On the necessity of human insight to improve natural adversarial robustness

Pierre-Louis Lemaire
Department of Applied Mathematics
Polytechnique Montréal
pierre-louis.lemaire@polymtl.ca

Tom Marty
UdeM - Department of Computer Science
Mila - Quebec AI Institute
tom.marty@mila.quebec

Zhenyu Yang
Department of Computer Engineering
Polytechnique Montréal
zhenyu.yang@polymtl.ca

Abstract

Over the past few years, adversarial attacks have been famous for exposing staggering vulnerabilities of state-of-the-art deep neural networks. This trend gave rise to a dynamic research community focused on discovering and addressing adversarial vulnerabilities. Over the year, this community has developed numerous methods to both exploit the vulnerabilities of vanilla prediction models and mitigate these vulnerabilities through post-hoc adversarial alignment techniques. Usually, adversarial attacks focus on atypical distributional shifts unlikely to occur in natural environments from which the data were sourced. We choose instead to focus on robustness against natural adversarial attacks, and leverage causality tools to better understand the functional mechanisms behind the success of this category of adversarial attacks. We evaluate causally inspired and sounded distribution alignment methods using natural adversarial samples and investigate interesting behaviors of deep neural networks.

1 Introduction & Related Works

Adversarial attacks [5, 7] are sophisticated methods that can successfully degrade state-of-the-art deep neural networks (DNNs) performance [8] by performing subtle perturbations on the input data. While deceiving DNNs, adversarially-corrupted data often remain imperceptible to human eyes. We believe that our difficulty in comprehending these unexpected failures from DNNs stems from a persistent anthropomorphic perspective towards learning algorithms: we assume that a machine's perception of the world, also known as its umwelt [3], mirrors our own, when in reality, decision rules can be highly different. Costa et al. [7] provide a comprehensive taxonomy of adversarial attacks, including a crucial distinction between white-box and black-box attacks. In white-box scenarios, attackers have full knowledge of the model, including its architecture and parameters, and exploit it through gradient-ascent breaches, whereas in black box scenarios, the attackers have no such information. The following adversarial white-box attacks: FGSM (Goodfellow et al., 2015 [12]), PGD (Madry et al., 2019 [18]), and CW attack (Carlini & Wagner, 2017 [4]) often serve as reference in the field. In a red team/blue team dynamic, researchers have also tried to mitigate DNNs' adversarial vulnerabilities by turning adversarial attacks against themselves using different kind of post-hoc learning procedures and/or regularization techniques. This field of study is often cited as *adversarial alignment* or *adversarial training*. Most method rely on data-augmentation techniques with adversarial samples and/or a whole arsenal of inductive bias [12, 18, 24], embedded inside a specifically-tailored loss signal.

From a probabilistic perspective, adversarial attacks can be interpreted as a type of deliberate distributional shift on the natural data distribution. These shifts are designed to expose a DNN reliance on *unintended*

Preprint. Under review.

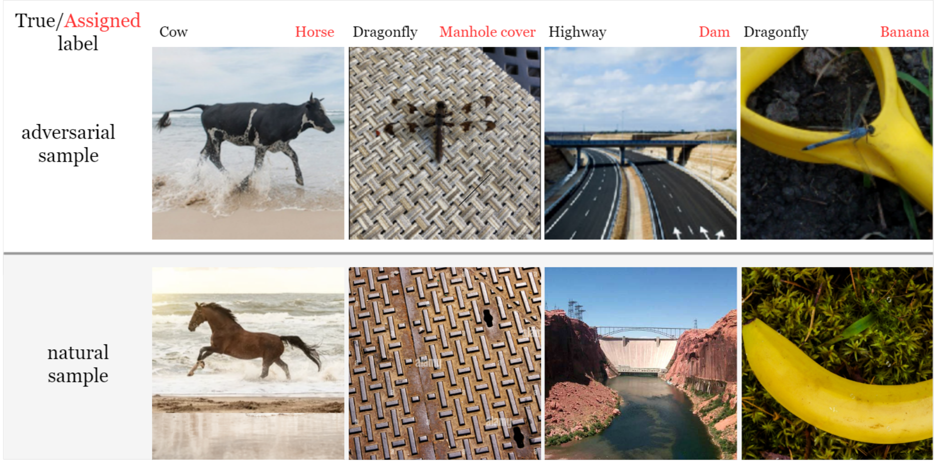


Figure 1: Example of natural adversarial samples from [15, 2] usually misclassified by vanilla learners due to the presence of specific non-semantic features strongly correlated with other labels. For example, in the third example, the retaining walls of the **highway** resembles the banks of a river on which **dams** are usually built, thus creating the confusion.

shortcut features, also referred to as non-causal features, as opposed to *intended features* [25, 11]. By definition, a shortcut is an efficient decision rule that performs particularly well within the training distribution, but fails to generalize to more challenging out-of-distribution (O.O.D) scenarios. Most of the time, shortcuts exploit strong correlational links between non-causal features and the output label, making them effective predictors within the distribution. It is exactly this in-distribution illusion of reliability that explains why they are naively learned by vanilla optimizer. However, their performance deteriorates rapidly O.O.D due to potential arbitrary covariate shifts in the test distribution between non-semantic features and label. Figure 1 illustrates notable examples of this phenomenon.

Zhang et al. [25] proposed to leverage this new perspective on adversarial attacks to design a causally-inspired *adversarial training* method: `CausalAdv`. Clarifying the distinction between intended features and unintended features, they introduced a causal graph for adversarial attacks and developed an alignment strategy that triggers problematic spurious correlations.

While standard adversarial attacks are almost imperceptible to humans, Hendrycks et al. [15] extend the taxonomy by considering a new sub-category refereed as *Natural Adversarial Attacks*. These attacks exploit specific distributional shifts that can occur in natural environments, creating likely yet particularly deceptive sample that can sometimes even fool human labelers. Notably, the well-known scenario involving cows and backgrounds introduced in [2] falls under this category. Figure 1 shows several example of natural adversarial samples. We consider these samples as expensive to collect for two reasons: (1) they require human insight to identify the problematic spurious correlations (i.e. style features that could be used as shortcuts by DNNs); (2) sampling from the appropriate de-correlated distribution is often complex and require additional data collection and labeling.

We believe that DDNs vulnerability to Natural Adversarial Attacks [15] is additional evidence that neural networks are strong *shortcut learners*, and by extension, that directly addressing this issue should be a strong priority in order to remedy multiple of their vulnerabilities. In this work, we evaluate and compare the performance of multiple training strategies when combined with natural adversarial samples: vanilla learning, `CausalAdv` adversarial training, and a contrastive learning alignment method we call `NatCL`. We justify the validity of experimenting on the latter as such contrastive learning methods have been proven to isolate semantic from non-semantic features [21], showing strong links with the causal perspective of adversarial vulnerability used for `CausalAdv`. Furthermore, we highlight a possible cost-efficiency tradeoff regarding the use of natural adversarial examples.

Also, while we recognize that shortcut learning can arise in every existing modalities like Natural Language Processing [23], time series forecasting [9], or Optimal Control [10]. We decided to focus our study on Computer Vision because this setting is the most convenient one as it allows for a visual explanation proposal of the observed failure. For further discussion around shortcut learning, we refer the following survey [11].

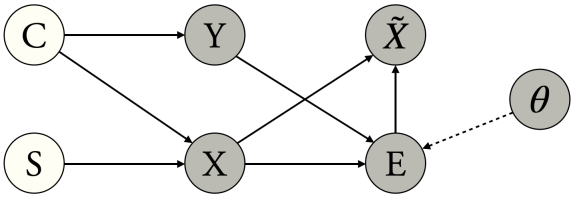


Figure 2: Causal graph $\mathcal{G}$ introduced by Zhang et al. [25]. The class Y of an image X is supposed to be solely dependent from the content C, while the image is the output of a mixing between content and style information. Content C and style S variables are supposed to be independent. A perturbation E from a certain adversarial attack would be caused by the original (or natural) image X, its class Y and the weights θ of the model it is trying to fool. Finally, the adversarial (or perturbed) image $\tilde{X}$ will depend on the natural image X and the perturbation E.

This report is structured as follows: Section 2 presents theoretical details for `CausalAdv` as well as contrastive learning and *NatCL*. Then, Section 3 covers our evaluation methodology, experiments conducted and dataset used. Finally, our findings are presented and discussed in Section 4.

2 Theory & Intuition

Through this work, we focus on comparing two fundamentally different learning strategies that both have roots in causality. `CausalAdv` is an adversarial training method directly derived from a causal perspective of adversarial attacks. `NatCL` uses a regularized loss function to guide the model to learn a style-free representation space in a self-supervised way with contrastive learning, which was proven to hold strong causal properties [21].

In this section, we first give a short overview of the theoretical details and the intuition behind of `CausalAdv`. We then introduce contrastive learning methods and summarize the theory and intuition from [21] that led to the `NatCL` loss.

2.1 `CausalAdv`: Adversarial Robustness through the lens of causality

2.1.1 Deducting adversarial vulnerability with causality

In their work, Zhang et al. [25] introduce `CausalAdv` : a *Causal*-inspired *Adv*ersarial distribution alignment method. They make the following assumption: as humans, we are less vulnerable to adversarial attacks as we base our reasoning on causality, and therefore are robust to imperceptible or non-semantic changes in the data we observe. Hence, introducing causality into the adversarial training framework should help improve adversarial robustness.

Following this intuition and by using domain expertise, we can analyze the process of adversarial attacks by considering the causal graph depicted in Figure 2. Using this newly introduced graph formalization, they elaborate on the distinction between natural and adversarial distribution. As an example, for an image of a trotting cow at the beach (as seen in Figure 1), the content variable C should encode distinctive morphological features of a *general* cow (e.g. the location and size of its horns) while the style variable S would encode all other non-causal features unique to the image X (e.g. the beach background).

An adversarial attack is designed to fool a certain classifier $h(.)$ with weights θ, by maximizing a loss function $l(.)$. For a sample X and its corresponding label Y, one can obtain the optimal adversarial perturbation according to Equation 1, which shows how E is caused by X, Y and θ:

$$E = \underset{E' \in \mathcal{P}}{\operatorname{argmax}} l(h(X+E';\theta),Y) \quad \text{with } \mathcal{P} \text{ the set of possible perturbations.} \tag{1}$$

According to the causal formalism introduced in the causal graph $\mathcal{G}$, content and style are assumed be statistically independent. However, as variable X is a collider between C and S, conditioning on X opens a backdoor-path between style and label. Therefore, we have that $Y \not\perp\!\!\!\perp S|X$ and since the backdoor-path opened by conditioning on X allowed style features to flow from S to Y and then from E to $\tilde{X}$, we also have that $\tilde{X} \not\perp\!\!\!\perp S|X$. The first result indicates that Y is not independent of the latent style variable S given the image X. This means that a model might implicitly learn to extract S from X and use it to predict Y. It's important to note that this dependence does not necessarily lead to shortcut learning, this phenomenon occurs only if the style variable S is strongly correlated to Y. The second result shows that solely aligning natural and adversarial distributions is not sufficient to alienate the impact of spurious correlations.

Considering that standard adversarial attacks should not generate visible changes of the style variable S, we can derive Equation 2 and deduct that these attacks successfully fool deep neural networks (DNNs) by leveraging spurious correlations between the class variable Y and non-causal features S encoded in $\mathbb{P}_\theta(Y|X,s), \forall s \in \mathcal{S}$ the set of possible styles:

$$P_\theta(Y|X)=\sum_{s\in\mathcal{S}}P(s|X)P_\theta(Y|X,s)\neq\sum_{s\in\mathcal{S}}P(s|X)P_\theta(Y|\tilde{X},s)=P_\theta(Y|\tilde{X}) \tag{2}$$

For the authors of `CausalAdv` [25], these results match empirical observations that adversarial examples are directly linked to the presence of 'non-robust' features in the natural data distribution as pointed out by Ilyas et al. [16].

2.1.2 Distribution alignment with `CausalAdv`

In order to mitigate the adversarial vulnerability that stems from the *shortcuts* learned by DNNs, one could follow two approaches. The most fundamental approach would be to regularize the model to make the impact of spurious correlations between Y and S negligible. Zhang et al. chose to follow the second approach, that is trying to align $P_\theta(Y|\tilde{X},s)$ with $P(Y|X,s)$, treating the latter as an anchor. The adversarial alignment is introduced in [25] as:

$$\min_\theta d(P(Y|X),P_\theta(Y|\tilde{X}))+\lambda\mathbb{E}_s[d(P(Y|X,s),P_\theta(Y|\tilde{X},s))] \tag{3}$$

with $d(.)$ a divergence metric and $\lambda > 0$ a hyperparameter that weighs the alignment incentive. As the optimization objective in Equation 3 is intractable, we can instead minimize a tractable upper-bound that corresponds to the realization of the `CausalAdv` distribution alignment method:

$$\begin{aligned}\min_{\theta,W_g}\mathbb{E}_{(X,Y)\sim P(X,Y)}&CE(h(X+E;\theta),Y)+\gamma CE(h(X;\theta),Y)\\ &+\lambda[CE(\bar{g}(\mu(X+E);W_g),Y)+\beta CE(\mu(X);W_g),Y)]\end{aligned} \tag{4}$$

with $h(X;\theta)$ modeling $Q_\theta(Y|X)$ which is an estimator of $P(Y|X)$, $\bar{g}$ parameterized by W_g and $\mu(X)$ and defined as an upper-bound of $P_\theta(Y|X,s)$, considering $s \sim \mathcal{N}(\mu(x),\sigma^2 I)$ and $\mu(X)$ to be the mean of the styles defined as an affine mapping between style weights W_s and representation r learned by the model h. Note that W_s is instantiated as the orthogonal complement of the weights that connect r to the logits of model h, making sure estimated style and content variables are statistically independent.

More details about the theoretical aspect of the alignment method can be found in the original paper.

2.2 `NatCL`: a style-free alignment strategy with contrastive learning

Self-supervised learning (SSL) is an unsupervised learning strategy that leverages unlabeled data and data augmentations techniques to learn a meaningful feature space. Generally, a pair of feature extractors (DNNs), **g** and **g'** (in practice we usually have **g = g'**) map two inputs **x** and **x'** to the corresponding features **z** and **z'**. Then, the feature extractors are trained such that if the inputs were "close", the similarity of **z** and **z'** should be high. For example, the features similarity of two images of the same cat in two different scenes should be high, and the features similarity of two different cats should be higher than the one of an image of a cat and an image of a dog. However, only maximizing similarity brings the problem of collapsed representations, as the model can achieve minimal loss by mapping all inputs to the same representation, leading to a degenerate optimal solution.

In order to avoid collapsed representations, a straightforward solution would be to penalize the model for bringing negative pairs of inputs (e.g. a cat and a dog) together. This strategy is called *Contrastive Learning* and is at the core of numerous famous SSL methods [6, 14].

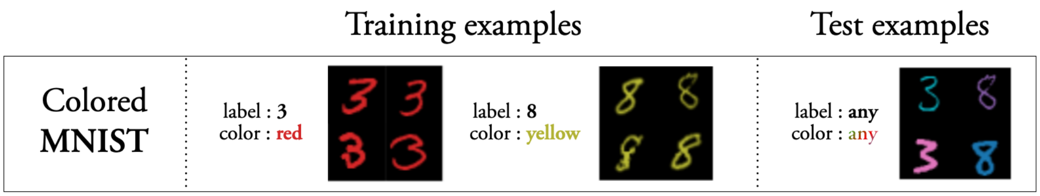


Figure 3: Example of training and test observations for `Colored-MNIST`. The `color` spuriously correlates with the digit label at train time, while the correlation doesn't hold for test examples.

Van den Oord et al. introduced a popular loss function based on noise-contrastive estimation (NCE) [13] for contrastive learning: `InfoNCE` [20]. Given a feature extractor g, p_t the distribution over a set of augmentations $\mathcal{T}$, $\tilde{z}=g(t(x))$ and $\tilde{z}'=g(t'(x))$ two augmented views of an input x with $t,t'\sim p_t$, K input samples from a distribution p_X and a temperature τ, `InfoNCE` writes as follows:

$$\mathcal{L}_{InfoNCE}(g;\tau;K)=\mathbb{E}_{\{x_i\}_{i=1}^{K}\sim p_X}[-\sum_{i=1}^{K}\frac{\exp\{\mathtt{sim}(\tilde{z}_i,\tilde{z}_i{}')/\tau\}}{\sum_{j=1}^{K}\exp\{\mathtt{sim}(\tilde{z}_i,\tilde{z}_j{}')/\tau\}}] \tag{5}$$

The numerator of the `InfoNCE` loss holds two properties: (1) *alignment* of the features through the numerator and (2) *uniformity* of the resulting distribution through its denominator. Moreover, as $K\rightarrow\infty$, the *uniformity* objective reformulates as an entropy estimator of $\tilde{z}$ [22]. From this result, the `InfoNCE` loss can be interpreted "as alignment of positive pairs with (approximate) entropy regularisation" [21]. Assuming augmentations leave the *content* information invariant and only affect the *style* variables, and considering an image x is induced by the underlying SCM from a graph similar to $\mathcal{G}$ from Figure 2, Von Kügelgen et al. show that SSL with the contrastive learning `InfoNCE` loss asymptotically identifies true *content* variables and isolates *style* variables (in their theory, C and S don't need to be statistically independent).

By definition, natural adversarial samples are augmented views of the original samples that leave the *content* information invariant. With ground-truth knowledge of the content-style partition of the data generating process of interest, we can consider natural adversarial samples as ideal augmentations. Compared to synthetic augmentations used in practice (e.g. horizontal flip, color distortion, etc.), we hypothesize that for a given number of samples, optimizing the `InfoNCE` loss with original and natural adversarial samples as the two views would perform better. As the scope of this work remains on supervised learning methods, we need a loss signal that also takes into account prediction accuracy. To this end, we derive the following `NatCL` loss that combines the style isolation property of unsupervised contrastive learning with a traditional Cross-Entropy loss for prediction accuracy:

$$\mathcal{L}_{NatCL}=\mathcal{L}_{InfoNCE}+\alpha*\mathbb{E}_{\{x_i,y_i\}_{i=1}^{K}\sim p_{X,Y}}CE[h(x_i),y_i] \tag{6}$$

with h being the full DNN, α a hyperparameter, g used in the `InfoNCE` loss corresponding to h without its last layer, $\tilde{z}=g(x)$ (i.e. original view) and $\tilde{z}'=\mathbb{E}_{\hat{s}\sim p_s}[g(x_{\hat{s}})]$ with $x_{\hat{s}}$ being the natural adversarial sample associated to style $\hat{s}$ with p_s the distribution of styles over the set of styles $\mathcal{S}$.

3 Experiments

Natural adversarial samples are notably expensive to collect as they (1) require human-insight to identify potential shortcuts (spurious correlations between the style and the label) used by DNNs, and then (2) imply collecting or generating samples from the resulting de-correlated distribution, which can be expensive to construct as style-content partition gets more complex. For this reason, an important metric to consider when benchmarking different adversarial alignment methods is **adversarial-data efficiency**. Said differently, a good alignment method should achieve high natural adversarial accuracy by training on the least amount of natural adversarial samples.

3.1 Dataset

`Colored MNIST` This dataset was first introduced in [1] and derives from the popular `MNIST` dataset [17] composed of 70000 images (32×32 pixels) of hand-written digits collected from approximately 250 writers. The goal is to predict the correct digit label for each image. Unlike standard grayscale

MNIST images, in the Colored-MNIST training dataset, *each digit* is colored such that the color perfectly correlates, albeit spuriously, with the class label. However, this color-to-label correlation does not hold true for the test dataset where color are assigned *randomly*. A natural adversarial sample consist in changing the color associated to a given digit to any other color in the predefined set of 10 colors. Figure 3 shows examples of training and test examples from Colored-MNIST.

The strength of the correlation between this spurious feature and the label exacerbates shortcut-learning in accordance with the least effort principle. A ResNet-18 trained on Colored-MNIST for 20 epochs achieves an accuracy of 100% on the training set while its test accuracy, 10.31%, remains equivalent to a random baseline. This result shows that a standard DNN solely learns spurious correlations between style and label on Colored-MNIST instead of learning more robust but noisy semantic features, like the recurrent geometric pattern of each digits.

3.2 Baselines

After discussing CausalAdv and NatCL thoroughly, one might ask to what extent these methods really outperform a vanilla supervised learner trained on a dataset containing both natural and natural adversarial samples. Therefore, in addition to comparing CausalAdv and NatCL trained with natural adversarial samples (as adversaries for the former and as augmented views for the latter), we also evaluate a vanilla ResNet-18 with different proportion of samples per batch switched to natural adversarial samples.

We generate natural adversarial samples synthetically by randomly changing the color of the sample pixels using the data generating process introduced in [1]. Doing so, we corrupt the *perfect* correlation between color and label. We evaluate 11 proportions of natural adversarial samples per batch, linearly increasing from 0% to 100% and monitor the number of natural adversarial samples seen during training.

The ResNet-18 model is train with 64 samples per batch, for 20 epochs, using a SGD optimizer with momentum $=0.9$, lr $=0.1$ and weight decay $=0.0002$.

3.3 Alignment methods for natural adversarial robustness

We train *CausalAdv* on the same ResNet-18 used for vanilla baselines. Recalling the objective function from Equation 4, we set $\beta = 1.0$ as suggested in [25] and performed a hyperparameter search over γ and λ. We found $\gamma = 0$ and $\lambda = 0.5$ achieve the best performance. For each batch of natural samples, we generate a corresponding natural adversarial batch that acts as $\tilde{X}$.

Naturally, *NatCL* is also trained on the same ResNet-18, however we set lr = 0.001 with the AdamW optimizer associated with CosineAnnealer learning rate scheduler. We performed a hyperparameter search over the temperature τ and α (see Equation 6) and set $\tau = 1.0$ and $\alpha = 0.1$. As for *CausalAdv*, for each batch of natural samples, we generate a corresponding batch of natural adversarial samples and treat them as augmented views, as explained in Section 2.2.

As with vanilla baselines, we evaluate the performance of these alignment methods on the natural adversarial test set while monitoring the number of natural adversarial samples seen by the models during training.

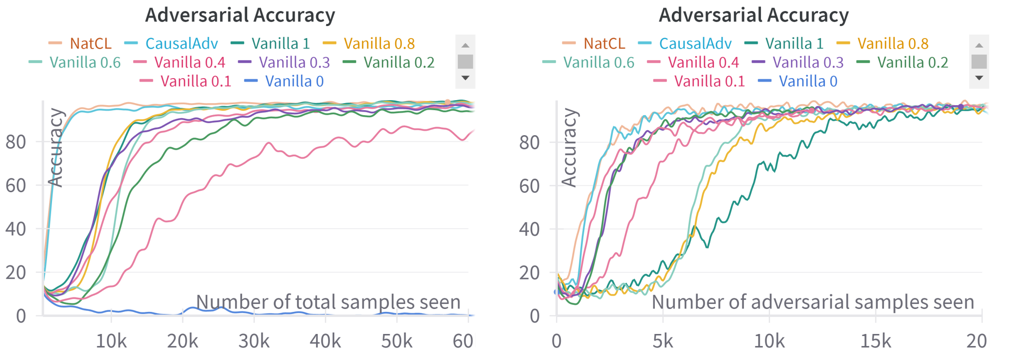
4 Results & Discussion

Figure 4a displays natural adversarial accuracy for vanilla models and alignment methods per amount of samples seen during training. Natural adversarial accuracy is evaluated after each batch step on natural adversarial examples where the digit color is different from the original digit color.3.

Firstly, Figure 4a shows that a Vanilla learner trained solely on natural samples reaches an asymptotic natural adversarial accuracy of 0%, demonstrating by itself naive learning method's reliance on simple, yet spurious, features (e.g. Color/Digits *perfect* correlation).

Secondly, both CausalAdv and NatCL exhibit significantly faster convergence in training compared to the Vanilla approach with respect to the total number of samples seen. However, augmenting the latter with even a few natural adversarial samples is enough for it to asymptotically reach similar natural adversarial accuracy that former methods [1]. As expected, the rate of this convergence is proportional to the natural adver-

[1] Note that this result is only valid for Colored-MNIST. This minimal ratio might change with the task difficulty. We let this consideration for future works



(a) Accuracy on natural adversarial samples for increasing number of samples seen during training

(b) Accuracy on natural adversarial samples for increasing number of natural adversarial samples seen during training

Figure 4: Accuracy on natural adversarial examples for `CausalAdv`, `NatCL` and Vanilla with various natural adversarial proportions.

sarial sample ratio used. Also, in term of natural adversarial data-efficiency, it seems that neither `CausalAdv` or `NatCL` offer a substantial advantage compared to a Vanilla learner with a reasonable proportion of natural adversarial samples (e.g 10%). These two results raise serious questions about the practical utility of these first two complex methods regarding the natural adversarial data efficiency. However, we should be cautious about the generalizability of this conclusion considering the complexity of `Colored-MNIST`.

Thirdly, Vanilla learners trained by considering an increasing proportion of natural adversarial samples show a reduced data-efficiency with respect to natural adversarial samples. This inefficiency is attributable to the misuse of natural adversarial samples as an expensive way to learn semantic feature (e.g. the shape of a digit), while they should mostly serve as a sparse regularizing signal. Ideally, their use should only serve as a mechanism to corrupt the spurious correlation, making it a less reliable predictor, thereby encouraging the model to prioritize robust semantic features.

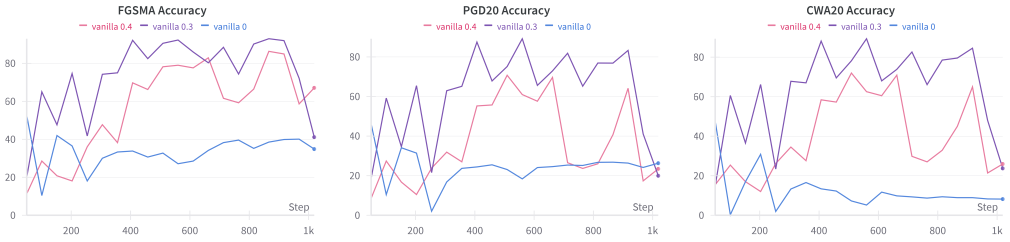
Finally, looking again at Vanilla learners, Figure 4b shows a notable *elbow* pattern in early-stage learning, which diminishes with a higher proportion of natural adversarial samples. This may be attributable to a gradient signal biased toward *obvious* features learning, good-enough for most (e.g. natural) samples, while later training step may refine the decision rule toward handling those set-aside adversarial samples.

Vanilla: Learning rate and Robustness When we set the learning rate as 0.1, a Vanilla training without any natural adversarial examples (i.e., $p_{adv} = 0$) achieves low accuracy for white-box attacks (i.e., FGSM: 40%, PGD: 20%, and CW: 10%), which means that the model is not robust against these gradient-based attacks. Interestingly, we can make the model robust (i.e., around 90% for both three attacks) by introducing an appropriate amount of natural adversarial examples, as displayed in Figure 5. This result suggests that we can potentially obtain a robust model against gradient-based attacks through training with natural adversarial examples.

However, we observed a reversed result if we applied the same Vanilla training with a learning rate 0.01. The ordinal Vanilla training without any natural adversarial examples already achieves a high robustness level against gradient-based attacks (e.g., around 94% for FGSM), and introducing natural adversarial examples only maintains the robustness level or makes it a little worse, as displayed in Figure 6. These two contradictory results reveal that the relation between the robustness against white-box attacks and the adversarial training with natural adversarial examples needs further investigation.

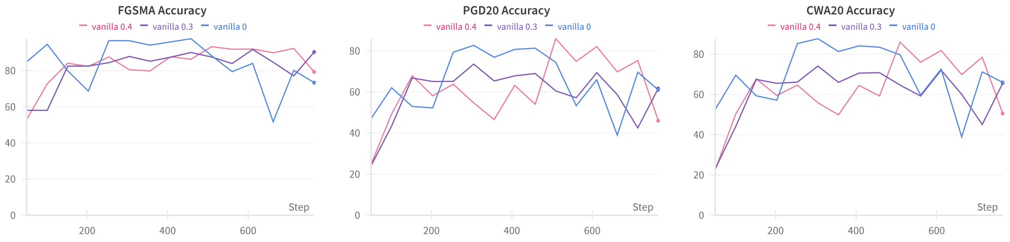
5 Conclusion & future work

In this work, we propose to evaluate multiple alignment methods to alleviate shortcut learning in a real-world scenario. We believe that in an attempt to teach a model, what makes an image of a digit refer to a digit falls under our perception of it as a human. There is no reason that a model prefers by itself a more noisy correlation pattern between a human-intended yet very noisy feature (e.g. geometric shapes



(a) FGSM Accuracy (b) PGD Accuracy (c) CW Accuracy

Figure 5: Accuracy by training steps for white-box attacks with Vanilla adversarial training of learning rate = 0.1.



(a) FGSM Accuracy (b) PGD Accuracy (c) CW Accuracy

Figure 6: Accuracy by training steps for white-box attacks with Vanilla adversarial training of learning rate = 0.01.

in the image) and the label rather than a clear signal between an unintended feature (e.g. the color) and the label. For this reason, we **have** to provide this human-feedback that gives insights about the underlying pseudo-causal graph. In our study, these insights take the form of natural adversarial samples. As these samples are expensive to collect, we focus our evaluation on a cost-efficiency point-of-view relative to the amount of natural adversarial samples needed. We introduce `NatCL`, a distribution alignment method derived from contrastive learning and compare its performance on natural adversarial samples with `CausalAdv`, a causally-inspired adversarial training method, as well as vanilla baselines. For a simple toy dataset, we found that using an appropriate proportion of natural adversarial samples with a vanilla model achieves almost similar performance to complex alignment methods.

For future work, we would like to investigate the potential for gradient-based adversarial robustness using natural adversarial samples. Furthermore, it would be crucial to conclude the preliminary findings of this work to extend our bench-marking to more complex datasets like the `WaterBirds` dataset [19].

References

[1] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization, 2020.

[2] S. Beery, G. van Horn, and P. Perona. Recognition in terra incognita, 2018.

[3] J. Bridle. *Ways of being*. Allen Lane, Apr. 2022.

[4] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks, 2017.

[5] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay. Adversarial attacks and defences: A survey, 2018.

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations, 2020.

[7] J. C. Costa, T. Roxo, H. Proença, and P. R. M. Inácio. How deep learning sees the world: A survey on adversarial attacks & defenses, 2023.

[8] A. Fawzi, H. Fawzi, and O. Fawzi. Adversarial vulnerability for any classifier, 2018.

[9] W. E. Ferson, S. Sarkissian, and T. T. Simin. Spurious regressions in financial economics? *The Journal of Finance*, 58(4):1393–1413, 2003. ISSN 00221082, 15406261. URL `http://www.jstor.org/stable/3648215`.

[10] M. Gasse, D. GRASSET, G. Gaudron, and P.-Y. Oudeyer. Using confounded data in latent model-based reinforcement learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=nFWRuJXPkU`.

[11] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2 (11):665–673, Nov. 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL `http://dx.doi.org/10.1038/s42256-020-00257-z`.

[12] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples, 2015.

[13] M. Gutmann and A. Hyv¨arinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models, 2010.

[14] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning, 2020.

[15] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song. Natural adversarial examples, 2021.

[16] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/e2c420d928d4bf8ce0ff2ec19b371514-Paper.pdf`.

[17] Y. LeCun and C. Cortes. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

[18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks, 2019.

[19] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization, 2020.

[20] A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding, 2019.

[21] J. von Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello. Self-supervised learning with data augmentations provably isolates content from style, 2022.

[22] T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere, 2022.

[23] A. Williams, N. Nangia, and S. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In M. Walker, H. Ji, and A. Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL `https://aclanthology.org/N18-1101`.

[24] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan. Theoretically principled trade-off between robustness and accuracy, 2019.

[25] Y. Zhang, M. Gong, T. Liu, G. Niu, X. Tian, B. Han, B. Schölkopf, and K. Zhang. Causaladv: Adversarial robustness through the lens of causality, 2022.

Appendix A Technical details

Architecture For all alignment methods, we employed the `ResNet18` architecture [], which is well-suited for various image recognition tasks. ResNet18 is composed of a *feature extractor* and a *prediction head*. The *feature extractor* includes multiple convolution layers and projects observations into a latent space of size `512`. The *prediction head* consists of a simple fully connected layer, with the output dimension equal to the number of output classes for the task, 10 for MNIST. The network incorporates batch normalization and ReLU activations following each convolution operation.

Hyperparameter Optimization To identify adequate hyperparameters for the different alignment procedures, we implemented a simple grid search strategy. This approach involved varying key hyperparameters value specific to each alignment method, within predefined ranges and evaluating the performance of each configuration in term of asymptotic adversarial accuracy and rate of convergence. This systematic exploration of the hyper-parameter space ensured that each method was evaluated in optimal conditions.

- `CausalAdv`
  1. **adv_beta:**
     - Values: [0.0, **0.5**, 1, 2]
     - Description: Adjusts the weighting of the adversarial CE loss relative to the natural CE loss. A higher value increases the influence of the adversarial CE loss term.
  2. **adv_alpha:**
     - Values: [0.0, **0.5**, 1, 2]
     - Description: Adjusts the weighting of the CE loss when only using style from adversarial sample to predict the label. Corresponds to λ in Equation 4.
- `NatCL`
  1. **learning_rate**:
     - Values: [0.0001, **0.0005**, 0.001, 0.005]
     - Description: Determines the step size at each iteration of ADAM. Smaller values lead to slower convergence but can provide more accurate results by avoiding overshooting the local minimum.
  2. **NatCL_natural_weight**:
     - Values: [**0.1**, 0.5, 1.0, 5, 10]
     - Description: Adjusts the weighting of the unsupervised contrastive loss term alongside the standard supervised Cross-Entropy loss term.
  3. **temperature**:
     - Values: [0.1, 0.5, **1.0**, 5, 10]
     - Description: Modifies the sharpness of the distribution used in the softmax term of the contrastive loss, influencing how the model discriminates between similar and dissimilar images.
- `Vanilla`
  1. **adv_proportion:**
     - Values: [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, **0.7**, 0.8, 0.9, 1]
     - Description: The proportion of adversarial examples used in the training set.