Superclass Learning with Representation Enhancement Supplemental Material

Zeyu Gan[1,2] Suyun Zhao[1,2,*] Jinlong Kang[2] Liyuan Shang[2]
Hong Chen[1,2] Cuiping Li[1,2]
[1]Key Lab of Data Engineering and Knowledge Engineering of MOE Renmin University of China
[2]Renmin University of China, Beijing, China
{zygan, zhaosuyun, kangjinlong, shangliyuan4032, chong, licuiping}@ruc.edu.cn

A. The Concept of Superclass

Superclass describes a situation where various categories share a common label. To be more specific, images under a superclass are not composed of the same thing with different perspectives, different poses or different breeds, just as the basic level class does, but is composed of similar or even completely distinct things according to the actual classification needs. The number of subclasses under one superclass may be extremely large. Therefore, one main characteristic of superclasses is their subclasses are huge and various.

In fact, superclass problems often occur in real-world or even in laboratories. The most classic scenario is refuse sorting, in which thousands of kinds of waste share only a handful of labels and it is resource-consuming or even impossible to label every kind of waste in training. And when sorting refuse, it is not wise and necessary to accurately identify what exact object they are. Fig. 1 gives a brief overview of refuse sorting problems and it shows the key differences between superclasses and basic-level classes. Another common scenario is to distinguish between pedestrians, animals, obstacles and vehicles. There are 4 superclasses and it is not necessary to identify objects accurately, such as cars, trucks and motorcycles.

B. Dataset Details

To simulate the superclass problems with known datasets, we select three benchmark datasets and two real-world datasets, and reorganize them into superclass datasets. To avoid contingency, we adopt a variety of different classification perspectives and numbers of categories. Tab. 1 shows our classification criteria in detail. The column *Superclass Labels* lists all the new labels of each reorganized dataset, and we control the balance of the number of samples per superclass when constructing. Taking CIFAR100-4 as an example, the whole structure and the specific correspondence between the superclass labels and the original labels are shown in Tab. 2.

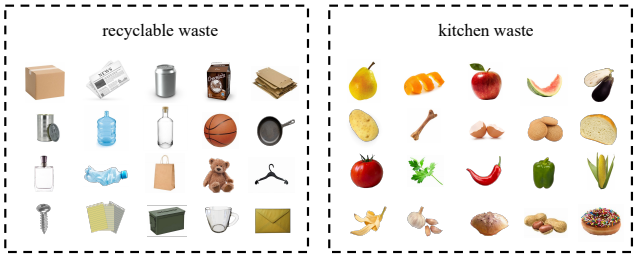

Figure 1. **Illustration of superclass.** Superclass describes a situation where various categories share only one common label. The number of subclasses under one superclass can be extremely large and it will be unrealistic to label every one of them.

C. Relationship between Superclass and Coarse-grained Classification

In past research, when it comes to discussing category hierarchy, most studies tend to focus on fine-grained classification [1, 4, 5], while coarse-grained classification received little attention. On the whole, coarse-grained classification aims to use only coarse-grained labels to directly complete the classification, which is a kind of weakly supervised learning. Though several existing works explored some coarse-grained situations [8, 10–12], we noticed that the so-called coarse- and fine-grained are just a pair of relative concepts [9, 13], and it depends on the research topic. To precisely define the superclass problem proposed in our work and distinguish it from previous research, Fig. 2 illustrates the hierarchy of classification. In our work, the superclass problem is at a higher level than previous topics, and the difficulty is about the various and distinct visual features in one superclass, while existing works do not.

Here we present the performance of an existing study on coarse-grained problem [12] on our constructed superclass

*Corresponding Author

Datasets	Superclass Labels	
CIFAR100-3	aquatic stationary	land-motional
CIFAR100-4	artifacts natures	mammals non mammals
CIFAR100-7	aquatics non small animals plants vehicles	household items outdoor scenes small animals
mini-ImageNet	animals	non animals
VOC	abiotic	biotic
FMoW	Africa Asia Oceania	Americas Europe Others
Adience	Age 0-2 Age 8-13 Age 25-32 Age 48-53	Age 4-6 Age 15-20 Age 38-43 Age 60+

Table 1. **Dataset details.** We reorganized CIFAR-100 in 3 different perspectives and adopt mini-ImageNet, VOC, FMoW and Adience as a supplement, forming 7 datasets to simulate the real superclass problems.

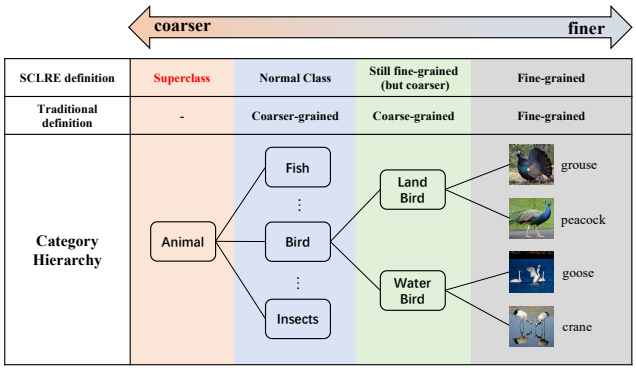

Figure 2. **An example of category hierarchy.** Since coarse and fine are relative concepts, previous studies tend to define them differently. The superclass we proposed in this paper is actually at a higher level in the hierarchy tree.

dataset, the result is shown in Tab. 3. Due to the difference between superclass problems and traditional coarse-grained problems, they fail to learn valid superclass-aware representations and then end up with bad performance.

D. Understanding SCLRE

The Relationship between CIA and Existing Attention Method. CIA explores the attention across different instances and thus learns a high-level superclass concept. Specifically, Eq.(1) in the manuscript shows that the attention is acted on the feature representation embedded from all the batch instances Z. Thus, a high-level representation crossed instances are enhanced by leveraging the attention relationships. While existing attention methods mainly explore the attention on representations from each single instance, then they learn an instance-level normal class concept.

An Analysis about Superclasses. To better understand SCLRE, We further calculated each superclass's accuracy and proportion-weighted total accuracy. The results are listed in Tab. 4. Since the datasets are constructed in a balanced way, we observed overall effectiveness across superclasses. However, there is still a gap in accuracy among different superclasses. For example, results of superclasses *Aquatics* and *Outdoor Scenes* are less attractive, while superclasses like *Stationary*, *Non-Mammals* and *Non-small Animals* claim higher accuracy.

To draw a conclusion, SCLRE performs better on superclasses that contain richer contents. Raw classes of *Aquatics* and *Outdoor Scenes* are more similar than that of *Stationary* and *Non-Mammals* (e.g. outdoor scenes are visually similar while items of various shapes can all be stationary). This phenomenon indicates that CIA explores and constructs the superclass concept more accessible when the superclass itself is informative enough. If the contents are quite similar, homogeneity across the contents will result in less effective representation enhancement. On the contrary, if the content is rich enough, the enhancement across instances can better explore the concept of superclass.

E. The Isotropy of Self-Supervised Contrastive Pretraining

When it comes to superclass problems, an obvious anomaly is why self-supervised contrastive models are still workable even without superclass labels. This is because the learning process of conventional self-supervised contrastive models is isotropic. In physics and mathematics, isotropy means exhibiting the property of being independent of direction. Here we use isotropy to express that self-supervised contrastive pretraining has no directivity because of the absence of supervised information.

Without any essential information about superclasses and/or basic-level classes, the self-supervised contrastive loss function just pulls instances and augmentations together and pushes other instances away in the embedding space. Accordingly, the self-supervised model is effective in the basic-level categories, as they can form natural clusters by themselves, but it is not the case for the superclass. Once the task becomes coarser than usual, self-supervised contrastive pretraining will show obvious uniformity and isotropy. Fig. 3 demonstrates this kind of isotropy of self-supervised contrastive pretraining in visualization. As shown in Fig. 3, self-supervised contrastive model treats both flowers and bottles as negative, pushing both of them away from the positive pair. Without directivity, the model

classes	CIFAR100-4	CIFAR-100	classes	CIFAR100-4	CIFAR-100	classes	CIFAR100-4	CIFAR-100	classes	CIFAR100-4	CIFAR-100
baby	0	2	squirrel	0	80	apple	2	0	bus	3	13
bear	0	3	tiger	0	88	cloud	2	23	can	3	16
beaver	0	4	whale	0	95	forest	2	33	castle	3	17
boy	0	11	wolf	0	97	maple_tree	2	47	chair	3	20
camel	0	15	woman	0	98	mountain	2	49	clock	3	22
cattle	0	19	aquarium_fish	1	1	mushroom	2	51	couch	3	25
chimpanzee	0	21	bee	1	6	oak_tree	2	52	cup	3	28
dolphin	0	30	beetle	1	7	orange	2	53	house	3	37
elephant	0	31	butterfly	1	14	orchid	2	54	keyboard	3	39
fox	0	34	caterpillar	1	18	palm_tree	2	56	lamp	3	40
girl	0	35	cockroach	1	24	pear	2	57	lawn_mower	3	41
hamster	0	36	crab	1	26	pine_tree	2	59	motorcycle	3	48
kangaroo	0	38	crocodile	1	27	plain	2	60	pickup_truck	3	58
leopard	0	42	dinosaur	1	29	poppy	2	62	plate	3	61
lion	0	43	flatfish	1	32	rose	2	70	road	3	68
man	0	46	lizard	1	44	sea	2	71	rocket	3	69
mouse	0	50	lobster	1	45	sunflower	2	82	skyscraper	3	76
otter	0	55	ray	1	67	sweet_pepper	2	83	streetcar	3	81
porcupine	0	63	shark	1	73	tulip	2	92	table	3	84
possum	0	64	snail	1	77	willow_tree	2	96	tank	3	85
rabbit	0	65	snake	1	78	bed	3	5	telephone	3	86
raccoon	0	66	spider	1	79	bicycle	3	8	television	3	87
seal	0	72	trout	1	91	bottle	3	9	tractor	3	89
shrew	0	74	turtle	1	93	bowl	3	10	train	3	90
skunk	0	75	worm	1	99	bridge	3	12	wardrobe	3	94

Table 2. **Details of CIFAR100-4.** To introduce more detailed information of those datasets and their structures inside, we use CIFAR100-4 as an example to list all the superclasses corresponding to each subclass.

	CIFAR100-3	CIFAR100-4	CIFAR100-7
Baseline	72.8	76.0	68.9
GEORGE [12]	68.8	65.8	43.7
SCLRE	80.1	84.0	78.1

Table 3. **Results of GEORGE on superclass problem.** When traditional coarse-grained classification models are applied to superclass problems, they may get even worse results than the baseline.

CIFAR100-3			CIFAR100-7		
Superclass	**Prop.**	**Acc.(%)**	**Superclass**	**Prop.**	**Acc.(%)**
Aquatics	18%	61.7	Aquatics	10%	66.9
Land-Motional	42%	82.5	Household Items	15%	83.9
Stationary	40%	87.9			
Total	-	80.1	Non-small Animals	15%	86.3
CIFAR100-4					
Superclass	**Prop.**	**Acc.(%)**	Outdoor Scenes	18%	66.9
Artifacts	30%	82.7			
Mammals	20%	71.1	Plants	22%	78.3
Natures	20%	84.8	Small Animals	10%	83.1
Non-Mammals	30%	92.4	Vehicles	10%	83.4
Total	-	83.7	Total	-	78.1

Table 4. **Accuracy of each superclass in CIFAR-100 datasets.** We calculate the proportion(Prop.) and accuracy(Acc.) of each category. Categories with better results are marked in gray, while categories that perform less well are marked in light gray.

effect on different flowers is almost the same as that on bottles, which presents the redundancy of self-supervised contrastive pretraining. What's more, the redundant boundaries generated during this process will result in fake and irrational superclasses being also effective in classification.

We conduct a brief test to further verify this phenomenon. We select orchids, roses and tulips as 3 subclasses of superclass A and select bottles as another irrelevant superclass B. ResNet-50 is chosen as the baseline to conduct normal training on superclass A and B, which is a real task. Then we take out the tulips from superclass A, change their labels to superclass B, and conduct fake tasks of self-supervised pretraining to find anomalies. Generally, self-supervised models can greatly improve the accuracy in real tasks. So what we are interested in is how would self-supervised models perform if subclasses and clusters are labeled on the contrary. To eliminate the interference of normal subclasses, we only focus on the classification results of tulips.

Tab. 5 presents the results of this brief experiment. It is expected that the accuracy of the baseline will decrease significantly in the fake task. It is observed that 57% of tulips are identified as bottles by baseline. However, it is unexpected that self-supervised models still follow the

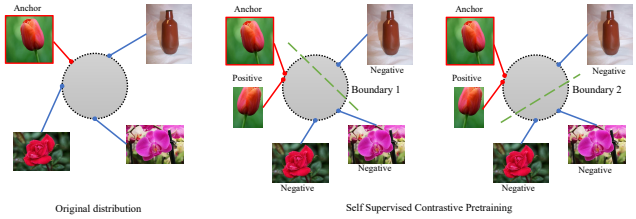

Figure 3. **The isotropy of self-supervised contrastive pretraining.** Due to the absence of supervised information, apart from different superclasses, self-supervised contrastive pretraining will push basic-level classes within the same superclass away, forming redundant boundaries, which interfere with the task.

	type	Percentage as flowers	Percentage as Bottles
Baseline	real task	94.0	6.0
Baseline	fake task	43.0	57.0
SimCLR [2]	fake task	25.0	75.0
MoCo v2 [3]	fake task	37.0	63.0

Table 5. **Classification result of tulips.** While self-supervised models generally improve accuracy in real tasks, We exchange the labels of tulips in downstream tasks and observe the accuracy changes. The result shows that even though tulips are labeled as irrelevant superclasses, self-supervised models still improve the task accuracy significantly according to their labels, but this improvement is fake and useless, which reveals the isotropy of self-supervised pretraining in superclass problems.

fake labels and continue to take effect. They significantly improve the percentage of classifying tulips into bottles, without any abnormality during the training process. In other words, self-supervised contrastive models are a mixed blessing. Whether tulips are regarded as flowers or bottles, self-supervised pretraining can improve accuracy. This is precisely because of the effect of isotropy in self-supervised contrastive training, which shows its limitations. Self-supervised models will not get any information about which superclass tulips are during the training process. What they do is simply complete the process of data augment and feature remapping at instance-level, and then hand it over to the downstream task. Therefore, once the task becomes coarser, no matter what label is given to the subclass, the self-supervised pretraining can improve the accuracy of the given label, which is unreasonable.

F. Target Generation

We explain the process of target generation in this subsection. The representations in superclass problems may have a very scattered distribution on the hypersphere embedding space. Pulling the representations close or away directly based on their superclass labels may fail due to the unclear class center. Because the class center can't be predicted, we first randomly generate several preset anchors for each category like [7], ensuring they're as far as possible from each other and then regarding them as the center of each category. In this way, each category has a predefined representation center, and we will be able to guide the scattered samples at the start of training to the same place. Given the category amount and the dimension of representations, we generate the anchors by minimizing the following loss function

$$l = \frac{1}{K}\sum_{i=1}^{K}\log\sum_{j=1}^{K} e^{t_i^T \cdot t_j/\tau},$$

where K demonstrates the category number and t_i demonstrates the representation of the i-th target anchor.

G. More Details of the Robustness of SCLRE

We design more experiments to verify the robustness of SCLRE by changing the backbone and calculate the accuracy on other datasets. We further apply this experiment to the CIFAR100-4 and CIFAR100-7 datasets.

As Tab. 6 illustrates, SCLRE can also make stable improvements on CIFAR100-4 and CIFAR100-7 datasets. On both smaller and larger convolutional neural networks, SCLRE improves performance in different degrees.

H. Proof of Generalization Ability

Proof of Lemma 1. Though we don't have a data augmentation process, the positive and negative pairs we create are still in a (σ, δ)-augmentation mode. As a conclusion in [6], if some basic conditions are matched, the generalization error of downstream classifier G_f has an upper bound:

$$Err(G_f) \leqslant (1-\sigma) + R_\epsilon. \tag{1}$$

When the encoder f is L-Lipschitz continuous, we have:

$$R_\epsilon^2 \leqslant \eta(\epsilon)^2 \cdot \mathbb{E}_v \mathbb{E}_{v_1, v_2 \in P(v)} ||v_1 - v_2||^2. \tag{2}$$

With the assumption that $||v|| = 1$, i.e. all the transformed feature vectors are L2-normalized, we have:

$$\mathbb{E}_v \mathbb{E}_{v_1, v_2 \in P(v)} v_1^T v_2 = \frac{1}{2}\mathbb{E}_{v, v'} \mathbb{E}_{\substack{v_1, v_2 \in P(v) \\ v^- \in P(v')}} ||v_1 - v_2||^2 - 1. \tag{3}$$

Therefore,

$$Err(G_f) \leqslant (1-\sigma) + \eta(\epsilon)\sqrt{2 - 2\mathbb{E}_v \mathbb{E}_{v_1, v_2 \in P(v)} v_1^T v_2}.$$

This finishes the proof.

The above lemma connects generalization error with L_1 and indicates that reducing the distances between features of positive samples helps to reduce the generalization error. But in superclass image recognition, the samples sharing the same coarse label don't necessarily share the same semantic information.

In this scenario, a traditional contrastive framework doesn't guarantee good alignment. Alternatively, our cross-instance attention module is shown to be successful in the experiment. In the following, we prove that the effect of the cross-instance attention module also limits the generalization error by constraining its upper bound.

Proof of Lemma 2. Noticing that in lemma 1, v_1, v_2 are samples pairs with same coarse label, we can further dived the condition into 2 parts:

1. v_1 and v_2 shares the same fine label and the same coarse label.

2. v_1 and v_2 have different fine label but the same coarse label.

	CIFAR100-4					CIFAR100-7				
	ResNet-18	**ResNet-34**	**ResNet-50**	**ResNet-101**	**ResNet-152**	**ResNet-18**	**ResNet-34**	**ResNet-50**	**ResNet-101**	**ResNet-152**
Baseline	78.3	80.9	76.0	78.6	78.2	73.3	74.9	70.1	73.5	71.9
SCLRE	82.0	82.9	84.0	83.1	83.3	76.1	77.0	78.1	77.0	77.0
Improve.(%)	**+3.7**	**+2.0**	**+8.0**	**+4.5**	**+5.1**	**+2.8**	**+2.1**	**+8.0**	**+3.5**	**+5.1**

Table 6. **Robustness on CIFAR100-4 and CIFAR100-7.** We verify the robustness of SCLRE on more datasets.

Assuming that there exists a ratio ρ that there are ρ pairs belong to occasion Item 1, and (1-ρ) pairs belong to occasion Item 2. Then the latter part of Eq. (4) can be diveded into:

$$\begin{aligned}\mathbb{E}_v\mathbb{E}_{v_1,v_2\in P(v)}v_1^Tv_2 = &\underbrace{\rho\cdot\mathbb{E}_v\mathbb{E}_{v_1,v_2\in F(v)}v_1^Tv_2}_{O_1}\\ &\underbrace{+(1-\rho)\cdot\mathbb{E}_v\mathbb{E}_{\substack{v_1,v_2\in P(v)\\ c(v_1)\neq c(v_2)}}v_1^Tv_2}_{O_2},\end{aligned} \tag{4}$$

where $F(\cdot)$ stands for all the samples share the same fine label.

In O_1, all the vectors share the same fine label, which means they have similar embeddings and semantic information. We can believe that majority of the samples in O_1 are close enough, which means for $(1-\epsilon_1)$ part of the sample, $s(v_1,v_2)\geqslant\varphi_1$. Then we have:

$$\mathbb{E}_v\mathbb{E}_{v_1,v_2\in F(v)}v_1^Tv_2\geqslant(1-\epsilon_1)\varphi_1. \tag{5}$$

In O_2, since the samples don't share any semantic information, we assume that $(1-\epsilon_2)$ part of the samples have different feature vector z. And for ϵ_2 part, considering that features z are not L2-normalized, we assume that $\inf_{i>j}(a_{1,i}a_{2,j}+a_{1,j}a_{2,i})z_i^Tz_j=\varphi_2$, where $a_{1,i}$ demonstrates the i-th value of vector a_1. Also, we can assume that the least L2-norm of z is K, which means $\forall z\in Z, ||z||^2\geqslant K$.

To analyze the O_2, we can bring the definition of cross-batch attention module into eq.(4), then the v_1 and v_2 in the O_2 part will be:

$$\begin{aligned}&M_1\cdot M_2\cdot v_1^Tv_2=\\ &\sum_{i=1}^n a_{1,i}a_{2,i}||z_i||^2+\sum_{i>j}(a_{1,i}a_{2,j}+a_{1,j}a_{2,i})z_i^Tz_j\\ \geqslant&\sum_{i=1}^n a_{1,i}a_{2,i}||z_i||^2+\binom{n}{2}\cdot\epsilon_2\cdot\varphi_2\\ \geqslant&\sum_{i=1}^n a_{1,i}a_{2,i}||z_i||^2+\frac{\epsilon_2\cdot n(n-1)}{2}\cdot\varphi_2\\ \geqslant&nK\cdot\sum_{i=1}^n a_{1,i}a_{2,i}+\frac{\epsilon_2\cdot n(n-1)}{2}\cdot\varphi_2\\ =&nK\cdot||a_1||\cdot||a_2||\cdot s(a_1,a_2)+\frac{\epsilon_2\cdot n(n-1)}{2}\cdot\varphi_2\\ \geqslant&K\cdot s(a_1,a_2)+\frac{\epsilon_2\cdot n(n-1)}{2}\cdot\varphi_2,\end{aligned} \tag{6}$$

where M_1, M_2 are L2-norm of v_1, v_2 before they are normalized, respectively. Since the attention vector a is an output after softmax-regulization, its leagth is at least $\sqrt{\frac{1}{n}}$.

Based on Eqs. (4) to (6), we can get the lower bound of $\mathbb{E}_v\mathbb{E}_{v_1,v_2\in P(v)}v_1^Tv_2$:

$$\begin{aligned}&\mathbb{E}_v\mathbb{E}_{v_1,v_2\in P(v)}v_1^Tv_2\geqslant\\ &\rho(1-\epsilon_1)\varphi_1+\frac{(1-\rho)\epsilon_2\cdot n(n-1)}{2M_1M_2}\varphi_2\\ &+\frac{(1-\rho)K}{M_1M_2}\mathbb{E}_{a_1,a_2\in O_2}s(a_1,a_2)\\ &=C_\varphi+\frac{(1-\rho)K}{M_1M_2}s(a_1,a_2)\end{aligned}$$

This finishes the proof.

The above lemma connects the alignment of positive samples with the similarity of their attention vector $s(a_1,a_2)$. We summarize the above ideas into the following theorem.

References

[1] Guy Bukchin, Eli Schwartz, Kate Saenko, Ori Shahar, Rogerio Feris, Raja Giryes, and Leonid Karlinsky. Fine-grained angular contrastive learning with coarse labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8730–8740, 2021. 1

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709, 2020. 4

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 4

[4] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4476–4484, 2017. 1

[5] Timnit Gebru, Judy Hoffman, and Li Fei-Fei. Fine-grained recognition in the wild: A multi-task domain adaptation approach. In *Proceedings of the IEEE international conference on computer vision*, pages 1349–1358, 2017. 1

[6] Weiran Huang, Mingyang Yi, and Xuyang Zhao. Towards the generalization of contrastive self-supervised learning. *arXiv preprint arXiv:2111.00743*, 2021. 4

[7] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogério Schmidt Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6908–6918, 2022. 4

[8] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pages 151–159, 2020. 1

[9] Zeyu Qiu, Minjie Hu, and Hong Zhao. Hierarchical classification based on coarse-to fine-grained knowledge transfer. *International Journal of Approximate Reasoning*, 149:61–69, 2022. 1

[10] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018. 1

[11] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 1

[12] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020. 1, 3

[13] Xinqi Zhu and Michael Bain. B-cnn: branch convolutional neural network for hierarchical classification. *arXiv preprint arXiv:1709.09890*, 2017. 1