FOCUS: FAMILIAR OBJECTS IN COMMON AND UNCOMMON SETTINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard training datasets for deep learning often contain objects in common settings (e.g., *"a horse on grass"* or *"a ship in water"*) since they are usually collected by randomly scraping the web. Uncommon and rare settings (e.g., *"a plane on water", "a car in snowy weather"*) are thus severely under-represented in the training data. This can lead to an undesirable bias in model predictions towards common settings and create a false sense of accuracy. In this paper, we introduce FOCUS (**F**amiliar **O**bjects in **C**ommon and **U**ncommon **S**ettings), a dataset for stress-testing the generalization power of deep image classifiers. By leveraging the power of modern search engines, we deliberately gather data containing objects in common *and* uncommon settings in a wide range of locations, weather conditions, and time of day. We present a detailed analysis of the performance of various popular image classifiers on our dataset and demonstrate a clear drop in performance when classifying images in uncommon settings. By analyzing deep features of these models, we show that such errors can be due to the use of spurious features in model predictions. We believe that our dataset will aid researchers in understanding the inability of deep models to generalize well to uncommon settings and drive future work on improving their distributional robustness.

(a) *deer at night* (b) *plane in snowy weather* (c) *ship indoors*

(d) *bird at night* (e) *car in rain* (f) *frog in snow*

Figure 1: Some uncommon images in the FOCUS dataset. The images in the first column depict uncommon *time of day*, those in the second column depict uncommon *weather*, and the ones in the third column depict uncommon *locations*.

1 INTRODUCTION

Since the remarkable success of AlexNet (Krizhevsky et al., 2012) in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2015), deep learning models have

been used in a variety of applications ranging from robotics and self-driving cars to stock trading and computational biology. Undoubtedly, large scale datasets such as ImageNet (Deng et al., 2009) deserve much of the credit for the success of deep learning as they allow training robust, large-scale models with good generalization behavior (especially in the absence of test time distribution shifts).

Although there have been a number of recent innovations in terms of deep architecture designs (Ioffe & Szegedy (2015), He et al. (2016), Vaswani et al. (2017)), non-convex optimization solvers (Kingma & Ba (2015)), etc., not much has changed in terms of how datasets are collected. To this day, the Internet, thanks to its ease of use and availability of huge amounts of data, remains to be the predominant source for building a dataset where typically, the data is mined through a search engine. However, this means that the data distribution in the acquired dataset is subject to biases in the search results. If the search queries are not constructed carefully, the resulting dataset could potentially suffer from various types of biases. Consequently, models trained on these datasets may have poor generalization and may not perform as well as one would expect.

Natural images of objects are at the heart of many datasets for computer vision tasks. These images often capture an object of interest in some environment. For our purposes, the environment in an image includes all the contextual information surrounding the object in the image. Evidently, objects do not occur independently of their environments. In other words, objects are more likely to be found in some environments than in others (we call these *common settings*). For example, ships are often on water; cars are usually on streets; birds are usually on trees, etc. Search engines are more likely to return images with objects in their common settings when queried for an object alone, i.e., without any additional qualifiers (e.g., just *"deer"* or *"frog"*). As a result, objects in *uncommon settings* are severely under-represented in many of the popular datasets in use today. Therefore, evaluating a classifier's performance on these datasets can be unreliable in novel environments.

Training deep models on datasets containing mostly objects in common settings can create biases and security risks since models may rely heavily on *spurious* visual attributes (Geirhos et al., 2020) in their predictions and thus may suffer from a severe performance degradation in test samples with uncommon settings. Arjovsky et al. (2019) pose a simple thought experiment that demonstrates this issue: consider a binary classification between cows and camels. Since the training dataset contains mostly images in common settings (i.e., cows in pastures and camels in deserts), trained models perform poorly in classification of images in uncommon settings (e.g., cows on sandy beaches.)

To address this issue, we introduce a new dataset containing images both in common and uncommon settings called *FOCUS* (**F**amiliar **O**bjects in **C**ommon and **U**ncommon **S**ettings). Our key idea is that modern search engines often return many relevant results even for qualified queries of objects in uncommon settings. For example, searching "bird indoors" still returns a few relevant images even though this is an uncommon setting. Building on this idea, we collect images of objects in various common and uncommon environments explicitly. Our FOCUS dataset has around 24K images of ten objects along with annotations for different aspects of the environment in the images including a wide range of locations, weather conditions, and time of day. Depending on the class, we further annotate these environmental settings as *common* or *uncommon*.

Using the FOCUS dataset, we assess the performance of some popular deep learning models with high standard accuracy on ImageNet, namely ResNet50 (He et al., 2016), Wide-ResNet50-2 (Zagoruyko & Komodakis, 2016), MobileNet-v3-large (Howard et al., 2019), EfficientNet-b4 (Tan & Le, 2019), and EfficientNet-b7 (Tan & Le, 2019) on uncommon settings. We observe that all of these popular models show significant drop in accuracy when tested on objects in uncommon settings with EfficientNet-b4 doing the best in terms of generalization and Wide-ResNet50-2 doing the worst. We find that generalizing to uncommon time is easier than generalizing to uncommon weather or locations. We also analyze the deep features (neurons in the penultimate layer) of ResNet50 to find that the most likely reason for this drop is that the model relies on spurious features in its predictions, mounting additional evidence to this undesirable behavior of deep neural networks.

To the best of our knowledge, FOCUS is the first large-scale dataset of *natural* images with explicit environmental annotations such as locations, weather conditions, and time of day for *both* common and uncommon settings. As we demonstrate in this work, FOCUS can help stress-test deep models and evaluate their generalization power to uncommon settings. We believe richly annotated datasets such as FOCUS can pave the way to develop models that not only have high accuracy in common settings but are reliable in rare and uncommon settings as well.

2 RELATED WORK

Realistic corruptions such as blur, noise, etc., can occur in the real world, for instance, due to camera shake, low light, etc. Thus, training on high-quality clean images may cause classifiers to perform poorly on corrupted images. Hendrycks & Dietterich (2019) propose ImageNet-C, a dataset of 15 types of *artifically* corrupted images to systematically study robustness of deep learning models against (synthetic) corruptions. Hendrycks et al. (2021a) propose Real Blurry Images, a dataset of 1000 real world blurry images. Ever since when deep neural networks surpassed human level performance on ImageNet (He et al., 2015), there has been an increasing need for more challenging datasets. Recht et al. (2019) carefully replicate the methodology through which ImageNet was originally built to create a new test set for ImageNet. They find a 11-14% drop in accuracy on this new test set. However, Engstrom et al. (2020) have found statistical biases in the replicated test set and found the accuracy drop to be about 3.6% after correcting for the biases. ImageNet-A (Hendrycks et al., 2021b) is a dataset of natural, adversarial images which yield drastically low performance on classifiers trained on ImageNet.

Hendrycks et al. (2021a) propose three datasets, namely, ImageNet-Renditions, DeepFashion Remixed, StreetView StoreFronts that are designed to test the generalization ability of classifiers to unseen rendition styles, camera view points, and geography, respectively. Beery et al. (2018) introduce a dataset of camera trap images. Since, the traps are fixed, the backgrounds in these images are also more or less fixed. This gives researchers an ideal opportunity to test the generalization ability of neural networks to unseen locations in animal classification tasks. Geirhos et al. (2020) have observed that deep learning models often fail to perform as well in the real world as they do in standard benchmarks because they rely on *shortcuts*. Many works have sought to find these shortcuts and/or circumvent them. Singla et al. (2021) propose a method for identifying the visual attributes that cause classification failures using the features of an adversarially robust model. Xiao et al. (2020) propose the Backgrounds Challenge to evaluate how robust models are to (synthetic) changes in backgrounds. Wong et al. (2021) show that training sparse linear models with deep features as inputs results in improved debuggability of neural networks. In a similar vein, 3DB (**?**) uses photorealistic simulations to test and debug computer vision models.

3 FOCUS: A DATASET WITH COMMON AND UNCOMMON SETTINGS

3.1 BUILDING FOCUS

We choose to work with the same 10 object classes in CIFAR-10 (Krizhevsky et al., 2014). We use the time of day, weather and the locations depicted in an image to characterize environment in it. Using the capability of modern search engines in returning relevant results even for uncommon qualified queries of objects (e.g., *"frog indoors"*), we collect images of objects in various common and uncommon environments explicitly. Concretely, we query the Microsoft Bing Image Search API[1] with statements of the form `<object> <preposition> <attribute>` (e.g., *"ship on grass"*). This ensures that we have sufficient number of uncommon images and alleviates the issue of bias towards common settings. We also use synonyms of the object categories and the attributes to increase the number of samples we collect. We only query for images which have a license that permits sharing allowing us to release the FOCUS dataset for the research community.

We collect a total of around 37K images using the above procedure. But the search results are not always accurate; a significant fraction of them do not have the relevant object or if they do, the object is not in the environment mentioned in the search query. In addition, because we query images based on only one attribute, we do not have any information about the other attributes in the images. For instance, we do not know the time of day or the locations in a search result for *"car in rain"*.

We conduct an Amazon Mechanical Turk study both to improve the accuracy of annotations derived from the search queries and to collect missing annotations. Images are shown to workers in a series of Human Intelligence Tasks (HITs), and they are asked to annotate the image with the appropriate choice for the different attributes. See the appendix C for more details about the design of our HITs.

[1]Google does not provide a publicly accessible API for its image search.

3.2 THE DATA IN FOCUS

The FOCUS dataset is a collection of around 24K images, each annotated with the time of day, the weather condition and the locations in the image. Concretely, our dataset is as follows:

$$\{(\mathbf{x}_i, y_i, t_i, w_i, l_i)\}_{i=1}^{n}$$

where

$\mathbf{x}_i$ is the image

y_i is the object label $\in \{$*truck, car, plane, ship, cat, dog, horse, deer, frog, bird*$\}$

t_i is the time of day $\in \{$*day, night, none*$\}$

w_i is the weather $\in \{$*cloudy, foggy, partly cloudy, raining, snowing, sunny, none*$\}$

l_i are the locations $\subset \{$*forest, grass, indoors, rocks, sand, street, snow, water, none*$\}$

The rationale behind our choices for different attributes is as follows:

1. **Time of day:** Most images in standard datasets are captured during the day, when the objects are well lit. In contrast, nighttime images often lack a lot of details and are corrupted by high levels of noise.
2. **Weather:** Our choices of weather are fairly comprehensive and include the raining, snowing and foggy conditions which often produce natural corruptions in images.
3. **Locations:** We choose a wide range of locations with a healthy mix between common and uncommon *(object, location)* pairs. Since images are often likely to include a combination of locations, we let the locations attribute of an image to be a *subset* of the above set instead of being exactly one element out of it.

"none" is assigned to an attribute if its ambiguous or impossible to determine from the image. In addition, *"none"* is also assigned to l_i of an image if none of the considered locations is in the image. Table 1 summarizes the number of FOCUS samples for each object in various environments.

3.3 COMMON VS. UNCOMMON SETTINGS

We consider two sources of uncommon *(object, environment)* pairs:

1. The pair is uncommon in the real world (e.g., *"ship on grass"*). On the Internet, searching for *"ship"* alone is extremely unlikely to return any images of a ship on grass in the top results. In other words, the rarity of a pair in the real world is reflected in the dataset.
2. The pair is uncommon due to the choice of labels and queries used to construct a particular benchmark. E.g., consider the *"plane"* class. ImageNet has two labels —*"warplane, military plane", "airliner"* — corresponding to an airplane. Neither of these are planes that are usually found on water making *("plane in water")* an uncommon, if not a non-existent pair in ImageNet. Seaplanes, however, are not that uncommon in the real world.

We declare an *(object, attribute)* uncommon if the number of samples corresponding to that pair is low (case 1 above). Additionally, we also declare the *("plane", "water")* pair as uncommon (case 2 above). Our final choices for uncommon settings are highlighted in orange, in table 1. Figure 1 shows some uncommon images from our dataset.

Obviously categorizing environments into common and uncommon settings for various objects can be subjective. We used a combination of the two criteria described above for this categorization and we acknowledge that this is by no means the one true way. We facilitate other studies opting to do it in other ways by providing all annotations for the collected samples our dataset, FOCUS.

4 EVALUATING DEEP MODELS ON FOCUS

It is crucial to have reliable machine learning models even in rare and uncommon settings especially when they are deployed in safety-critical applications. As an example, consider a self-driving car

		Truck	Car	Plane	Ship	Cat	Dog	Horse	Deer	Frog	Bird
Time of Day	Day	1217	2798	1740	1731	2109	2766	2260	1838	1096	2526
	Night	66	304	136	167	84	66	56	57	201	58
	None	32	266	99	33	997	566	88	44	370	131
Weather	Cloudy	189	400	314	377	79	195	248	175	18	187
	Foggy	22	115	47	106	6	44	90	55	3	48
	Partly Cloudy	161	363	340	307	98	184	282	159	50	170
	Raining	16	161	27	10	15	11	14	3	3	21
	Snowing	13	66	3	5	29	49	49	53	0	39
	Sunny	662	1252	850	837	768	1266	1183	880	503	1297
	None	252	1011	394	289	2195	1649	538	614	1090	953
Locations	Forest	185	473	179	75	90	262	519	835	117	434
	Grass	355	738	397	97	532	874	1171	1254	293	578
	Indoors	37	311	119	3	1451	897	76	12	37	41
	Rocks	57	119	52	67	130	114	83	86	265	192
	Sand	305	596	247	181	137	451	625	186	162	315
	Street	702	1811	331	96	376	386	235	74	32	124
	Snow	113	308	129	157	140	357	222	306	8	192
	Water	65	225	354	1728	102	433	331	135	404	611
	None	90	188	673	64	524	342	245	93	555	804

Table 1: A frequency breakdown of the various categories and attributes in the FOCUS dataset. Uncommon settings are highlighted in orange.

that infers various attributes about its surroundings using deep learning models. The deployed models may be accurate in 99.99% of cases that occur in common settings (e.g., pedestrian crossing on a sunny day). However, given the vast complexity of the real world, uncommon and corner cases, although rare, are still possible (e.g., a heavily snow covered car cutting in). If a model is not reliable in those uncommon settings, it could make a grave error resulting in loss of life and/or property.

In spirit of the above mentioned, we stress test the generalization power of various deep learning models to uncommon settings using the FOCUS dataset. Specifically, we are considering models that are trained using images close to the mode of the *(object, environment)* distribution (i.e., *common images*) and evaluating them on images that fall more on the tail of the *(object, environment)* distribution (i.e., *uncommon images*).

4.1 Experimental setup

Model architectures. We select some of the most popular deep learning models that have high test accuracy on ImageNet, namely ResNet50 (He et al., 2016), Wide-ResNet50-2 (Zagoruyko & Komodakis, 2016), MobileNet-v3-large (Howard et al., 2019), EfficientNet-b4 (Tan & Le, 2019), and EfficientNet-b7 (Tan & Le, 2019). For the first three models, we use the pretrained weights provided by PyTorch and for both variations of EfficientNet, we obtain the weights from this GitHub repository: `https://github.com/lukemelas/EfficientNet-PyTorch`.

Total (23902)							
P_0 (16692)	P_1 (6530)			P_2 (647)			P_3 (35)
	$P^{(t)}$ (752)	$P^{(w)}$ (597)	$P^{(l)}$ (5181)	$P^{(t,w)}$ (51)	$P^{(w,l)}$ (440)	$P^{(t,l)}$ (156)	$P^{(t,w,l)}$ (35)

Table 2: Sizes of different partitions in FOCUS. P_i is the set of images with i uncommon attributes and P^A, $A \subseteq \{t, w, l\}$ is the set of images where the attributes in A are uncommon. Note that P_0 constitutes *common* images.

Evaluation metrics. Since the models we evaluate are pretrained on ImageNet, they output 1,000 probabilities (recall, ImageNet has 1,000 classes). On the other hand, our dataset has only 10 object categories. We resolve this apparent mismatch by first constructing a mapping (denoted by M) between the 1000 labels in ImageNet and those in our dataset. Concretely, a label l_I in ImageNet is assigned to a label l_F in our dataset if l_F is a semantic superclass of l_I. For example, all the different dog breeds in ImageNet are mapped to the *dog* label in FOCUS. Decidedly, some labels in ImageNet do not have any corresponding labels in our dataset (e.g., *"analog clock", "carton" etc.*). As our dataset has no images from these labels, we declare a misclassification whenever the network predicts a label that is not in the domain of M. We say a prediction is correct if the ImageNet label with the highest logit was assigned to the ground truth label in FOCUS. That is, for a sample $(\mathbf{x}_i, y_i, t_i, w_i, l_i)$ from the FOCUS dataset, let $g(\mathbf{x}_i)$ be the ImageNet label predicted by a trained network. Then, we have:

$$\text{correct prediction} \iff f(\mathbf{x}_i) \coloneqq M(g(\mathbf{x}_i)) = y_i.$$

To facilitate the evaluation of the effect of uncommon attributes, we first partition the dataset based on the number of uncommon attributes in images: P_i is a subset of FOCUS samples with i uncommon attributes for $i = 0, 1, 2, 3$. Note that P_0 denote *common* samples while $\bigcup_{i \geq 1} P_i$ denotes *uncommon* samples that have at least one uncommon attribute.

We further subdivide P_i into P^A, $A \subseteq \{t, w, l\}$, $|A| = i$ where the attributes in A are uncommon (for instance, $P^{(w,l)}$ is the set of all images with two uncommon attributes: weather and location). Table 2 shows the sizes of the different partitions in our dataset.

We then evaluate the classification accuracy of different models in different partitions (referred to as $Acc(P)$). In an attempt to measure the effect of a single attribute on the accuracy of a model f, we define the following generalization gap with respect to an attribute a:

$$G_a = \frac{|\{\mathbf{x}_i \in \mathrm{C}(a) \mid f(\mathbf{x}_i) = y_i\}|}{|\mathrm{C}(a)|} - \frac{|\{\mathbf{x}_i \in \mathrm{UC}(a) \mid f(\mathbf{x}_i) = y_i\}|}{|\mathrm{UC}(a)|}, \tag{1}$$

where $\mathrm{C}(a)$ and $\mathrm{UC}(a)$ are the subsets of images in which attribute a is common and uncommon, respectively. Succinctly, G_a is the difference in the classification accuracy between images with a common choice for a and those with an uncommon choice for the same. The larger the G_a, the worse the generalization performance of the model on uncommon choices for a.

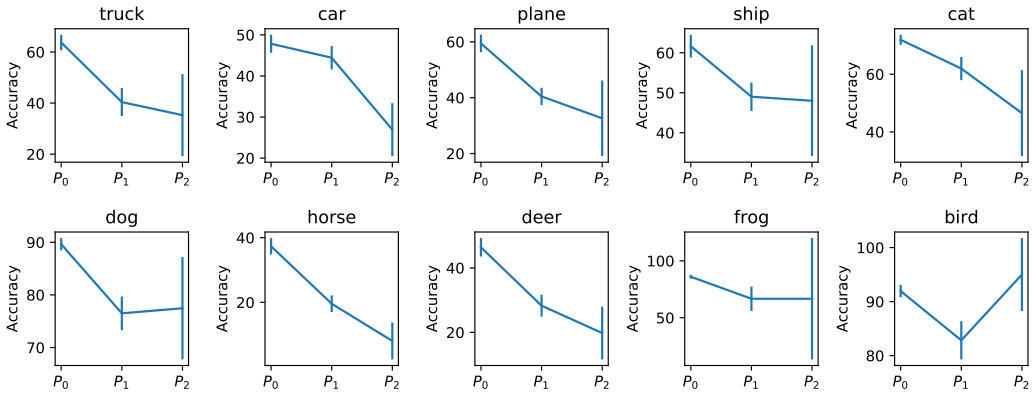


Figure 2: Classwise top-1 accuracy for ResNet50. The large error bars at P_2 are due to insufficient number of samples in this partition. Similar plots for other models are in appendix D.

4.2 Results

Figure 3 shows the accuracy of different model architectures on different partitions of the FOCUS dataset. We observe that for all the models, the accuracy falls as the number of uncommon attributes

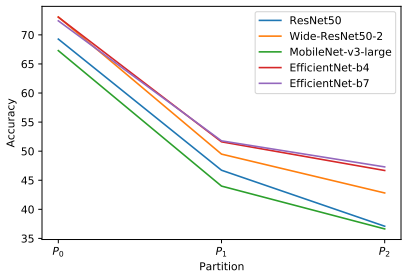


Figure 3: Top-1 classification accuracy for different models on the FOCUS dataset as a function of the partitions P_i

Model	G_t	G_w	G_l
ResNet50	8%	16%	19%
Wide-ResNet50-2	8%	18%	18%
MobileNet-v3-large	9%	14%	19%
EfficientNet-b4	**5%**	**12%**	17%
EfficientNet-b7	8%	13%	**15%**

Table 3: Generalization gap (as in Equation 1) per attribute for various models. The best gap on each attribute is in boldface.

Model	P_0	$P^{(t)}$	$P^{(w)}$	$P^{(l)}$	$P^{(t,w)}$	$P^{(w,l)}$	$P^{(t,l)}$
ResNet50	69.22	58.57	48.33	44.81	39.22	37.33	35.26
Wide-ResNet50-2	**73.04**	60.42	46.82	48.19	41.18	41.86	46.15
MobileNet-v3-large	67.26	54.98	46.82	42.07	27.45	38.46	35.26
EfficientNet-b4	73.01	**64.01**	**54.68**	49.44	**47.06**	45.70	**50.00**
EfficientNet-b7	72.39	59.63	49.67	**50.89**	39.22	**48.42**	47.44

Table 4: Top-1 accuracies of different models on various partitions of the dataset. The best accuracy on each partition is in boldface. The models perform the best on the first column corresponding to the common images while the accuracy decreases as the number of uncommon attributes increases.

increases[2]. The overall generalization gap between common and uncommon images (i.e., $A(P_0) - A(\bigcup_{i\geq 1} P_i)$) is the highest (i.e., the worst generalization) for Wide-ResNet50-2 at 24.2% and the lowest (i.e., the best generalization) for EfficientNet-b7 at 21.00%.

Table 3 shows the generalization gap (as in Equation 1) per attribute for various models. We see that all the gaps are positive, clearly indicating poor generalization ability to uncommon settings. Note that G_t is smaller than both G_w and G_l for all the models. So, these models are not hurt as much by uncommon time (i.e., "night") as they are by uncommon weather or location. Additionally, we see that EfficientNet-b4 has the best generalization in uncommon *time of day* and *weather*, while EfficientNet-b7 has the best generalization in uncommon *locations*.

Table 4 shows the accuracy for various combinations of uncommon attributes. Though Wide-ResNet50-2 has the best accuracy on common images, it does not do as well as other models on uncommon images. Just like our observation from table 3, table 4 also shows EfficientNet-b4 and Efficient-b7 doing the best on different partitions.

Finally, figure 2 shows the *classwise* top-1 accuracy of ResNet50 on the partitions P_i. We observe a declining trend here as well in almost all cases. The uptick from P_1 to P_2 for the *"dog"* and *"bird"* classes could be due to their small number of P_2 samples (i.e., the dog class has 71 P_2 samples while the bird class has only 40) as their accuracies show high variances. We observe similar trends for other models as well (the plots for them are included in Appendix D).

4.3 Neural Feature Analysis of Common and Uncommon Samples

In this subsection, we take a close look at the deep features (i.e., neurons in the penultimate layer) for common and uncommon images. We also make an effort to gain insight into the aspects of uncommon images that affect the model's classification performance on them.

[2]Note that we have not included P_3 in this analysis as it has very few samples (only 35).

	truck	car	plane	ship	cat	dog	horse	deer	frog	bird
forest	70.74	77.61	76.36	76.00	81.25	82.31	70.58	65.29	64.18	72.66
grass	67.21	74.49	80.69	77.66	86.17	83.01	77.50	74.65	67.90	78.69
indoors	75.00	**91.57**	**96.38**	**100.00***	82.05	**89.26**	**94.23**	66.67	70.00	86.36
rocks	63.33	77.54	69.23	76.47	77.50	61.72	72.73	68.06	73.95	78.52
sand	74.51	79.67	74.87	79.39	82.76	77.06	73.87	63.24	72.77	82.45
street	71.49	67.90	71.00	58.70	67.48	64.88	82.43	**87.50**	62.50	68.24
snow	**80.95**	86.24	85.44	78.97	**87.78**	85.50	86.63	86.52	**100.00***	88.03
water	58.82	73.43	78.62	75.49	72.12	87.21	83.39	78.82	81.44	**88.50**

Table 5: Test accuracies of a linear classifier that predicts the presence of a location solely from the deep features of images. It does significantly better than random implying that deep features contain information about the location. For each class, the location attribute that is most easily identifiable is in boldface. * - have high statistical uncertainty because there are only 3 samples of *"ships indoors"* and 8 samples of *"frogs in snow"*.

4.4 RESNET50 USES SPURIOUS FEATURES

We hypothesize that the information about the environment (that is spurious with respect to the true object) may have been encoded in the deep features of the model, causing drops in model accuracy in uncommon settings. To test our hypothesis, we conduct an experiment where we attempt to predict the presence (or absence) of an environmental attribute from the deep features of a ResNet50 pretrained on ImageNet. Concretely, we pick an environmental attribute (say, *"water"*) and an object class (say, *"plane"*) and train a simple binary linear classifier that uses the deep features of the input images to predict if the image has the attribute in it (is there *"water"* in this image of a *"plane"*?). If we do not use images from only one class then the linear classifier could inadvertently use the information about the object in the image to exploit any imbalances in the *(object, attribute)* distribution and do well in this task. We also ensure that both the train and test sets in the experiment contain as many images with the attribute as those without it.

Table 5 shows the test accuracy of the linear classifier for various instances of this experiment, each with a different choice for the location and the object class. The accuracies are well above 50% (the accuracy of a random guess) highlighting that environmental attributes have been encoded in deep features of the model, conditioned on a particular class. In addition, we see that "snow" and "indoors" are by far the most easily identifiable locations for all object classes except "deer". Ignoring *"ships indoors"* and *"frogs in snow"* as they have very few samples (3 and 8, respectively), we see that identifying *"'indoors"* in *"plane"* images is the easiest among all object, location pairs, while identifying *"street"* in images of *"ship"* is the hardest.

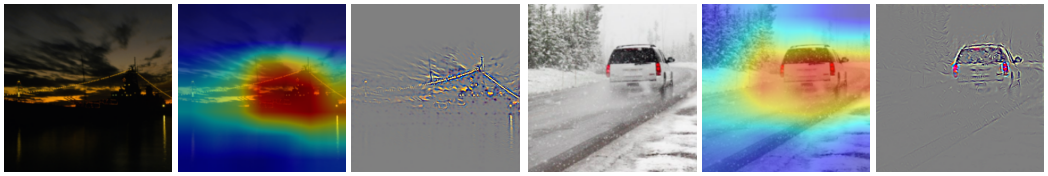
We go further in the above experiment; we randomly pick various object, location pairs and we identify the ImageNet labels for which the 2049-dimemsional (including bias) weight vector in the fully connected layer of ResNet-50 is most similar to the linear classifier of that location. This is to reveal with which labels the object is most likely to be confused with when it is in that particular location. Table 6 summarizes the results from this experiment. We see that the model has falsely associated some object labels to the location in many cases (*'hay'* being similar to *'grass'* is perhaps acceptable). This is more evidence that neural networks may rely on spurious features in their predictions, especially when there is not enough diversity in the dataset.

4.5 MODES OF FAILURE

Finally, we show the Grad-CAM localization maps for some *misclassified* uncommon images in figure 4. All visualizations were generated with the predicted class as the target class for Grad-CAM so as to highlight the parts of the images that caused their misclassification. Figures 4a, 4b and 4c show that the model was able to localize more or less correctly on the corresponding objects. However, they were misclassified because: (a) the striking lack of details (as is common in low light photographs) in the middle of the image in figure 4a makes it hard to identify the warship; (b) the model seems to be falsely correlating the presence of snow in the surrounding air with the presence of a snowplow; (c) the frog is sticking out from between tiles of the pavement/street; its hind legs are

Object class	Location	ImageNet labels
deer	forest	'suspension bridge' (0.18), 'centipede' (0.15), 'coral fungus' (0.15)
car	snow	'snowplow' (0.35), 'dogsled' (0.26), 'snowmobile' (0.26)
plane	water	'sandbar' (0.30), 'lakeside' (0.29), 'promontory' (0.27)
cat	street	'manhole cover' (0.22), 'sundial' (0.18), 'go-kart' (0.17)
dog	sand	'sandbar' (0.26), 'seashore' (0.26), 'baboon' (0.18)
truck	grass	'hay' (0.20), 'harvester' (0.20), 'rapeseed' (0.17)
frog	rocks	'marmot' (0.21), 'rock crab' (0.20), 'European fire salamander' (0.19)

Table 6: Each row shows the most similar ImageNet labels to the linear classifier for the location in the row. The numbers in the brackets are the cosine similarities between the linear classifier and the MLP weights (biases included) of the corresponding labels.



(a) *ship at night* (prediction: 'pier')

(b) *car on a snowy day* (prediction: 'snowplow')

(c) *frog on street* (prediction: 'sidewinder')

(d) *plane in water* (prediction: 'submarine')

Figure 4: Localization maps on some uncommon images misclassified by ResNet50. Each subfigure shows the original image, the Grad-CAM overlay on the image, and the Guided Grad-CAM image (in that order); all Grad-CAM and Guided Grad-CAM images use the predicted class as their target. See text for possible explanations for these misclassifications. More examples are in the appendix.

hidden and the model seems to be ignoring its front legs (most noticeable in the Guided Grad-CAM image). As a result, the model incorrectly sees a snake ('sidewinder') in the image.

On the other hand, in figure 4d, we see that the model is incorrectly zeroing in on the floats of the plane which look like a submarine breaking the surface of water from afar. We postulate that this is because of the water in the image which the model has incorrectly learned to associate with the presence of watercraft such as boats, ships, submarines etc. We acknowledge that these explanations of model failures in uncommon settings may not be complete and there may be other confounding factors for such errors.

5 Conclusion

In this work, we introduced FOCUS, a dataset that contains images both in common and uncommon settings. FOCUS has around 24K samples annotated by their various environmental attributes such as locations, weather conditions, and time of day. Using FOCUS, we evaluated the performance of several popular ImageNet classifiers such as ResNet50, Wide-ResNet50-2, MobileNet-v3-large, EfficientNet-b4 and EfficientNet-b7. These models showed a clear drop in performance when classifying images in uncommon settings. By analyzing deep features of ResNet50, we found that this accuracy drop in uncommon settings is partially due to the model's reliance on spurious features in its predictions. We believe that richly annotated datasets such as FOCUS open new directions for the development of deep models that are reliable both in common and uncommon settings.

REPRODUCIBILITY STATEMENT

We give more details (search queries, licenses, filters etc.,) about the procedure we followed to collect the images in appendix B. Appendix C describes the design of our Human Intelligence Tasks (HITs) on Amazon Mechanical Turk. In addition, as our images are licensed to be shareable, we will share the entire dataset as well as the code to access the images, annotations and reproduce the various experiments presented here, upon acceptance of the paper. For now, we include some more samples from the FOCUS dataset in Appendix A.

ETHICS STATEMENT

This paper introduces a dataset that can assess potential biases of deep models against uncommon settings. We have made the following efforts to mitigate harm in the collection of our dataset:

1. We paid the workers on Amazon Mechanical Turk 40% more than the minimum wage on average.
2. We use the same object labels as CIFAR-10 (Krizhevsky et al., 2014), none of which, to the best of our knowledge are stereotypes or slurs. In addition the environmental attributes we choose are quite generic and are not targeted against any people subgroups.
3. That said, some of the images in our dataset have identifiable humans in them even though it was not intended. However, all the images we gathered are either in the public domain or are free to be used and shared. We ensure that by using the appropriate license keywords in the Bing Image Search API. See `https://help.bing.microsoft.com/#apex/18/en-us/10006/0` for more details. We believe that this license constraint filters out the images taken or shared without the consent of the subjects.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473, 2018.

Kate Crawford and Trevor Paglen. Excavating ai: The politics of images in machine learning training sets, 2019. URL `https://excavating.ai`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Jacob Steinhardt, and Aleksander Madry. Identifying statistical bias in dataset replication. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2922–2932. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/engstrom20a.html`.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15262–15271, June 2021b.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1314–1324, 2019.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456, Lille, France, 07–09 Jul 2015. PMLR. URL `https://proceedings.mlr.press/v37/ioffe15.html`.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pp. 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. The cifar-10 dataset. *online: http://www. cs. toronto. edu/kriz/cifar. html*, 55(5), 2014.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

Sahil Singla, Besmira Nushi, Shital Shah, Ece Kamar, and Eric Horvitz. Understanding failures of deep networks via robust feature extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12853–12862, 2021.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pp. 6105–6114. PMLR, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

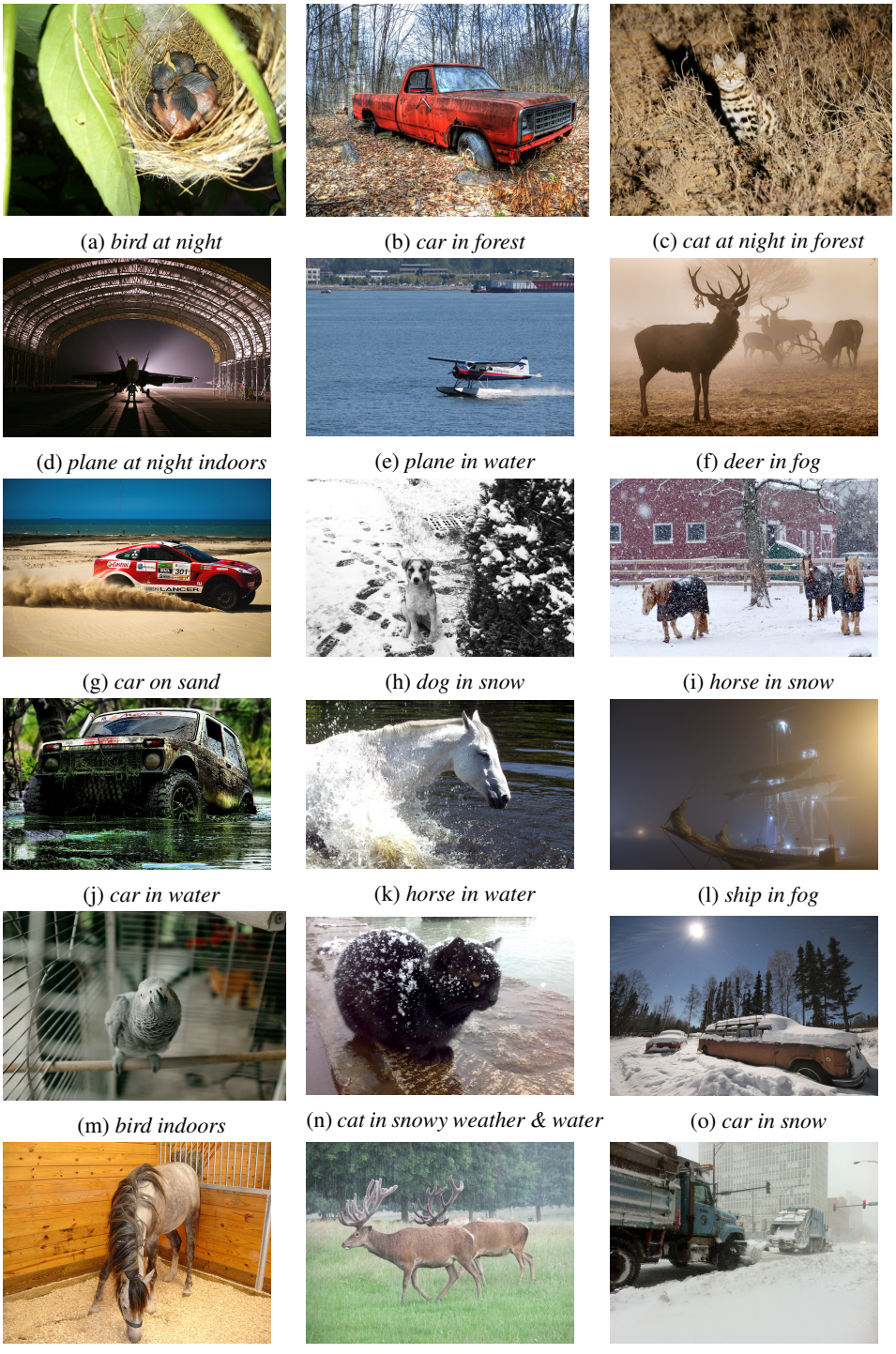
Eric Wong, Shibani Santurkar, and Aleksander Madry. Leveraging sparse linear layers for debuggable deep networks. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11205–11216. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/wong21b.html`.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *ArXiv preprint arXiv:2006.09994*, 2020.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Appendix

A VISUALIZATIONS OF SOME FOCUS IMAGES IN UNCOMMON SETTINGS



(a) *bird at night* (b) *car in forest* (c) *cat at night in forest*

(d) *plane at night indoors* (e) *plane in water* (f) *deer in fog*

(g) *car on sand* (h) *dog in snow* (i) *horse in snow*

(j) *car in water* (k) *horse in water* (l) *ship in fog*

(m) *bird indoors* (n) *cat in snowy weather & water* (o) *car in snow*

(p) *horse indoors* (q) *deer in rain* (r) *truck in snow*

Figure 5: Visualizations of some uncommon images in the FOCUS dataset.

B IMAGE SEARCH

The images for our dataset were collected using queries formed as a concatenation of an object label and one of the phrases from below:

Attribute	Phrases used in queries
"raining"	"in rain"
"foggy"	"in fog"
"snow"	"on snow"
"sand"	"in a desert", "on sand"
"forest"	"in forest"
"water"	"on water"
"night"	"at night"
"grass"	"on grass"
"street"	"on a street", "on a road"

In addition, we use some class specific queries: *"ship on ice"*, *"ship on a dock"*, *"dog on a couch"*, *"dog on a bed"*, *"dog on the floor"*, *"cat on a couch"*, *"cat on a bed"*, *"cat on the floor"*, *"horse in a stable"*, *"car in a garage"*, *"truck in a garage"*, *"plane in a hangar"*.

C HUMAN INTELLIGENCE TASKS (HITS)

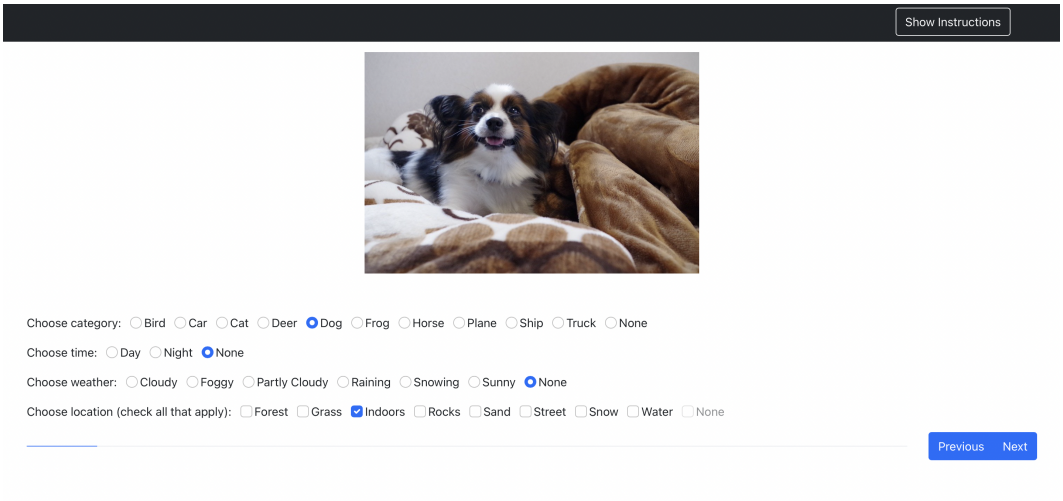


Figure 6: Our UI for annotating images in FOCUS.

To gather high quality annotations, we first vet the workers through a qualification process; workers are shown a series of 10 images (in the UI shown in figure 6) from our dataset for which the ground truth is known (these images were annotated by us manually). We qualify workers who have done well on this qualification test. Worker's annotations were checked manually in this stage instead of using a strict threshold as there is an element of subjectivity to the annotations. In the second stage, our HITs have 25 images each; 23 of which are unannotated and 2 are from the subset that were annotated by us. We use these 2 images as a way to track workers' annotation accuracy. Each image is annotated by two workers and we pick annotations of the worker who has the higher annotation accuracy on the 2 "check" images in that HIT. Workers received a base pay of $0.67 per HIT (25 images) which takes an average of around 5 minutes to annotate. A bonus of $30 was paid for completion of 100 HITs. This payment structure has created an incentive for workers to annotate a large number of images.

D CLASSWISE CLASSIFICATION ACCURACIES FOR VARIOUS MODELS

In the main text, we presented classwise classification accuracies for the ResNet-50 model (figure 2). For completeness, we present similar plots for other models used in evaluations here. We observe similar trends to the ones reported in the main text.

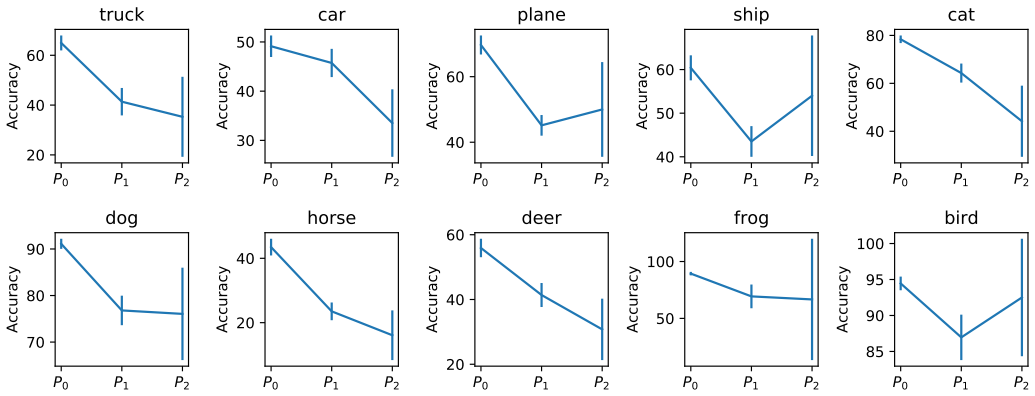


Figure 7: Classwise top-1 accuracy for Wide-ResNe50-2.

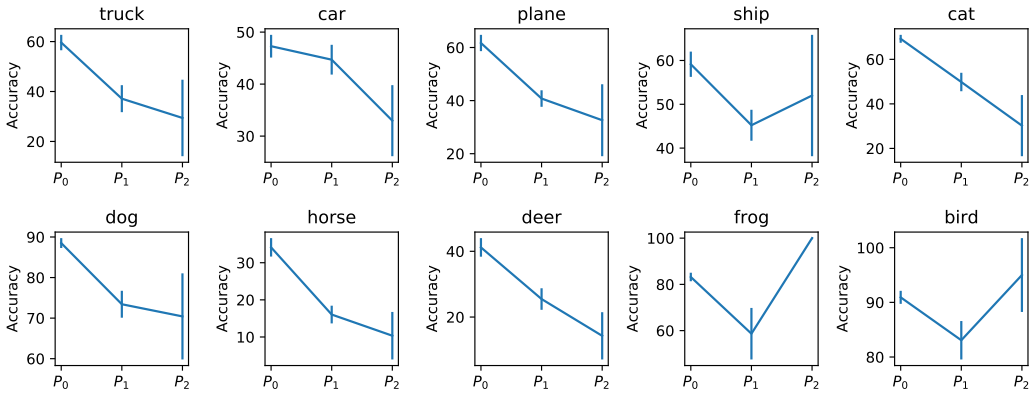


Figure 8: Classwise top-1 accuracy for MobileNet-v3-large.

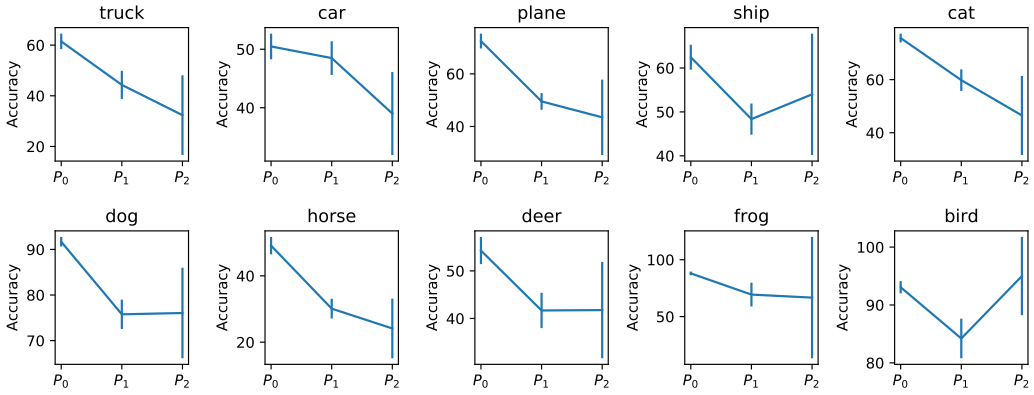


Figure 9: Classwise top-1 accuracy for EfficientNet-b4.

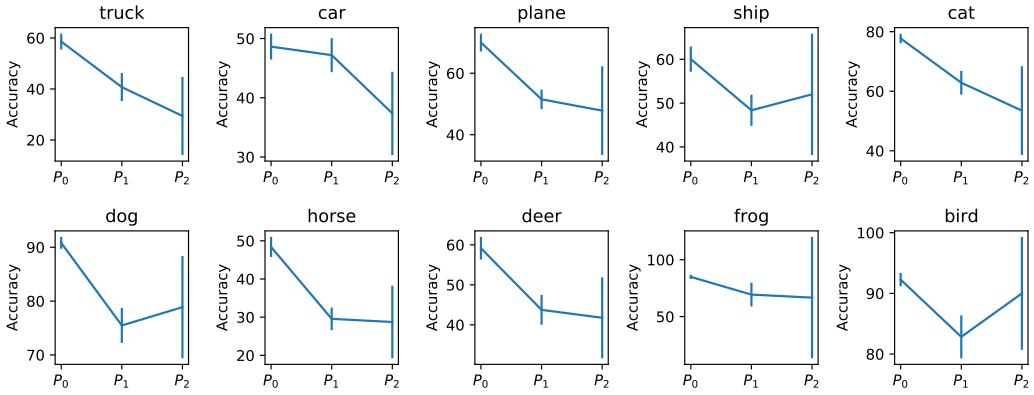
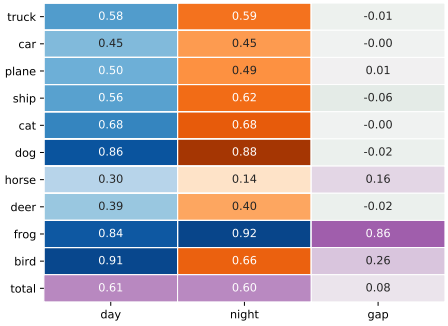


Figure 10: Classwise top-1 accuracy for EfficientNet-b7.

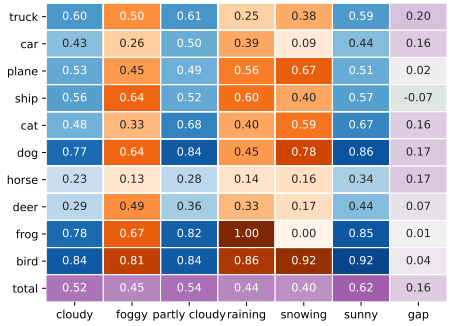
E ACCURACY AS A FUNCTION OF CLASS AND ATTRIBUTES

This section shows the accuracy of different models for each class and attribute. Uncommon attributes are highlighted in shades of orange, while common attributes are highlighted in shades of blue. The last row (except the bottom right-most value) shows the overall accuracy for the corresponding attribute. Additionally, the last column is the generalization gap with respect to the corresponding attribute. The first 10 values in the last column are class specific generalization gaps, while the last (i.e., bottom rightmost) value is the aggregate generalization gap defined in Equation 1 (and reported in table 3).

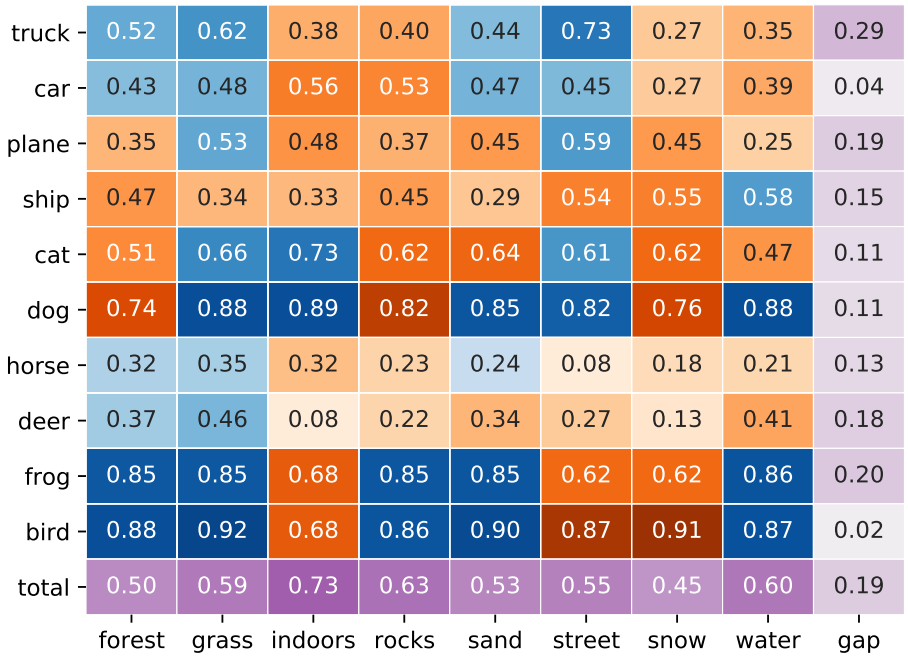
An image can have a combination of common and uncommon attributes (e.g., *"cat on street at night"* — uncommon time but common location). Such images may appear in a "blue" cell in one of the tables while in a different table they may appear in "orange". This explains why for some classes the models seem to do better at "night" than in day. Note that majority of the uncommon weather conditions: *("raining", "snowing", "foggy")* occur during the day and they are potentially decreasing the classification accuracy so much that it outweighs the drop due to "night" and leads to this apparently paradoxical result.



(a) Category vs Time of Day

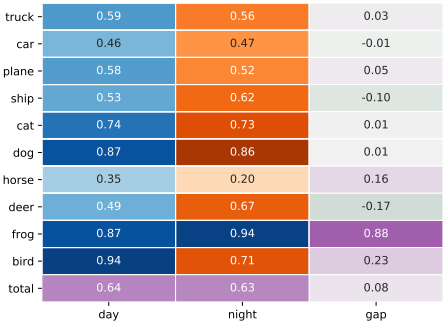


(b) Category vs Weather

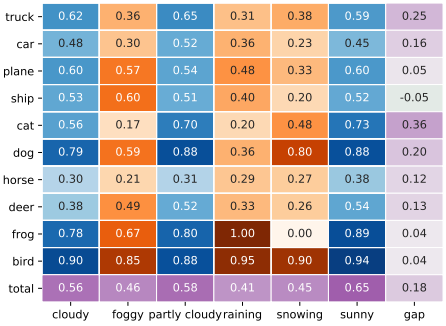


(c) Category vs Location

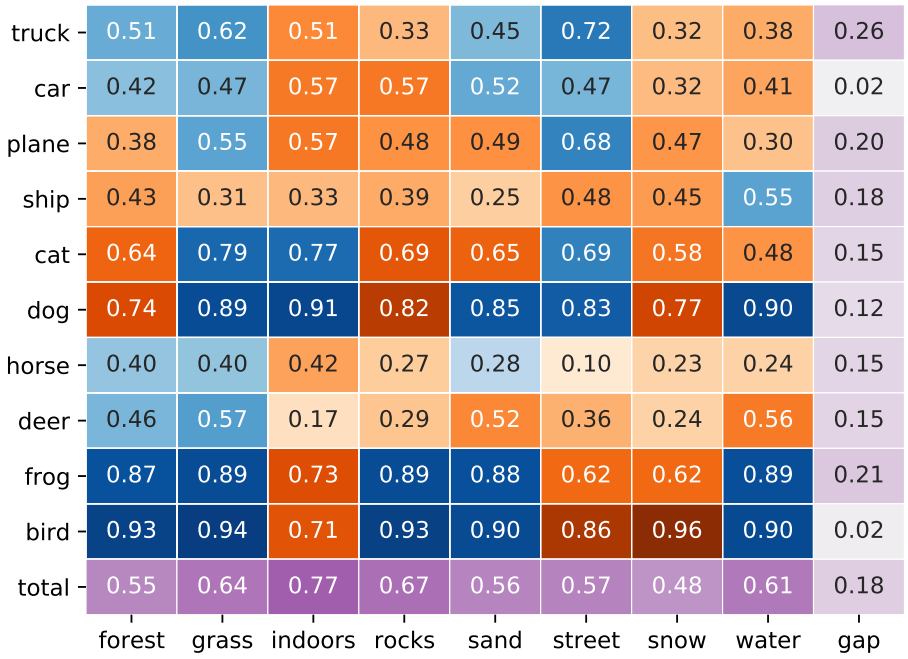
Figure 11: Accuracy of ResNet50 for all combinations of classes and attributes.



(a) Category vs Time of Day

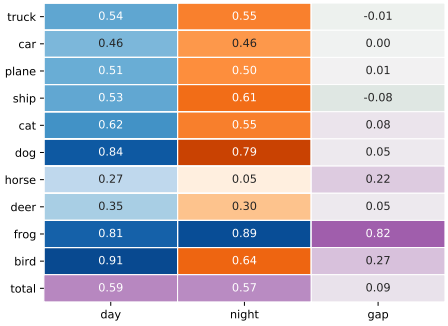


(b) Category vs Weather

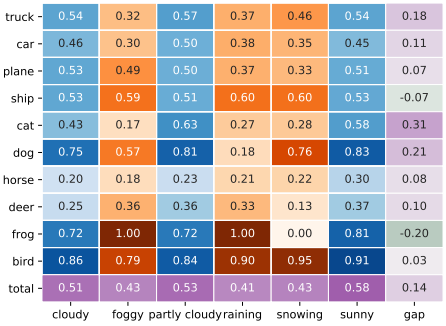


(c) Category vs Location

Figure 12: Accuracy of Wide-ResNet50-2 for all combinations of classes and attributes.



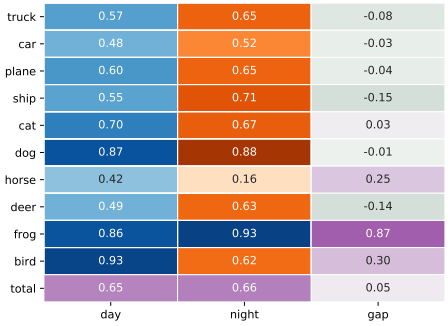
(a) Category vs Time of Day



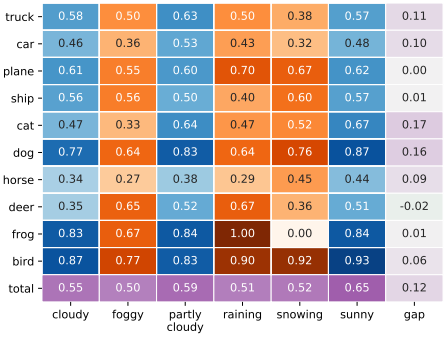
(b) Category vs Weather

(c) Category vs Location

Figure 13: Accuracy of MobileNet-v3-large for all combinations of classes and attributes.



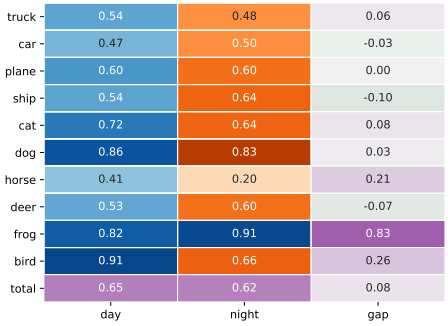
(a) Category vs Time of Day



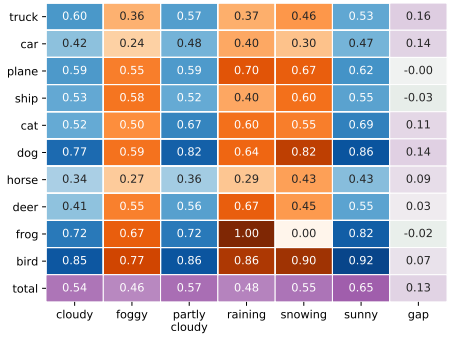
(b) Category vs Weather

(c) Category vs Location

Figure 14: Accuracy of EfficientNet-b4 for all combinations of classes and attributes.



(a) Category vs Time of Day



(b) Category vs Weather

(c) Category vs Location

Figure 15: Accuracy of EfficientNet-b7 for all combinations of classes and attributes.