This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Towards Explaining Image-Based Distribution Shifts

Sean Kulinski, David I. Inouye
School of Electrical and Computer Engineering
Purdue University
{skulinsk, dinouye}@purdue.edu

Abstract

Distribution shift can have fundamental consequences such as signaling a change in the operating environment or significantly reducing the accuracy of downstream models. Thus, understanding such distribution shifts is critical for examining and hopefully mitigating the effect of such a shift. Most prior work has focused on either natively handling distribution shift (e.g., Domain Generalization) or merely detecting a shift while assuming any detected shift can be understood and handled appropriately by a human operator. For the latter, we hope to aid in these manual mitigation tasks by explaining the distribution shift to an operator. To this end, we suggest two methods: providing a set of interpretable mappings from the original distribution to the shifted one or providing a set of distributional counterfactual examples. We provide preliminary experiments on these two methods, and discuss important concepts and challenges for moving towards a better understanding of image-based distribution shifts.

1. Introduction

Most real-world environments are constantly changing and understanding how a specific operating environment has changed is crucial to making decisions respective to such a change. Such a change might be a new data distribution seen in deployment which causes a machine learning model to begin to fail. When these changes are encountered, the burden is often placed on a human operator to investigate the shift and determine the appropriate reaction, if any, that needs to be taken. In this work, our goal is to aid these operators by providing an explanation of such a change.

This ubiquitous phenomena of having a difference between related distributions is known as distribution shift. Much prior work focuses on *detecting* distribution shifts; however, there is little prior work on *understanding* or *characterizing* a detected distribution shift. A naïve baseline in analyzing an image-based distribution shift is to compare a grid of samples from the original, i.e., *source*, distribution to a grid of samples from the new, i.e., *target*, distribution. However, due to the complexity of image-based shifts, this approach can be uninterpretable or even misleading to an operator (e.g., the left parts of Figure 1 and Figure 2).

Therefore, we propose two preliminary methods for explaining image-based distribution shifts and discuss open challenges. The first is a novel framework which provides an operator with interpretable mappings which shows how latent features have changed or how latent groups have shifted between the distributions. The second approach is similar to that of unpaired Image-to-Image Translation (I2I) [14] such as CycleGAN [24], and explains the shift to the operator as pairs of a real example and its corresponding counterfactual example. These counterfactuals are generated by mapping samples from one domain to the other domain such that the distributions become indistinguishable. We summarize our contributions as follows:

- We introduce high-dimensional interpretable transport maps for explaining image-based shifts if an interpretable latent space is known.
- We also leverage prior I2I work to explain image-based distribution shifts via counterfactual examples if an interpretable latent space is unavailable.
- We provide preliminary results and interpretations.
- We discuss open questions for explaining image-based distribution shifts.

2. Explaining Image Distribution Shifts via Transportation Maps

The underlying assumption of distribution shift is that there exists a relationship between the source and target distributions. From a distributional standpoint, we can view distribution shift as a *movement*, or transportation, of samples from the source distribution P_{src} to the target distribution P_{tgt}. Thus, we can capture this relationship between the distributions via a transport map T from the source distribution to the target, i.e., if $\boldsymbol{x} \sim P_{src}$, then $T(\boldsymbol{x}) \sim P_{tgt}$. Additionally, if an interpretable representation of the map

T can be formed, this representation can be provided to an operator to aid in understanding and reacting to shifts more effectively. However, an interpretable representation likely requires interpretable (latent) features, which may not be available for some image domains. In this case, we can represent the map by merely showing pairs of inputs $\boldsymbol{x}$ and "counterfactual" outputs $T(\boldsymbol{x})$. Therefore, we define a shift explanation to be: *a (possibly interpretable) transport map T that maps a source distribution P_{src} onto a target distribution P_{tgt} such that $T_\sharp P_{src} \approx P_{tgt}$.*

2.1. Interpretable Transportation Maps

In order to find such a mapping between distributions, it is natural to look to Optimal Transport (OT) due to it allowing for a rich geometric structure on the space of distributions and having extensive prior work in this field [1, 5, 15, 21]. An OT mapping is originally defined by Monge [15, 22] as a method of aligning two distributions in a minimal cost way given a transport cost function c. To find interpretable transport maps, we build upon the OT framework by restricting the candidate transport maps to belong to a set of user-defined interpretable mappings Ω. Additionally we use a Lagrangian relaxation on the full alignment constraint seen in OT, giving us an *Interpretable Transport* mapping T_{IT}:

$$T_{IT} \coloneqq \underset{T \in \Omega}{\arg\min}\ \mathbb{E}_{P_{src}}\left[c(\boldsymbol{x}, T(\boldsymbol{x}))\right] + \lambda\, \phi(P_{\mathrm{T}(\boldsymbol{x})}, P_{tgt}) \tag{1}$$

where $\phi(\cdot,\cdot)$ is a divergence function, which, unless otherwise stated, is assumed to be the squared Wasserstein-2 metric, W_2^2.

An example of an set of interpretable mappings Ω is k-cluster mappings. Where given a $k \in \{1, \ldots, d\}$ we define k-cluster transport to be a mapping which moves each point $\boldsymbol{x}$ by constant vector which is specific to $\boldsymbol{x}$'s cluster. More formally, we define a labeling function $\sigma(\boldsymbol{x}; M) \triangleq \arg\min_j \|\boldsymbol{m}_j - \boldsymbol{x}\|_2$, which returns the index of the column in M (i.e., the label of the cluster) which $\boldsymbol{x}$ is closest to. With this, we define $\Omega^{(k)}_{\text{cluster}} = \left\{T : T(\boldsymbol{x}) = \boldsymbol{x} + \delta_{\sigma(\boldsymbol{x};M)}, M \in \mathbb{R}^{d\times k}, \Delta \in \mathbb{R}^{d\times k}\right\}$, where δ_j is the j^{th} column of Δ. For another set of interpretable mappings (k-sparse transport) and methods for solving for these mappings in practice, please see section Appendix C.

In order to find interpretable transport mappings for high dimensional spaces like images, we can project P_{src} and P_{tgt} onto an *interpretable* latent space (e.g., a space which has disentangled and semantically meaningful dimensions) which is learned by some (pseudo-)invertible function $g : \mathbb{R}^d \to \mathbb{R}^k$ where $k < d$ (e.g., an autoencoder). Then we can solve for an interpretable mapping such that it aligns the distributions in the latent space, $P_{T(g(\boldsymbol{x}))} \approx P_{g(\boldsymbol{y})}$. For counterfactual purposes, we can use g^{-1} to project $T\left(g(\boldsymbol{x})\right)$ back to $\mathbb{R}^d$ in order to display the transported image to an operator. With this, we can define our set of high dimensional interpretable transport maps: $\Omega_{\text{high-dim}} \coloneqq \left\{T : T = g^{-1}\left(\tilde{T}\left(g(\boldsymbol{x})\right)\right), \tilde{T} \in \Omega, g \in \mathcal{I}\right\}$ where Ω a the set of interpretable mappings and $\mathcal{I}$ is the set of (pseudo-)invertible functions with a interpretable (i.e., semantically meaningful) latent space. Given an interpretable $g \in \mathcal{I}$, we define our problem as:

$$\underset{\tilde{T} \in \Omega^{(k)}}{\arg\min}\ \mathbb{E}_{P_{src}}\left[c\left(g(\boldsymbol{x}), \tilde{T}\left(g(\boldsymbol{x})\right)\right)\right] + \lambda\phi(P_{\tilde{T}(g(\boldsymbol{x}))}, P_{g(\boldsymbol{y})}) \tag{2}$$

which results in an interpretable map $\tilde{T}$ which approximately shows how images from P_{src} shifted to P_{tgt} in a semantically meaningful way (e.g., how the H&E staining in histopathology images changes across hospitals).

2.2. Counterfactuals via Unpaired Image-to-image Translation

In some cases, a shift cannot be expressed by an interpretable mapping function because an interpretable latent space is not known. Thus, we can remove the interpretability constraint, and leverage methods from the unpaired Image-to-Image translation (I2I) literature to translate between the source and target domain while preserving the content. For a comprehensive summary of the recent I2I works and methods, please see [14]. Once a mapping is found, to serve as an explanation, we can provide an operator with a set of counterfactual pairs $\{(\boldsymbol{x}, T(\boldsymbol{x})) : \boldsymbol{x} \sim P_{src}, T(\boldsymbol{x}) \sim P_{tgt}\}$. Then, by determining what commonly stays invariant and what commonly changes across the set of counterfactual pairs, this can serve as an explanation of how the source distribution shifted to the target distribution. While more broadly applicable, this approach could put a higher load on the operator than the interpretable mapping approach.

3. Experiments

In this section we provide preliminary results showing the advantages and shortcomings of explaining shifts via interpretable transportation maps and via counterfactual pairs. We begin with explaining a shifted Color MNIST dataset via cluster-based transportation maps using a semi-supervised VAE [18]. Next, we use StarGAN [4] to generate counterfactual examples to explain the shift in histopathology images across five hospitals as seen in the Stanford Wilds [9] variant of the Camelyon17 dataset [2]. [1]

[1] Code to recreate all experiments can be found at https://github.com/inouye-lab/towards-explaining-image-distribution-shifts.

3.1. Explaining a Colorized-MNIST shift via High-dimensional Interpretable Transport

This experiment consists of using k-cluster maps to explain a shift in a colorized-version of MNIST, where the source environment has more yellow digits with a light gray background while the target environment consists of more red digits and/or darker gray backgrounds. The data is created by randomly red/yellow colorizing the foreground and grayscale coloring the background of 60,000 grayscale MNIST digits [6]. The source distribution P_{src} is set to be be any images where colorized digits that had over 40% of the green channel visible (thus yielding a yellow color) *and* a background at least 40% white, and the target environment P_{tgt} is all other images. Informally, this split can be thought of as three heterogeneous sub-shifts: a shift which is only reddens the foreground digit, a second shift which only a darkens the background, and a third shift which both reddens the digit reddening and darkens the background. The environments can be seen in Figure 3 in the Appendix.

We follow the framework presented in Equation 2, where g is a semi-supervised VAE [18] with a latent dimension of 50. The SSVAE was trained for 200 epochs on a concatenation of both P_{src} and P_{tgt} with 80% of the labels available per environment, and a batch size of 128 and otherwise followed the training details in [18]. To explain the shift, we use Algorithm 1 in the appendix to learn $k = 3$ cluster maps because there are 3 subshifts.

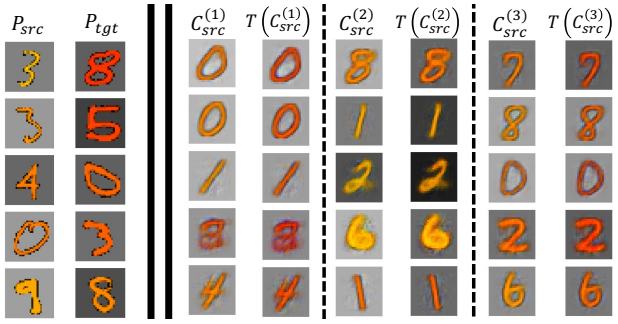

Figure 1. The baseline of unpaired source and target samples (left) is unable to distinguish between the three subshifts. Our cluster-based transport (right) separates the shift into 3 subshifts: $C^{(1)}$ clearly reddens the digit color but maintains the background color, $C^{(2)}$ clearly darkens the background color but maintains the digit color, and $C^{(3)}$ changes both the digit color and background color.

While the cluster map is inherently simple because each map merely translates points by a constant vector, the latent features are not disentangled into semantically meaningful features. Thus, to represent the cluster map, we merely show input and output pairs for each cluster map. The goal is for the operator to discern the meaning of each cluster's shift by finding the invariances for each cluster. The cluster based explanations can be seen in Figure 1. Our preliminary results demonstrate that k-cluster transport can explain this heterogeneous shift by separating at least two distinct shifts in the data. However, we acknowledge that this is a relatively simple example and expect more work will be needed to improve this idea for real-world image shifts.

3.2. Explaining Shifts in H&E Images Across Hospitals via Counterfactual Examples

This experiment explores the alternative for explaining image-based distribution shifts by supplying an operator with a set of translated images (i.e., a set of images from the source distribution which have been altered to look like they belong to the target distribution), with the notion that the operator would resolve which semantic features are distribution-specific. We apply this approach the Camelyon17 dataset [2] which is a real-world distribution shift dataset that consists of whole-slide histopathology images from five different hospitals. We use the Stanford WILDS [9] variant of the dataset which converts the whole-slide images into over 400 thousand patches. Since each hospital has varying hematoxylin and eosin (H&E) staining characteristics, this, among other batch effects, leads to heterogeneous image distributions across hospitals as can be seen in Figure 2.

To generate the counterfactual examples, we treat each hospital as a domain and train a StarGAN model [4] to translate between each domain. For training, we followed the original training approach seen in [4], with the exception that we perform no center cropping. After training, we can generate image counterfactual examples via inputting a source image and the label of the target hospital domain to the model.

Counterfactual generation was done for all five hospitals and can be seen in the right-hand side of Figure 2. It can be seen that the StarGAN model captures the different staining characteristics across the hospitals. For example, hospital 1 (P_1) consists of mostly light staining and thus transporting to this domain usually involves a lightening of image while P_3 seems to have more hematoxylin stain thus leading to deeper purple images when pushing onto this domain. We can also see that the model tends to respect the content of the image where patches which contain tumor cells (e.g., the P_5 sample on the right-hand side) still contain tumor cells in the counterfactual cases and likewise for lymphocyte cells (e.g., the P_4 sample on right-hand side).

4. Open Questions for Explaining Image-based Distribution Shifts

In this section, we introduce a series of open questions which we hope will help move towards developing the foundations for explaining image-based distribution shifts including defining exemplar tasks, metrics, datasets, and

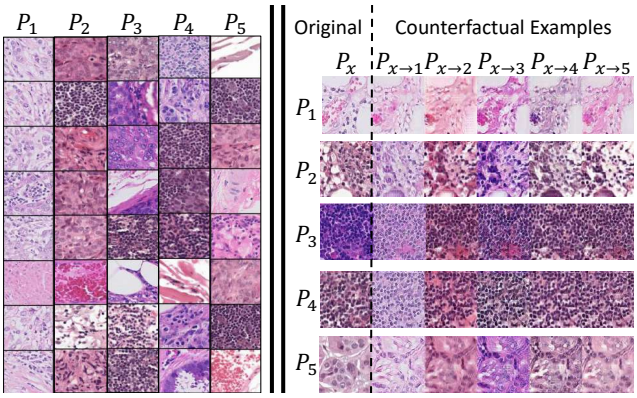

Figure 2. The baseline method of unpaired samples (left) which requires many samples to begin to understand the differences across the hospitals domains (represented as P_1, P_2, ...). Our explanation approach (right) of showing paired counterfactual images translated between the hospital domains (where the (i, j) row, column pair represents the pushforward of the i$^{\text{th}}$ domain onto the j$^{\text{th}}$ domain) quickly makes it clear how the staining/coloring differs across the hospital domains.

baseline methods. We begin with introducing tasks where an operator would need to *understand* a distribution shift and give criteria for finding exemplar datasets which can serve as benchmarks for the tasks. Then we discuss possible other approaches for explaining distribution shift and close with suggesting criteria for evaluating and comparing such methods.

We (non-exhaustively) envision several possible tasks: **Knowledge discovery** - This would entail helping an operator extract knowledge by characterizing the differences between distributions (e.g., finding important differences in nanostructure imaging with different experimental conditions), and would focus on complex distribution shifts that would not be easy to understand using conventional visualization or dataset inspection tools. **Post-hoc explanations of model failure due to shifts** - This would involve finding the qualitative differences between the training environment and this new testing environment that caused the model to fail. It would help an operator answer the question: Can we determine how to alleviate this problem? Should we collect more labeled data, adjust the instrument, or robustify the model? **Detecting adversarial shifts** This would help an operator determine if the distributional changes are due to benign effects or due to an adversary (i.e., an enemy compromises a surveillance camera). Due to the highly context-dependent nature of distribution shift, it would be beneficial to have exemplar datasets on which to train and evaluate methods for each of these tasks. Ideally, these distribution shift examples would be complex distribution shifts—not something that can be easily explained by a simple plot or by looking at the difference in mean statistics—, have real-world use cases where understanding the distribution shift is important, and has some form of known oracle explanation(s) that could be used to validate a predicted explanation against.

In this paper we introduced a novel way for explaining image-based distribution shifts via interpretable transport maps; however, there are other ways characterize and explain an image-based distribution shift. For example, we also discuss and show how image translation works can be used to explain distribution shift via providing an operator with sets of counterfactual pairs. However, we are not sure if the current work in I2I can directly be applied to explain distribution shifts. For example, the problem of style-transfer focuses on transferring the "style" of an image to between two domains while keeping the "content" constant, but what is considered "content" likely needs to be specified by an operator for their specific context in order to be directly actionable (e.g., ensuring road features are considered constant when analyzing human-driving data). Another approach would be to find a causal model of the semantic content between the two distributions, and characterizing the causal differences between them (e.g., the approach of [3] applied to images). In addition to finding methods for explaining image-based distribution shifts, we need ways to evaluate and compare methods. For transport maps, we suggest that a natural metric is to determine how well the transported source distribution aligns with the target distribution via distributional divergences such as Wasserstein distance or KL divergence. However, the interpretability or actionability of a shift explanation is more challenging to define. A proxy method for evaluating this would likely be task specific (but ideally not dataset specific) and should not require expensive human evaluation. For mapping-based methods, the measurement of interpretability could be a function of the complexity of the mapping, however, how to systematically measure the interpretability of counterfactual approaches is currently unclear.

5. Conclusion

In this paper, we introduced a novel framework for explaining image-based distribution shift using transport maps T between a source and target distribution. If a semantically meaningful latent space is known, we can constrain T to be relatively simple. If a meaningful latent space is unavailable, we show how prior image-to-image translation work can explain such shifts via sets of counterfactual examples. We demonstrate both approaches on two distribution shift examples. We then initialized a discussion which hopefully will lead to a better foundation of explaining image distribution shifts. We ultimately hope this work lays the groundwork explaining and thus understanding image distribution shifts.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017. 2

[2] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 2018. 2, 3

[3] Kailash Budhathoki, Dominik Janzing, Patrick Bloebaum, and Hoiyi Ng. Why did the distribution change? In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1666–1674. PMLR, 13–15 Apr 2021. 4, 6

[4] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 3

[5] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013. 2

[6] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. 3, 10

[7] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. 6

[8] Maximilian Ilse, Jakub M Tomczak, Christos Louizos, and Max Welling. Diva: Domain invariant variational autoencoders. In *Medical Imaging with Deep Learning*, pages 322–348. PMLR, 2020. 11

[9] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021. 2, 3, 6

[10] Sean Kulinski, Saurabh Bagchi, and David I Inouye. Feature shift detection: Localizing which features have shifted via conditional distribution tests. *Advances in Neural Information Processing Systems*, 33, 2020. 6

[11] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018. 6

[12] Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012. 6

[13] Wayne B Nelson. *Applied life data analysis*, volume 521. John Wiley & Sons, 2003. 6

[14] Yingxue Pang, Jianxin Lin, Tao Qin, and Zhibo Chen. Image-to-image translation: Methods and applications. *IEEE Transactions on Multimedia*, 2021. 1, 2

[15] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. 2

[16] Joaquin Quiñonero-Candela, Masashi Sugiyama, Neil D Lawrence, and Anton Schwaighofer. *Dataset shift in machine learning*. Mit Press, 2009. 6

[17] Stephan Rabanser, Stephan Günnemann, and Zachary C Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. *arXiv preprint arXiv:1810.11953*, 2018. 6

[18] N. Siddharth, Brooks Paige, Jan-Willem van de Meent, Alban Desmaison, Noah D. Goodman, Pushmeet Kohli, Frank Wood, and Philip Torr. Learning disentangled representations with semi-supervised deep generative models. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5927–5937. Curran Associates, Inc., 2017. 2, 3

[19] Amos Storkey. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, 30:3–28, 2009. 6

[20] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007. 6

[21] Luis Caicedo Torres, Luiz Manella Pereira, and M. Hadi Amini. A survey on optimal transport for machine learning: Theory and applications, 2021. 2

[22] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009. 2

[23] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827. PMLR, 2013. 6

[24] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1