Research Statement

Vihari Piratla (vihari@cse.iitb.ac.in)

Over the last decade, Machine Learning (ML) has transformed from a tool used on controlled data by experts to a product deployed for public use. Traditional ML models are trained and tested under the assumption that instances are independent and identically distributed. They often under-perform when test and train distributions do not match. Due to their increasing model complexity, ML models usually are (i) trained once and released publicly or (ii) hosted as a service and accessed through APIs. When ML models are released publicly, we cannot control the distribution of the test queries. Therefore, it is crucial that a released model (a) generalize to test distributions, (b) adapt on the fly with labelled or unlabeled data, (c) declare their accuracy under distribution shifts. In what follows, we use the terms distribution and domain interchangeably to mean probability distribution of input instances.

Standard training methods fail to generalize to unseen domains because the available training data under-represent the distribution of domains from which test domains are sampled. ML models, particularly high capacity deep networks, can overfit train domains and fail to generalize to new domains. Section 1 summarizes our work on training methods for domain generalization.

Often labeled or unlabeled data is available for the target domain and we wish to use it for adaptation. The existing research literature on domain adaptation follows batched parameter update. However, neither (hardware deficient) users nor (heavily subscribed) servers can afford traditional fine-tuning in a server-client architecture. Section 2 summarizes our attempts at devising scalable adaptation methods that perform on-the-fly adaptation.

ML models must provide more formal accuracy specifications when released publicly. Today's ML services come with few accuracy specifications, perhaps only numbers from benchmarks that may not be closely related to the domain of most users. Ideally, they should go beyond reporting a few aggregated accuracy numbers and characterize domains where the model performs well and, more importantly, fails. Section 3 describes in more detail our work in this direction.

1 Training for Domain Generalization

Standard training methods fail to generalize to unseen domains for two reasons: (1) Under-representation: the training data is only a sparse and incomplete representation of the distribution of domains; (2) Overfitting: the training model entangles what is expressed (label) with how it is expressed (domain), failing to recognize labels when expressed in a new domain. Since acquiring more data is often impractical, we study training algorithms that generalize to new test domains by exploiting multiple domains during training called the domain generalization problem.

A popular approach for domain generalization is to assume and recover a domain invariant label classifier. A pioneering algorithm called Domain Adversarial Networks learns representations that do not reveal domain information. When domain and label are closely intertwined, domain invariant representations could also suppress the label information.

Alternatively, we can learn representations that share the same optimal classifier. We proposed CSD (Common Specific Decomposition) [1], which jointly learns a common component (which generalizes to new domains) and a domain-specific component (which overfits on training domains). We discard domain-specific components after the training and retain only the common component. We analyzed identifiability of the common component, and studied the effect of the number of domain-specific components on domain generalization. We found CSD beats or matches existing state-of-art approaches despite being simple.

A limitation of CSD is that, due to the linear decomposition assumption, it cannot extend the parameter decomposition beyond the simple linear classification layer, potentially limiting the gains. Subsequently, we proposed the CGD (Common Gradient Descent) [2] algorithm, which recovers the common component of parameter gradients without any decomposition assumption. Existing work established that standard training overfits and therefore generalizes poorly on minority domains when domain distribution during training is highly skewed. CGD avoids such minority overfitting by updating parameters with only gradients from common domains that enable better performance to all domains. Theoretically, we showed that the proposed algorithm is a descent method and finds first-order stationary points of smooth nonconvex functions.

Training algorithms guided by domain invariance are suitable when there exists a hypothesis that can perform well on all domains. Otherwise, a domain-invariant hypothesis can render uniform but arbitrarily low performance on all domains, including well-represented train domains. In such cases, when the training data sufficiently span the domain space, we can generalize better by augmenting with new domains that interpolate train domains. Based on this observation, we proposed CrossGrad [3]. We jointly train, along with the label classifier, a classifier that maps an example to its domain. We condition the label classifier with the continuous domain representation vector from the domain classifier, thereby explicitly encouraging the label classifier to use the domain information for label classification. We interpolated train domains by augmenting with domain-adversarial perturbation of the original example as additional training data for the label classifier.

2 Adaptation

Traditional adaptation methods that require parameter fine-tuning do not scale when deploying ML as a service. We researched a new breed of adaptation methods that can adapt on the fly without requiring parameter updates, revisiting source or target side labelled data. Our core insight is to leverage the additional context provided by target side unlabelled data.

We proposed KYC (Know-Your-Client) [4] designed for natural language applications. It recovers and conditions the main labelling network for each domain with a corpus-based sketch from unsupervised data. Further, we encourage the model to exploit the sketch for labelling. When a new domain registers with its sketch, KYC gets benefits immediately on various text tasks. However, KYC only used a simple corpus sketch to correct the final linear classification layer, limiting model expressivity and performance gains.

Recovering additional complementary information from the input and conditioning a pre-trained model on the auxiliary information is challenging. We examine end-to-end training

methods for aggregating and exploiting context information from any available unsupervised data in ongoing work to an even larger effect.

3 Evaluation

The performance of ML models can vary when the domain shifts, rendering non-uniform accuracy across different users of the model. The varying accuracy across users calls for prescribing the model accuracy on every user, without extensive computation requiring user-specific labeled or unlabeled data.

We characterize a user as any combination of pre-specified interpretable attributes. For example, the attributes for a speech recognizer task could be a combination of background noise and the speaker's gender, emotion, accent. We propose that a publicly released model should report accuracy as a surface over the interpretable attribute space. The user can then obtain accuracy over their data by querying the accuracy surface with user-specific attribute value assignment. Mapping accuracy for a combinatorially large number of users without access to per-user label data is challenging. We addressed this by smoothing related users using Gaussian-Process based estimator: AAA [5]. We sample the accuracy of the user from their associated Beta density. We showed that the obvious application of GPs cannot address the challenge of heteroscedastic uncertainty over a huge attribute space that is sparsely and unevenly populated. In response, we presented two enhancements: pooling sparse observations and regularizing the scale parameter of the Beta densities. We established the effectiveness of AAA in terms of its estimation accuracy and exploration efficiency in real-world scenarios.

Characterizing the domain shifts, like in AAA, with human-interpretable attributes is challenging, especially for high dimensional unstructured data such as text applications. Future studies could focus on creative new ways for declaring, diagnosing and detecting failures during deployment.

4 Future Work

I am passionate about improving ML models' reliability, interpretability and safety. I worked on three broad themes that addressed some of these challenges. The research on evaluating failures and on-the-fly adaptation requires further exploration. Moreover, despite the research community's sustained effort, the domain generalization problem is far from being solved, i.e. there is still a large gap between performance on seen and unseen domains. We need creative ideas to address these issues. I am excited about the potential impact of ML applications, but we have a long way to go before ML achieves the safety and trust levels of other engineering disciplines like civil or chemical engineering. I discuss some directions for future research below.

Guided Data Acquisition: The root cause and bottleneck for the lack of domain generalization is because training data does not densely cover the domain space and the underspecification of the hypotheses that it causes. We could use Bayesian methods to track and guide the acquisition of sparsely represented domains from large unlabeled data, which would require further study to ensure robust uncertainty estimate.

Generalization to anticipated domain shifts: The standard domain generalization problem does not specify the data shifts in train or test data. Practical systems, however, can identify the expected domain shifts during deployment. Therefore, a more natural variant of the problem is training and evaluating for robustness on task-specific and anticipated domain shifts. Defining such expected shifts is easier for structured data with interpretable random variables , but when modelling unstructured data, we need techniques that attune the high-level specification to the underlying data shift, such as interpreting data shifts that would have caused accent shift in a speech application.

Explicit label supervision: Recent work exposed how models exploited spurious correlations in data and raised alarm over the difficulty in detecting and avoiding them. A model trained using limited data is under-specified, and we risk learning several such spurious correlations along every axis of under-specification. Acquiring diverse examples and their labels to obliterate every such spurious correlation is inefficient, if not impractical. Humans do not simply learn from a notebook of example, label pairs. Instead, we also seek explanations on what constitutes a label and constant feedback from interactions that reconcile our reality with the truth. Exploration of explicit forms of label supervision is a promising direction for training robust models.

Editable models: A single model can never cater to the broad population, and the need for personalization and model editing is inevitable. The standard adaptation techniques can only digest instances (labelled or unlabelled), ignoring any thoughtful feedback that the user can readily provide. We need innovative forms of model editing that can engage the user in conversation on why the prediction is incorrect for data-efficient customization.

References

[1] Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In *International Conference on Machine Learning*, pages 7728–7738. PMLR, 2020.

[2] Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Focus on the common good: Group distributional robustness follows. *arXiv preprint arXiv:2110.02619*, 2021.

[3] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018.

[4] Sahil Shah, Vihari Piratla, Soumen Chakrabarti, and Sunita Sarawagi. Nlp service apis and models for efficient registration of new clients. *arXiv preprint arXiv:2010.01526*, 2020.

[5] Vihari Piratla, Soumen Chakrabarti, and Sunita Sarawagi. Active assessment of prediction services as accuracy surface over attribute combinations. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.