Ananth Balashankar - Research Statement

Research Summary

My research has fostered positive socio-technical outcomes by building robust machine learning (ML) systems that faithfully incorporate domain specific structure. Although deep learning has fastened the speed of adoption of ML in various application domains like image recognition, natural language processing (NLP), robotics, etc., when it comes to deploying ML in high-stakes businesses like health, socio-economic inference and content moderation, practitioners find that models' predictions diverge from established domain knowledge in the real world. Through my research on **Domain Faithful Deep Learning Systems**, I have translated expert-understandable domain knowledge and business logic to be faithfully incorporated in training deep learning models. I have deployed these ML systems for consequential socio-technical and business scenarios by collaborating with domain experts, and addressed critical research questions such as, "What data distributions do domain practitioners care about?", "How to faithfully convert domain knowledge into model constraints for better generalization?" and finally "How to evaluate whether the ML models we learn are grounded in the domain knowledge and in what ways do they deviate?". To this end, I have contributed novel methodologies through ML model design [14], constraints [3, 10, 11, 12], data augmentation [1, 2] and feature selection [7, 8], for real world applications. For instance, the causal-aware prediction models I have developed show that relying on data alone can lead to incorporating spurious correlations and low accuracy over counterfactual inputs, and hence, domain-specific structure is necessary. Broadly, my work has had demonstrable *research impact across 5 domains*:

- **Famine Forecasting:** Forecasting famine is critical for the mobilization of aid to millions of people, but hard to solve due to data scarcity in fragile countries. By building a news-based causal-aware forecasting framework that extracts *causal features* from 11.2 million news articles across 2 decades in 15 fragile countries [4], we have improved forecasting accuracy by 32% compared to state-of-the-art predictive models. This work is accepted at *IC2S2 '21*, the premiere computational social science conference, and in review at a Science sub-journal. The tool will be used by the World Bank Data Science group for aid allocation on food security. Based on this research, a few co-authors have founded a socio-economic inference start-up Velai, Inc.
- **Toxic Comments Detection:** Automated detection of online toxic comments improves the quality of interaction in social media. However, the variations in context of comments make it hard to protect specific demographic groups from disparate impact. By explicitly modeling such nuances through *counterfactual data augmentation*, we improved the accuracy of detecting toxicity by 6% [2, 13]. Through this publication at *EMNLP '21*, a premiere NLP conference, I have fostered deep engagements with Google's Responsible ML team. I have also deployed ML models that optimize business objectives like diversity at Google Play.
- **Causal Question Answering:** Causality based question answering lies at the core of customer support tools like chatbots. Prior ML models fail to capture the directed nature of causality, for example rain causes traffic delay, and not vice versa. By learning asymmetric causal embeddings faithful to causal graphs [5], we improved accuracy on Yahoo! Answers by 21% in this paper at *ACL '21*, a premiere NLP conference.
- **Health Recommendations:** Trustworthy ML models in health recommendations need to be faithful to medical concepts over unseen patient data, while traditional ML models focus only on optimizing accuracy over the observed but limited test data. By incorporating trust through doctor specified mapping rules between diagnoses and medications through data augmentation [1], we have improved accuracy of state-of-the-art end-to-end neural models by 12% in this publication at *WSDM '21*, the premiere data mining conference.
- **Privacy Enforcement:** Corporate privacy compliance policies are legally prescriptive, but not directly enforceable in computer systems. By using the theory of contextual integrity through post-processing mappings [6, 9], we have improved the accuracy of BERT-based deep learning models by 6% to extract privacy parameters for SQL-based enforcement in this paper at WWW' 19, the premiere web research conference.

Research Contributions

My dissertation research has focused on applying *domain faithful deep learning* in the domains of toxicity detection, socio-economic inference, health, causal-aware deep learning, and privacy. Each of these domains pose unique challenges on how to incorporate structure and the diverse techniques required to execute them. Below, I discuss how by developing domain faithful deep learning systems, I have improved outcomes in each one of them.

Causal-Aware Predictive Models for Socio-Economic Inference: In socio-economic inference, the motivation is to have a broader positive societal impact using data-driven machine learning tools. Many applications which relied purely on data have faced issues as they did not incorporate domain-specific causal structure. For example, in the Flu prediction model based on Google Search Trends, it was shown that the model deviates over-time as compared to a one which incorporates signals derived from the Center for Disease Control (CDC). In the problem of predicting food insecurity [4] task, we overcome the challenge of data sparsity in fragile states which are often encumbered with infrastructural and conflict-based issues that makes the task of data collection harder. As traditional indicators like rainfall, vegetation index, etc are often delayed, we aim to use the news streams [7] published by reputed sources like BBC, Reuters, AP, etc. to automatically extract and construct causally grounded indicators. Our contributions extend beyond the methodologies and have implications on the ethical and operational trade-offs a domain practitioner needs to make in a socio-technical system. In the famine prediction task, by extracting causes from scientific literature using Semantic Frame Parsing and then constructing time-series indicators by expanding to tokens with low Word-Mover distances, we are able to *reduce the food insecurity forecasting errors by 32%*. Additionally, alignment of models to domain expertise provides an additional incentive to practitioners - counterfactual reasoning: Not all episodes of famine are the same, and our methodology allows us to model what is the implication of each of the causes in improving the prediction accuracy at a fine-grained level of districts in 15 of the most fragile countries in the world over two decades.

Faithful Neural Embeddings of Causal Graphs for Question Answering: Question Answering tasks power technologies like chatbots for customer support in businesses. Recent advances in machine learning for processing natural language text has broadly relied on large neural language models like Transformers which capture the relationships between the word tokens in long sequences. The fine-tuning of these language models for multiple tasks have demonstrated state-of-the-art performance on benchmarks like GLUE. However, these fine-tuned models perform poorly on counterfactual sentences or inconsistently on downstream tasks which have specific structure like graphical causal models or domain-specific theory. In the causal-QA dataset [5], questions of the form “What causes X?” are posed, where X can be a disease, phenomenon and a real-world event. Neural Network models have been modified to predict causal links, but lack the consistency required, i.e undirected paths in a graph are still considered causal, whereas causal graphs are strictly directional. On the other hand, traditional Information Retrieval (IR) techniques that mine such causal information from knowledge graphs are limited in their generalizability to new and related terms mentioned in questions, i.e “flood” and “deluge” may have similar causes, but if “deluge” is not in the graph, then we have no way of estimating its cause. To overcome the limitations of using either an end-to-end model or domain knowledge as-is in its limited scale, we provide a way to incorporate the constraints imposed by the domain-specific structure - causal graphs in this case into BERT-like transformer based models. We demonstrate that when proximity between the embeddings of two nodes is modeled using a pseudo quasi-metric, we are able to capture the directedness of causal graphs. Specifically, we measure three properties of *faithfulness* namely the uniformity of the embeddings, correlation between distances of any two random nodes in the graph and link prediction accuracy. In each of these graph-specific indicators, by imposing a regularization loss which penalizes inconsistencies in how the embeddings satisfy these two properties over two large causal graphs with 800K nodes, we obtain a fine-tuned embedding that not only achieves causal faithfulness better, but also improves the area under the Precision-Recall curve over the *Yahoo! Answers causal-QA dataset by 21%.*

Medically Concordant Neural Recommenders in Health: Recent advances in applying AI for healthcare have often relied purely on data, but fail categorically when patients with differing characteristics than the ones present in training data are presented. Specifically, in the medication recommendation task [1], learning end-to-end neural models based on historical electronic health records might prove to be accurate, but may not inculcate trust in doctors, unless the ontologies of medicine used as standards by trusted medical associations are incorporated. In the medication recommendation task, since all possible diagnoses that may be relevant might not be present in the training data, we improve the neural network model - G-BERT’s *domain-specific concordance* based on expert-specified medical ontologies like medication and diagnostic code hierarchies and the mapping rules between them. By incorporating causal structure into machine learning models through categorical counterfactual data augmentation and regularization, we guard against predictions that violate the domain knowledge over categories, and improve the *categorical robustness of prediction models by 1.2x* and accuracy by 12% on the MIMIC-III dataset, as we rely less on spurious correlations in the data.

Additionally, domain practitioners have often minimal guidance on the choice of parameters that AI tools in healthcare operate over. For example, in the angiographic disease status prediction task [3, 10], the variability of diagnostic features in different demographic groups is well studied. Here, practitioners need to carefully evaluate the trade-offs between the per-group accuracy across demographic groups, when an end-to-end jointly trained model is used. When we analyze the performance of ML models on specific demographic groups, we outline the choice of parameters of fairness and accuracy trade-offs that practitioners have, based on Pareto Efficiency. For example, how accurate an ML model should be over patients with darker skin tone than lighter skin tone in a heart disease status prediction model is a choice that cannot be made blindly, but with careful consideration of the medical diagnostic equipment's characteristics and the Pareto optimality of the model's performance across demographic groups [3,10]. Through the principle of Pareto Efficiency, we can potentially *improve group-level accuracies by 9.6%* on UCI datasets. Acting blindly based on the neural model's decisions in high-stakes scenarios might be sub-optimal and using our methodology, experts can now justify their choice, in case they were to be contested [3, 4].

Counterfactually Robust Toxicity Detection: In the domain of toxicity detection in online social media comments, social-science experts have long advocated for incorporating how specific demographic groups are susceptible to specific types of toxic comments. It is important to model secondary attributes that are relevant to the toxicity of a sentence explicitly when we aim to be fair based on demographic groups. In this scenario, one needs to be aware of group-specific language, idioms, quirks and background history to ascertain the toxicity of a comment. But this nuance was never captured explicitly in BERT-based neural network models. At Google, I incorporated this domain knowledge through counterfactual data augmentation [2,13] that model secondary variables and was able to improve the ability to detect toxic comments for all demographic groups, specifically black women, who were susceptible to more directed toxic comments. By augmenting examples of directed toxicity in a weighted manner to demographic groups that are more exposed to such comments, we are able to classify toxicity better on all demographic groups. Without this nuance of how toxic comments vary, and just optimizing for overall absolute error, the toxicity detection model would disparately perform poorer on specific demographic groups unintentionally. Through intervention on secondary attributes through counterfactual data augmentation, we not only improved the model's understanding of what constitutes toxicity, but also improved the accuracy on all demographic groups by 7%. This application clearly demonstrates that as a text classification model is scaled to be applicable to all demographic groups in a society, the secondary effects of covariates and how they impact the performance of a ML system depends on domain knowledge, and needs carefully expert supervision. Such business decisions and design choices have the capacity to influence the product experience for billions of users.

Privacy-policy Faithful Information Systems: Recently, since the adoption of GDPR, firms have invested in data compliance systems and privacy policy enforcement. In this work, we aim to automate the enforcement by relying on established theories of privacy like Contextual Integrity [6, 9], that dictate that 5 parameters of an information flow need to be present in privacy norms of policies. This theory imposes a specific structure of the legal language containing 5 parameters: sender, receiver, attribute, subject and transmission principle in each sentence of any binding privacy policy. The extraction of these domain-specific parameters from unstructured natural language text is important to reason about the soundness and completeness of any privacy policy. Fine-tuned end-to-end neural models like BERT-based Semantic Role Labeling and Dependency Parsers significantly performed poorly by 6% as compared to a combined approach where the output parameters of these models were re-mapped post-hoc based on the theory of Contextual Integrity. These examples show that incorporating domain-specific structure is important, but also how we do it matters, when we work with a diverse set of application domains.

Future Research Agenda

My long term goal is to develop a framework where domain experts and ML practitioners can collaborate on mutually beneficial abstractions for fairness, concordance, causal models, etc., that is interpretable for the practitioners and operable for the ML researchers. With this framework, as part of future work, I will be exploring the following domains of research:

- **Faithful ML specification language for healthcare**: A common specification language for practitioners and researchers to collaborate. For example, in the medication recommendation task, we are working towards automating the process of data augmentation, regularization [1], into a specification language for the domain expert. This not only improves the transparency of ML design, but allows researchers more flexibility in choosing among a set of techniques applicable for the health domain.
- **ML-based climate change modeling:** While my current research has analyzed the performance of the ML models while keeping domain specific constraints in mind, it can also pave the way for scientific discovery in the future. For example, to measure and combat climate change, we are working with atmospheric and ocean scientists to model the Earth's atmospheric pressure by parameterizing gravity waves and it's impact

on the atmospheric pressure changes through message passing graph neural networks. Using ML models for generating these hypotheses can further improve the pace of scientific experiments.

- **Physics-based pollution forecasting models:** Incorporating domain structure into ML models can be an integral part of design choices in future socio-technical systems. For example, in the domain of pollution monitoring [14, 15], we are working towards incorporating the knowledge of pollution scientists in building fine-grained urban sensing that have the ability to forecast air quality in the next 2 hours in your neighborhood.
- **Automated decision-making for privacy:** Many domains currently have complex unstructured data, that can benefit from using ML to interpret its structure. For example, in the domain of privacy, we are working on automatically translating complex legal language in privacy policies to enforceable logic and database queries that can be directly deployed in information retrieval systems.
- **ML for ethical management and policy-making:** How domain expertise gets translated into statistical constraints and concepts can have ethical implications and is precisely the type of conversations we need to be having with domain experts. For example, in the angiographic disease status prediction task [3], how we balance the error rates across demographic groups can unearth historical biases in the measurements and calibrations of medical diagnostic tools. This way, we endeavor to incorporate socio-economic inference models as part of participatory policy making and business decision making.

Through my research vision, I will work towards enabling domain experts and ML researchers to work together. We need to converge to a common understanding of how the ML models that power our decision making systems operate; as enabling that interaction will have mutually beneficial outcomes. The challenges of the future like climate change, pollution, health and toxicity in social media need our concerted efforts. Through my research on incorporating domain structure into end-to-end ML models, I have opened the doors for domain experts like economists, doctors, physicists, gene biologists, earth scientists, linguists, lawyers and social scientists to provide inputs based on their domain knowledge to help build robust ML models. In the future, I endeavor to work on bringing robust ML to more high-stakes domains like finance, transportation, conservation and development to perform more efficient and safe decision making.

References

1. Enhancing Neural Recommender Models through Domain-Specific Concordance. Ananth Balashankar, Alex Beutel, Lakshminaryanan Subramanian. **WSDM ‘21**.
2. Can We Improve Model Robustness through Secondary Attribute Counterfactuals?. Ananth Balashankar, Xuezhi Wang, Ben Packer, Nithum Thain, Ed Chi and Alex Beutel. **EMNLP ‘21**.
3. Predicting Angiographic Disease Status: Where to draw the line between demographically decoupled and jointly trained models?. Ananth Balashankar, Alyssa Lees, Srikanth Jagabathula, Lakshminaryanan Subramanian. Working Paper.
4. Fine-grained prediction of food crisis using news. Ananth Balashankar, Samuel Fraiberger, Lakshminaryanan Subramanian. **Under Review at Science Advances, Accepted at INFORMS ’21, IC2S2 ‘21**.
5. Learning Faithful Representations of Causal Graphs. Ananth Balashankar, Lakshminarayanan Subramanian. **ACL’ 21**.
6. VACCINE: Using Contextual Integrity for Data Leakage Detection. Yan Shvartzshnaider, Zvonimir Pavlinovic, Ananth Balashankar, Thomas Wies, Lakshminarayanan Subramanian, Helen Nissenbaum and Prateek Mittal. **WWW ‘19**.
7. Identifying Predictive Causal Factors from News Streams. Ananth Balashankar, Sunandan Chakraborty, Samuel Fraiberger, Lakshminarayanan Subramanian. **EMNLP ‘19**
8. Learning Overlap-Aware Temporal Prediction Models. Ananth Balashankar, Srikanth Jagabathula, Lakshminarayanan Subramanian. Working Paper.
9. Beyond The Text: Analysis of Privacy Statements through Syntactic and Semantic Role Labeling. Yan Shvartzshnaider, Ananth Balashankar, Vikas Patidar, Thomas Wies, Lakshminarayanan Subramanian. Under Review.
10. Pareto Efficient Fairness for Skewed Subgroup Data. Ananth Balashankar, Alyssa Lees, Chris Welty, Lakshmi Subramanian. **ICML ’19-** AI for Social Good Workshop.
11. Reconstructing the MERS Disease Outbreak from News. Ananth Balashankar, Aashish Dugar, Lakshmi Subramanian, Samuel Fraiberger. **ACM COMPASS ‘19**.
12. The need for transparent demographic group trade-offs in Credit Risk and Income Classification. Ananth Balashankar, Alyssa Lees. Working Paper. CHI ’19 - Bridging the Gap Between AI and HCI Workshop.
13. Improving Robustness through Pairwise Generative Counterfactual Data Augmentation. Ananth Balashankar, Xuezhi Wang, Yao Qin, Ben Packer, Nithum Thain, Ed Chi and Alex Beutel. Under Review.

14. Spatio-temporal modeling of urban air quality using low-cost monitors. Shiva Iyer, Ananth Balashankar, William Aeberhard, Ulzee An, Sameeksha Jain, Sujoy Bhattacharya, Guiditta Rusconi, Anant Sudarshan, Rohini Pande, Lakshmi Subramanian. **Under Review at Nature Sustainability.**
15. Localized Pollution Hotspots: Inferences from a Three-year Fine-grained Air Quality Monitoring Study in Delhi. Shiva Iyer, Ananth Balashankar, Rohini Pande, Anant Sudarshan, Lakshminarayanan Subramanian.