Extended Proposal: Energy-Based Domain Generalization for Machine Learning Models in Healthcare

Lev Paciorkowski LP2663@NYU.EDU *Center for Data Science*
New York University
New York, New York, USA

Joseph Edell JRE6163@NYU.EDU *Department of Computer Science*
New York University
New York, New York, USA

1. Problem Description

For a model to be effective in a real-world healthcare setting, it must be able to make accurate predictions in new situations. How a model performs in development versus how it performs in production can vary based on its stability: a model's robustness in response to data set shifts. We want to examine methods of generalization to determine the most effective techniques to deploy useful, production-ready models. We plan to focus on deep learning methods involving energy-based models to produce shift-stable representations of data that can generalize to different medical settings.

It is useful to provide a more formalized definition of our problem: the following mathematical framework is adopted from Blanchard, Lee and Scott (2011). Given an arbitrary input space $\mathcal{X}$ and outcome space $\mathcal{Y}$, we define the notion of a **domain** as a probability distribution P_{XY} over $\mathcal{X} \times \mathcal{Y}$. Without loss of generality, we can break apart P_{XY} into the pair $(P_X, P_{Y|X})$.

Define $\Upsilon_{\mathcal{X}}$ to be the set of all possible P_X, and $\Upsilon_{\mathcal{Y}|\mathcal{X}}$ to be the set of all possible $P_{Y|X}$. Then suppose we have probability distributions μ_X and $\mu_{Y|X}$ over $\Upsilon_{\mathcal{X}}$ and $\Upsilon_{\mathcal{Y}|\mathcal{X}}$, respectively. We then assume our data generating process is as follows:

1. Choose a $P_X \sim \mu_X$ and a $P_{Y|X} \sim \mu_{Y|X}$ (this defines the domain).
2. Generate samples, each sample being an $S^{(k)} = \{(x^{(i)} \sim P_X^{(k)}, y^{(i)} \sim P_{Y|X}^{(k)})\}_{i=1}^{n^{(k)}}$

We assume that we have multiple training samples S, coming from different domains, for which the labels y are observed. We further assume that there is some other target (or test) domain for which we observe x with or without any labels y. The ultimate goal is then to find some prediction function $f(x, \cdot)$ which outputs a $\hat{y}$ that persists as a "good" prediction in the test domain. As is detailed in section 2, methodologies for structuring and searching for f vary, hence the additional optional arguments as indicated.

While we would like to focus on the unsupervised case with no target labels observed, we also plan to include experimental results for the semi-supervised case. Both scenarios are realistic for healthcare machine learning tasks: frequently a model may be deployed in a new environment from which it has not seen data during training; however, it is also not

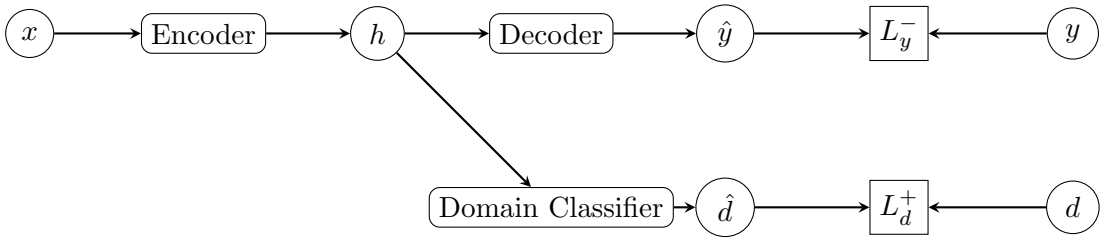
uncommon for models to undergo a trial period in deployment - this could be an opportunity to observe new data from the target domain. This is overall a pervasive but difficult problem to address; hence the interest in working on it.

2. Prior Work

The modern theoretical paradigm of domain generalization largely rests on the seminal works of Ben-David, et al. (2006, 2010) and Blanchard, Lee and Scott (2011), who all together provide us with a formalized mathematical framework (as outlined in Section 1) for thinking about the problem.

Early work tended to structure the prediction function as an $f(x, P_X)$ (or more generally, $f(x, \hat{P}_X)$) that would consider not only the input but also the input's domain (or estimated domain) for making predictions. For example, in Muandet, et al. (2013), the authors use kernel methods to devise an algorithm that searches for such a prediction function $f : \Upsilon_X \times \mathcal{X} \to \mathbb{R}$. They accomplish this by first mapping input features to a reproducing kernel Hilbert space, $\mathcal{H}$, and then using an optimization algorithm to find some transformation $\mathcal{B}$ in $\mathcal{H}$ that minimizes the distances between the training samples (within $\mathcal{H}$) while still preserving the functional relationship between x and y.

More recent innovations have focused on deep learning models. A particularly groundbreaking architecture was proposed in Ganin, et al. (2016), building on much of the rigorous theoretical findings of the previously mentioned Ben-David, et al. (2006, 2010). Their neural network architecture is depicted below:



In this novel architecture, there is an adversarial loss function (L_d^+) on the predicted domain from the domain classifier which is *maximized*, while the label loss L_y^- can be any standard loss function to be minimized. During backpropagation, the gradients from the domain classifier are reversed to achieve maximization of L_d^+. This ultimately has the effect of learning an encoder such that the decoder can accurately predict y from x, but the domain classifier can *not* accurately predict which domain x came from. In other words, this encourages the model to develop a hidden representation h that focuses only on the most critical elements of x for the task of predicting y.

Even more recent work, which we will attempt to develop further, focuses on the use of energy-based models for domain generalization. A good overview of EBMs in general may be found in LeCun, et al. (2006). In Xie, et al. (2022), the authors take advantage of the inherent "free-energy" artifact of EBMs to devise a learning algorithm centered on selectively incorporating small annotated samples from the test domain during each training

round. We suspect that it may be possible to combine a similar EBM architecture with an architecture like that of Ganin, et al. (2016) to devise a model that is capable of even fully unsupervised domain generalization.

3. Proposed Methodology

We propose to create an energy-based model with an adversarial loss component similar to the one demonstrated in Ganin, et al. (2016). Our end goal is to have a model which is able to discover the functional relationship (which we assume is at least reasonably consistent across domains) between x and y without becoming distracted by the domain shifts in the training data it sees during learning.

However, taking into account a particular input's domain is likely to be at least *somewhat* useful for a prediction function, if this information is used in the right way. For example, if there is substantial variation in P_{XY} across domains, then it may be worthwhile to see if a model can profitably incorporate some information about P_Y or P_X into its prediction function. If we assume it is reasonable to have estimates of such meta-information about our different domains, then this could sensibly be incorporated into an EBM architecture. Then, when testing on a new domain, we could provide the model with our best external estimates available for the underlying P_{XY}. As an analogy, this could be thought of as being akin to alerting a physician changing practice to a different hospital (in a different domain) that the underlying environment is different and can be encoded in this estimate of P_{XY}. We think this idea has some intuitive justification and would like to at least experiment with it to see if it can be made viable.

4. Experimental Setup

Set up will involve obtaining access to the proposed data sets, implementing and collaborating on our model architecture, and splitting our data sets into cohorts based on a task while preserving the original domain. Both MIMIC-CXR and the CheXpert dataset require credentialed access with MIMIC-CXR additionally requiring that users complete a training as well. The PadChest and Openi dataset are freely available for download. Once we have obtained access to our data, it can be stored in a cloud persistent storage location.

We plan to use Github to manage our code versioning and collaboration, Pytorch to build our model, and will rely on a dedicated compute cluster to train. Additionally, we estimate that we may need up to two terabytes worth of persistent storage to house the data.

Our goal is to produce accurate predictions across differing data domains, thus our model will be trained on subsets of available data domains and tested on an entirely held out dataset. Additionally, the data will be separated into multiple cohorts, each dedicated to a different binary classification task (i.e. identifying if pneumonia is present or not). Each cohort will consist of items that share the same view position and diagnoses labels across all data sets. This step will require some standardization of diagnosis labels across each

data set using ICD-10 codes.

Within each task cohort, a subset of available domains will be selected for training and validation data, while the items from remaining domains will be used for tests. MIMIC-CXR contains the most volume of data and thus will be used for training and validation. Using unseen domains as test sets will allow us to measure the domain-generalisation capabilities of our model.

5. Evaluation Plan

Our model will be evaluated by comparing the model's accuracy on the validation set using MIMIC-CXR data with the accuracy of the external domain test set. If our model generalizes well, we should see similar AUC calculations for each external domain when compared to our models performance on the MIMIC-CXR validation set. These results will be compared to a standard deep CNN model following the same methodology for training and testing.

6. Questions

- How should we choose baseline models for comparison? Should we try to directly compare with an architecture from a previous work, or is it OK if we compare with simple architectures like a basic SVM or CNN?

7. References

1. Adarsh Subbaswamy, Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health AI, Biostatistics, Volume 21, Issue 2, April 2020, Pages 345–352, https://doi.org/10.1093/biostatistics/kxz041.

2. Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks, 2017; arXiv:1706.06083.

3. Ben-David, Shai, et al. "Analysis of representations for domain adaptation." Advances in neural information processing systems 19 (2006).

4. Ben-David, Shai, et al. "A theory of learning from different domains." Machine learning 79.1 (2010): 151-175.

5. Blanchard, Gilles, Gyemin Lee, and Clayton Scott. "Generalizing from several related classification tasks to a new unlabeled sample." Advances in neural information processing systems 24 (2011).

6. Bustos, Aurelia, et al. PadChest: A Large Chest x-Ray Image Dataset with Multi-Label Annotated Reports. Medical Image Analysis, vol. 66, Dec. 2020, p. 101797. Crossref, https://doi.org/10.1016/j.media.2020.101797.

7. Eche T, Schwartz LH, Mokrane FZ, Dercle L. Toward Generalizability in the Deployment of Artificial Intelligence in Radiology: Role of Computation Stress Testing to Overcome Underspecification. Radiol Artif Intell. 2021 Oct 27;3(6):e210097. doi: 10.1148/ryai.2021210097. PMID: 34870222; PMCID: PMC8637230.

8. Ganin, Yaroslav, et al. "Domain-adversarial training of neural networks." The journal of machine learning research 17.1 (2016): 2096-2030.

9. Irvin, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. arXiv, 2019. https://doi.org/10.48550/arxiv.1901.07031.

10. Johnson, A., Pollard, T., Mark, R., Berkowitz, S., Horng, S. (2019). MIMIC-CXR Database (version 2.0.0). PhysioNet. https://doi.org/10.13026/C2JT1Q.

11. LeCun, Yann, et al. "A tutorial on energy-based learning." Predicting structured data 1.0 (2006).

12. Muandet, Krikamol, David Balduzzi, and Bernhard Schölkopf. "Domain generalization via invariant feature representation." International Conference on Machine Learning. PMLR, 2013.

13. Weicheng Zhu and Narges Razavian. Variationally Regularized Graph-based Representation Learning for Electronic Health Records, 2019; arXiv:1912.03761. DOI: 10.1145/3450439.3451855.

14. Xie, Binhui, et al. "Active learning for domain adaptation: An energy-based approach." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. No. 8. 2022.

15. Yingtao Luo, Zhaocheng Liu and Qiang Liu. Deep Stable Representation Learning on Electronic Health Records, 2022; arXiv:2209.01321.