EECS 598 Project Draft

Anonymous Author(s)

Abstract

Transfer learning is where a source model trained on one domain is adapted for a downstream task on another domain. Recently, it has been shown that the unfair behaviors of the source model can persist even after it has been adapted for a downstream task. In this work, we propose a solution to this problem by using causally-motivated regularization schemes for creating fair source models through using auxiliary labels. Our regularization schemes work by enforcing independences with respect to the causal DAG. Our approach only requires having auxiliary labels at the time of source model training and it promotes adapted downstream models that don't make predictions based off of sensitive attributes. We show empirically and theoretically that source models that use our proposed causally-motivated regularization schemes lead to fairer downstream models and require less data to adapt to other tasks.

1 Introduction

Transfer learning is where a model that was trained for some arbitrary task is repurposed for a downstream task. It is a widely used technique in machine learning as models that utilize transfer learning require far less downstream domain data in order to perform well on a downstream task [22, 20]. The increase in finite sample efficiency is often necessary for models to learn in situations where little data is available for a downstream task. For instance, a person may want to train a model to predict a disease from chest X-rays, but if that disease is rare, their may not be enough samples to train a model from scratch [2]. To tackle this problem, Google has developed CXR Foundation [16], a CXR-specific model that generates feature-rich embeddings of chest X-rays that can be used for transfer learning. By using embeddings from CXR Foundation, they showed that an accurate model can be trained in as little as 10-100 samples. This is important as data may be sparse due to the rarity of a disease or it may take time for large datasets to become available if a disease is new.

Although transfer learning is very effective for learning a downstream task with little data, it also has serious problems related to fairness. Recently, it has been shown that unfair behavior can transfer during transfer learning from the source model to the downstream task [14]. This can lead to severe consequences when a source model learns sensitive attributes. For instance, Google's CXR Foundation model has evidence of producing embeddings that encode unfair racial and sex-related attributes [5]. Such a model can lead to unfair and potentially harmful results if a production model were to ever utilize its embeddings to learn a downstream task. Therefore, it is ideal that source models should produce embeddings that contain no information about sensitive attributes.

In this paper, we propose using supervised and unsupervised causally-motivated regularization schemes to train fair source models, i.e., models that produce feature embeddings invariant to sensitive attributes that may be spuriously associated with a label. We do not consider the case of fine-tuning the source model – we only utilize the source models feature embeddings. To help disentangle and remove the sensitive attributes from the feature embeddings, we utilize auxiliary labels similary to what is done in Makar et al. [11]. Learning a fair source model has two major benefits. First, a model trained for some downstream task will not utlize controlled sensitive attributes to make predictions, even if the data used to train the downstream task is biased. Second, there will be an increase in finite sample efficiency for learning downstream tasks compared to other source models.

Our main contributions in this paper are as follows:

- We show that causally-motivated regularization schemes can be applied to create fair source models for transfer learning.
- We show that fair source models adapted for some downstream task will still be fair with respect to controlled sensitive attributes, even if the data used to train the downstream task has unfair spurious associations.
- We show theoretically that less data is required to adapt fair source models that were trained with causally-motivated regularization schemes.

2 Related Work

Transfer Learning Much work has covered the benefits of using transfer learning to learn a downstream task [8, 13]. Other work has shown that when the source dataset is closer to the dataset of the downstream task, finite sample efficiency can be improved further [17]. This has led to source models that are used for specific applications, such as for various types of medical imaging [16].

Researchers have also looked at transfer learning from a causal perspective. For instance, Yang et al. [21] developed a Causal Autoencoder (CAE) for domain adaption. This autoencoder attempts to seperate causal representations from task-irrelevant representations for more robust domain adaption. Schölkopf et al. [15] discussed the possible implications of how causal representation learning could impact transfer learning.

Causally-motivated invariance Our work is most similar to Makar et al. [11], where the authors present a causally-motivated shortcut removal regularization scheme that utilizes auxiliary labels. Their work prevents shortcut learning through a reweighting scheme followed by MMD regularization that enforces independences implied by the causal DAG. Similarly, Veitch et al. [18] presents an approach that uses auxiliary labels to create a counterfactually invariant predictor.

Other approaches do not rely on auxiliary labels, but instead focus on creating invariant predictors across different environments. For instance, Invariant Risk Minimization (IRM) [1] splits the training set into separate environments, and seeks a data representation that elicits a classifier that is optimal across all environments. Another approach, proposed by Wang et al. [19], uses data augmentations that only modify non-causal features, called causal invariant transformations. They introduce a regularization to promote a model that learns similar representations of the original sample and samples produced by causal invariant transformations.

Fairness Literature A lot of research has gone into showing that machine learning models can learn sensitive attributes to make unfair predictions [10, 12]. Work by Gajane et al. [4] has looked at different ways to formalize fairness in machine learning predictions. The formalization of fairness that most aligns with this paper is counterfactual fairness, which as introduced by Kusner et al. [9].

3 Approach

Here, we describe our approach to learning a fair source model. We assume the data is generated by the causal DAG in Figure 1, where Y is the main label, V is an auxiliary label, and X is the input. We assume Y only affects X through X^*. We consider two different settings for training a source model: an unsupervised approach where we only have access to auxiliary labels of sensitive attributes and a supervised approach, where we also have access to the main label.

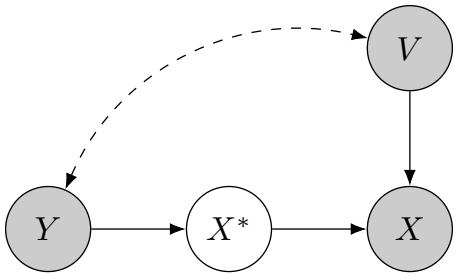

Figure 1: Causal DAG that encodes the assumptions of our setting.

Supervised base model For this approach, we will train the base model $f = h(\phi(X))$ using a similar causally motivated regularization scheme by Makar et al. [11] – we will denote this approach as HSIC-S. First, given the source distribution $D \sim P_s$, we map the learning problem to the unconfounded distribution P° using re-weighting. We define the weights as

$$u(y, v) = \frac{P_s(Y = y)P_s(V = v)}{P_s(Y = y, V = v)} \tag{1}$$

for each sample $u_i := u(y_i, v_i)$. Each weight u_i is then normalized, denoted as $\tilde{u_i}$, such that $\sum_i \tilde{u_i} = 1$.

Next, to reduce the variance of our estimator, we will enforce the conditional independence $\phi(X) \perp V$ as is implied by the causal DAG. Instead of using MMD to enforce $\phi(X) \perp V$, we will use the Hilbert-Schmidt Independence Criterion (HSIC) [6]. Although the MMD approach used in [11] is more minibatch-friendly than the equivalent approach used by Veitch et al. [18], it is still not ideal. For instance, if we wanted our feature embedding to be invariant to both sex and race, we would have to split our dataset into twelve different sex-race groups for each minibatch (assuming there is only six different races). This will require very large minibatch sizes in order to make sure there is a sufficient amount of samples in each group. In contrast, HSIC will not require any minibatch to be split.

Let $D \sim P_s$, let P^u_ϕ be the weighted distribution of ϕ, and let there be some $\alpha > 0$. Then the objective for the source model is:

$$h^*, \phi^* = \arg\min_{h,\phi} \sum_i u_i \cdot l(h(\phi(x_i)), y_i) + \alpha \cdot \widehat{HSIC}(P^u_\phi, V)$$

Unsupervised base model The supervised approach may fail to produce embeddings that transfer effectively to other settings. For instance, consider the setting of where a source model is learning whether or not a person is healthy given a chest X-ray. If the majority of the unhealthy base training data are X-rays of patients with pneumonia, the learned embedding may only contain information needed to classify pneumonia. If we then wanted to adapt the source model for a downstream task that predicts a disease with symptoms unrelated to pneumonia, it may fail to produce an effective downstream model. Therefore, we will test an unsupervised approach, which is denoted as HSCIC-UN.

For this approach, we train an autoencoder $f(x) = h(\phi(x))$, where ϕ is the encoder and h is the decoder. Ideally, we want to find ϕ that produces samples in the latent space that are invariant to sensitive attributes V. To do this, we enforce the conditional independence $\phi(X) \perp V$ in conjunction with the reconstruction error. Once the autoencoder is trained, we will only utilize ϕ for the source model.

Let $D \sim P_s$ and let there be some $\alpha > 0$. Then the objective for the source model is:

$$h^*, \phi^* = \arg\min_{h,\phi} \sum_i \|x_i - h(\phi(x_i))\|_2^2 + \alpha \cdot \widehat{HSIC}(P^u_\phi, V)$$

4 Theory and Experiments

For the theory section, we will show that both the supervised and unsupervised approaches lead to source models that have increased finite sample efficiency for adapting to downstream tasks compared

to models that only utilize L2 regularization, denoted L2-S and L2-UN. I will work towards proving two different Propositions to show that this is true. For the final paper, I will write and prove these Propositions using more rigorous mathematical notation.

Proposition 4.1. *Less data is required to adapt supervised source models that were trained with HSIC-S and L2 than with only L2 regularization given some conditions hold.*

Proposition 4.2. *Less data is required to adapt AE source models that were trained with HSIC-UN than with only L2 regularization given some conditions hold.*

Proposition 1 should follow easily from combining Proposition A5 in [11] and from Theorem 3 in [17]. Proving Proposition 2 will not be a trivial task. I will have to find the upper bound of the Rademacher complexity of the function space of $\mathcal{F}_{HSIC-UN}$. If the upper bound of the complexity of $\mathcal{F}_{HSIC-UN}$ is less than the complexity of $\mathcal{F}_{L2-UN}$, I should be able to prove Proposition 2.

For our experiements, we consider the setting of creating a chest X-ray specific source models that can be adapted for a variety of chest X-ray prediction tasks. Our approach of pretraining is similar to Google's CXR Foundation model: we first pretrain a ResNet-50 [7] on ImageNet [3] and then we train the model over a dataset of chest X-rays.

The dataset used to train the models was derived from MIMIC-III and was manipulated to create unfair spurious associations. We sampled the data such that the majority of women were healthy and the majority of men were unhealthy, i.e., P(Y = Healthy | V = Women) = P(Y = Unhealthy | V = Men) = 0.9. Furthermore, we removed all instances of pneumonia in the training set to see how well each method performs on unseen diseases.

For HSIC-S and HSIC-UN, we will train the models with four different kernel bandwidths and four different costs using 5-fold cross-validation. The best HSIC-UN model is chosen is the 5-fold cross validation and the best HSIC-S model is chosen using two-step cross validation [11]. To see if the regularization schemes are working, I will compare these models to a supervised and unsupervised models that only use L2 regularization. Furthermore, I will run the following experiments:

- Test the models over a dataset with a similar distribution to the biased training dataset and test the models where the test dataset distribution is flipped, i.e., P(Y = Healthy | V = Women) = P(Y = Unhealthy | V = Men) = 0.1.
- Perform the previous tests where the models are trained on data with no association between health and sex.
- Test the models over a datasets only including pneumonia samples.

5 Completed Steps and Next Steps

Step 0 and 1 (Completed) I have retrieved the MIMIC-III dataset and finished the data pipeline.

Step 2 (Deadline 11/10) Finish creating the code to create and train the source models in PyTorch. I am mostly done with this step, but I still have some stuff I need to finish up. This code should be able to train the models using different kernel bandwidths, different regularization costs, and different learning rates. It should save the models to be used later, and it should also have Tensorboard implemented so that I can verify that the models have converged when training.

Step 3 (Deadline 11/17) Train the models on the GreatLakes cluster. Based on prior experience, I think that training will probably take about a week or so. I will train four types of models: two supervised source models with L2 and HSIC regularization, and two unsupervised base models with L2 and HSIC regularization. The models trained only with L2 regularization will be used as a baseline. The training data will under-sample women who are healthy and over-sample men who are not healthy so that sex will be associated with who is unhealthy.

Step 4 (Deadline 11/24) Run each of the experiments described at the end of Section 4. Create plots to analyze and interpret the results. Make conclusions based off of the empirical results; explain why the proposed methods either worked or did not work.

Step 5 (Deadline 11/27) Prove Proposition 1 and prove or disprove Proposition 2.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Salman Ul Hassan Dar, Muzaffer Özbey, Ahmet Burak Çatlı, and Tolga Çukur. A transfer-learning approach for accelerated mri using deep neural networks. *Magnetic resonance in medicine*, 84(2): 663–685, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*, 2017.

Ben Glocker, Charles Jones, Melanie Bernhardt, and Stefan Winzeck. Risk of bias in chest x-ray foundation models. *arXiv preprint arXiv:2209.02965*, 2022.

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. *Advances in neural information processing systems*, 20, 2007.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf`.

Nicol Turner Lee. Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society*, 16(3):252–260, 2018.

Maggie Makar, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D'Amour. Causally motivated shortcut removal using auxiliary labels. In *International Conference on Artificial Intelligence and Statistics*, pages 739–766. PMLR, 2022.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32, 2019.

Hadi Salman, Saachi Jain, Andrew Ilyas, Logan Engstrom, Eric Wong, and Aleksander Madry. When does bias transfer in transfer learning? *arXiv preprint arXiv:2207.02842*, 2022.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Andrew B Sellergren, Christina Chen, Zaid Nabulsi, Yuanzhen Li, Aaron Maschinot, Aaron Sarna, Jenny Huang, Charles Lau, Sreenivasa Raju Kalidindi, Mozziyar Etemadi, et al. Simplified transfer learning for chest radiography models using less data. *Radiology*, 305(2):454–465, 2022.

Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in neural information processing systems*, 33:7852–7862, 2020.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv preprint arXiv:2106.00545*, 2021.

Ruoyu Wang, Mingyang Yi, Zhitang Chen, and Shengyu Zhu. Out-of-distribution generalization with causal invariant transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 375–385, 2022.

Liu Yang, Steve Hanneke, and Jaime Carbonell. A theory of transfer learning with applications to active learning. *Machine learning*, 90:161–189, 2013.

Shuai Yang, Kui Yu, Fuyuan Cao, Lin Liu, Hao Wang, and Jiuyong Li. Learning causal representations for robust domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 35(3): 2750–2764, 2023. doi: 10.1109/TKDE.2021.3119185.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.