Identifying and Benchmarking Natural Out-of-Context Prediction Problems

Anonymous Author(s)
Affiliation
Address
email

Abstract

Deep learning systems frequently fail at out-of-context (OOC) prediction, a problem of paramount importance in practical applications. However, most benchmarks for OOC prediction feature restricted or synthetically-created shifts, thus only measuring performance on a narrow range of realistic OOC problems. In this work, we introduce novel methods for leveraging annotations to automatically identify candidate OOC examples in existing computer vision datasets. Utilizing two contrasting notions of context, we present NOOCH: a suite of naturally-occurring "challenge sets", intended for evaluating a model's OOC performance. Experimentally, we evaluate and compare several robust learning methods on these challenge sets, identifying opportunities for further research in OOC prediction.

1 Introduction

People often find context useful for prediction, both for improving accuracy and processing efficiency [10]. However, deep learning systems frequently over-rely on context cues [17, 18, 34], which can lead to poor performance on out-of-context (OOC) examples, when contextual information is misleading. Indeed, the OOC prediction problem is a challenging one to solve and define: in a natural image, "context" can be complex and high-dimensional, including aspects of object relationships [61], image backgrounds [61], experimental settings [43], or world models [6].

However, machine learning researchers rarely evaluate OOC prediction methods in ways that capture the full complexity of practical context shifts. Many standard benchmarks rely on sharp, synthetically-created shifts in one salient feature, e.g., image background or color [28, 50, 61]. Meanwhile, other datasets represent more natural OOC settings but are small [8], do not contain a training set [4], or entail significant domain knowledge [29]. While these are significant and valuable works, a closer analogue to the OOC problems that arise in the standard model development cycle might more scalably provide insights into the challenges faced by current approaches.

In this work, we introduce methods for identifying sets of OOC problems which naturally arise in existing computer vision datasets, and use these methods to expose OOC "challenge sets" [27] within an existing benchmark. We focus on the task of determining object presence in an image, a problem where relationships between various objects naturally provide helpful context. By identifying candidate OOC examples at scale, we enable quantitative and qualitative analysis of the OOC problems that are *already present*, gaining a clearer window into both the performance of current approaches, as well as opportunities for further improvement. Our contributions are as follows:

- We present NOOCH (Naturally-Occurring Out-of-context Challenge sets), a suite of "challenge sets" for evaluating performance on naturally-arising OOC problems;

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

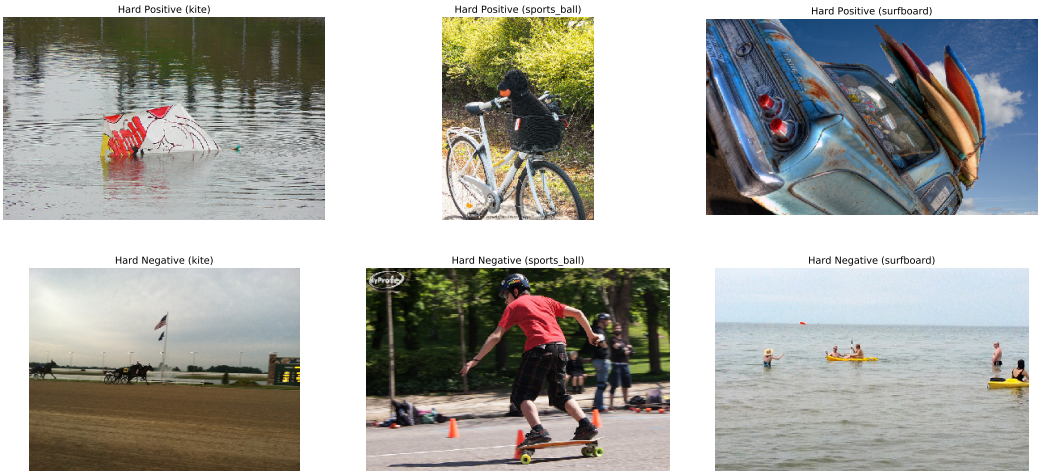

Figure 1: Using the co-occurrence/extractibility (CE) criterion, examples of (0.05, 0.1)-hard positives (top row) and negatives (bottom row) for the classes (L to R): `kite, sports_ball, surfboard`.

- We introduce novel methods for automatically identifying OOC challenge sets from existing annotated data, and define concepts of "hard positives" and "hard negatives" in the OOC setting; and
- We analyze the performance of several methods from the robust learning literature on these challenge sets, and explore the tradeoffs inherent in different approaches.

2 Finding Naturally-Occurring OOC Problems

Throughout this paper, we consider the binary prediction task of determining object presence — given an image X, is an object of class Y present or not? In order to study how context shifts arise naturally within an existing computer vision benchmark, we propose two methods which leverage annotations to automatically identify candidate OOC examples in the test set.

2.1 COCO and COCO-Stuff

The Microsoft Common Objects in COntext dataset (COCO) [32] is a computer vision dataset consisting of images of natural scenes. Each image is annotated with instance labels and bounding boxes for every "thing" in the image, as well as several captions describing the content of the scene. Images usually contain multiple items and as such usually have multiple labels.

However, for the purposes of investigating OOC prediction, many relevant objects are not labelled in COCO; for instance, background objects such as "sky" or "grass" are not COCO classes. Fortunately, the COCO-Stuff dataset [7] provides labels and bounding boxes for all of the "stuff" in the images from COCO. The "thing" vs "stuff" distinction is a technical one: a thing is an object with a specified size and shape, whereas stuff has no "defined spatial extent" [15]. Having both thing and stuff labels is essential for understanding model behaviour on OOC examples, since it is exactly these "stuff" classes which often (but not always) provide important context cues. Taken together, the thing and stuff annotations yield a rich sandbox for queries about the role of context in prediction. For our purposes, COCO-Stuff contains 171 binary tasks for determining object presence (81 thing classes and 90 stuff classes; we ignore the 1 "unlabelled" class).

2.2 Automatically Identifying OOC Examples: Hard Positives and Negatives

As mentioned in Sec. 1, context is hard to define. In this work, we consider two contrasting notions of context: 1. frequently co-occurring, easily extractible objects; and 2. the "gist" of a scene. Based on these, we present two criteria for identifying naturally-occurring OOC prediction problems within

(a) hard positive (CE) (b) hard positive (Gist) (c) hard negative (CE) (d) hard negative (Gist)

Figure 2: To contrast the CE and Gist criteria, we show samples from the `airplane` (L) and `bowl` (R) tasks.

the existing COCO(-Stuff) dataset. We identify both *hard positives*, where the class is present despite an unusual context, and *hard negatives*, where the class is *not* present, despite the usual context.

2.2.1 Defining Context Using Co-Occurrences and Extractibility

For an object class Y, context cues often come in the form of another object class C that has two properties (cf. [34]). First, C and Y co-occur frequently [5]. Second, C is more extractible than Y — easier to detect. If C were *less* extractible than Y it would not be a useful cue for detecting Y, as a model could detect Y directly.

We can utilize these properties to create candidate context cues C for a class of interest Y. Given bounding box annotations, we can use an object's size within an image as a proxy for extractibility (larger objects tend to be more extractible). Let the $Area$ operator take the sum of the areas of all bounding boxes of instances of that object, or return 0 if the object is not present. Then, to estimate from the training set how important of a context variable C is, we can computer $A(C, Y) = \mathbb{E}[Area(C) - Area(Y)|Y = 1]$. When $A(C, Y)$ is larger, this means that when Y is present, C is usually also present, and on average, takes up more of the image than Y does. When $A(C, Y) > \alpha$, we say that C is an α-strong context cue for Y (or just α-context for brevity). We find that many of the contexts returned using this method for large enough α are intuitive. Some examples of (label, 0.05-context) pairs are: (`car`, `road`), (`bowl`, `dining_table`), (`cow`, `grass`).

Using this notion of context, we can then define hard positive and hard negative examples. We make the simplifying *noisy-or* assumption: that each context cue provides evidence for Y, so that the presence of any cue supports Y being present, while the absence of all provides evidence against Y's presence. Given some image, we can define an (α, β)-hard positive or negative. If $Y = 1$, and for all α-context cues C, we have $Area(C) < \beta$ in this image (and there is at least one α-context cue), then the example is an (α, β)-hard positive. Alternatively, if $Y = 0$, and there exists some α-context variable C such that $Area(C) > \beta$, then the example is an (α, β)-hard negative. We will call this method the co-occurrence/extractibility (CE) criterion (see Fig 1 for examples). Throughout, we use $\alpha = 0.05, \beta = 1$ unless otherwise noted; these parameters were chosen since they approximately equalize $P[(X, Y)$ is (α, β)-hard $|Y = y]$ across $y = 0, 1$. See Appendix B for more details.

2.2.2 Defining Context Using Gist

We now turn to a broader notion of context, that of the "gist" of a scene [57], or its overall semantic content. This is something that humans are able to recognize easily [31], but goes beyond object frequency and extractibility. When an object is present in a scene whose gist is very different from the scenes the object was present in at training time, this may create a large challenge for OOC prediction.

We here describe our method for estimating gist shift, which we call the gist criterion. We leverage caption annotation data in COCO (each image has 5 caption sentences), making the assumption that information in a caption captures the gist of a scene. We then take an SBERT embedding [44] of each available image caption for an image, and average these embeddings to get a single embedding for each image. Then, for a given image and some target label Y, we can look at the cosine similarity between that image's embedding and the average embedding across all training images with $Y = 1$. If this similarity is below some threshold τ, and $Y = 1$ for the test image, it is a τ-hard positive; if this similarity is above τ, and $Y = 0$ for the test image, it is a τ-hard negative. Note, we do not look at distance to $Y = 0$ examples; we assume that captions for $Y = 0$ images may have little in common, whereas the mean caption for $Y = 1$ is a prototypical description of Y "in context".

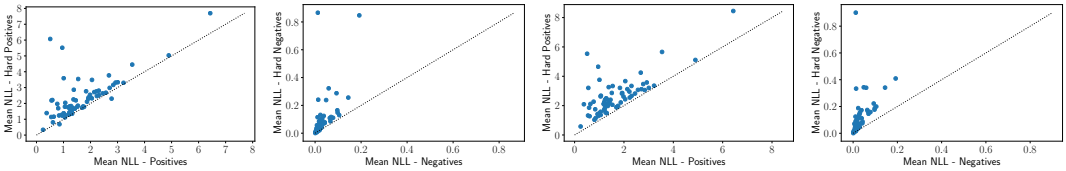

Figure 3: We find that hard positives/negatives induce higher average loss for both criteria. Each point is a task in COCO-Stuff: the x- and y-axis values show the average loss achieved by an ERM model on all positives (negatives) and the average loss on hard positives (negatives). From L to R: CE criterion (positives), CE criterion (negatives), gist criterion (positives), gist criterion (negatives). The diagonal line represents where the hard example losses match marginal losses.

Throughout, we set the threshold τ for each task so that the number of hard positives and negatives is the same as the CE criterion for that task, to facilitate comparisons. See Figure 2 for examples of hard positive and negatives chosen by this criterion in comparison to the CE criterion, and Appendix B for more details.

2.3 NOOCH-CE and NOOCH-Gist: Selecting Challenge Sets

We train binary classifiers to minimize average NLL on each of the 171 classes in COCO-Stuff. We find that for nearly all tasks, the hard positives and negatives defined by our methods incur higher average loss than positive and negative examples respectively (Fig. 3), for both the CE and Gist criteria. This provides some evidence that our criteria are, in fact, identifying examples which are more difficult to classify correctly.

To select candidate OOC tasks for our challenge sets, we select the 12 tasks with the largest difference between average NLL on hard examples (by the CE criterion) and average NLL on all examples, filtering out those tasks without at least 50 hard positives and hard negatives. We call these tasks the NOOCH (Naturally-Occurring Out-of-context Challenge) suite (Table 1). We then identify two groups of challenge sets: NOOCH-CE, which consists of the hard positive and negative examples on each of the 12 tasks in NOOCH as identified by the CE criterion; and NOOCH-Gist, which is the analogous set for the gist criterion. This gives us 24 total challenge sets on which to evaluate an ML model's OOC performance.

Task	# Hard +	# Hard -	# +
car	1539	949	2585
bowl	944	1084	1463
boat	361	756	649
fire-hydrant	189	1577	332
airplane	258	3622	635
cow	194	2360	445
backpack	607	6413	1183
cup	743	6852	1926
surfboard	147	4011	750
tie	136	6347	778
sports-ball	151	6574	860
kite	55	6725	451

Table 1: Counts of hard positive, hard negative & positive examples per NOOCH task.

Contrasting the Two Criteria (Fig. 2). The left two images in Fig. 2 are hard positives on the `airplane` task. On the far left, we see the inside of an airplane: this is selected as a hard positive by the CE criteria because there is no sky visible. On the second left, the sky is visible but the overall scene is unusual; this is selected as a hard positive by the Gist criteria. The right two images are hard negatives on the `bowl` task. On the second right, we see a large dinner with many plates: this is selected as a hard negative by the CE criteria because there is a prominent dining table but no bowls. On the far right, we see a paper plate on a desk (not a dining table): this is selected as a hard negative by the Gist criteria since there is no bowl but it is similar to images where you might expect a bowl.

3 Methods

We choose to focus on five methods which provide a useful contrast between different approaches to OOC prediction. Expected risk minimization (**ERM**) is the standard loss function in ML systems; as shown by Gulrajani and Lopez-Paz [19], it can be difficult to beat on domain generalization problems.

Group DRO (**GDRO**) [50] and invariant risk minimization (**IRM**) [2] are popular examples of environment-based learning methods. In contrast to these, we also look at two *adaptive* methods, which do not require auxiliary information at training time and instead attempt to adaptively determine the most difficult examples at each step. Conditional value at risk minimization (**CVaR**) [45], like GDRO, can be thought of as an example of distributionally robust optimization, a common approach in the robustness literature [14] — however, in CVaR we minimize the loss for a worst case distribution over examples at each step, and in GDRO we minimize the worst-case loss among pre-defined groups. **Focal** loss [33] is not necessarily commonly used in the OOC setting; however, it is a widely used reweighting technique in computer vision; a class of approaches common in OOC prediction [54]. See App. A for information on other methods we experimented with but did not include here, including methods that adapt parameters to context and methods which aim to predict context. We found the methods we included here had the best results overall, on the standard and hard examples.

Throughout, we use the following notation. We are given a dataset $\{x_i, y_i\}_{i=1}^n$ with inputs and target labels respectively, and possibly some side information $\{c_i\}_{i=1}^n$ as well (e.g. environment variables). If side information is available, we assume it is available at training time but not necessarily test time. We are aiming to learn a function $f : \mathbb{X} \longrightarrow \mathbb{Y}$. We assume ℓ to be the the example-wise cross-entropy loss function.

3.1 Expected Risk Minimization

Expected risk minimization (ERM) is the standard paradigm for training ML models. In ERM, we minimize the mean loss ℓ on the training set, plus potentially some data-independent regularizer $\mathcal{R}$:

$$\mathcal{L}_{ERM}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i) + \mathcal{R}(f) \tag{1}$$

3.2 Environment-Based Methods

One common setting is where the side information c_i for each example is a one-hot variable, representing an *environment*. The notion of environments for robust learning was originally given with causal motivation [43], where each value of c represents a different intervention on the underlying causal graph. However, this type of grouping procedure can be used completely separately from the causal context, and so we just consider it to represent some informative partition of our data, which is defined as an input at training time. Here, $c \in \{1 \dots C\}$ is an integer or a one-hot encoded variable. We will use the following shorthand to describe the average loss for a group:

$$\ell_c(f) = \frac{1}{n_c}\sum_{i=1}^{n} \ell(f(x_i), y_i)\mathbb{1}\{c_i = c\} \tag{2}$$

Group DRO. The first environment-based method we consider is Group Distributionally Robust Optimization (GDRO) [50], which aims to minimize the loss on the worst of C partitions, with partition c of size n_c, and group adjustment hyperparameter K:

$$\mathcal{L}_{GDRO}(f) = \max_{c \in \{1 \dots C\}} \left\{ \ell_c(f) + \frac{K}{\sqrt{n_c}} \right\} + \mathcal{R}(f) \tag{3}$$

This loss aims to ensure that no group's loss is that bad, and the group adjustment term ensures greater focus on smaller groups, which may otherwise be ignored.

IRM. The second environment-based method we consider is invariant risk minimization (IRM) [2]:

$$\mathcal{L}_{IRM}(f) = \sum_{c=1}^{C} \ell_c(f) + \lambda \|\nabla_{w|w=1} \ell_c(w \cdot f)\| + \mathcal{R}(f) \tag{4}$$

The second term here is a gradient penalty on the output of f, w is a constant multiplier on the output of f, and λ is a hyperparameter. The intuition for this is somewhat involved [2]; the overarching motivation is that each environment should learn a representation such that the same predictive classifier is optimal across environments.

3.3 Adaptive Methods

Finally, we consider a class of loss functions which, rather than using an injection of side information to specify which examples are OOC, focuses dynamically on the hardest examples at each step. The first is conditional variance-at-risk optimization (CVaR). This can be cast as a type of distributionally robust optimization (DRO) [14], where we aim to maximize the loss over, rather than the whole training set, a worst-case distribution over training examples. For some $p \in (0, 1)$, CVaR(p) is defined as the average loss of the p-percent worst-loss examples.

We also consider focal loss [33], which dynamically upweights high loss examples using a parameter $\gamma \geq 0$. With binary y, $q(x, y) = f(x)$ if $y = 1$ and $q(x) = 1 - f(x)$ if $y = 0$, with $f(x) \in [0, 1]$ as the continuous-scored output of f, not the binary prediction. Then, we have focal loss as a modification of the cross-entropy loss — at $\gamma = 0$ this reduces to cross-entropy; as γ increases, it focuses more loss on the examples which have higher loss already:

$$\mathcal{L}_{Focal}(f) = \frac{1}{n} \sum_{i=1}^{n} -(1 - q(x_i, y))^{\gamma} \log(q(x_i, y)) + \mathcal{R}(f) \tag{5}$$

4 Related Work

A range of datasets have been proposed for the purposes of benchmarking OOC prediction. Several focus on realistic OOC prediction: Koh et al. [29] contains problems from across a range of applications; Barbu et al. [4] focuses on object recognition and shows objects varied by a range of attributes; and Choi et al. [8] isolates 26 images of objects in unusual contexts from a larger dataset. Our work functions well as a compliment to any of these datasets; we believe it is novel due to our ideas for scalably identifying challenge sets from annotated data, as well as its delineation of hard positives and hard negatives. The notion of "challenge sets" [27], "stress tests" [38], or "contrast sets" [16] from the NLP literature is an inspiration for our work as well. A range of primarily semi-synthetic datasets include those that center around: image corruption [25, 35]; object hierarchy [51]; synthetic shifts in background [50, 61]; color [28]; a group attribute [1]; or purely synthetic data [3].

Several works have discussed explicit examples where deep models failed to perform OOC prediction in practice. Oakden-Rayner et al. [40] and Winkler et al. [59] discuss the risk of this occurring in the medical domain, and Shetty et al. [53] in the autonomous driving domain. Other works have detailed the challenge of OOC prediction for deep models, using frames of "shortcuts" [18], "simplicity" [52], "extractibility" [34], texture biases in CNNs [17], or the challenge of out-of-place objects [46].

A range of work outside deep learning considers the OOC prediction problem from a different direction, focusing on how to improve prediction by taking context into account [9, 24, 37, 63]. Other work looks at the idea of using a latent variable to represent scene gist [41, 60]. A number of newer methods not discussed elsewhere in the paper also aim to solve the OOC problem, including those that involve side information [26, 30, 55, 62], those that involve side information through causal underpinnings [23, 48], and those that ignore side information altogether [11–13].

5 Experiments

For all experiments we use a ResNet-50 [22], finetuned from ImageNet-pretrained features [49]. See App. B and C for further experimental details. For the environment-based methods, we follow Sagawa et al. [50] and create 4 environments: 1 for each element of the cross-product of the label and its highest-α context class. Many of the robust baselines from Sec. 3 come with a hyperparameter which aims to trade off between average performance and OOC performance; we choose the hyperparameter which minimizes the maximum loss of hard positives and hard negatives on the validation set.

5.1 Quantitative Analysis

In Table 2, we show AUC (area under the ROC curve) on hard examples for each of the five methods we focus on, for each of the 12 NOOCH tasks individually — AUC is a particularly informative metric since it is sensitive to label imbalance, and tasks in NOOCH contain 1-10% positive examples. In Figures 4, 5, 6, and 7, we show AUC, classification error, mean negative log-likelihood (NLL), and

Task	ERM	GDRO	IRM	CVAR	Focal
car	0.769	**0.923**	0.868	0.780	0.773
bowl	0.749	**0.846**	**0.828**	0.701	0.751
boat	0.869	**0.945**	0.909	0.894	0.854
fire-hydrant	0.913	**0.942**	0.902	0.933	0.925
airplane	0.986	**0.995**	0.986	0.984	0.983
cow	0.935	**0.963**	0.943	0.922	0.939
backpack	0.812	**0.849**	**0.811**	0.755	0.816
cup	**0.870**	0.825	0.820	0.794	**0.868**
surfboard	0.939	**0.960**	**0.957**	0.937	0.935
tie	0.742	**0.781**	0.640	**0.696**	**0.743**
sports-ball	0.867	**0.909**	**0.859**	0.856	0.865
kite	0.932	**0.960**	**0.953**	0.931	0.940
Average	0.865	**0.908**	0.873	0.849	0.866

Task	ERM	GDRO	IRM	CVAR	Focal
car	0.766	**0.850**	**0.817**	0.780	0.770
bowl	0.642	**0.674**	0.631	0.620	**0.655**
boat	0.826	**0.880**	0.818	**0.855**	0.806
fire-hydrant	0.840	0.798	0.788	**0.865**	0.847
airplane	0.977	**0.981**	0.961	0.979	0.975
cow	**0.901**	**0.906**	0.845	**0.900**	0.895
backpack	**0.716**	**0.732**	**0.729**	0.604	**0.731**
cup	**0.733**	**0.742**	**0.705**	0.664	**0.729**
surfboard	**0.913**	0.875	0.836	**0.918**	0.903
tie	**0.822**	**0.820**	0.696	0.783	**0.821**
sports-ball	0.830	**0.880**	0.848	0.826	0.824
kite	**0.942**	**0.947**	**0.943**	**0.940**	**0.947**
Average	0.826	**0.841**	0.801	0.811	0.825

Table 2: AUC on hard test examples for all 12 NOOCH stress tests, after hyperparameter selection. Bold numbers have overlapping standard deviations. L: NOOCH-CE. R: NOOCH-Gist.

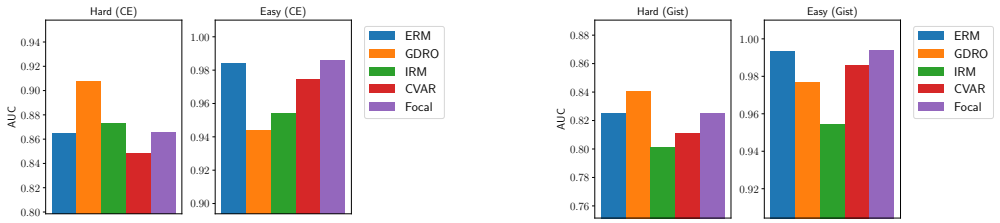

Figure 4: AUC (area under the ROC curve) achieved on hard and easy examples. Higher is better.

expected calibration error (ECE) [20] respectively, averaged across the 12 tasks. For AUC and ECE, we display results for hard and easy examples separately (where an "easy" example is defined as one which is not "hard", i.e. not a hard positive or hard negative). For error and NLL, we further break out the results on hard examples into hard positives and hard negatives. For all, we show NOOCH-CE and NOOCH-Gist separately.

Gist Shift is Difficult. We note that performance on NOOCH-Gist is worse than NOOCH-CE. This suggests that the more complex notion of context embodied by the gist yields a more difficult OOC task. In some ways, would expect models to find these more holistic shifts harder, as these gist shifts go well beyond "shortcuts" [18]. We also note that the gap between the environment-based methods and ERM is much smaller on NOOCH-Gist, indicating that ERM is a strong baseline on this task, making gist shift a useful direction for OOC prediction research.

Tradeoffs Exist Between Harder and Easier Examples. We note two tradeoffs across AUC, error, and NLL results: models that are better on hard examples tend to be worse on easy examples (and vice versa), and models that are better on hard positives tend to be worse on hard negatives. ERM is better on easy examples and hard negatives: since the dataset is imbalanced, the hard negatives are the easier "hard" examples, and we expect ERM to perform better here since negatives are the majority class. We also note that environment-based methods beat ERM on hard positives (the harder "hard" examples) suggesting that the extra information from environments helps.

Environments Matter. GDRO's performance falls the most between NOOCH-CE and NOOCH-Gist. We interpret GDRO's performance on NOOCH-CE as unusually high, as it optimizes over a set of groups similar to the ones that calculate this metric. However, this seems to overstate its ability to handle more complex OOC problems, such as NOOCH-Gist. The choice of environment variables is crucial in general; IRM might improve with the addition of environment variables which vary more independently from the context variables of concern. The effect of environments and how to choose or create them is an open question, and an active area of research.

ERM Makes Confident OOC Errors. We find that the AUC and NLL/Error plots tell different stories. When looking at AUC, we see that the environment-based methods perform the best on NOOCH-CE, and things are more even on NOOCH-Gist, with GDRO falling off quite a bit and IRM

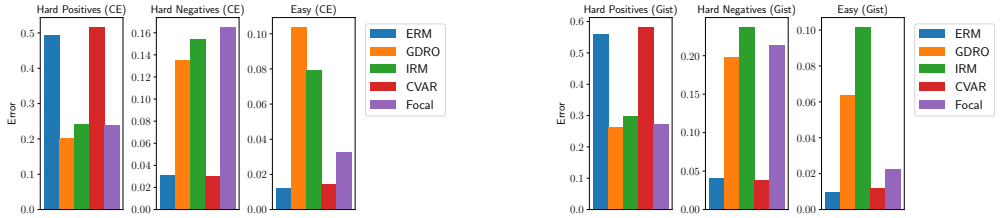

Figure 5: Classification error achieved on hard positive, hard negative, and easy examples.

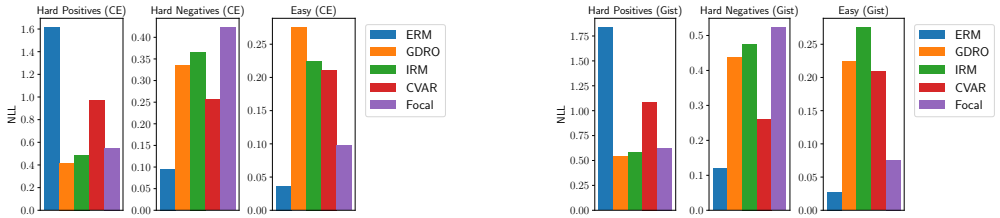

Figure 6: Negative log-likelihood (NLL) achieved on hard positive, hard negative, and easy examples.

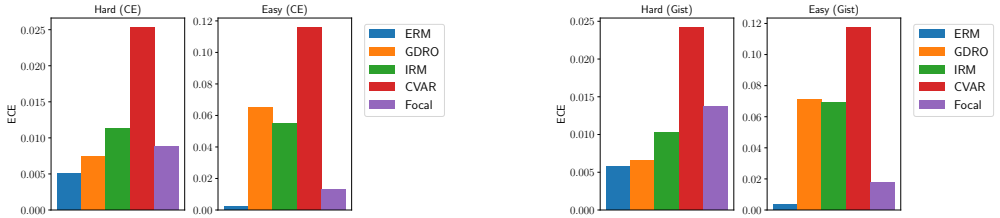

Figure 7: Expected Calibration Error (ECE) achieved on hard and easy examples. Lower is better.

recording the worst AUC on hard examples. However, when looking at NLL (or error), we see that on both challenge sets, GDRO and IRM have much lower NLL than ERM on hard positives (and higher on hard negatives/easy examples). This suggests that while the relative ordering of predictions on hard examples from ERM models may be fairly good (as suggested by the AUC), when it is wrong, it is sometimes extremely wrong, incurring high NLL on its worst mistakes.

Adaptive Methods Find Different Tradeoffs. We find that CVaR and focal loss find interesting tradeoffs between ERM and environment-based methods. CVaR performs slightly worse than ERM on both hard and easy examples by AUC and error measures; and by ECE it is by far the worst of any method. However, CVaR's NLL on hard positives is between ERM and the environment-based methods. Focal loss, on the other hand, performs similarly to ERM by AUC on both hard and easy examples, but it performs similarly to the environment-based methods on hard positives and negatives by error and NLL, providing a strong tradeoff between overall and OOC performance.

Two Surprising Calibration Results. First, we found that there was *not* the same tradeoff with calibration as there was with the other metrics. In particular, ERM had the best calibration on both hard and easy examples, suggesting there may be a path forward to avoid some of the tradeoffs between OOC and average performance by thinking more about calibration. Second, we found that focal loss had much better calibration on easy examples than any other method (except ERM). The well-calibrated nature of focal loss is known [36]; interestingly, it was the only method with low NLL/Error on hard positives *and* good calibration on easy examples.

5.2 Qualitative Analysis

Where Environment-based Learning Improves Over ERM. In Fig. 11, we show the images with the largest gaps in NLL on the `car` task among two groups: hard positives and hard negatives

(in NOOCH-CE). The hard positive and hard negative where GDRO most overperformed (top row) are both images we might expect, where the object label and context do not match: a living room with a tiny toy car on the shelf, and a normal street scene that happens to not have any cars. See App. D for analysis on where GDRO underperforms ERM.

Figure 8: Test set examples where GDRO most improves over ERM (L) hard positives and (R) hard negatives from the NOOCH-CE `car` category. Titles show NLL for each method on that image.

Contrasting NOOCH-CE and NOOCH-Gist. In Fig. 9, we use examples of hard positives from the `cow` task to contrast the CE and gist criteria. On the left, we show a hard positive from NOOCH-CE, with cows standing on pavement: this is a hard positive in NOOCH-CE since there is no grass; GDRO and IRM outperform ERM on this image. However, it is *not* a hard positive in NOOCH-Gist, since cows are the central focus of the image. On the right, we show a hard positive from NOOCH-Gist, where the image focuses on a giraffe, but there are several white cows in the background standing on a grassy field. This is *not* a hard positive in NOOCH-CE, due to the large field, but it *is* a hard positive in NOOCH-Gist, since the giraffe is the focus of the image. ERM outperforms GDRO and IRM on this image — the environment-based objectives do not encourage as strong performance where the context (grass) and object (cow) align.

6 Discussion

Figure 9: Examples of hard positives from NOOCH-CE (L) and NOOCH-Gist (R) for the `cow` task. Titles show NLL for several methods on that image.

Limitations & Impacts. The idea of computationally identifying OOC examples is necessarily limited: by their nature, OOC examples are exceptions to rules, and there will always be OOC examples which can only be discovered through qualitative examination. Further, the task chosen for our benchmark was chosen for simplicity of analysis in a research context rather than real world significance — outputting segmentations or many-way classification may be more applicable to most applications. We hope that the impact of our work is to enable better evaluation of model performance on atypical or under-represented examples, both through usage of these challenge sets and ones inspired by the ideas presented here. However, the usage of benchmarks can have downsides where standardized performance metrics are prioritized over fundamental advances in modelling which are not captured in those metrics.

Conclusion & Looking Forward. Here, we hope to accent several aspects of this work. Firstly, we think that the notion of gist shift is an important direction for OOC research — we found it to be a very challenging problem with no one method standing out above the others. Fundamentally, a model which maintains performance even when the gist of the scene is unusual will be solving one of the hardest aspects of the OOC problem. Another is the relationship between calibration and robustness: both are properties what we generally desire of "trustworthy" models, but our results show that these two goals are not always practically aligned.

Finally, we would like to highlight the idea of using annotations to automatically discover OOC examples. This not only enables the scalable creation of challenge sets, but it allows for new hypotheses about the type of context shifts that models struggle on. We hope that methods of this type can be applied on other annotated datasets to create new challenge sets for probing OOC performance.

References

[1] R. Adragna, E. Creager, D. Madras, and R. Zemel. Fairness and robustness in invariant learning: A case study in toxicity classification. *arXiv preprint arXiv:2011.06485*, 2020.

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

[3] B. Aubin, A. Słowik, M. Arjovsky, L. Bottou, and D. Lopez-Paz. Linear unit-tests for invariance discovery. *arXiv preprint arXiv:2102.10867*, 2021.

[4] A. Barbu, D. Mayo, J. Alverio, W. Luo, C. Wang, D. Gutfreund, J. Tenenbaum, and B. Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32:9453–9463, 2019.

[5] I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive psychology*, 14(2):143–177, 1982.

[6] A. Bobick and C. Pinhanez. Using approximate models as source of contextual information for vision processing. In *Proc. of the ICCV*, volume 95, pages 13–21. Citeseer, 1995.

[7] H. Caesar, J. Uijlings, and V. Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018.

[8] M. J. Choi, A. Torralba, and A. S. Willsky. A tree-based context model for object recognition. *IEEE transactions on pattern analysis and machine intelligence*, 34(2):240–252, 2011.

[9] M. J. Choi, A. Torralba, and A. S. Willsky. Context models and out-of-context objects. *Pattern Recognition Letters*, 33(7):853–862, 2012.

[10] M. M. Chun. Contextual cueing of visual attention. *Trends in cognitive sciences*, 4(5):170–178, 2000.

[11] E. Creager, J.-H. Jacobsen, and R. Zemel. Environment inference for invariant learning. In *ICML Workshop on Uncertainty and Robustness*, 2020.

[12] N. Dagaev, B. D. Roads, X. Luo, D. N. Barry, K. R. Patil, and B. C. Love. A too-good-to-be-true prior to reduce shortcut reliance. *arXiv preprint arXiv:2102.06406*, 2021.

[13] J. Duchi and H. Namkoong. Variance-based regularization with convex objectives. *arXiv preprint arXiv:1610.02581*, 2016.

[14] J. C. Duchi, P. W. Glynn, and H. Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 2021.

[15] D. A. Forsyth, J. Malik, M. M. Fleck, H. Greenspan, T. Leung, S. Belongie, C. Carson, and C. Bregler. Finding pictures of objects in large collections of images. In *International workshop on object representation in computer vision*, pages 335–360. Springer, 1996.

[16] M. Gardner, Y. Artzi, V. Basmov, J. Berant, B. Bogin, S. Chen, P. Dasigi, D. Dua, Y. Elazar, A. Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1307–1323, 2020.

[17] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[18] 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, 2020.

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

[20] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[21] R. Hamon, H. Junklewitz, and I. Sanchez. Robustness and explainability of artificial intelligence. *Publications Office of the European Union*, 2020.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[23] C. Heinze-Deml and N. Meinshausen. Conditional variance penalties and domain shift robustness. *Machine Learning*, 110(2):303–348, 2021.

[24] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *European conference on computer vision*, pages 30–43. Springer, 2008.

[25] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[26] W. Hu, G. Niu, I. Sato, and M. Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR, 2018.

[27] P. Isabelle, C. Cherry, and G. Foster. A challenge set approach to evaluating machine translation. *arXiv preprint arXiv:1704.07431*, 2017.

[28] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019.

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

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

[31] A. M. Larson, T. E. Freeman, R. V. Ringer, and L. C. Loschky. The spatiotemporal dynamics of scene gist recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 40(2):471, 2014.

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[33] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[34] C. Lovering, R. Jha, T. Linzen, and E. Pavlick. Predicting inductive biases of pre-trained models, 2021.

[35] C. Michaelis, B. Mitzkus, R. Geirhos, E. Rusak, O. Bringmann, A. S. Ecker, M. Bethge, and W. Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019.

[36] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. H. Torr, and P. K. Dokania. Calibrating deep neural networks using focal loss. *arXiv preprint arXiv:2002.09437*, 2020.

[37] K. Murphy, A. Torralba, W. Freeman, et al. Using the forest to see the trees: a graphical model relating features, objects and scenes. *Advances in neural information processing systems*, 16: 1499–1506, 2003.

[38] A. Naik, A. Ravichander, N. Sadeh, C. Rose, and G. Neubig. Stress test evaluation for natural language inference. *arXiv preprint arXiv:1806.00692*, 2018.

[39] A. Noack, I. Ahern, D. Dou, and B. Li. An empirical study on the relation between network interpretability and adversarial robustness. *SN Computer Science*, 2(1):1–13, 2021.

[40] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pages 151–159, 2020.

[41] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

[42] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[43] J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016.

[44] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[45] R. T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of banking & finance*, 26(7):1443–1471, 2002.

[46] A. Rosenfeld, R. Zemel, and J. K. Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

[47] A. Ross and F. Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[48] D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: heterogeneous data meets causality. *arXiv preprint arXiv:1801.06229*, 2018.

[49] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[50] 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. *arXiv preprint arXiv:1911.08731*, 2019.

[51] S. Santurkar, D. Tsipras, and A. Madry. Breeds: Benchmarks for subpopulation shift. *arXiv preprint arXiv:2008.04859*, 2020.

[52] H. Shah, K. Tamuly, A. Raghunathan, P. Jain, and P. Netrapalli. The pitfalls of simplicity bias in neural networks. *arXiv preprint arXiv:2006.07710*, 2020.

[53] R. Shetty, B. Schiele, and M. Fritz. Not using the car to see the sidewalk–quantifying and controlling the effects of context in classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8218–8226, 2019.

[54] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[55] M. Srivastava, T. Hashimoto, and P. Liang. Robustness to spurious correlations via human annotations. In *International Conference on Machine Learning*, pages 9109–9119. PMLR, 2020.

[56] M. A. Tahir, J. Kittler, and F. Yan. Inverse random under sampling for class imbalance problem and its application to multi-label classification. *Pattern Recognition*, 45(10):3738–3750, 2012.

[57] A. Torralba. Contextual priming for object detection. *International journal of computer vision*, 53(2):169–191, 2003.

[58] E. Triantafillou, H. Larochelle, R. Zemel, and V. Dumoulin. Learning a universal template for few-shot dataset generalization. *arXiv preprint arXiv:2105.07029*, 2021.

[59] J. K. Winkler, C. Fink, F. Toberer, A. Enk, T. Deinlein, R. Hofmann-Wellenhof, L. Thomas, A. Lallas, A. Blum, W. Stolz, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA dermatology*, 155(10):1135–1141, 2019.

[60] K. Wu, E. Wu, and G. Kreiman. Learning scene gist with convolutional neural networks to improve object recognition. In *2018 52nd Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2018.

[61] K. Xiao, L. Engstrom, A. Ilyas, and A. Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

[62] S. M. Xie, A. Kumar, R. Jones, F. Khani, T. Ma, and P. Liang. In-n-out: Pre-training and self-training using auxiliary information for out-of-distribution robustness. *arXiv preprint arXiv:2012.04550*, 2020.

[63] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 17–24. IEEE, 2010.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We discuss how our work complements existing benchmarks; see the Supplemental Material (App. A) for a continuation of discussion around limitations of our evaluations.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Sec. 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] These are included in the supplemental material: App. B.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] These are included in the supplemental material: App. B and C.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] These are reported in the supplemental material (C) where not shown in the main body.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] We were not actively tracking amount of compute throughout the process; we used an internal cluster and discuss some of this information in the supplemental material (App. C).
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] COCO and COCO-Stuff, ResNets
   (b) Did you mention the license of the assets? [Yes] This is included in the supplementary material, App. B, C.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] There will be code and data splits included in the supplemental material or as a URL.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] The data is already publicly available for research use.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] The data is already publicly available for research use.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]