Fine-Grained ImageNet Classification in the Wild

Maria Lymperaiou[1], Konstantinos Thomas[1] and Giorgos Stamou[1]

[1] *AILS Lab, School of Electrical and Computer Engineering, National Technical University of Athens*

Abstract

Image classification has been one of the most popular tasks in Deep Learning, seeing an abundance of impressive implementations each year. However, there is a lot of criticism tied to promoting complex architectures that continuously push performance metrics higher and higher. Robustness tests can uncover several vulnerabilities and biases which go unnoticed during the typical model evaluation stage. So far, model robustness under distribution shifts has mainly been examined within carefully curated datasets. Nevertheless, such approaches do not test the real response of classifiers in the wild, e.g. when uncurated web-crawled image data of corresponding classes are provided. In our work, we perform fine-grained classification on closely related categories, which are identified with the help of hierarchical knowledge. Extensive experimentation on a variety of convolutional and transformer-based architectures reveals model robustness in this novel setting. Finally, hierarchical knowledge is again employed to evaluate and explain misclassifications, providing an information-rich evaluation scheme adaptable to any classifier.

Keywords

Image Classification, Knowledge Graphs, Robustness, Explainable Evaluation

1. Introduction

ImageNet [1] has been one of the most popular image classification datasets in literature, inspiring a variety of popular model implementations for over a decade. The first significant breakthrough in ImageNet classification was marked with AlexNet [2], a convolutional neural network (CNN) for image classification that greatly outperformed its competitors. Ever since various CNN-based implementations continued pushing accuracy scores even higher [3].

The local nature of convolutional filters that cannot capture long-range visual dependencies was suspected to hinder further improvements in performance, demanding the exploration of alternative architectural choices. To this end, attention mechanisms that have successfully served Natural Language Processing [4] appear as a promising substitute to convolutions, as they are able to detect spatially distant concepts and assign appropriate importance weights to them. Indeed, the adaptation of the Transformer [4] for visual tasks, led to the introduction of the Visual Transformer (ViT) [5], which divides the image into visual patches and processes them similarly to how the original Transformer handles words. Consequently, transformer-based image classifiers emerged [6, 7, 8, 9], reaching unprecedented state-of-the-art results.

In A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.), Proceedings of the AAAI 2023 Spring Symposium on Challenges Requiring the Combination of Machine Learning and Knowledge Engineering (AAAI-MAKE 2023), Hyatt Regency, San Francisco Airport, California, USA, March 27-29, 2023.

marialymp@islab.ntua.gr (M. Lymperaiou); kthomas@islab.ntua.gr (K. Thomas); gstam@cs.ntua.gr (G. Stamou)

0000-0001-9442-4186 (M. Lymperaiou); 0000-0002-7489-7776 (K. Thomas); 0000-0003-1210-9874 (G. Stamou)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

Even though so much effort is invested to perpetually improve model performance by employing more and more refined architectures and techniques, inevitably increasing the demand for computational resources necessary for training, there are still some open questions regarding the ability of such models to properly handle distribution shifts. Distribution shifts refer to testing an already trained model on a data distribution that diverges from the one the model was trained on. The analysis of distribution shifts has gained interest in recent years [10, 11, 12, 13, 14], as a crucial step towards enhancing model robustness. Most of these endeavors apply pixel-level perturbations to artificially influence the distribution under investigation. Nevertheless, the highly constrained setting of artificial distribution shifts excludes various real-world scenarios, impeding robust generalization of image classifiers. In this case, natural shifts [15, 16, 17, 18] are more representative. They usually require the creation of a *curated* dataset containing image variations such as changes in viewpoint or object background, rotations, and other minor changes. Both synthetic and natural shifts can comprise data augmentation techniques, which aid the development of robust models when incorporated during training [19, 20, 21, 22].

So far, there is no approach testing image classification 'in the wild', where *uncurated* images corresponding to pre-defined dataset labels are encountered. We argue that this is a real-world user-oriented scenario, where totally new images corresponding to ImageNet labels need to be appropriately classified. For example, an image of a cat found on the web may significantly differ from ImageNet cat instances, even when popular distribution shifts are taken into account. Even though a human can identify a cat present in an image with satisfactory confidence, we question whether an image classifier can do so; the unrestricted space of possible variations of uncurated images demands advanced generalization capabilities to properly understand the real discriminative characteristics of an ImageNet class without getting distracted from extraneous features.

The problem of classification 'in the wild' becomes even more difficult when fine-grained classification needs to be performed, as distinguishing between closely related categories relies on detailed discriminative characteristics, which may be less prevalent in uncurated settings. For example, siamese and persian cat races present many visual similarities, increasing the potential risk of learning and reproducing dataset biases, especially when distribution shifts are present. We can attribute this risk to the fact that existing classifiers lack *external* or *domain knowledge*, which can help humans discriminate between closely related categories.

To sum up, in our current paper we aspire to answer the following questions:

1. How do different models, pre-trained on ImageNet or web images, behave on uncurated image sets crawled from Google images (given ImageNet labels as Google queries)? We target this question by producing a novel natural *distribution shift* based on uncurated web images upon which we evaluate various image classifiers.
2. How does hierarchical knowledge help with evaluating classification results since several ImageNet categories are hierarchically related? We attempt to verify to which extent the assumption that *the lack of external knowledge limits the generalization capabilities of classifiers* holds. Thus, we leverage WordNet [23] to discover neighbors of given terms and test whether classifiers struggle with discriminating between closely related classes.
3. Can evaluation of classification be *explainable*? Knowledge sources, such as WordNet can reveal the semantic relationships between concepts (ImageNet classes), providing

possible paths connecting frequently confused classes.

Our code can be found at https://github.com/marialymperaiou/classification-in-the-wild.

2. Related work

Image classifiers With the outburst of neural architectures for classification tasks, Computer Vision has been one of the fields most benefited from recent developments. Convolutional classifiers (CNN) is a well-established backbone, with first successful endeavors [2] already paving the way for more refined architectures, such as VGG [24], Inception [25], ResNet [26], Xception [27], InceptionResnet [28] and others [3]. There is some criticism around the usage of CNNs for image classification, even though some contemporary endeavors such as ConvNext [29] revisit and insist on the classic paradigm, providing advanced performance. The rapid advancements that the Transformer framework [4] brought via the usage of self-attention mechanisms, widely replacing prior architectures for Natural Language Processing applications, inspired the usage of similar models for Computer Vision as an answer to the aforementioned criticism [30]. Thus, Vision Transformers (ViTs) [5] built upon [4] set a new baseline in literature; ever since, several related architectures emerged. In general, transformer-based models rely on an abundance of training data to ensure proper generalization. This requirement was relaxed in DeiT [31], enabling learning on medium-sized datasets. Further development introduced novel transformer-based architectures, such as BeiT [9], Swin [32] and RegNets [33], which realize specific refinements to boost performance. Overall, it has been proven that ViTs are more robust compared to classic CNN image classifiers [34]. In our work, we verify the degree this claim holds by testing CNN and transformer-based classifiers on the uncurated fine-grained setting.

Robustness under distribution shifts Generalization capabilities of existing image classifiers have been a crucial problem [35], currently addressed from a few different viewpoints. Artificial corruptions [36, 14, 37, 16, 11] or natural shifts [15, 38] on *curated* data have already exposed biases and architectural vulnerabilities. Adversarial robustness [39, 40, 41, 42, 43] is a related field where models are tested against adversarial examples, which introduce imperceptible though influential perturbations on images. Contrary to such attempts, we concentrated around naturally occurring distribution shifts stemming from *uncurated* image data. Regarding architectural choices, many studies perform robustness tests attempting to resolve the CNN vs Transformer contest [34, 44, 45], while other ventures focus on interpreting and understanding model robustness [46, 47, 48]. In our approach, by experimenting with both CNN and transformer-based architectures we adopt such research attempts to the *uncurated* setting.

3. Method

The general workflow of our method (Figure 1) consists of three stages. First, the dataset should be constructed by gathering common terms (queries) and their subcategories which exist as ImageNet classes. Images corresponding to those terms are crawled from Google search. In the second stage, various pre-trained classifiers are utilized to classify crawled images. The hierarchical relationships between the given classes are reported to enrich the evaluation process. Finally, all semantic relationships between misclassified samples are gathered to extract explanations and quantify how much, falsely predicted classes, diverge from their ground truth.

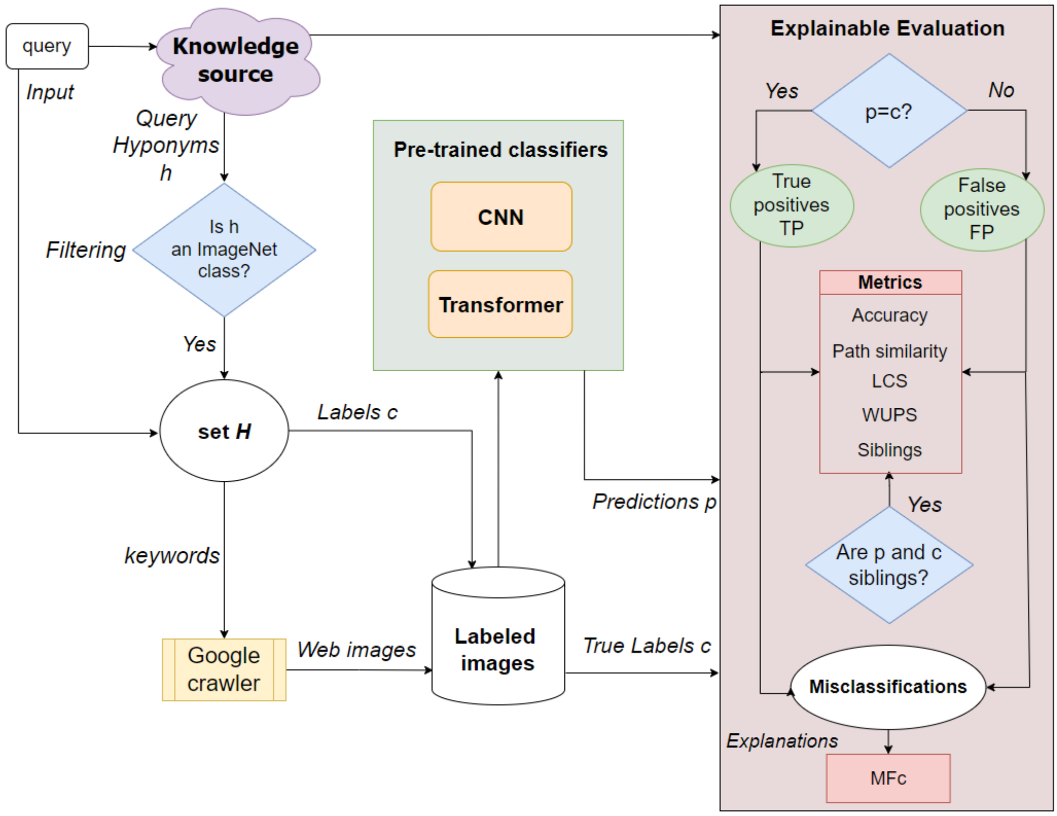

Figure 1: Outline of our method.

Dataset creation We start by gathering user-defined common words regarding visual concepts as *queries*, which will act as starting points towards extracting subcategories. The WordNet hierarchy [23] is used to provide the subcategories, via the hypernym-hyponym (IsA) relationships, which refer to more general or more specific concepts respectively. For example, given the query 'car', its hypernym is 'motor vehicle' ('car' IsA 'motor vehicle'), while its hyponyms are 'limousine' ('limousine' IsA 'car'), 'sports car' ('sports car' IsA 'car') and other specific car types. Therefore, we map queries on WordNet to obtain all their immediate hyponyms, constructing a *hyponyms set* H. We then filter out any hyponyms not belonging to ImageNet class labels.

The filtered categories of H among the initial query are provided as search terms to a web crawler suitable for searching Google images. We set a predefined threshold k for the number of Google images returned so that we evaluate classifiers on categories containing almost equal numbers of samples. This is necessary since some popular categories may return way more

Google images compared to others. We will experiment with several values of *k*, thus influencing the tradeoff between relevance to the keyword and adequate dataset size. The retrieved images comprise a labeled dataset D, with the keywords as labels.

Classification We consider a variety of image classifiers to test their ability for fine-grained classification on uncurated web images. We commence our experimentation with convolutional-based models as baselines, which have generally been considered to be less robust against distribution shifts and other perturbations [34], and we proceed with recent transformer-based architectures. We perform no further training or fine-tuning on the selected models.

For each model, we perform inference on the crawled images that constitute our dataset, as explained in the previous paragraph. We implement a rich evaluation scheme to capture various insights of the classification process. Accuracy is useful as a benchmark metric to compare our findings with expected classification results. WordNet similarity functions offer valuable information about misclassifications; for example, let's assume that the true label of a sample is 'cat' and the classifier predicts the label 'dog' in one case and the label 'airplane' in another case. Intuitively, we hypothesize that a 'cat' is more closely related to a 'dog' than an 'airplane' since they are both animals. This human intuition is reflected in the WordNet hierarchy, thus assigning a different penalty depending on the concept relevance within the hierarchy.

This concept-based evaluation can be realized using the following WordNet functions: path similarity, Leacock-Chodorow Similarity (LCS), and Wu-Palmer Similarity (WUPS). **Path similarity** evaluates how similar two concepts are, based on the shortest path that connects them within the WordNet hierarchy. It can provide values between 0 and 1, with 1 denoting the maximum possible similarity score. **LCS** also seeks for the shortest path between two concepts but additionally regards the depth of the taxonomy. Specifically, equation 1 mathematically describes LCS between two concepts c_1 and c_2:

$$LCS = -\log \frac{path(c_1, c_2)}{2 \cdot d} \tag{1}$$

where $path(c_1, c_2)$ denotes the shortest path connecting c_1 and c_2 and d refers to the taxonomy depth. Higher LCS values indicate higher similarity between concepts. **WUPS** takes into account the depth that the two concepts c_1 and c_2 appear in WordNet taxonomy and the depth of their most specific common ancestor node, called Least Common Subsumer. Higher WUPS scores refer to more similar concepts. For each of the path similarity, LCS, and WUPS metrics we obtain an average value over the total number of samples of the constructed dataset D.

Moreover, we report the percentage of *sibling concepts* among misclassifications. Two concepts are considered to be siblings if they share an immediate (1 hop) parent. For example, the concepts 'tabby cat' and 'egyptian cat' share the same parent node ('domestic cat'). It is highly likely that a classifier is more easily confused between two sibling classes, thus providing false positive (FP) predictions closely related to the ground truth (GT) label. Therefore, a lower number of siblings denotes reduced classification capacity compared to models of higher siblings percentage.

Explanations are provided during the evaluation stage, aiming to answer *why a pre-trained classifier cannot correctly classify uncurated images belonging to a class c.*

FP predictions contain valuable information regarding which classes are confused with the GT. The per-class misclassification frequency (MF) refers to the percentage of occurrences of each false positive class f within the total number of false positive instances. Thus, given a dataset with N classes, c as the ground truth class and f as one of the false positive classes, the misclassification frequency for the $c \rightarrow f$ misclassification is:

$$MF_c = \frac{FP_{i=f}}{\sum_{i=0}^{i=N} FP_i} \cdot 100\% \tag{2}$$

MF scores can be extracted for all $f \neq c$ FP classes so that the most influential misclassifications are discovered. Higher MF scores denote some classifier tendency to choose the FP class over the GT one, therefore indicating either a classifier bias or an annotation error in the dataset. Specifically, a classifier bias refers to consistently classifying samples from class c as samples of class f, given that the annotation is the best possible. Of course, such a requirement cannot be always satisfied, especially when expert annotators are needed, as may happen in the case of fine-grained classification. On the other hand, since our explainable evaluation approach is able to capture such misclassification patterns, it is not necessary to attribute the source of misclassification beforehand. Human annotators can be employed at a later stage, identifying and verifying the source of misclassifications.

4. Experiments

In all following experiments, we selected a threshold of T=50 crawled images per class. We will present results on a random initial query as a proof-of-concept to demonstrate our findings. For this reason, we provide the query 'cat', which returns the following WordNet hyponyms (also corresponding to ImageNet labels):

H={'angora cat', 'cougar cat', 'egyptian cat', 'leopard cat', 'lynx cat', 'persian cat', 'siamese cat', 'tabby cat', 'tiger cat'}

The same experimentation can be replicated for other selected queries, as long as they can be mapped on WordNet.

4.1. Convolutional classifiers

We leveraged the following CNN classifiers: VGG16/19, [24], ResNet50/101/152 [26], InceptionV3 [25], InceptionResnetV2 [28], Xception [27], MobileNetV2 [49], NasNet-Large [50], DenseNet121/169/201 [51], EfficientNet-B7 [52], ConvNeXt [29]. We present results for CNN classifiers in Table 1. Bold instances denote lower accuracy than the best ImageNet accuracy of each model, as reported by the authors of each model respectively[1]. Underlined cells indicate best accuracy/sibling percentage scores for each category. The absence of models or keywords from Table 1 means that they correspond to zero accuracy scores. For example, we observe the complete absence of models such as InceptionV3, InceptionResNetV2, Xception, NASNetLarge,

[1]https://paperswithcode.com/sota/image-classification-on-imagenet

Table 1
Classification results using CNNs. Bold entries denote lower accuracy compared to best model accuracy.

Model	Label	Accuracy↑	Siblings↑	Label	Accuracy↑	Siblings↑
ResNet50	tabby cat	**50.00**%	24.00%	siamese cat	90.00%	0.00%
ResNet101		**52.00**%	41.67%		88.00%	16.67%
ResNet152		**50.00**%	12.00%		90.00%	20.00%
VGG16		**38.00**%	38.71%		82.00%	11.11%
VGG19		**50.00**%	32.00%		88.00%	16.67%
MobileNetV2		**2.00**%	2.04%		**4.00**%	0.00%
EfficientNet		**10.00**%	33.33%		96.00%	100.00%
ConvNext		**60.00**%	15.00%		92.00%	75.00%
ResNet50	lynx cat	82.00%	0.00%	cougar cat	84.00%	0.00%
ResNet101		84.00%	0.00%		78.00%	0.00%
ResNet152		86.00%	0.00%		88.00%	0.00%
VGG16		82.00%	0.00%		86.00%	0.00%
VGG19		80.00%	0.00%		78.00%	0.00%
EfficientNet		90.00%	0.00%		98.00%	100.00%
ConvNext		92.00%	0.00%		98.00%	0.00%
ResNet50	tiger cat	**18.33**%	0.00%	persian cat	92.00%	25.00%
ResNet101		**23.33**%	0.00%		88.00%	16.67%
ResNet152		**26.67**%	0.00%		88.00%	33.33%
VGG16		**20.00**%	0.00%		86.00%	14.29%
VGG19		**28.33**%	0.00%		80.00%	10.00%
MobileNetV2		**1.67**%	0.00%		**8.00**%	2.17%
EfficientNet		**36.67**%	0.00%		98.00%	100.00%
ConvNext		**26.67**%	0.00%		98.00%	100.00%
ResNet50	leopard cat	**12.00**%	18.18%	angora cat	**12.00**%	50.00%
ResNet101		**12.00**%	15.91%		**20.00**%	50.00%
ResNet152		**4.00**%	14.58%		**20.00**%	62.50%
VGG16		**10.00**%	6.67%		**10.00**%	46.67%
VGG19		**10.00**%	6.67%		**8.00**%	54.35%
EfficientNet		**2.00**%	16.33%		**4.00**%	95.83%
ConvNext		**16.00**%	16.67%		**10.00**%	88.89%
ResNet50	egyptian cat	**24.00**%	2.63%	cat	82.05%	0.00%
ResNet101		**30.00**%	2.86%		82.05%	0.00%
ResNet152		**34.00**%	6.06%		79.49%	0.00%
VGG16		**28.00**%	0.00%		87.18%	0.00%
VGG19		**26.00**%	2.70%		76.92%	0.00%
MobileNetV2		**0.00**%	0.00%		**2.56**%	0.00%
EfficientNet		**70.00**%	0.00%		92.31%	0.00%
ConvNext		**52.00**%	0.00%		94.87%	0.00%

DenseNet121/169/201 meaning that they are completely unable to properly classify the crawled images, even those belonging to categories that show satisfactory accuracy when other classifiers are deployed. MobileNetV2 also shows deteriorated performance for all categories. We will investigate later if hierarchical knowledge can help extract any meaningful information regarding this surprisingly low performance.

Other results that can be extracted from Table 1 is that some categories can be easily classified ('siamese cat', 'lynx cat', 'cougar cat', 'persian cat', 'cat') contrary to others ('tabby cat', 'tiger cat', 'egyptian cat', 'leopard cat', 'angora cat'). Since we have no specific knowledge of animal species, we will once again leverage WordNet to obtain explanations regarding this behavior. Sibling percentages offer a first glance at the degree of confusion between similar classes in the fine-grained setting. For example, even though 'siamese cat' and 'cougar cat' classes demonstrate high accuracy scores, we observe a completely different behavior regarding the sibling percentages: most CNN classifiers return some sibling false positives for 'siamese cat' ground truth label, while the opposite happens for the 'cougar cat' ground truth label, which mostly receives zero sibling misclassifications. This behavior indicates that for 'siamese cat' if a sample is misclassified, it is likely that it belongs to a conceptually similar class, while for 'cougar cat' misclassifications, false positives belong to more semantically distant categories.

Regarding model capabilities, we observe that for both 'siamese' and 'cougar cat' classes, all ResNet50 false positives belong to non-sibling classes, contrary to EfficientNet false positives, which all belong to sibling classes. By also looking to other categories, we observe that in general, EfficientNet achieves a higher sibling percentage compared to ResNet50, meaning that EfficientNet misclassifications are more justified compared to ResNet50 misclassifications.

4.2. Transformer-based classifiers

The following transformer-based image classifiers were used: ViT [5], Regnet-x [33], DeiT [31], BeiT [9], CLIP [53], Swin Transformer V2 [32]. Results for Transformer-based classifiers are provided in Table 2. We spot a similar pattern regarding the categories upon which models struggle to make predictions: instances belonging to 'tabby cat', 'tiger cat', 'egyptian cat' categories are classified with low accuracy compared to 'siamese cat', 'lynx cat', 'cougar cat', 'persian cat', 'cat', 'angora cat' and 'leopard cat'. We suspect that there is a common reason behind this behavior, probably attributed to unavoidable intra-class similarities present in the fine-grained classification setting.

As for model performance, we examine sibling percentage apart from exclusively evaluating accuracy. The behavior of transformer-based models regarding sibling misclassification is harder to be interpreted compared to CNN models, because models that return high sibling percentages for some categories may present low sibling percentages on other categories and vice versa. For example, BeiT scores low on sibling percentages for 'tabby cat' (3.45%), 'siamese cat' (0%) and 'persian cat' (10%) compared to other models for the same classes; on the other hand, it returns *best* sibling scores for 'leopard cat' (78.72%), 'tiger cat' (22.45%) and 'egyptian cat' (22.50%). More results about the explainability of results are provided in Section 4.3.

4.3. Explaining misconceptions

In Tables 3,4 & 5 we report the top-3 misclassifications per ground truth (GT) category and per model, as well as the misclassification frequency (MF) for each false positive (FP) label. GT column refers to cat species exclusively, even if the word 'cat' is omitted (for example, 'tiger' GT

Table 2
Classification results using Transformers. Bold entries denote lower accuracy compared to best model accuracy, underlined metrics indicate best metric performance per class.

Model	Label	Accuracy↑	Siblings↑	Label	Accuracy↑	Siblings↑
ViT	tabby cat	**44.00%**	<u>42.86%</u>	siamese cat	92.00%	50.00%
BeiT		**42.00%**	3.45%		94.00%	0.00%
DeiT		<u>**60.00%**</u>	30.00%		94.00%	33.33%
Swin		**48.00%**	30.77%		94.00%	<u>100.00%</u>
xRegNet		**52.00%**	25.00%		92.00%	50.00%
CLIP		**30.00%**	28.57%		<u>96.00%</u>	50.00%
ViT	lynx cat	90.00%	0.00%	cougar cat	<u>96.00%</u>	0.00%
BeiT		**26.00%**	0.00%		92.00%	0.00%
DeiT		<u>92.00%</u>	0.00%		<u>96.00%</u>	0.00%
Swin		86.00%	0.00%		<u>96.00%</u>	0.00%
xRegNet		90.00%	0.00%		<u>96.00%</u>	<u>50.00%</u>
CLIP		86.00%	0.00%		92.00%	0.00%
ViT	tiger cat	**18.33%**	0.00%	persian cat	92.00%	75.00%
BeiT		**18.33%**	<u>22.45%</u>		80.00%	10.00%
DeiT		**15.00%**	0.00%		96.00%	50.00%
Swin		**21.67%**	0.00%		96.00%	50.00%
xRegNet		<u>**35.00%**</u>	0.00%		<u>98.00%</u>	<u>100.00%</u>
CLIP		46.67%	0.00%		96.00%	50.00%
ViT	leopard cat	**12.00%**	2.27%	angora cat	**6.00%**	89.36%
BeiT		**6.00%**	78.72%		<u>**62.00%**</u>	52.63%
DeiT		**10.00%**	11.11%		**0.00%**	94.00%
Swin		<u>**14.00%**</u>	9.30%		**8.00%**	<u>95.65%</u>
xRegNet		**6.00%**	21.28%		**0.00%**	76.00%
CLIP		**10.00%**	<u>55.56%</u>		**8.00%**	91.30%
ViT	egyptian cat	**38.00%**	3.23%	cat	89.74%	0.00%
BeiT		**20.00%**	<u>22.50%</u>		**53.85%**	0.00%
DeiT		**36.00%**	3.12%		<u>94.87%</u>	0.00%
Swin		<u>**52.00%**</u>	0.00%		89.74%	0.00%
xRegNet		**48.00%**	0.00%		92.31%	0.00%
CLIP		70.00%	6.67%		69.23%	0.00%

entry refers to 'tiger cat'). We highlight with red irrelevant FP classes, which are semantically distant compared to the GT label, while misconceptions involving sibling classes are highlighted with blue. Moreover, magenta indicates that an FP is actually an immediate (1 hop) hypernym of the GT. Due to space constraints, we present here all transformer-based models, but only a subset of the CNN models tested in total; more results can be found in the Appendix.

Interestingly, we can spot some surprising frequent misconceptions, such as confusing cat species with the 'mexican hairless' dog breed. For CNN classifiers, we spot this peculiarity for all models under investigation: 10.53% of ResNet50 FP for 'egyptian cat' GT label belong to the 'mexican hairless' class; the same applies to 14.29% of ResNet101 FP, 18.18% of ResNet152 FP

Table 3
Common misclassifications for selected GT cat classes and misclassification frequency (CNNs).

Model	GT	Top-1 FP	Top-1 MF	Top-2 FP	Top-2 MF	Top-3 FP	Top-3 MF
Res Net50	tabby	tiger cat	32.00%	egyptian cat	24.00%	web site	8.00
	angora	persian cat	34.00%	arctic fox	11.36%	lynx	9.09%
	lynx	coyote	22.22%	tabby cat	11.11%	egyptian cat	11.11%
	siamese	great dane	20.00%	hare	20.00%	american egret	20.00%
	tiger	tabby cat	40.82%	egyptian cat	20.41%	tiger	14.29%
	persian	old English sheepdog	25.00%	siamese cat	25.00%	hatchet	25.00%
	cougar	lynx	25.00%	malinois	25.00%	wallaby	25.00%
	leopard	egyptian cat	30.00%	tiger cat	16.00%	jaguar	12.00%
	egyptian	mexican hairless	10.53%	mask	5.26%	comic book	5.26%
	cat	fur coat	14.29%	carton	14.29%	book jacket	14.29%
Res Net 101	tabby	egyptian cat	41.67%	tiger cat	29.17%	web site	8.33%
	angora	persian cat	32.50%	egyptian cat	12.50	lynx	10.00%
	lynx	tabby cat	12.50%	egyptian cat	12.50%	cheetah	12.50%
	siamese	Boston bull	16.67%	egyptian cat	16.67%	hare	16.67%
	tiger	tabby cat	34.78%	tiger cat	17.39%	egyptian cat	15.22%
	persian	keeshond	16.67%	guinea pig	16.67%	collie	16.67%
	cougar	lynx	45.45%	meerkat	9.09%	dhole	9.09%
	leopard	egyptian cat	36.00%	tiger cat	14.00%	leopard	12.00%
	egyptian	mexican hairless	14.29%	mask	8.57%	sea lion	5.71%
	cat	macaque	14.29%	barbershop	14.29%	Pembroke	14.29%
Res Net 152	tabby	tiger cat	40.00%	egyptian cat	12.00%	lynx	12.00%
	angora	persian cat	35.00%	siamese cat	10.00%	shower curtain	10.00%
	lynx	tabby cat	42.86%	coyote	14.29%	norwich terrier	14.29
	siamese	whippet	20.00%	egyptian cat	20.00%	angora cat	20.00%
	tiger	tabby cat	34.09%	egyptian cat	18.18%	tiger	15.91%
	persian	siamese cat	33.33%	collie	16.67%	fur coat	16.67%
	cougar	menu	16.67%	wild boar	16.67%	wallaby	16.67%
	leopard	egyptian cat	28.00%	lynx	22.00%	jaguar	16.00%
	egyptian	mexican hairless	18.18%	web site	9.09%	tabby cat	6.06%
	cat	macaque	12.50%	Pembroke	12.50%	chihuahua	12.50%
VGG 16	tabby	egyptian cat	38.71%	tiger cat	22.58%	wood rabbit	3.23%
	angora	persian cat	26.67%	egyptian cat	15.56%	lynx	8.89%
	lynx	coyote	33.33	egyptian cat	22.22%	madagascar	11.11%
	siamese	mexican hairless	22.22%	whippet	11.11%	fur coat	11.11%
	tiger	tabby cat	33.33%	egyptian cat	20.83%	tiger	16.67%
	persian	arctic fox	14.29%	angora cat	14.29%	lynx	14.29%
	cougar	lynx	42.86%	coyote	28.57%	menu	14.29%
	leopard	egyptian cat	42.00%	lynx	18.00%	jaguar	10.00%
	egyptian	mexican hairless	8.33%	lynx	5.56%	sombrero	5.56%
	cat	norwich terrier	20.00%	schipperke	20.00%	kit fox	20.00%

Table 4
Common misclassifications for selected GT cat classes and misclassification frequency (Transformers).

Model	GT	Top-1		Top-2		Top-3	
		FP	MF	FP	MF	FP	MF
CLIP	tabby	madagascar	40.00%	egyptian cat	22.86%	tiger cat	11.43
	angora	persian cat	78.26%	madagascar	6.52%	siamese cat	6.52%
	lynx	madagascar	14.29%	leopard cat	14.29%	grey fox	14.29%
	siamese	polecat	50.00%	persian cat	50.00%	-	-
	tiger	egyptian cat	30.77%	madagascar	19.23%	leopard cat	15.38%
	persian	madagascar	50.00%	siamese cat	50.00%	-	-
	cougar	lynx	75.00%	madagascar	25.00%	-	-
	leopard	tiger cat	55.56%	madagascar	17.78%	egyptian cat	11.11%
	egyptian	mexican hairless	26.67%	madagascar	26.67%	armadillo	6.67%
	cat	madagascar	66.67%	orange	16.67%	bib	8.33%
BeiT	tabby	tiger cat	17.24%	cat	13.79%	domestic	13.79%
	angora	persian cat	42.11%	domestic	21.05%	quadruped	5.26%
	lynx	common lynx	59.46%	Canada lynx	16.22%	bobcat	5.41%
	siamese	kitten	66.67%	feline	33.33%	-	-
	tiger	tabby cat	23.40%	margay	14.89%	domestic	6.38%
	persian	domestic	20.00%	angora cat	10.00%	breadwinner	10.00%
	cougar	feline	25.00%	big cat	25.00%	cub	25.00%
	leopard	margay	42.55%	ocelot	21.28%	spotted lynx	8.51%
	egyptian	Abyssinian	15.00%	mexican hairless	10.00%	mouser	5.00%
	cat	feline	33.33%	kitten	22.22%	caterer	11.11%
DeiT	tabby	tiger cat	35.00%	egyptian cat	30.00%	web site	15.00%
	angora	persian cat	62.00%	egyptian cat	28.00%	tabby cat	2.00%
	lynx	tabby cat	75.00%	coyote	25.00%	-	-
	siamese	egyptian cat	33.33%	mexican hairless	33.33%	lynx	33.33%
	tiger	tabby cat	37.50%	egyptian cat	27.50%	leopard cat	12.50%
	persian	soft-coated wheaten terrier	50.00%	siamese cat	50.00%	-	-
	cougar	web site	50.00%	dingo	50.00%	-	-
	leopard	egyptian cat	48.89%	lynx	22.22%	tiger cat	11.11%
	egyptian	mexican hairless	15.62%	comic book	9.38%	kelpie	3.12%
	cat	fur coat	50.00%	chihuahua	50.00%	-	-
xReg Net	tabby	tiger cat	62.50%	egyptian cat	20.83%	menu	4.17%
	angora	persian cat	48.00%	egyptian cat	18.00%	lynx	8.00%
	lynx	tabby cat	40.00%	tiger cat	20.00%	egyptian cat	20.00%
	siamese	egyptian cat	50.00%	polecat	25.00%	lynx	25.00%
	tiger	tabby cat	40.00%	egyptian cat	25.71%	lynx	11.43%
	persian	siamese cat	100.0%	-	-	-	-
	cougar	tiger cat	50.00%	lynx	50.00%	-	-
	leopard	egyptian cat	40.43%	tiger cat	21.28%	lynx	17.02%
	egyptian	mexican hairless	15.38%	mask	7.69%	comic book	7.69%
	cat	comic book	33.33%	tub	33.33%	drake	33.33%

Table 5
(Continuation of Tab 4). Common misclassifications and misclassification frequency.

Model	GT	Top-1 FP	Top-1 MF	Top-2 FP	Top-2 MF	Top-3 FP	Top-3 MF
Swin	tabby	tiger cat	57.69%	egyptian cat	30.77%	web site	7.69%
	angora	persian cat	58.70%	egyptian cat	26.09%	tabby cat	10.87%
	lynx	tabby cat	57.14%	fur coat	14.29%	timber wolf	14.29%
	siamese	egyptian cat	100.0%	-	-	-	-
	tiger	tabby cat	35.00%	egyptian cat	32.50%	leopard cat	12.50%
	persian	siamese cat	50.00%	hand blower	50.00%	-	-
	cougar	web site	50.00%	Irish wolfhound	50.00%	-	-
	leopard	egyptian cat	44.19%	lynx	37.21%	tiger cat	9.30%
	egyptian	mexican hairless	20.83%	comic book	16.67%	table lamp	8.33%
	cat	fur coat	25.00%	jersey	25.00%	chihuahua	25.00%
ViT	tabby	egyptian cat	42.86%	tiger cat	32.14%	web site	10.71%
	angora	egyptian cat	48.94%	persian cat	38.30%	tabby cat	2.13%
	lynx	tabby cat	40.00%	egyptian cat	40.00%	timber wolf	20.00%
	siamese	egyptian cat	50.00%	chihuahua	25.00%	mexican hairless	25.00%
	tiger	egyptian cat	48.78%	tabby cat	26.83%	leopard cat	14.63%
	persian	plastic bag	25.00%	egyptian cat	25.00%	siamese cat	25.00%
	cougar	egyptian cat	50.00%	malinois	50.00%	-	-
	leopard	egyptian cat	86.36%	snow leopard	2.27%	web site	2.27%
	egyptian	mexican hairless	16.13%	pedestal	12.90%	vase	6.45%
	cat	washer	25.00%	fur coat	25.00%	mexican hairless	25.00%

and 8.33% of VGG16 FP. More animals such as 'wallaby', 'jaguar', 'sea lion', 'cheetah', 'arctic fox', 'coyote' etc appear as frequent FPs.

For transformer models, the 'egyptian cat' → 'mexican hairless' abnormality is observed for all classifiers when 'egyptian cat' GT label is provided, resulting in the following 'mexican hairless' FP percentages: 26.67% for CLIP, 10% for BeiT, 15.62% for DeiT, 15.38% for xRegNet, 20.83% for Swin, and 16.33% for ViT. Obviously, regardless of whether the CNN or transformer classifier is being used, images of 'egyptian cats' are often erroneously perceived as 'mexican hairless dogs'. A qualitative analysis between 'egyptian cat' images and 'mexican hairless dog' images indicates that these animals are obviously distinct, even though they present similar ear shapes and rather hairless, thin bodies. Therefore, we can assume that the transformer-based classifiers are biased towards texture, verifying relevant observations reported for CNNs [11]. Also, ear shape acts as a confounding factor, overshadowing other actually distinct animal characteristics. There are more misclassifications involving animals, such as 'armadillo', 'chihuahua', 'soft-coated wheaten terrier', 'kelpie', and others.

Even more surprising are misclassifications not including animal species. For example, CNN classifiers predict 'web site' instead of 'tabby cat', 'hatched' instead of 'persian cat', 'barbershop' instead of 'cat', 'menu' instead of 'cougar' etc. All ResNet50/101/152 and VGG16 make at least one such misclassification, something that highly questions which features of cat species contribute to such predictions.

Misclassifications involving non-animal classes using transformers (Tables 4, 5) provide the following interesting abnormalities: 'cat' is classified as 'fur coat' for 50% of the FP instances using DeiT. This non-negligible misclassification rate once again verifies the aforementioned texture bias. In a similar sense, xRegNet classifies 'egyptian cat' images as 'mask' and as 'comic book' 7.69% of the FPs respectively. Such categories had also appeared in CNN misclassifications. We cannot provide a human-interpretable explanation about the 'mask' misclassification, since the term 'mask' may refer to many different objects. We hypothesize that 'mask' ImageNet instances may contain carnival masks looking similar to cats, therefore the lack of context confused xRegNet. 'Comic book' appears 9.38% of the times an 'egyptian cat' image is misclassified by DeiT, 33.33% of the times a 'cat' photo is misclassified by xRegNet, and 16.67% of the times an 'egyptian cat' is misclassified by Swin. This can be attributed to the fact that crawled images may contain cartoon-like instances, which cannot be clearly regarded as cats. Other interesting misclassifications involving irrelevant categories are 'cat'→'washer' (25% of FPs using ViT), 'leopard cat'→'web site' (2.27% of FPs using ViT, 15% of FPs using DeiT), 'persian cat'→'plastic bag' (25% of FPs using ViT), 'cat'→'jersey' (25% of FPs using Swin), 'egyptian cat'→'table lamp' (8.33% of FPs using Swin), 'cat'→'tub (33.33% of FPs using xRegNet), and others.

An interesting observation revolves around the 'egyptian cat' label. For CNN models, almost all top-3 FP of 'egyptian cat' GT label correspond to irrelevant ImageNet categories. On the contrary, 'tabby cat', 'angora cat', and 'tiger cat' present more sensible FPs, which usually involve sibling categories (highlighted with blue). As for transformer models, we observe that 'egyptian cat' label is always being confused with at least one irrelevant ImageNet category, while 'angora cat' is only confused with other cat species, and not with conceptually distant classes. Thus, 'egyptian cat' crawled images seem to contain some misleading visual features that frequently derail the classification process. Indeed, when viewing 'egyptian cat' crawled images, some of them are drawings or photos of cat souvenirs; however, misconceptions such as 'table lamp' or 'armadillo' cannot be visually explained by human inspectors, unraveling more questions on the topic. A comparison between CNN classifiers (Table 3) and transformer-based classifiers (Table 4, 5) denotes that transformers are more capable of retrieving similar categories to the GT; this becomes obvious by observing the higher number or irrelevant misclassifications highlighted with red for CNNs, compared to transformer results.

By combining Tables 3, 4 & 5 with Tables 1& 2, we obtain some very interesting findings: how are low classification metric scores connected to the relevance between misclassified categories? We start with categories presenting low accuracy scores ('tabby cat', 'tiger cat', 'egyptian cat'), and we compare them with categories offering frequent extraneous misclassifications ('egyptian cat' and 'cat', followed by 'tabby cat' and 'lynx'). Classifying 'egyptian cat' images both yields low classification scores and returns irrelevant false positives. On the other hand, even though 'cat' images present high accuracy scores, misclassifications are highly unrelated when they happen. 'Tiger cat' scores low in accuracy, however, misclassifications are rather justified, since other cat species are returned. Surprisingly, 'tiger cat' also scores low in siblings percentage, indicating that false positives are not immediately related to the GT 'tiger cat' class. In this case, we assume that false positives ('egyptian cat', 'tabby cat', 'leopard cat' etc) belong to more distant relatives of the 'tiger cat' concept class, even though bearing some similar features.

Overall, throughout this analysis we prove that classification accuracy is unable to reveal the whole truth behind the way classifiers behave; to this end, knowledge sources are able to shed some light on the inner workings of this process. By analyzing a constraint family of related ImageNet labels (cat species) we already disentangled the classification accuracy from the classification *relevance*: false positives can be highly relevant to the ground truth (such as 'tiger cat' misclassifications) or not ('cat' misclassifications). We, therefore, argue that fine-grained classification also demands *fine-grained evaluation*, which can provide insightful information when driven by knowledge. The human interpretable insights of Tables 4, 5 are

Table 6
Conceptual metrics based on WordNet distances using CNN classifiers.

Model	Label	Path sim↑	LCH↑	WUPS↑	Label	Path sim↑	LCH↑	WUPS↑
ResNet50	tabby cat	0.18	1.79	0.69	siamese cat	0.10	1.25	0.57
ResNet101		0.22	1.99	0.75		0.15	1.60	0.72
ResNet152		0.16	1.59	0.62		0.17	1.71	0.67
VGG16		0.21	1.85	0.70		0.16	1.65	0.65
VGG19		0.18	1.68	0.63		0.17	1.73	0.70
MobileNetV2		0.09	1.13	0.39		0.09	1.15	0.40
EfficientNet		**0.24**	**2.17**	**0.86**		**0.33**	**2.54**	**0.88**
ResNet50	lynx cat	0.05	0.53	0.11	cougar cat	0.08	0.97	0.49
ResNet101		0.05	**0.62**	**0.13**		0.08	0.90	0.41
ResNet152		0.05	0.56	0.09		0.08	1.01	0.49
VGG16		0.04	0.46	0.08		0.08	0.88	0.37
VGG19		0.04	0.48	0.08		0.07	0.86	0.38
EfficientNet		0.05	0.54	0.09		**0.33**	**2.54**	**0.94**
ResNet50	tiger cat	**0.15**	**1.61**	**0.70**	persian cat	0.16	1.59	0.60
ResNet101		0.14	1.47	0.64		0.17	1.73	0.66
ResNet152		0.13	1.43	0.61		0.17	1.57	0.56
VGG16		0.14	1.51	0.65		0.12	1.27	0.50
VGG19		0.13	1.43	0.61		0.13	1.45	0.57
MobileNetV2		0.07	0.90	0.41		0.09	1.19	0.43
EfficientNet		0.13	1.41	0.60		**0.33**	**2.54**	**0.88**
ResNet50	leopard cat	0.17	1.62	0.67	angora cat	0.22	1.94	0.72
ResNet101		0.17	1.62	0.67		0.22	1.93	0.71
ResNet152		0.16	1.55	0.64		0.24	2.05	0.73
VGG16		0.15	1.49	0.62		0.21	1.90	0.72
VGG19		0.15	1.53	0.64		0.23	2.01	0.75
EfficientNet		**0.18**	**1.68**	**0.68**		**0.32**	**2.48**	**0.86**
ResNet50	egyptian cat	0.11	1.32	0.51	cat	0.11	1.29	0.56
ResNet101		0.11	1.32	0.50		0.11	1.34	0.63
ResNet152		**0.12**	**1.40**	**0.55**		0.11	1.35	0.61
VGG16		0.09	1.15	0.41		**0.14**	**1.67**	**0.80**
VGG19		0.10	1.21	0.45		0.12	1.42	0.64
MobileNetV2		0.08	1.06	0.36		0.07	0.88	0.34
EfficientNet		0.10	1.24	0.49		0.12	1.52	0.74

going to be quantified and verified in the next Section.

4.4. Knowledge-driven metrics

The aforementioned claim regarding the need for *fine-grained evaluation* is supported by demonstrating results using *knowledge-driven metrics* based on conceptual distance as provided by WordNet (Tables 6& 7). Since higher path similarity/LCH, WUPS scores are better, we denote with bold best (higher) scores for each category.

By comparing path similarity, LCH, and WUPS metrics across categories, we observe that categories having a large number of irrelevant FP (marked in red in Tables 4, 5), such as 'cougar

Table 7
(Continuation of Tab 6). Conceptual metrics based on WordNet distances using transformers.

Model	Label	Path sim↑	LCH↑	WUPS↑	Label	Path sim↑	LCH↑	WUPS↑
ViT	tabby cat	0.23	2.02	0.76	siamese cat	0.23	2.01	0.71
BeiT		**0.24**	2.02	0.75		0.18	1.88	0.77
DeiT		0.20	1.83	0.70		0.19	1.72	0.58
Swin		0.23	**2.07**	**0.82**		**0.33**	**2.54**	**0.88**
xRegNet		0.22	2.00	0.79		0.21	1.84	0.66
CLIP		0.18	1.78	0.75		0.24	2.12	0.84
ViT	lynx cat	0.05	0.55	0.09	cougar cat	0.15	1.63	**0.79**
BeiT		0.04	0.33	0.07		**0.21**	**1.94**	**0.79**
DeiT		0.05	0.54	0.09		0.09	1.13	0.54
Swin		0.05	**0.56**	0.09		0.07	0.97	0.51
xRegNet		0.04	0.50	0.08		0.19	1.44	0.50
CLIP		0.04	0.51	**0.10**		0.06	0.62	0.24
ViT	tiger cat	0.15	1.60	0.69	persian cat	0.27	2.19	0.76
BeiT		**0.21**	**1.89**	**0.76**		0.22	1.87	0.66
DeiT		0.13	1.42	0.61		0.24	2.12	0.80
Swin		0.14	1.47	0.63		0.21	1.85	0.65
xRegNet		0.14	1.50	0.64		**0.33**	**2.54**	**0.88**
CLIP		0.12	1.39	0.62		0.22	1.99	0.80
ViT	leopard cat	0.17	1.76	0.75	angora cat	0.31	2.39	0.83
BeiT		**0.31**	**2.47**	**0.93**		0.30	2.22	0.73
DeiT		0.15	1.47	0.60		**0.32**	2.48	**0.86**
Swin		0.13	1.26	0.50		**0.32**	**2.49**	**0.86**
xRegNet		0.18	1.70	0.69		0.28	2.22	0.78
CLIP		0.22	1.87	0.74		0.31	2.44	**0.86**
ViT	egyptian cat	0.11	1.33	0.50	cat	0.09	1.15	0.50
BeiT		**0.16**	1.63	0.60		**0.23**	**1.81**	**0.68**
DeiT		0.11	1.31	0.50		0.05	0.72	0.29
Swin		0.11	1.34	0.52		0.05	0.73	0.29
xRegNet		0.10	1.24	0.46		0.06	0.89	0.39
CLIP		0.15	**1.65**	**0.70**		0.12	1.42	0.66

cat' and 'lynx cat', followed by 'egyptian cat' and 'cat' also present low knowledge-driven metric scores in Tables 6, 7, as expected. Other categories such as 'angora cat', 'leopard cat', and 'tiger cat' that present misclassifications of related (sibling or parent) categories also present higher knowledge-driven metric scores. Therefore, we can safely assume that employing knowledge-driven metrics for evaluating fine-grained classification results is highly correlated with human-interpretable notions of similarity and therefore trustworthy.

Model performance is rather clear when examining CNN classifiers. EfficientNet achieves predicting more relevant FP images compared to other classifiers for the majority of the categories. On the other hand, it is harder to draw a similar conclusion for Transformer-based classifiers, as different models perform better for different categories; however, compared to CNN classifiers the results of knowledge-driven metrics are the same or higher for most categories. Even though this difference is not impressive, transformer-based models showcase an improved capability of predicting more relevant classes, when failing to return the GT one.

5. Conclusion

In this work, we implemented a novel distribution shift involving uncurated web images, upon which we tested convolutional and transformer-based image classifiers. Selecting closely related categories for classification is instructed by hierarchical knowledge, which is again employed to evaluate the quality of results. We prove that accuracy-related metrics can only scratch the surface of classification evaluation since they cannot capture semantic relationships between misclassified samples and ground truth labels. To this end, we propose an explainable, knowledge-driven evaluation scheme, able to quantify misclassification relevance by providing the semantic distance between false positive and real labels. The same scheme is also used to compare the classification capabilities of CNN vs transformer-based models on the implemented distribution shift. As future work, we plan to extend our analysis to more query terms in order to examine the extend of our current findings, and also combine the uncurated image classification setting with artificial corruptions to enhance our insights.

Acknowledgments

The research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the 3rd Call for HFRI PhD Fellowships (Fellowship Number 5537).

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.

[2] A. Krizhevsky, I. Sutskever, G. 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, Curran Associates Inc., Red Hook, NY, USA, 2012, p. 1097–1105.

[3] Z. Li, W. Yang, S. Peng, F. Liu, A survey of convolutional neural networks: Analysis, applications, and prospects, arXiv, 2020. URL: https://arxiv.org/abs/2004.02806. doi:10.48550/ARXIV.2004.02806.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. URL: https://arxiv.org/abs/1706.03762. doi:10.48550/ARXIV.1706.03762.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, 2020. URL: https://arxiv.org/abs/2010.11929. doi:10.48550/ARXIV.2010.11929.

[6] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, Y. Wu, Coca: Contrastive captioners are image-text foundation models, 2022. URL: https://arxiv.org/abs/2205.01917. doi:10.48550/ARXIV.2205.01917.

[7] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, L. Schmidt, Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 23965–23998. URL: https://proceedings.mlr.press/v162/wortsman22a.html.

[8] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, B. Guo, Swin transformer v2: Scaling up capacity and resolution (2021). URL: https://arxiv.org/abs/2111.09883. doi:10.48550/ARXIV.2111.09883.

[9] H. Bao, L. Dong, S. Piao, F. Wei, Beit: Bert pre-training of image transformers, 2021. URL: https://arxiv.org/abs/2106.08254. doi:10.48550/ARXIV.2106.08254.

[10] J. Cohen, E. Rosenfeld, Z. Kolter, Certified adversarial robustness via randomized smoothing, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 1310–1320. URL: https://proceedings.mlr.press/v97/cohen19c.html.

[11] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, W. Brendel, Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness, 2018. URL: https://arxiv.org/abs/1811.12231. doi:10.48550/ARXIV.1811.12231.

[12] F. Yang, Z. Wang, C. Heinze-Deml, Invariance-inducing regularization using worst-case transformations suffices to boost accuracy and spatial robustness, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/1d01bd2e16f57892f0954902899f0692-Paper.pdf.

[13] X. Zhai, J. Puigcerver, A. Kolesnikov, P. Ruyssen, C. Riquelme, M. Lucic, J. Djolonga, A. S. Pinto, M. Neumann, A. Dosovitskiy, L. Beyer, O. Bachem, M. Tschannen, M. Michalski, O. Bousquet, S. Gelly, N. Houlsby, A large-scale study of representation learning with the visual task adaptation benchmark, 2019. URL: https://arxiv.org/abs/1910.04867. doi:10.48550/ARXIV.1910.04867.

[14] D. Hendrycks, T. Dietterich, Benchmarking neural network robustness to common corrup-

tions and perturbations, 2019. arXiv:1903.12261.
[15] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, L. Schmidt, Measuring robustness to natural distribution shifts in image classification, 2020. URL: https://arxiv.org/abs/2007.00644. doi:10.48550/ARXIV.2007.00644.
[16] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. L. Zhu, S. Parajuli, M. Guo, D. X. Song, J. Steinhardt, J. Gilmer, The many faces of robustness: A critical analysis of out-of-distribution generalization, 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2020) 8320–8329.
[17] A. Barbu, D. Mayo, J. Alverio, W. Luo, C. Wang, D. Gutfreund, J. Tenenbaum, B. Katz, Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/97af07a14cacba681feacf3012730892-Paper.pdf.
[18] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, D. Song, Natural adversarial examples, 2019. URL: https://arxiv.org/abs/1907.07174. doi:10.48550/ARXIV.1907.07174.
[19] T. DeVries, G. W. Taylor, Improved regularization of convolutional neural networks with cutout, 2017. URL: https://arxiv.org/abs/1708.04552. doi:10.48550/ARXIV.1708.04552.
[20] S. Zheng, Y. Song, T. Leung, I. Goodfellow, Improving the robustness of deep neural networks via stability training, 2016. URL: https://arxiv.org/abs/1604.04326. doi:10.48550/ARXIV.1604.04326.
[21] L. Taylor, G. Nitschke, Improving deep learning using generic data augmentation, 2017. URL: https://arxiv.org/abs/1708.06020. doi:10.48550/ARXIV.1708.06020.
[22] S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, T. Mann, Data augmentation can improve robustness, 2021. URL: https://arxiv.org/abs/2111.05328. doi:10.48550/ARXIV.2111.05328.
[23] C. Fellbaum, Wordnet: An electronic lexical database (1998).
[24] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR abs/1409.1556 (2014).
[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818–2826. doi:10.1109/CVPR.2016.308.
[26] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.
[27] F. Chollet, Xception: Deep learning with depthwise separable convolutions, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 1800–1807.
[28] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17, AAAI Press, 2017, p. 4278–4284.
[29] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, CoRR abs/2201.03545 (2022). URL: https://arxiv.org/abs/2201.03545. arXiv:2201.03545.
[30] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, M. Shah, Transformers in vision: A survey, ACM Computing Surveys 54 (2022) 1–41. URL: https://doi.org/10.1145%2F3505244.

doi:10.1145/3505244.
[31] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jegou, Training data-efficient image transformers & distillation through attention, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 10347–10357.
[32] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, B. Guo, Swin transformer v2: Scaling up capacity and resolution, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11999–12009. doi:10.1109/CVPR52688.2022.01170.
[33] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, P. Dollár, Designing network design spaces, 2020.
[34] S. Paul, P.-Y. Chen, Vision transformers are robust learners, in: AAAI Conference on Artificial Intelligence, 2021.
[35] B. Recht, R. Roelofs, L. Schmidt, V. Shankar, Do imagenet classifiers generalize to imagenet?, in: International Conference on Machine Learning, 2019.
[36] R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge, F. A. Wichmann, Generalisation in humans and deep neural networks, 2018. URL: https://arxiv.org/abs/1808.08750. doi:10.48550/ARXIV.1808.08750.
[37] A. Laugros, A. Caplier, M. Ospici, Using synthetic corruptions to measure robustness to natural distribution shifts, ArXiv abs/2107.12052 (2021).
[38] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, D. Song, Natural adversarial examples, 2019. URL: https://arxiv.org/abs/1907.07174. doi:10.48550/ARXIV.1907.07174.
[39] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Š rndić, P. Laskov, G. Giacinto, F. Roli, Evasion attacks against machine learning at test time, in: Advanced Information Systems Engineering, Springer Berlin Heidelberg, 2013, pp. 387–402. URL: https://doi.org/10.1007%2F978-3-642-40994-3_25. doi:10.1007/978-3-642-40994-3_25.
[40] Y. Dong, Q.-A. Fu, X. Yang, T. Pang, H. Su, Z. Xiao, J. Zhu, Benchmarking adversarial robustness on image classification, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 318–328. doi:10.1109/CVPR42600.2020.00040.
[41] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, A. Madry, Robustness may be at odds with accuracy, 2018. URL: https://arxiv.org/abs/1805.12152. doi:10.48550/ARXIV.1805.12152.
[42] U. Ozbulak, E. T. Anzaku, W. D. Neve, A. V. Messem, Selection of source images heavily influences the effectiveness of adversarial attacks, ArXiv abs/2106.07141 (2021).
[43] Y. Wang, E. Ullah, P. Mianjy, R. Arora, Adversarial robustness is at odds with lazy training, ArXiv abs/2207.00411 (2022).
[44] F. Pinto, P. H. S. Torr, P. K. Dokania, An impartial take to the cnn vs transformer robustness contest, in: European Conference on Computer Vision, 2022.
[45] Z. Wang, Y. Bai, Y. Zhou, C. Xie, Can cnns be more robust than transformers?, ArXiv abs/2206.03452 (2022).
[46] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, A. Veit, Understanding robustness of transformers for image classification, 2021.
[47] W. Deng, S. Gould, L. Zheng, On the strong correlation between model invariance and generalization, ArXiv abs/2207.07065 (2022).

[48] E. Mintun, A. Kirillov, S. Xie, On interaction between augmentations and corruptions in natural corruption robustness, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems, volume 34, Curran Associates, Inc., 2021, pp. 3571–3583. URL: https://proceedings.neurips.cc/paper/2021/file/1d49780520898fe37f0cd6b41c5311bf-Paper.pdf.
[49] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, 2018, pp. 4510–4520. doi:10.1109/CVPR.2018.00474.
[50] B. Zoph, V. Vasudevan, J. Shlens, Q. Le, Learning transferable architectures for scalable image recognition, 2018, pp. 8697–8710. doi:10.1109/CVPR.2018.00907.
[51] G. Huang, Z. Liu, L. van der Maaten, K. Weinberger, Densely connected convolutional networks, 2017. doi:10.1109/CVPR.2017.243.
[52] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, 2019.
[53] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, 2021.

A. More CNN misclassifications

In Table 8, we present the continuation of the results present in Table 3 for the rest of the CNN models presenting non-zero accuracy. It becomes evident that the capacity of the classifier plays an important role in identifying relevant FP: MobileNetV2, which already demonstrated low accuracy scores, also fail to retrieve semantically related FP classes. This can be easily observed from the numerous red entries corresponding to this model.

Other than that, the results agree with the observations analyzed in Table 3, where ’egyptian cat’ label demonstrated many irrelevant FP, contrary to ’tabby cat’ or ’tiger cat’ labels.

Table 8
(Continuation of Tab. 3). Common misclassifications for selected GT cat classes and misclassification frequency for CNNs.

Model	GT	Top-1 FP	Top-1 MF	Top-2 FP	Top-2 MF	Top-3 FP	Top-3 MF
VGG 19	tabby	egyptian cat	28.00%	tiger cat	20.00%	lynx	16.00%
	angora	persian cat	34.78%	arctic fox	10.87%	egyptian cat	10.87%
	lynx	egyptian cat	20.00%	coyote	20.00%	timber wolf	20.00%
	siamese	whippet	16.67%	fur coat	16.67%	egyptian cat	16.67%
	tiger	tabby cat	34.88%	egyptian cat	16.28%	tiger	13.95%
	persian	lynx	20.00%	pekinese	25.00%	fur coat	10.00%
	cougar	lynx	45.45%	coyote	18.18%	timber wolf	9.09%
	leopard	egyptian cat	46.00%	lynx	16.00%	jaguar	12.00%
	egyptian	lynx	8.11%	mask	5.41%	book jacket	5.41%
	cat	fur coat	11.11%	snow leopard	11.11%	mousetrap	11.11%
Mobile Net V2	tabby	comic book	14.29%	mask	10.20%	sock	8.16%
	angora	shower curtain	20.00%	window screen	16.00	spotlight	8.00%
	lynx	west highland white terrier	8.00%	tiger	6.00%	traffic light	6.00%
	siamese	shower curtain	14.58%	sock	14.58%	mask	14.58%
	tiger	zebra	11.86%	mask	6.78%	maze	6.78%
	persian	spotlight	10.87%	shower curtain	8.70%	ant	8.70%
	cougar	comic book	14.00%	mask	8.00%	theater curtain	8.00%
	leopard	knot	14.00%	tiger	12.00%	mask	6.00%
	egyptian	windsor tie	10.00%	theater curtain	10.00%	spotlight	8.00%
	cat	shower curtain	10.53%	window screen	7.89%	teddy	7.89%
Effic ient Net	tabby	tiger cat	62.22%	egyptian cat	28.89%	persian cat	4.44%
	angora	persian cat	72.92%	egyptian cat	20.83%	tabby cat	2.08%
	lynx	egyptian cat	60.00%	tiger cat	20.00%	tabby cat	20.00
	siamese	egyptian cat	100.0%	-	-	-	-
	tiger	egyptian cat	34.21%	tabby cat	18.42%	tiger	15.79%
	persian	tabby cat	100.0%	-	-	-	-
	cougar	tiger cat	100.0%	-	-	-	-
	leopard	egyptian cat	56.00%	lynx	22.00%	tiger cat	16.00%
	egyptian	comic book	13.33%	mexican hairless	13.33%	lampshade	6.67%
	cat	macaque	33.33%	mexican hairless	33.33%	indigo bunting	33.33%
Conv Next	tabby	tiger cat	75.00%	egyptian cat	15.00%	web site	5.00%
	angora	persian cat	46.67%	egyptian cat	35.56	tabby cat	6.67%
	lynx	tabby cat	75.00	tiger cat	25.00%	-	-
	siamese	egyptian cat	75.00%	golden retriever	25.00%	-	-
	tiger	tabby cat	31.82%	egyptian cat	31.82%	tiger	9.09%
	persian	siamese cat	100.0%	-	-	-	-
	cougar	web site	100.0%	-	-	-	-
	leopard	egyptian cat	32.00%	lynx	18.00%	leopard	16.00%
	egyptian	mexican hairless	20.83%	mask	12.50%	comic book	12.50%
	cat	fur coat	50.00%	mexican hairless	50.00%	-	-