a pivotal milestone in highlighting and addressing the potential vulnerabilities of data-free KD, especially concerning backdoor threats. We believe this work not only contributes significantly to enhancing the security of KD techniques but also paves the way for future research in developing more robust and trustworthy machine learning models.

3 Training Opportunities

At MSU, the PI and Co-PI organize a weekly study group with graduate students on large-scale optimization techniques to review and analyze state-of-the-art stochastic optimization techniques. The study group also held extensive discussions on the design of new deep learning algorithms and network interpretability. Co-PI Zhou has organized International Workshop on Federated Learning for Distributed Data Mining, Co-located with the 29th ACM SIGKDD Conference (KDD 2023).

4 Honors and Awards

The paper *MolSearch: Search-based Multi-objective Molecular Generation and Property Optimization* received the Best Paper Award at the Health Day Section, Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2022.

5 Deliverables

The results that came out of this support were published in journals and as conference papers [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27] and presented at premier data mining and machine learning conferences, such as SIGKDD Conference on Knowledge Discovery and Data Mining, International Conference on Machine Learning, Conference on Neural Information Processing Systems, Annual AAAI Conference on Artificial Intelligence.

The source code of the multi-task learning techniques developed in these papers is included in the open source software package MALSAR (www.malsar.org), which is widely used among researchers working in multiple application domains, including medical informatics, computer vision, chemoinformatics, and bioinformatics.

References

[1] Zhuangdi Zhu, Kaixiang Lin, Bo Dai, and Jiayu Zhou. Off-policy imitation learning from observations. *Advances in Neural Information Processing Systems*, 33:12402–12413, 2020.

[2] Junyuan Hong, Haotao Wang, Zhangyang Wang, and Jiayu Zhou. Learning model-based privacy protection under budget constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7702–7710, 2021.

[3] Zhaonan Qu, Kaixiang Lin, Zhaojian Li, and Jiayu Zhou. Federated learning's blessing: Fedavg has linear speedup. In *ICLR 2021 - Workshop on Distributed and Private Machine Learning (DPML)*, 2021.

[4] Boyang Liu, Ding Wang, Kaixiang Lin, Pang-Ning Tan, and Jiayu Zhou. Rca: A deep collaborative autoencoder approach for anomaly detection. In *IJCAI: proceedings of the conference*, volume 2021, page 1505. NIH Public Access, 2021.

[5] Boyang Liu, Mengying Sun, Ding Wang, Pang-Ning Tan, and Jiayu Zhou. Learning deep neural networks under agnostic corrupted supervision. In *International Conference on Machine Learning*, pages 6957–6967. PMLR, 2021.

[6] Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *International conference on machine learning*, pages 12878–12889. PMLR, 2021.

[7] Junyuan Hong, Zhuangdi Zhu, Shuyang Yu, Zhangyang Wang, Hiroko H Dodge, and Jiayu Zhou. Federated adversarial debiasing for fair and transferable representations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 617–627, 2021.

[8] Mengying Sun, Jing Xing, Huijun Wang, Bin Chen, and Jiayu Zhou. Mocl: data-driven molecular fingerprint via knowledge-aware contrastive learning from molecular graph. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3585–3594, 2021.

[9] Divyansh Aggarwal, Jiayu Zhou, and Anil K Jain. Fedface: Collaborative learning of face recognition model. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021.

[10] Jun Chen, Jieping Ye, Fengyi Tang, and Jiayu Zhou. Automatic detection of alzheimer's disease using spontaneous speech only. In *Interspeech*, volume 2021, page 3830. NIH Public Access, 2021.

[11] Fengyi Tang, Jun Chen, Hiroko H Dodge, and Jiayu Zhou. The joint effects of acoustic and linguistic markers for early identification of mild cognitive impairment. *Frontiers in digital health*, 3:702772, 2022.

[12] Haotao Wang, Junyuan Hong, Aston Zhang, Jiayu Zhou, and Zhangyang Wang. Trap and replace: Defending backdoor attacks by trapping them into an easy-to-replace subnetwork. *Advances in neural information processing systems*, 35:36026–36039, 2022.

[13] Junyuan Hong, Lingjuan Lyu, Jiayu Zhou, and Michael Spranger. Outsourcing training without uploading data via efficient collaborative open-source sampling. *Advances in neural information processing systems*, 35:20133–20146, 2022.

[14] Shuyang Yu, Zhuangdi Zhu, Boyang Liu, Anil K Jain, and Jiayu Zhou. Robust unsupervised domain adaptation from a corrupted source. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 1299–1304. IEEE, 2022.

[15] Guangliang Liu, Owen Yuan, Lifeng Jin, and Jiayu Zhou. Dynamic augmentation data selection for few-shot text classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, page 4870. NIH Public Access, 2022.

[16] Mengying Sun, Jing Xing, Han Meng, Huijun Wang, Bin Chen, and Jiayu Zhou. Molsearch: search-based multi-objective molecular generation and property optimization. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 4724–4732, 2022.

[17] Zhuangdi Zhu, Junyuan Hong, Steve Drew, and Jiayu Zhou. Resilient and communication efficient learning for heterogeneous federated systems. *Proceedings of machine learning research*, 162:27504, 2022.

[18] Junyuan Hong, Zhangyang Wang, and Jiayu Zhou. Dynamic privacy budget allocation improves data efficiency of differentially private gradient descent. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 11–35, 2022.

[19] Junyuan Hong, Haotao Wang, Zhangyang Wang, and Jiayu Zhou. Efficient split-mix federated learning for on-demand and in-situ customization. *ICLR*, 2022.

[20] Zhuangdi Zhu, Kaixiang Lin, Bo Dai, and Jiayu Zhou. Self-adaptive imitation learning: learning tasks with delayed rewards from sub-optimal demonstrations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9269–9277, 2022.

[21] Boyang Liu, Pang-Ning Tan, and Jiayu Zhou. Unsupervised anomaly detection by robust density estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4101–4108, 2022.

[22] Haotao Wang, Junyuan Hong, Jiayu Zhou, and Zhangyang Wang. How robust is your fairness? evaluating and sustaining fairness under unseen distribution shifts. *Transactions on machine learning research*, 2023, 2023.

[23] Shuyang Yu, Junyuan Hong, Haotao Wang, Zhangyang Wang, and Jiayu Zhou. Turning the curse of heterogeneity in federated learning into a blessing for out-of-distribution detection. In *The Eleventh International Conference on Learning Representations*, 2022.

[24] Junyuan Hong, Lingjuan Lyu, Jiayu Zhou, and Michael Spranger. Mecta: Memory-economic continual test-time model adaptation. In *The Eleventh International Conference on Learning Representations*, 2022.

[25] Junyuan Hong, Haotao Wang, Zhangyang Wang, and Jiayu Zhou. Federated robustness propagation: sharing adversarial robustness in heterogeneous federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7893–7901, 2023.

[26] Guangliang Liu, Zhiyu Xue, Liang Zhan, Hiroko H Dodge, and Jiayu Zhou. Detection of mild cognitive impairment from language markers with crossmodal augmentation. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2023: Kohala Coast, Hawaii, USA, 3–7 January 2023*, pages 7–18. World Scientific, 2022.

[27] Junyuan Hong, Yi Zeng, Shuyang Yu, Lingjuan Lyu, Ruoxi Jia, and Jiayu Zhou. Revisiting data-free knowledge distillation with poisoned teachers. *ICLM 2023*, 2023.