Xin Du

Research Statement

My research is motivated by the challenge of evaluating the performance of machine learning models. Recent concerns in trustworthy AI suggest that machine learning systems need to be rigorously tested before being deployed in safety-critical applications. This problem can be theoretically regarded as the generalization of machine learning models, which is the primary goal of learning theories. The general methodology for testing the generalization ability of a model is to use a small set of hold-out data. Even though this methodology is useful, but due to the limited test data only include part of the modes which may appear in deployment environments, it may also introduce biases such as over-confident performance. This problem indicates that a single aggregated statistical measure is not enough to evaluate the complex performance of machine learning models.

Exceptional Model Mining

Exceptional Model Mining (EMM) is a local pattern mining approach that cares about how differently a model would perform in subpopulations, as compared to the same model but fitted on the whole data. Here are two study cases from my previous work: Fairness in network representation: we argue that the structural heterogeneity in networks can bias the network representation models across subgroups, which will prevent us from building fair decision making models for downstream tasks like node classification or link prediction. We evaluate the statistical significance of the discovered subgroups by applying a kernel two-sample test. Spatio-temporal behavior on collective social media: behavior in this setting can be exceptional in three distinct ways: in terms of spatial locations, time, and texts. We develop a Bayesian Non-Parametric Model (BNPM) to automatically identify spatio-temporal behavioral patterns on the subgroup level, explicitly modeling the three exceptional behavior types. This allows us to provide an effective evaluation method to measure the exceptionality of a behavioral pattern and to employ it in finding exceptional subgroups with collective social behavior.

Causal Inference

Learning causal effects from observational data greatly benefits a variety of domains such as health care, education, and sociology. For instance, one could estimate the impact of a new drug on specific individuals to assist clinical planning and improve the survival rate. In this study, we focus on studying the problem of estimating the Conditional Average Treatment Effect (CATE) from observational data. We propose a neural network framework ABCEI, based on recent advances in representation learning. To ensure the identification of the CATE, ABCEI uses adversarial learning to balance the distributions of covariates in the treatment and the control group in the latent representation space, without any assumptions on the form of the treatment selection/assignment function.

Trustworthy Autonomous Systems

✉ x.du@ed.ac.uk

Local decision rules could provide both high predictive performance and explainability. However, the stability of those explanations given by the rules is still underexplored. We propose two regularization terms to improve the robustness of decision rule ensembles. The graph-based term is built by decomposing invariant features using a given causal graph; the variance-based term relies on an additional artificial feature that can restrict the model's decision boundary within groups. In another work, we propose to evaluate the performance of image classification models. Specifically, we are interested in investigating the family of vision transformer based image models like ViT and DeiT. Considering their success and high-stake application of image models like clinical health care, it is necessary to develop a systemic model test framework to the evaluation of vision transformer based model families. We propose to develop a set of tools which could generate a bunch of test samples for different test tasks to evaluate the model, and formulate a systemic report to provide a comprehensive understanding for the model's performance.

Cognitive Computing

My vision about the future research problems on Cognitive Computing is two-fold: on the one hand, due to the high cost or ethical problems, the data at hand for Cognitive Computing are usually not complete, or biased. For instance, clinical test data are unbalanced with regard to demographic groups. It is necessary to focus on the study of the data rather than the model. On the other hand, to deploy a socioeconomic policy, usually large amount of trials are needed to get a comprehensive understanding about the potential effects of that policy. Machine learning methods are employed to help make predictions about those effects using limited data. To prevent overfitting and overconfidence problems, it is necessary to study the uncertainty quantification problem so that the model can answer 'I do not know' for some given samples.

Data Augmentation

Using historical data to predict the potential outcome of a policy is of high importance in Cognitive Computing due to the high-cost of randomized controlled trials. Deep learning models are well-known for their superb performance based on large amount of training data. However, they are also known to be vulnerable to copy or amplify the biases in training data. Hence, instead of focusing on developing new models, one of my future research is to focus on the data. I would like to develop methods to analyze the potential biases that the data could introduce to the model, and what the outcome would be biased according to the characteristics of the data. Based on that, I would like to study different data augmentation methods, such as generative model, to generate more data by debiasing algorithms.

Robustness

In general deep learning applications, the aims are to learn a model that for a given sample, e.g. an image, the model can give an answer about the semantic label. This problem is usually treated as the image recognition problem. Current image recognition models suffer from generalization problems. Models learned from training set cannot perform well on data collected from other sources, e.g. another group of people. This is partly due to that the model learns spurious features in the data instead of the true regions that indicate the semantic label. In this direction, I would like to explore the uncertain quantification methods to let the model report the confidence level for the prediction they made. The aim is to build a system that can answer 'I'm not sure' to the out-of-distribution samples, and let the human take over when it is necessary. By doing this, a safe autonomous / policy making system can be built to improve the efficiency and reduce the risk.