Out-of-Distribution Failure through the Lens of Labeling Mechanisms: An Information Theoretic Approach

Soroosh Shahtalebi [1] Zining Zhu [2 1] Frank Rudzicz [2 1 3]

Abstract

Machine learning models typically fail in deployment environments where the distribution of data does not perfectly match that of the training domains. This phenomenon is believed to stem from networks' failure to capture the invariant features that generalize to unseen domains. However, we attribute this phenomenon to the limitations that the labeling mechanism employed by humans imposes on the learning algorithm. We conjecture that providing multiple labels for each datapoint where each could describe the existence of particular objects/concepts on the data point, decreases the risk of capturing non-generalizable correlations by the model. We theoretically show that learning over a multi-label regime, where K labels for each data point are present, tightens the expected generalization gap by a factor of $1/\sqrt{K}$ compared to a similar case where only one label for each data point is in hand. Also, we show that learning under this regime is much more sample efficient and requires a fraction of training data to provide competitive results.

1. Introduction

Machine learning models often fail to generalize to unseen domains where the data distribution is shifted with respect to the training distribution. The shift in the distribution can stem from changes in the correlation of data and labels, or the covariates of the input data. To be more specific, the correlation shifts are typically caused by spurious correlations in training domains, which are either induced by selection bias or anti-causal correlations between data and label. Although the failure of machine learning models under distribution shifts has received significant attention in the recent years (Shen et al., 2021), an optimal solution has not yet been achieved. The body of literature on this problem typically incorporate training data from different environments that can potentially highlight the globality or locality of correlations within the data, and then employ regularizing techniques to enforce the discovery and learning of global correlations. The enforcement can be satisfied by minimizing the average/interpolation/extrapolation of risks across training domains, matching the representations of data across domains, or matching the loss landscape of model across domains.

Regardless of the progress on this problem, we believe that machine learning models capturing spurious correlations as global/invariant correlations is not a failure, rather is the effect of human's mistake in translating an environment into machine's language. In fact, the labeling mechanisms employed by humans directly mislead and confuse the learning algorithms with a mixture of local and invariant correlations, without no extra information for the model to discriminate them. Although some correlations in the data can be spurious to a specific task, it is again misleading to render a correlation as "*globally spurious*". In fact, a model would ultimately be able to generalize to unseen domains if it has correctly and concretely learned all the correlations in the training domains. For instance, in the current labeling scheme, an image of a dog that includes a number of concepts such as the dog itself together with several other objects/concepts such as sky, grass, and different colors is labeled as "Dog". Thus, it is no surprise that the model confuses any other concept/object in the image with the Dog label. While the model is not receiving any extra information on the image other than the one word label, Dog, it is unrealistic to expect the model to autonomously distinguish between different objects. In the "*Cow-Camel*" example introduced by (Arjovsky et al., 2019), while the model does not have a general knowledge about the set of concepts/objects it can expect to see in an image, it is no surprise that easier to learn features, i.e., the green and khaki colors, are mistakenly learned instead of the true correlation between the objects cow and camel and the label of the image. In this work, we make the following contributions: (i) We employ auxiliary labels for each datapoint to provide our learning model with extra information that might help

[*] Equal contribution [1] Vector Institute for Artificial Intelligence [2] University of Toronto [3] Unity Health Toronto. Correspondence to: Soroosh Shahtalebi <shahtalebi@cs.toronto.edu>.

Published at the ICML 2022 Workshop on Spurious Correlations, Invariance, and Stability. Baltimore, Maryland, USA. Copyright 2022 by the author(s).

disentangling the correlations and building more reliable decision rules; (ii) We investigate the effect of considering the compositionality of a problem into account in order to learn high-complexity tasks; (iii) We theoretically prove that the proposed multi-label learning scheme with K labels tightens the generalization gap of a single-label learning scheme by a factor of $1/\sqrt{K}$, if equal number of training samples across the two scenarios are provided.

2. Method

Pleas note that throughout the paper, capital letters, lowercase letters, and caligraphic letters correspond to random variables, their values, and their domain. If X is a random variable, $\bar{X}$ is its independent copy, and if $(X, Y) \sim P_{X,Y}$ is a pair of data and label, then the joint distribution of $(\bar{X}, \bar{Y})$ is $P_{\bar{X},\bar{Y}} = P_{\bar{X}} \bigotimes P_{\bar{Y}} = P_X \bigotimes P_Y$. Conventionally in supervised learning, for each data point $z = (x, y)$, where $X \in \mathcal{X} \subset \mathbb{R}^d$ and $Y \in \mathcal{Y} \subset \mathbb{R}$ only one label is used to form a labeled dataset, i.e., $\mathcal{S} = \{(x_i, y_i); i = 1, ..., N\}$. However, in this work, we consider a labeled dataset such that for each data point x, a binary vector, $\boldsymbol{y}$, of length K identifies the presence of K different objects/concepts in the data point, i.e., $y^K = [y^1, ..., y^k, ..., y^K] : Y^K \in \mathcal{Y}^{\mathcal{K}} \subset \{0, 1\}^K$. Thus, we can form a dataset of n data points as $\mathcal{S} = \{(x_i, y_i^K); y_i^K = [y_i^1, ..., y_i^K]; i = 1, ..., n\}$. Next, we show that the expected generalization bound of a network under the multi-label scheme is $1/\sqrt{K}$ times tighter than the one of a network trained on single-label data, if both are given equal number of training samples.

2.1. Expected Generalization Bound

In this part, we build on the results provided by Harutyunyan et al. (2021); Xu & Raginsky (2017), which derive an upper bound for the expected generalization gap of a learning algorithm in terms of the mutual information between an input dataset and the output of a learning algorithm, i.e., the trained parameters. To start, let $S = (z_1, z_2, \dots z_n) \sim \mathcal{D}^n$ be a dataset of n i.i.d sampled training examples, and $R \perp\!\!\!\perp \mathcal{R}$ be a random variable representing the stochasticity of the data ($R \in S$). The learning algorithm A is defined as $A : \mathcal{Z}^n \times R \rightarrow \mathcal{W}$, where $\mathcal{W}$ are the parameters of the model. Assuming that $W = A(S, R)$ and a loss function as $\ell : \mathcal{W} \times \mathcal{Z} \rightarrow \mathbb{R}$, we define the following items

$$\text{empirical risk:} \quad \mathcal{L}_{emp}(A, S, R) = \tfrac{1}{n}\textstyle\sum_{i=1}^n \ell(W, Z_i), \tag{1}$$

$$\text{population risk:} \quad \mathcal{L}(A, S, R) = \mathbb{E}_{Z'\sim D}\ell(W, Z'), \tag{2}$$

where Z' denotes the random variable of test samples independent from S. Then, we define the expected generalisation gap as $\mathbb{E}_{S,R}\big[\mathcal{L}(A, S, R) - \mathcal{L}_{emp}(A, S, R)\big]$. The goal in generalization is to design algorithm A such that a model closes the gap between empirical and population risks. In the following theorem, (Xu & Raginsky, 2017) provide an upper bound for this gap, given the stochasticity of dataset.

Theorem 2.1 (Xu & Raginsky (2017))**.** *If $\ell(w, Z')$, where $Z' \sim \mathcal{D}$, is $\sigma -$ subgaussian for all $w \in \mathcal{W}$, then*

$$|\mathbb{E}_{S,R}[\mathcal{L}(A, S, R) - \mathcal{L}_{emp}(A, S, R)]| \leq \sqrt{\frac{2\sigma^2 I(W; S)}{n}} \tag{3}$$

In addition to this result, (Harutyunyan et al., 2021) in Theorem 2.2 have shown that the above theorem still holds when only a subset of size m from the whole dataset of size n is used for training.

Theorem 2.2 (Harutyunyan et al. (2021))**.** *Let U be a random subset of $[n]$ with size m, independent of S and R. If $\ell(w, Z')$, where $Z' \sim \mathcal{D}$, is $\sigma -$ subgaussian for all $w \in \mathcal{W}$, then*

$$\begin{aligned} &|\mathbb{E}_{S,R}[\mathcal{L}(A, S, R) - \mathcal{L}_{emp}(A, S, R)]| \\ &\qquad \leq \mathbb{E}_{u\sim U}\sqrt{\frac{2\sigma^2 I(W; S_u)}{m}} \end{aligned} \tag{4}$$

Since in the problem setting of this paper, K different labels are associated with a datapoint, $n = mK$ multi-label datapoints can be reused K times, each time for a different label, as if n different datapoints are available in the dataset. Thus, by capitalizing on Theorems 2.1 and 2.2, we can state that $m = n/K$ multilabel datapoints provide the same upper bound for expected generalization gap as n single-label datapoints would achieve. In addition, it is inferred that given the same number of samples across the single-label and multi-label scenarios, the upper bound of the expected generalization gap of the multi-label one is $1/\sqrt{K}$ times tighter than the one of the single-label one.

Assumption 2.3. Given a dataset S of n i.i.d sampled examples collected from $P(\boldsymbol{X})$ where each sample has K labels, $k \in \mathcal{Z}, K > 1$, we assume that n/K is still large enough such that n/K samples drawn from $P(\boldsymbol{X})$ are still i.i.d with respect to each other.

Theorem 2.4. *Let $\ell(w, Z')$ is $\sigma -$ subgaussian for all $w \in \mathcal{W}$, and $Z' \sim \mathcal{D}$. Given a dataset S of n samples where each sample has K labels, for all $w \in \mathcal{W}$, the expected generalization bound is tighter by a factor of $\frac{1}{\sqrt{K}}$ than the case where each sample of a dataset with the same size has only 1 label. In other words,*

$$\begin{aligned} &|\mathbb{E}_{S,R}[\mathcal{L}(A, S, R) - \mathcal{L}_{emp}(A, S, R)]| \\ &\leq \sqrt{\frac{2\sigma^2 I(W; S)}{n}} = \sqrt{\frac{2\sigma^2 I(W; S)}{Km}}. \end{aligned} \tag{5}$$

Please note that across the single-label and multi-label scenarios, we assume that the number of parameters and the stochasticity of dataset does not change. What can be inferred from Theorem 2.4 is that $m = n/K$ number of multilabel training samples provide the same upper bound on the expected generalization gap that n number of single-label datapoints from the same distribution would do. In other words, given equal number of training examples from both scenarios, the upper bound of expected generalization gap for the multi-label scheme is $1/\sqrt{K}$ times tighter than the one of single-label case.

3. Experiments

To evaluate the efficacy of the concept models over OoD generalization problems, we mainly focused on datasets that exhibits a pure case of correlation shifts, i.e., $P_{train}(Y \mid X) \neq P_{test}(Y \mid X)$, that are Waterbirds, CelebA, and Colored-MNIST (Ye et al., 2021). The reason for this choice is that correlation shifts typically occur due to selection bias in the data sampling process or presence of anti-causal features in the dataset. Thus, disentangling the concepts in an environment is expected to help the model capture the true concepts that describe a label.

3.1. Experimental Setup

Our experimental setup has two stages: (i) for the CelebA and Waterbirds datasets in which the final label of the data is among the concepts, we train a model in a multi-label setting, where the labels are the concepts to be learned, and the evaluation is done by looking at the accuracy of the model in predicting a specific label or concept; and (ii) for the Colored-MNIST dataset where the label is *not* among the concepts but can be inferred based on them, we train a network in a multi-label setting to learn the concepts, which is then topped by another layer to do the inference. To train the modular architecture in the latter case, we follow the three training strategies as described by (Koh et al., 2020a). Let $\boldsymbol{x} \in \mathbb{R}^d$ is the input, $\boldsymbol{y} \in \mathbb{R}^K$ is the target vector identifying the existence of K concepts, and $l \in \mathbb{R}$ is the final label of the data point. We define the concept bottleneck as $g : \mathbb{R}^d \to \mathbb{R}^K$ and the inference module as $f : \mathbb{R}^K \to \mathbb{R}$. Also, let $L_y : \mathbb{R}^K \times \mathbb{R}^K \to \mathbb{R}_+$ and $L_l : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ be the loss functions at the concept level and the label level, respectively. To train this modular architecture, the following strategies are employed:

- **Independent Bottleneck,** where the modules are trained independent from each other, i.e., $\hat{g} = \arg\min_g \sum_{i,j} L_y(g^j(x_i); y_i^j)$ and $\hat{f} = \arg\min_f \sum_i L_l(f(y_i); l_i)$.
- **Sequential Bottleneck,** where the concept bottleneck is trained first based on $\hat{g} = \arg\min_g \sum_{i,j} L_y(g^j(x_i); y_i^j)$, and then the inference module is trained on the outputs of the concept bottleneck, i.e., $\hat{f} = \arg\min_f \sum_i L_l(f(\hat{g}(x_i)); l_i)$.
- **Joint Bottleneck,** where the two modules are trained simultaneously based on a weighted sum of the loss for the two modules, i.e., $\hat{g}, \hat{f} = \arg\min_{g,f} \sum_i \left[L_l(f(g(x_i)); l_i) + \sum_j \lambda L_y(g^j(x_i); y_i^j)\right]$.

It is worth noting that, to train the concept networks, binary cross entropy for each of the individual concepts is applied as the loss function, which is in contrast with the typical training strategy for multi-class classification problems in which a Softmax layer is followed by cross-entropy loss. In the multi-class setting, the Softmax layer maximizes the probability of a certain class by minimizing the probability for other classes while, in the multi-label setting, each of the concepts are learned independent of other labels through a binary cross-entropy loss.

3.2. Results

Please note that the results reported here are averaged over 10 random seeds to mitigate any effect of initialization in the final accuracy. Also, the hyper-parameters of the deployed models are fine-tuned over a validation set sampled from training domains.

Concept Bottleneck Network over Colored-MNIST. The results are provided in Table 1. Since for the Colored-MNIST dataset our model follows a modular architecture where the first module predicts the concepts and the second one predicts the final binary label of samples, Table 1 shows the performance of the two modules separately. In this table, please note the the columns denote the source of training data, which we name them as $+90\%$, and $+80\%$, and the combination of the two. An important outcome of this table is that learning the concepts effectively, is a necessary condition for a model to generalize to unseen domains. The evidence is the ERM method that fails to capture the concepts, thus fails to generalize to test domain. However, as the results for the "Independent" and "Joint" methods suggest, a good accuracy at the concept level is not sufficient to guarantee a good OoD performance. The "Sequential" method, on the other hand, offers significantly important results as it achieves 57.09% classification accuracy over a test domain whose correlation is completely opposite of the training domain. Please note that the best achievable test accuracy theoretically is 75%. These results can be seen as an evidence that learning generalizable features does not necessarily rely on collecting data from different domains.

Concept Bottleneck Network over CelebA and Waterbirds. The results of this analysis are provided in table 2. Since for these two datasets, the final label is among the concepts, we only consider the prediction accuracy for the original labels given for each classification problem. Due to the nature of the two datasets, the correlation shift we

Table 1: Accuracy of concept-based learning in OoD generalization over the Colored-MNIST dataset.

	Concept Accuracy			Label Accuracy		
Method	+90%	+80%	$\{+90\%\}\bigcup\{+80\%\}$	+90%	+80%	$\{+90\%\}\bigcup\{+80\%\}$
Independent	98.98	98.87	99.24	10.95	26.90	11.82
Sequential	98.82	98.89	99.35	**57.09**	**54.09**	**57.59**
Joint	98.93	99.07	99.16	12.93	27.01	13.00
ERM	50.55	26.18	74.32	17.08	29.82	28.51

Table 2: Accuracy of concept learning in OoD generalization over Waterbirds and CelebA datasets.

	Waterbirds		CelebA	
Model	**Worst group**	**Average**	**Worst group**	**Average**
GDRO (Sagawa et al., 2019)	83.80	89.40	88.30	91.80
ERM	60.00	97.30	41.10	94.80
VIB (Alemi et al., 2016)	75.31	95.39	78.13	91.94
CIM (Taghanaki et al., 2021)	73.35	89.78	81.25	89.24
CIM+VIB (Taghanaki et al., 2021)	77.23	95.60	83.59	90.61
Ours	**88.99**	91.85	**97.65**	**98.13**

face in these problems is induced by selection bias. Thus, it is imperatively important to validate the worst group accuracy as well as the average accuracy of the model in test domains. As the results suggest, a concept-based learning approach not only improves the average accuracy of the network, by providing more accurate predictions on the over-presented group, but also significantly enhances the prediction accuracy of the network over less-represented group. This behaviour is crucially required to ensure a fair operation for a model.

Sample Efficiency. Another important factor that the proposed concept-based learning method needs to be tested on is its sample efficiency, i.e., the minimum number of training data required to achieve a certain level of performance. The reason is that the growing trend in collecting more data to train better models in terms of generalizability on one hand, and the extra effort required for providing auxiliary labels for each datapoint, on the other hand, might render our proposed technique as impractical. Thus, it is imperative to check if the sample efficiency of the method compensates for the extra effort in labeling the datapoints (we have already shown that our technique outperforms its counterparts by large margins). The results of this analysis are shown in Figs. 1 and 2, where the former shows the classification accuracy of the model, and the latter shows the loss of the model during the training process. As the results suggest, a model through the proposed learning strategy manages to outperform any existing baseline by using only 10% of the whole dataset, achieving almost the same degree of performance compared to a case where the entire training samples are deployed. Although our theoretical results suggest a tighter generalization bound for the case that a dataset of single-labeled samples be transformed to a multi-label dataset and be entirely employed in training, the current results suggest that almost the same performance can be achieved by using only a fraction of dataset.

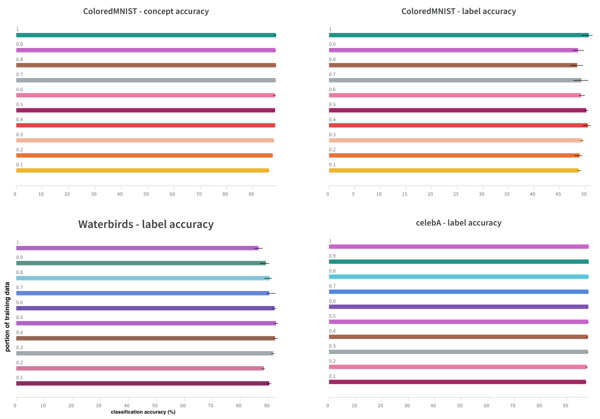

Figure 1: The accuracy of the concept-based learning over different datasets. $x-$axis is the classification accuracy $(\%)$ and $y-$axis is the portion of training dataset used for training. **Top left:** concept accuracy over Colored-MNIST. **Top right:** label accuracy over Colored-MNIST. **Bottom left:** label accuracy over Waterbirds. **Bottom right:** label accuracy over CelebA.

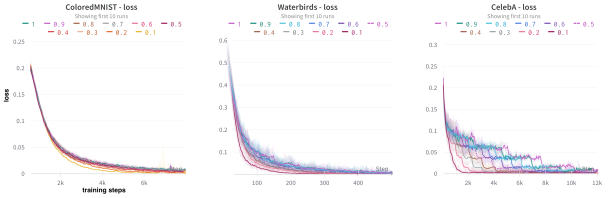

Figure 2: Training loss of a model for different portions of training data. $x-$axis is the training step, and $y-$axis is the loss. **Left:** training loss for Colored-MNIST. **Middle:** training loss for Waterbirds. **Right:** training loss for CelebA.

4. Conclusion

In this work, we investigated the efficacy of concept networks in generalizing to out of distributions. Ensuring that neural networks are not overfitted on spurious correlations requires that we make sure if the network has correctly learned the underlying concepts that define each class. Although providing multiple labels for a data point seems to require extra effort in gathering datasets, we have shown that the multi-label learning scheme requires a fraction of data to achieve the performance of ordinary neural networks. Moreover, we have theoretically proven that providing K labels for a data point rather than one, tightens the generalization bound of the intended label by a factor of $1/\sqrt{K}$. Our results over the three benchmarking datasets of correlation shifts, i.e., CelebA, Colored-MNIST, and Waterbirds, offer the state-of-the-art/competitive performances compared to its counterparts.

References

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Bai, H., Sun, R., Hong, L., Zhou, F., Ye, N., Ye, H.-J., Chan, S.-H. G., and Li, Z. DecAug: Out-of-distribution generalization via decomposed feature representation and semantic augmentation. *arXiv preprint arXiv:2012.09382*, 2020.

Bengio, Y., Deleu, T., Rahaman, N., Ke, R., Lachapelle, S., Bilaniuk, O., Goyal, A., and Pal, C. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019.

Bouchacourt, D., Tomioka, R., and Nowozin, S. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Bucher, M., Herbin, S., and Jurie, F. Semantic bottleneck for computer vision tasks. In *Asian Conference on Computer Vision*, pp. 695–712. Springer, 2018.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in $beta$-vae. *arXiv preprint arXiv:1804.03599*, 2018.

Caruana, R. Multitask learning. *Machine learning*, 28(1): 41–75, 1997.

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., and Ghassemi, M. Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science*, 4:123–144, 2021.

Chen, R. T., Li, X., Grosse, R. B., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.

Eastwood, C. and Williams, C. K. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

Földiák, P. Learning invariance from transformation sequences. *Neural computation*, 3(2):194–200, 1991.

Gao, Y., Ma, J., Zhao, M., Liu, W., and Yuille, A. L. NDDR-CNN: Layerwise feature fusing in multi-task cnns by neural discriminative dimensionality reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3205–3214, 2019.

Gong, T., Lee, T., Stephenson, C., Renduchintala, V., Padhy, S., Ndirango, A., Keskin, G., and Elibol, O. H. A comparison of loss weighting strategies for multi task learning in deep neural networks. *IEEE Access*, 7:141627–141632, 2019.

Goyal, A., Lamb, A., Hoffmann, J., Sodhani, S., Levine, S., Bengio, Y., and Schölkopf, B. Recurrent independent mechanisms. *arXiv preprint arXiv:1909.10893*, 2019.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Guo, P., Lee, C.-Y., and Ulbricht, D. Learning to branch for multi-task learning. In *International Conference on Machine Learning*, pp. 3854–3863. PMLR, 2020.

Harutyunyan, H., Raginsky, M., Ver Steeg, G., and Galstyan, A. Information-theoretic generalization bounds for black-box learning algorithms. *Advances in Neural Information Processing Systems*, 34, 2021.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-VAE: Learning basic visual concepts with a constrained variational framework. 2016.

Ke, N. R., Bilaniuk, O., Goyal, A., Bauer, S., Larochelle, H., Schölkopf, B., Mozer, M. C., Pal, C., and Bengio, Y. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.

Kim, H. and Mnih, A. Disentangling by factorising. In *International Conference on Machine Learning*, pp. 2649–2658. PMLR, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Klindt, D., Schott, L., Sharma, Y., Ustyuzhaninov, I., Brendel, W., Bethge, M., and Paiton, D. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint arXiv:2007.10930*, 2020.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In *International Conference on Machine Learning*, pp. 5338–5348. PMLR, 2020a.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Beery, S., et al. Wilds: A benchmark of in-the-wild distribution shifts. *arXiv preprint arXiv:2012.07421*, 2020b.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Priol, R. L., and Courville, A. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688*, 2020.

Kumar, A., Sattigeri, P., and Balakrishnan, A. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.

Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K. Attribute and simile classifiers for face verification. In *2009 IEEE 12th international conference on computer vision*, pp. 365–372. IEEE, 2009.

Lampert, C. H., Nickisch, H., and Harmeling, S. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 951–958. IEEE, 2009.

Liu, S., Johns, E., and Davison, A. J. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1871–1880, 2019.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Lu, Y., Kumar, A., Zhai, S., Cheng, Y., Javidi, T., and Feris, R. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5334–5343, 2017.

Mathieu, E., Rainforth, T., Siddharth, N., and Teh, Y. W. Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning*, pp. 4402–4412. PMLR, 2019.

Parascandolo, G., Neitz, A., Orvieto, A., Gresele, L., and Scholkopf, B. Learning explanations that are hard to vary. *arXiv preprint arXiv:2009.00329*, 2020.

Rame, A., Dancette, C., and Cord, M. Fishr: Invariant gradient variances for out-of-distribution generalization. *arXiv preprint arXiv:2109.02934*, 2021.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

Ruder, S. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

Ruder, S., Bingel, J., Augenstein, I., and Sogaard, A. Latent multi-task architecture learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4822–4829, 2019.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Schmidt, M., Niculescu-Mizil, A., Murphy, K., et al. Learning graphical model structure using l1-regularization paths. In *AAAI*, volume 7, pp. 1278–1283, 2007.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021.

Shahtalebi, S., Gagnon-Audet, J.-C., Laleh, T., Faramarzi, M., Ahuja, K., and Rish, I. Sand-mask: An enhanced gradient masking strategy for the discovery of invariances in domain generalization. *arXiv preprint arXiv:2106.02266*, 2021.

Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., and Cui, P. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

Shi, Y., Seely, J., Torr, P. H., Siddharth, N., Hannun, A., Usunier, N., and Synnaeve, G. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021.

Taghanaki, S. A., Choi, K., Khasahmadi, A. H., and Goyal, A. Robust representation learning via perceptual similarity metrics. In *International Conference on Machine Learning*, pp. 10043–10053. PMLR, 2021.

Träuble, F., Creager, E., Kilbertus, N., Locatello, F., Dittadi, A., Goyal, A., Schölkopf, B., and Bauer, S. On disentangled representations learned from correlated data. In *International Conference on Machine Learning*, pp. 10401–10412. PMLR, 2021.

Tripuraneni, N., Jordan, M., and Jin, C. On the theory of transfer learning: The importance of task diversity. *Advances in Neural Information Processing Systems*, 33: 7852–7862, 2020.

Vandenhende, S., Georgoulis, S., De Brabandere, B., and Van Gool, L. Branched multi-task networks: deciding what layers to share. *arXiv preprint arXiv:1904.02920*, 2019.

Vandenhende, S., Georgoulis, S., Van Gansbeke, W., Proesmans, M., Dai, D., and Van Gool, L. Multi-task learning for dense prediction tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

Xu, A. and Raginsky, M. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 30, 2017.

Ye, N., Li, K., Hong, L., Bai, H., Chen, Y., Zhou, F., and Li, Z. OoD-Bench: Benchmarking and understanding out-of-distribution generalization datasets and algorithms. *arXiv preprint arXiv:2106.03721*, 2021.

Yi, K. et al. Disentangling reasoning from vision and language understanding. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 1039–1050, 2018.

Zhang, M., Marklund, H., Dhawan, N., Gupta, A., Levine, S., and Finn, C. Adaptive risk minimization: A meta-learning approach for tackling group distribution shift. *arXiv preprint arXiv:2007.02931*, 2020.

Zhang, Y. and Yang, Q. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

Appendix

A. Details of Experiments

A.1. Datasets

Modified CelebA Dataset. In this version of the CelebA dataset (Liu et al., 2015), the task is to solve a binary classification task $y_i \in \{\texttt{male}, \texttt{female}\}$. The challenge is that models trained based on empirical risk minimization, due to the strong correlation between gender and hair colour in celebrity images, pick up hair colour as the discriminating feature for the male/female problem. We have enriched this dataset by adding two extra labels about hair colour and formulated a multi-label classification problem such that the presence of objects or concepts is determined by a binary vector

$$\boldsymbol{y} = [\texttt{male}, \texttt{female}, \texttt{black hair}, \texttt{blond hair}].$$

Waterbirds Dataset. This dataset also presents a binary classification problem where the task is to distinguish between waterbirds and landbirds (Sagawa et al., 2019). The challenge is that these two classes are strongly correlated with their background colour, and neural networks build strong decision rules based on the background colour instead of the birds' shape and colour. We reformulated this problem, again, as a multi-label problem where each image is labeled with a binary vector determining the presence of the following objects.

$$\boldsymbol{y} = [\texttt{waterbird}, \texttt{landbird}, \texttt{water background}, \texttt{land background}].$$

Colored-MNIST. This dataset that was originally introduced by (Arjovsky et al., 2019) and reformulates the original MNIST dataset as a binary classification problem where the digits less than 5 are labeled as "0" and the rest are labeled as "1". The binary labels are also corrupted by 25% label noise. The challenge with this dataset is that depending on the definition of environment and the binary label of each image, a background of red or green is added to MNIST images, which in fact, introduces a spurious correlation to the dataset. Following the environment definition in (Gulrajani & Lopez-Paz, 2020), there are two training environments where in each, background colours are $+90\%$ and $+80\%$ correlated with the binary label of an image (green with class"0" and red with class"1"), while in the test environment, the correlation is -90%. This dataset introduces a correlation shift problem caused by selection bias. To evaluate our proposed approach on this dataset, we have enriched it by introducing the following set of concept labels, that indeed are the ones that a human would use to address the binary classification problem. The concept labels are:

$$\boldsymbol{y} = [\texttt{0}, \texttt{1}, \cdots, \texttt{9}, \texttt{red colour}, \texttt{green colour}].$$

A.2. Models

To fairly evaluate and compare the performance of concept networks over the aforementioned OoD generalization problems, we apply a pretrained ResNet-18 model for the CelebA and Waterbirds datasets. The model is initialized on ImageNet pretraining parameters, and its final layer is replaced by two dense layers of 64 and 4 nodes each, yielding the set of concepts defined for each dataset. Although a pretrained model for the concept network is deployed, its parameters will still be updated in the training phase. We use an Adam optimizer with a weight decay of 0.0001 for the whole model with a learning rate of 0.0001 for the pretrained model and a learning rate of 0.001 for the dense layers.

For the Colored-MNIST dataset, on the other hand, the employed model is of two modules, i.e., a concept bottleneck module and an inference module. The former follows the same architecture and hyper-parameters as defined in the DomainBed suite (Gulrajani & Lopez-Paz, 2020) for the MNIST-based datasets, which is a four-layer CNN, respectively having 64, 128, 128, and 128 feature maps, followed by three layers of dense layers each having 64, 32, and 4 nodes. The latter module to infer the binary classification rule based on the outputs of the first module (the concepts) consists of three consecutive dense layers, each with 10, 4, and 2 nodes. Here we also use an Adam optimizer with a learning rate of 0.001 and a weight decay of 0.0001 to train the model.

A.3. Code base

An anonymous repository of the code and instructions needed to reproduce the results of this paper can be found here[1]. In this code base, the Colored-MNIST dataset is borrowed from the DomainBed suite (Gulrajani & Lopez-Paz, 2020) (MIT License), and Waterbirds and CelebA datasets are borrowed from the WILDS library (Koh et al., 2020b) (MIT License).

[1] https://anonymous.4open.science/r/BottleneckGeneralization-8274/README.md

A.4. Model Architectures and Hyperparameters

The architectures of the models employed in this paper are given in Tables 3 and 4. For both models, an Adam optimizer is used to optimize the network. The network is Table 3 uses a learning rate of 0.00001 for the ResNet model, and a learning rate of 0.0001 for the subsequent dense layers. The model described in Table 4, utilizes a learning rate of 0.001 for the whole network. Please note that in these tables, d is the dimension of of input data.

All the models and the hyperparameters employed in this work are fine-tuned over a validation set taken from the training data. The training data of each dataset is split into two groups of 80% and 20% portions, where the former is reserved for the training phase, and the latter is used for validation purposes. Please note that this scheme of forming the validation set imposes a true case of Domain Generalization problem, where no information about the test domain is available during the training process.

Table 3: Model Architecture for CelebA and Waterbirds datasets

	#	Layer
Concept Module	1	ResNet-18 (in=d, out=1024)
	2	Dense Layer (in=1024, out=64)
	3	Dense Layer (in=64, out=4)

Table 4: Model Architecture for Colored-MNIST dataset

	#	Layer
Concept Module	1	Conv2D (in=d, out=64)
	2	ReLU
	3	GroupNorm (groups=8)
	4	Conv2D (in=64, out=128, stride=2)
	5	ReLU
	6	GroupNorm (8 groups)
	7	Conv2D (in=128, out=128)
	8	ReLU
	9	GroupNorm (8 groups)
	10	Conv2D (in=128, out=128)
	11	ReLU
	12	GroupNorm (8 groups)
	13	Global average-pooling
	14	Dense Layer (in=128, out=64)
	15	ReLu
	16	Dense Layer (in=64, out=32)
	17	ReLu
	18	Dense Layer (in=32, out=12)
Inference Module	19	Dense Layer (in=12, out=10)
	20	BatchNorm (10)
	21	Dropout (0.25)
	22	ReLu
	23	Dense Layer (in=10, out=4)
	24	BatchNorm (4)
	25	ReLu
	26	Dense Layer (in=4, out=2)

A.5. Compute Power

The experiments presents in this paper are implemented in Python language, thanks to the PyTorch library. All the experiments take about one week to be executed on 2 NVIDIA Quadro RTX 6000.

B. Related Works

Concept Bottleneck Models. In these models, the downstream tasks are learned based on a number of human-identified concepts that underlie the class of labels in an environment/dataset. In fact, a model can be split in two modules where the first one predicts the concepts and the second one predicts the label of data point based on the identified concepts. The significance of earlier works on this approach (Kumar et al., 2009; Lampert et al., 2009) was overshadowed by the performance of end-to-end neural networks but more recently, deep neural networks with concepts bottlenecks have received growing attention for different applications. For instance, this method is employed for retinal disease diagnosis (De Fauw et al., 2018), visual question-answering (Yi et al., 2018), content-based image retrieval (Bucher et al., 2018), and healthcare applications (Chen et al., 2021).

Out-of-Distribution (OoD) Generalization. Invariance principle (Arjovsky et al., 2019; Shen et al., 2021) is the backbone of methods in OoD generalization which aims at, either explicitly or implicitly, leveraging the invariances among multiple training environments that can potentially enable the network to generalize to unseen domains. Invariances could be sought in the representation space where the goal is to minimize the discrepancy between different environments (Arjovsky et al., 2019; Bai et al., 2020; Zhang et al., 2020), or at the risk level where the goal is to have a network which performs equally well in different environments (Sagawa et al., 2019; Krueger et al., 2020), or at the loss surface level where to objective is to train a model such that it converges to minima common across different domains (Parascandolo et al., 2020; Shahtalebi et al., 2021; Rame et al., 2021; Shi et al., 2021). In parallel, there are works that render invariances as causal mechanisms that regardless of the environment, always cause a set of features to receive a specific label (Schölkopf et al., 2021; Bengio et al., 2019).

Multi-task Learning (MTL). In MTL the goal is to learn multiple number of tasks based on a shared nonlinear representation derived through a neural network (Vandenhende et al., 2021). Since in the MTL settings, different tasks are learned based on a shared representation, it is known that if different tasks share complementary information or act as regularizer for one another, MTL methods offer an enhanced performance compared to the case of single task learning (Caruana, 1997; Zhang & Yang, 2021; Ruder, 2017; Gong et al., 2019). The shared representation is typically obtained via *Hard Parameter Sharing* (Lu et al., 2017; Vandenhende et al., 2019; Guo et al., 2020) or *Soft Parameter Sharing* (Ruder et al., 2019; Gao et al., 2019; Liu et al., 2019) architectures.

Disentangling. Here the goal is to map data to a space where Factors of Variation (FoV) are represented independently (Träuble et al., 2021). The majority of recent works on disentanglement use variational autoencoders (VAEs) (Kingma & Welling, 2013; Rezende et al., 2014) as their core block and modify its objective function such that the notion of disentangled representations can be fulfilled, e.g., β-VAE, AnnealedVAE, FactorVAE, β-TCVAE, and DIP-VAE (Higgins et al., 2016; Burgess et al., 2018; Kim & Mnih, 2018; Chen et al., 2018; Kumar et al., 2017; Eastwood & Williams, 2018; Mathieu et al., 2019). To disentangle representations in unsupervised settings, inductive biases such as grouping information (Bouchacourt et al., 2018) or weak labels (Goyal et al., 2019; Földiák, 1991; Schmidt et al., 2007; Bengio et al., 2019; Ke et al., 2019; Klindt et al., 2020) are typically employed. Although impressive progress has been made in the past, such methods assume that FoVs are independent, which restricts their applicability in real world problems.

C. Extended Discussions

Our results on Waterbirds and CelebA datasets suggest the efficacy of our method in enhancing the prediction accuracy for the less-represented populations in the dataset, as well as improving the average classification accuracy over the entire dataset. Such accuracy gains are of significant importance for problems concerning the fairness of a model, where subpopulation shifts are induced due to selection bias. The results on Colored-MNIST dataset reveal an extraordinary capacity for the proposed technique, as a model trained on data only from the environment with $+90\%$ correlation, achieves 57.09% accuracy on an environment with a completely opposite correlation. In addition, the set of results over this dataset implicitly suggest the necessary and sufficient conditions for a model to generalize to unseen domains. What can be told confidently about the necessary condition is that for a model to generalize to unseen domains, it should capture, either implicitly or explicitly as in this work, the underlying concepts of an environment. The learned concepts not only provide better generalizability, but also allow for improved interpretability of the model, as well as faster transfer to new tasks.

Limitations. Despite the gain in the performance and generalizibility of neural networks that is achieved by the proposed multi-label training mechanism, this techniques requires datasets where multiple labels are provided for each datapoint where each identify the existence of an object/concept in a datapoint. This requirement renders the majority of existing datasets as obsolete, and requires collection of new datasets. Nonetheless, in the literature it is shown theoretically (Tripuraneni et al., 2020) and we have observed empirically that training a network in multi-label scheme is much more sample efficient, i.e., requires less training data, which compensates for extra labels needed. As an extension to this work, one can investigate unsupervised learning methods to detect and learn the underlying concepts of an environment, so that the extra effort in collecting multi-label datapoints can be avoided.

Societal impacts. Our proposed multi-label learning strategy provides neural networks with a higher capacity to generalize to unseen domains, which is a crucial behaviour when deploying such models in the wild. This feature is of utmost importance in applications like self-driving cars or AI-assisted diagnosis, where the deployment environments are divers and collecting training data from every single environment is not feasible. Besides, the multi-label learning strategy provides us with a network (the concept bottleneck) which can be deployed for a variety of different downstream tasks, without the need to retrain a model from scratch. The reusability of concepts not only helps with learning new tasks where abundance of training data is not available, but also saves a considerable amount of time and computational power.