PRETRAINED MODELS ARE ACTIVE LEARNERS

Anonymous authors
Paper under double-blind review

ABSTRACT

An important barrier to the safe deployment of machine learning systems is the risk of *task ambiguity*, where multiple behaviors are consistent with the provided examples. We investigate whether pretrained models are better active learners, capable of asking for example labels that *disambiguate* between the possible tasks a user may be trying to specify. Across a range of image and text datasets with spurious correlations, latent minority groups, or domain shifts, finetuning pretrained models with data acquired through simple uncertainty sampling achieves the same accuracy with **up to 6× fewer labels** compared to random sampling. Moreover, the examples chosen by these models are preferentially minority classes or informative examples where the spurious feature and class label are decorrelated. Notably, gains from active learning are not seen in unpretrained models, which do not select such examples, suggesting that the ability to actively learn is an emergent property of the pretraining process.

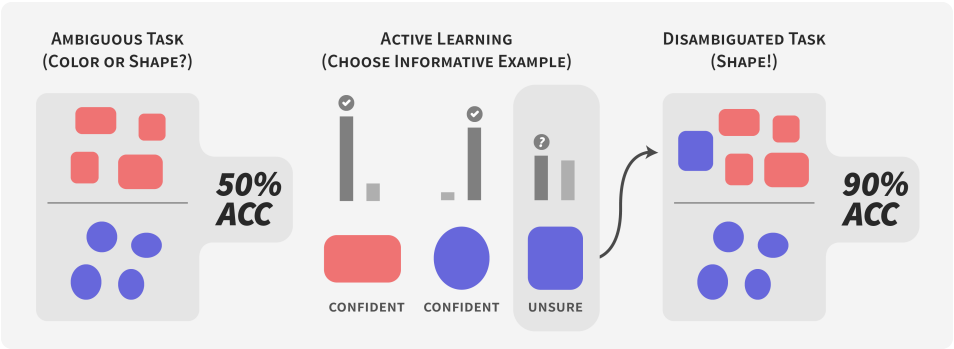


Figure 1: **Active learning can resolve task ambiguity, which is especially salient in few-shot settings.** Here, the provided training data leaves the model unsure of the task: is it to predict the shape or the color of the object? Pretraining enables models to disentangle and weigh various competing features, making them good active learners that can choose disambiguating examples (e.g. the blue square), resolving this task ambiguity.

1 INTRODUCTION

Modern pretrained models can be adapted to new tasks with remarkably little data, enabling downstream applications for tasks with only tens or hundreds of examples (Brown et al., 2020; Radford et al., 2021). However, an important but neglected challenge that is especially salient in few-shot settings is *task ambiguity*, when the desired behavior is not uniquely specified by the provided examples. Task ambiguity can manifest in different ways: For example, the class-relevant features of an input (e.g., the shape of an object in Figure 1) may be spuriously correlated with other features predictive of the training labels (e.g., the color of the object), making the desired task unclear. In addition, task ambiguity may arise from an underdiverse training set, causing models to be unsure of the desired behavior for minority groups or during distribution shifts that occur at test-time.

We consider whether the task ambiguity problem can be addressed through *active learning*, where models select informative examples for users to label. In principle, active learning allows models themselves to assist in resolving task ambiguity by identifying examples whose labels would be

informative; for example, in fig. 1, asking for the label of the blue square helps determine that the desired task is to predict the object shape not the object color. Since it may be challenging in general for users to intuit possible sources of task ambiguity, much less how to address them with example selection, an automated active learning approach may be desirable.

In traditional settings with smaller, unpretrained models, several challenges often prevent active learning from seeing success in practice, including label noise, unlearnable examples, and a lack of generalizability across models and tasks (Lowell et al., 2019; Karamcheti et al., 2021). However, using pretrained models in few-shot or low-data settings has several key differences from traditional ML pipelines which may make them particularly well suited for active learning. First, the effect of an individual data point (and the difference between an informative vs. an uninformative one) is magnified, compared to the thousands or millions of examples typically acquired through standard active learning pipelines. Second, pretrained models excel at learning high-level representations of inputs, which better surface relevant features and may encourage active learning to select examples that disambiguate between these high-level features (e.g. shape and color).

We consider the use of active learning on a range of spuriously correlated, imbalanced, or domain shifted datasets where task ambiguity is salient. We compare an active learning approach with a random-sampling baseline, and compare the difference in performance with and without the use of pretrained models. Our contributions are:

1. Identifying and motivating task ambiguity as a unified concern across a range of real-world datasets, and an especially salient problem for low-data settings.
2. Proposing active learning as a potential solution, along with a simple yet practical recipe that does not require adjustment for new datasets or tuning on validation data.
3. Showing that active learning can enable large gains (up to $6\times$ reduction in data points, +12% absolute gain for the same labeling budget) and presenting scaling trends demonstrating that this is an emergent property of pretraining.

2 METHOD

We study the pool-based active learning setup common in the literature (Settles, 2009), where we have a (possibly pretrained) model $\mathcal{M}$, a small *seed set* of training data $\mathcal{S} = \{(x_i, y_i)\}$, and a larger *pool* of unlabeled data $\mathcal{P} = \{x_i\}$. The active learning procedure proceeds as follows: first, finetune $\mathcal{M}$ on $\mathcal{S}$ until convergence; then, select points x_i from $\mathcal{P}$ that are deemed *most informative* according to an acquisition function $a(x; \mathcal{M})$, obtaining the corresponding labels y_i, until some budget k of data points is exhausted. This newly labeled batch $\mathcal{B} = \{(x_i, y_i)\}$ is then added to the existing data $\mathcal{S}$, and the model is retrained on $\mathcal{S}$ for the next acquisition step. This process is repeated for a fixed number of acquisition steps.

For our acquisition function, we adopt the classic *uncertainty sampling* approach to active learning, in particular the *least confidence* heuristic, where we acquire points for which our model is *least confident* in its predicted label (Settles, 2009). Specifically, treating the outputs of the model $\mathcal{M}$ as a probability distribution[1] over possible labels $p(y \mid x; \mathcal{M})$, we define the acquisition function to be

$$a(x; \mathcal{M}) = -\max_i p(y_i \mid x; \mathcal{M}) \tag{1}$$

Although we could use more complex measures such as the *entropy* over labels, the *least confidence* heuristic has shown to be simple and effective in a variety of settings (Settles, 2009; Hendrycks & Gimpel, 2017; Mussmann et al., 2020), and we similarly find good results here.

On top of this standard active learning pipeline, we propose the following change to improve the practical applicability of pretrained models in data-scarce settings:

Removing the need for a separate validation set The active learning cycle begins by finetuning the pretrained model on the seed set. If the size of the seed set is large enough, the seed set may

[1] While in general there is no guarantee that this probability distribution will be well-calibrated, recent work has found that pretraining improves model calibration across a variety of settings, including on out-of-domain data (Desai & Durrett, 2020; Hendrycks et al., 2019).

be partitioned into a training set and a validation set, and early stopping may be performed on the validation set. However, in few-shot settings, labeling costs may be high, and the seed set may be too small to meaningfully partition. This has led to a variety of ad-hoc solutions, e.g. finetuning for an arbitrary constant number of gradient steps depending on the task (Kolesnikov et al., 2020).

Instead of an arbitrary fixed number of finetuning steps, we propose an alternative method to terminate finetuning in the absence of a validation set. Specifically, we found that a simple but effective heuristic was to stop finetuning when the training loss decreases to 0.1% of the original training loss at the start of finetuning. In our experiments, this heuristic performed as well as early stopping on an actual validation set (see Appendix D for more details). By using a standardized recipe across tasks and removing the need for a separate validation set, our active learning pipeline is more robust to the real-world difficulties of deploying active learning in data-scarce settings where use of a validation set is impractical (Lowell et al., 2019; Perez et al., 2021), although further work is needed to capture the full extent of this recipe's generalizability.

3 Datasets

We consider a variety of datasets where task ambiguity is especially likely when only a few examples are provided, especially due to spurious correlations, latent minority classes, and domain shifts. These datasets provide an empirical testbed for the ability of pretrained models to choose disambiguating examples using active learning.

3.1 Distinguishing causal from spurious features

Spurious correlations arise when multiple features are predictive of the label in a training dataset, yet it is ambiguous which ones are causally linked to the task label (Geirhos et al., 2020). We consider two such datasets, and see whether active learning can choose the *disambiguating examples* where the spurious features are not copresent with the causal features:

Waterbirds The Waterbirds dataset (Sagawa et al., 2019) consists of photographs of landbirds or waterbirds digitally edited onto land or water backgrounds. The task is to classify whether the bird is a landbird or a waterbird. In the train set, 77% of the pictures feature landbirds and 23% waterbirds. 95% of both landbirds and waterbirds appear on land and water backgrounds, respectively. In the validation and test sets, this percentage is decreased to 50%, instead. Thus, the image background is a spurious feature the model may come to rely on when making the prediction.

Treeperson As the Waterbirds dataset was synthetically generated, we also consider a dataset where we perform classification over real, unedited images with spuriously correlated objects. We use the object annotations in Visual Genome (Krishna et al., 2016) to create a new dataset of 8,638 images called Treeperson, for which the task is to predict whether a person is in a given image. While 50% of the images contain a person in this dataset, each image also contains either a tree or a building, and the presence of these objects is spuriously correlated with the presence of people. At train time, 90% of training images with people contain a building, while 90% of training images without people contain a tree. Thus, a model may be incentivized to form representations that classify according to the presence of trees and buildings, rather than the presence of the actual causal variable of interest (people). These values are changed to 50% at test time, removing this correlation to evaluate how well the model learned the actual feature of interest. For more details on this dataset, see Appendix C.

3.2 Measuring robustness to distribution shift

Distribution shifts occur when algorithms are evaluated on different data distributions than the ones they were trained on. Examples include changing the location or time of day that photos were taken, or changing the topic or author of a particular textual source. These shifts can reduce performance, and we consider whether active learning can help choose diverse, informative examples that clarify how the model should behave over a range of natural distribution shifts.

iWildCam2020-WILDS This dataset considers the task of species classification from a database of photos taken from wildlife camera traps (Beery et al., 2020; Koh et al., 2021). The dataset is unbalanced, with most images containing no animal, and the distribution of camera locations and species changes between the in domain (ID) and out-of-domain (OOD) subsets.

Amazon-WILDS This dataset considers the task of predicting the number of stars associated with the text of a given Amazon review (Ni et al., 2019; Koh et al., 2021). The reviewers are different in the training set versus the test set, and the task is to perform as well as possible on this set of new reviewers. In addition to number of stars, we also consider model performance stratified by different product types, which highlights minority subgroups whose categorization is not visible to the model.

3.3 The role of data imbalance

Finally, we explore whether pretraining may improve active learning in *data imbalanced* settings, building off of work that shows the effectiveness of active learning for traditional models in such regimes (Ertekin et al., 2007). Here, *data imbalance* refers to an otherwise typical dataset where there are many more examples of some classes (*majority classes*) than others (*minority classes*). We compare active learning strategies on CIFAR-10, a typical balanced image-classification dataset, as well as an artificially-imbalanced variant where, at training time (but not test time), half of the classes have only 10% of their examples. See Appendix F for more details.

4 Experiments

4.1 Models and Training

Vision For computer vision datasets, we finetune BiT (Kolesnikov et al., 2020), a recently-proposed family of vision models which have achieved state-of-the-art performance on several vision tasks. We primarily consider the BiT-M-R50x1 model, pretrained on ImageNet-21k (Deng et al., 2009). To explore the effectiveness of larger architectures and pretraining sources, in Section 4.3.1 we also consider performance achieved by the same-size BiT-S-R50x1, trained on ImageNet-1k, and the deeper BiT-M-R101x1 model, also trained on ImageNet-21k. These models have been shown to have emergent few-shot learning abilities, where strong classifiers for new tasks can be obtained by simply finetuning on tens or hundreds of examples with typical gradient descent techniques (rather than meta-learning techniques, for example).

Text For the text dataset (Amazon), we use RoBERTa-Large (Liu et al., 2019), another pretrained model with similar properties as BiT, and a representative of the BERT (Devlin et al., 2019) family of models which together have obtained state-of-the-art scores on modern NLP benchmarks (Wang et al., 2018).

Other details, including hyperparameters and seed set/acquisition sizes are deferred to Appendix B.

Random acquisition baseline As a running baseline, we compare to the same model finetuned with a *random acquisition* function (equivalent to not doing active learning). That is, $a(x; \mathcal{M}) = \text{rand}(0, 1)$, so we simply sample a random batch of data from the pool at each acquisition step.

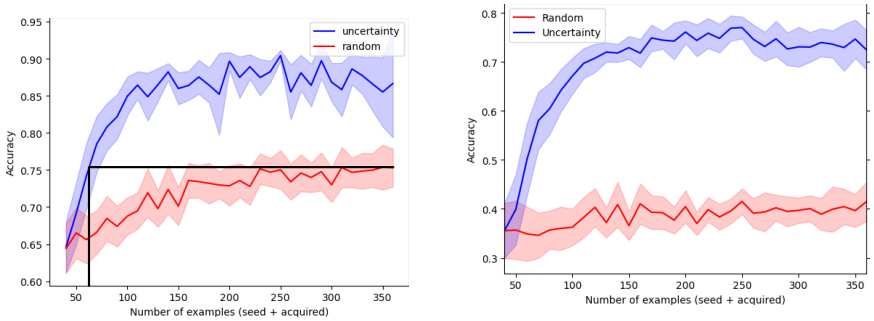
Comparison with unpretrained models To examine whether effective active learning is an emergent property of the pretraining process, we also compare to the performance observed when applying active learning to a *randomly initialized*, instead of pretrained, BiT-M-R50x1.

4.2 Accuracy per acquisition

For a general measure of success, we plot the accuracy of active learning versus random sampling on the validation datasets as a function of the number of samples acquired during training.

Waterbirds Waterbirds is evaluated on a balanced dataset where the foreground and background are not correlated. In this setting, uncertainty sampling achieves a **+12% improvement** in average validation accuracy over random sampling (Figure 2a). This comes primarily from a **+31% average**

increase across the landbird-on-water and waterbird-on-land images (i.e. those without the spurious correlation; Figure 2b). Uncertainty sampling required **5.8x fewer labels** than random sampling to achieve random sampling's final accuracy.



(a) Overall average accuracy

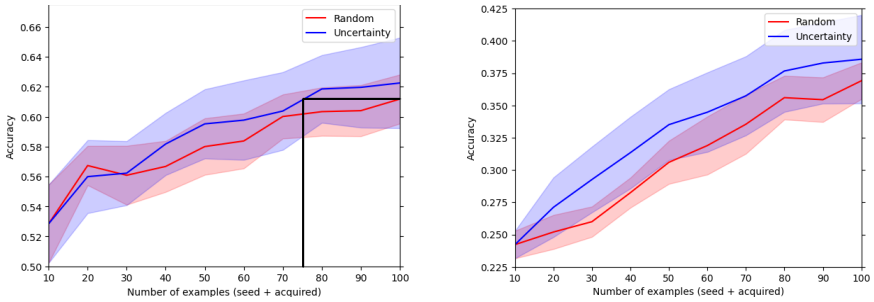
(b) Minority group accuracy (mismatched bird and background)

Figure 2: **Uncertainty sampling outperforms random sampling on Waterbirds, especially on minority subgroups.** Minority classes are waterbirds-on-land and landbirds-on-water. Shaded regions represent 95% CIs (Gaussian approx.).

Amazon In the Amazon dataset, we also see gains from active learning in the OOD setting, including **+1% on average** across reviewers, and **+2.5% on the worst 10th percentile** (Figure 3a). This result suggests that our active learning recipe may be of use outside of BiT or computer vision settings more broadly. The confidence not overlapping by the end of training is evidence that while the difference between uncertainty and random sampling is not large, it is statistically significant. Uncertainty sampling required **1.3x fewer labels** than random sampling to achieve random sampling's final accuracy.

iWildCam With the iWildCam dataset, uncertainty sampling achieved a **+8% improvement** upon random sampling. Uncertainty sampling also required **1.9x fewer labels** than random sampling to achieve random sampling's final accuracy (Figure 8b).

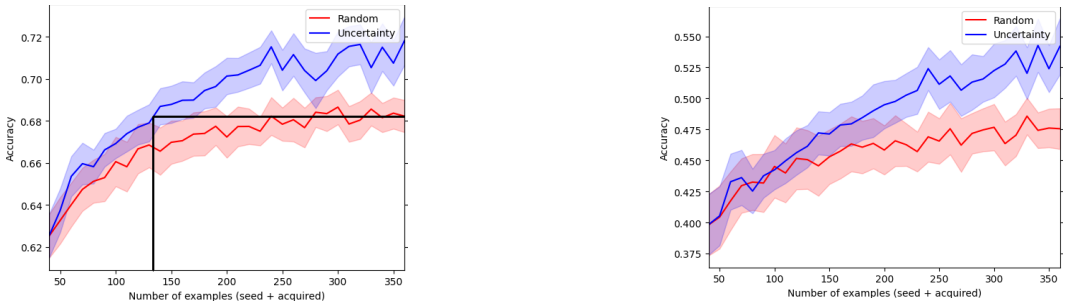
Treeperson In the Treeperson dataset, uncertainty sampling is **+4% improved** over random sampling by the end of training (Figure 4a). This gain is due to improvement on minority examples where the spurious and core features disagree; considering just these examples, uncertainty performs **+7% better** than random sampling at the end of training, with the gap growing over time (Figure 4b). Uncertainty sampling required **2.5x fewer labels** than random sampling to achieve random sampling's final accuracy.



(a) Overall average accuracy

(b) Worst 10th percentile group accuracy

Figure 3: **Uncertainty sampling outperforms random sampling on Amazon, especially on the worst 10th percentile of raters.** Shaded regions represent 95% CIs (Gaussian approx.).



(a) Overall average accuracy

(b) Minority group accuracy

Figure 4: **Uncertainty sampling outperforms random sampling on Treepereson, especially on minority subgroups.** Minority subgroups are people with trees and buildings without people. Shaded regions represent 95% CIs (Gaussian approx.).

4.3 What examples do models acquire?

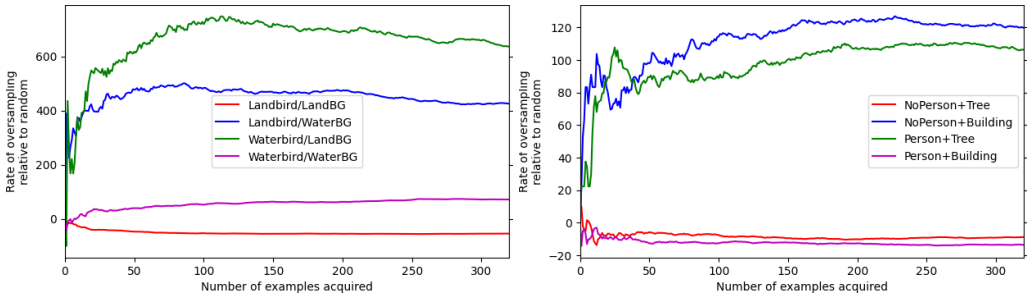
Overall, we attribute improved performance to pretrained models' ability to identify and preferentially sample disambiguating examples and latent minority examples.

Waterbirds Figure 5a depicts the rate at which uncertainty sampling is acquiring examples of each subgroup compared to the expected rate at which random sampling would acquire examples from those same subgroup. Waterbird/land-background and landbird/water-background, which take up approximately 1.1% and 3.8% of the training set, respectively, are being heavily oversampled. Even waterbird/water-background (21.8%) is being oversampled.

We emphasize that these minority examples are not simply members of the minority *class* (waterbirds). If active learning were simply upsampling minority classes, then all landbird images would be downsampled and all waterbird images would be upsampled. However, this is not the case—instead, the model identifies and preferentially upsamples *informative* examples where the spurious feature (background) and the core feature (bird type) disagree.

Treeperson For Treeperson we see the same pattern as in Waterbirds: the model identifies and upsamples examples where only one of the spurious or causal features is present (Figure 5b).

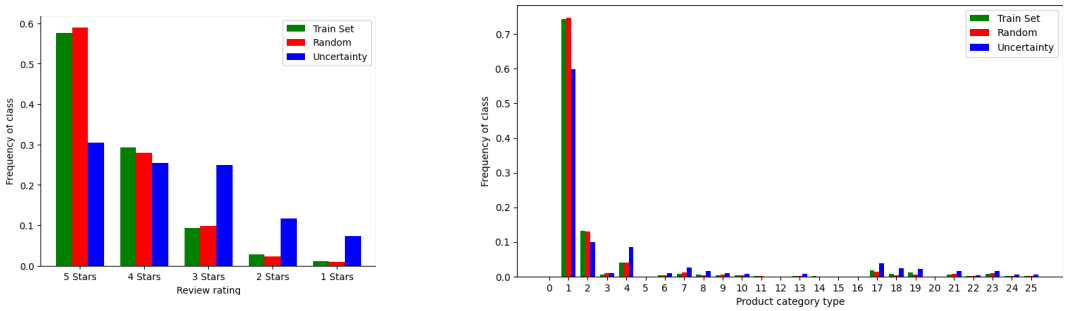
Amazon We also see similar behavior in the Amazon dataset, indicating our method's applicability to multiple modalities and pretrained models. Not only do we find that the model upsamples lower star ratings, which are less common in the data, it is also able to upsample rarer product categories—an unseen attribute.



(a) Waterbirds

(b) Treeperson

Figure 5: **Uncertainty sampling identifies and upsamples disambiguating examples.** For both Waterbirds and Treeperson, uncertainty sampling selectively acquires examples where the spurious and core features disagree. Y-axis: percentage oversampling of uncertainty over random sampling.



(a) Acquisitions by star rating

(b) Acquisitions by product category

Figure 6: **Uncertainty sampling upsamples both visible and latent minority subgroups.** Fraction of Amazon examples acquired by random and uncertainty sampling, stratified by star rating and product category. Uncertainty sampling preferentially acquires examples with lower star ratings and rarer product categories, despite the latter attribute not being visible to the model.

4.3.1 Scaling trends

We also consider the role of the pretraining dataset and model size. We conduct active learning experiments with an unpretrained BiT model, as well as models trained with different datasets or numbers of parameters.

Waterbirds We experimented with 3 pretrained models: BiT-S-R50x1, BiT-M-R50x1, BiT-M-R101x1, and their corresponding non-pretrained versions. For all pretrained models, uncertainty acquisition outperformed random acquisition (Figures 7a, 7b, 7c). Importantly however, we did not see this behavior with the non-pretrained models (Figure 7d). An analysis of acquisition trends on the Waterbirds dataset for pretrained versus non-pretrained models can be found in Appendix E.

iWildCam We experimented with 3 pretrained models: BiT-S-R50x1, BiT-M-R50x1, BiT-M-R101x1, and their corresponding non-pretrained versions. For all pretrained models, uncertainty acquisition outperformed random acquisition. Furthermore, the BiT-S-R50x1 model, which was pretrained on a smaller dataset than the BiT-M models (ImageNet-1k vs ImageNet-21k) achieves less of a boost from active learning (Figure 8a). This suggests that the gains from active learning may continue to grow as pretained models are trained for longer on more data. However, we did not see a difference between BiT-M-50x1 and BiT-M-101x1, which were trained on the same dataset but have different numbers of parameters (Figures 8b, 8c). We encourage future work that more thoroughly examines these behaviors and scaling laws across a range of models and datasets.

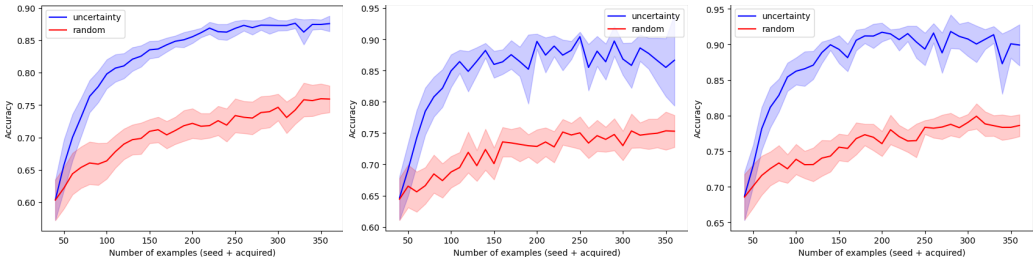
4.4 Failure Cases

To assess how similar the datasets needed to be to the training distribution of BiT, we also perform preliminary experiments on Camelyon17-WILDS (Bándi et al., 2019; Koh et al., 2021), which considers tumor identification from tissue patches, and FMoW-WILDS (Christie et al., 2018; Koh et al., 2021), which considers land-use classification from satellite images. Active learning performs comparably or worse than random sampling on these datasets, even when using a pretrained BiT model, suggesting that generalization to domains far from the training distribution (ImageNet-21k) may be challenging. However, Camelyon17-WILDS is also known to exhibit high variance across seeds,[2] which may also be a contributing factor.

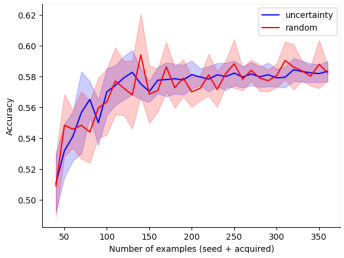
5 Related work

Task ambiguity and specification Several works address ambiguity or poor specification in machine learning problems. Taylor et al. (2020) describe the problem of "inductive ambiguity identification," and describe active learning as a promising potential solution that has failed to see practical

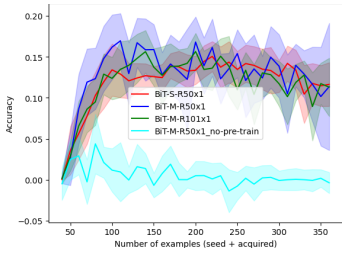
[2] https://wilds.stanford.edu/get_started/



(a) Pretrained BiT-S-R50x1 (b) Pretrained BiT-M-R50x1 (c) Pretrained BiT-M-R101x1



(d) Non-pretrained ResNet R50x1



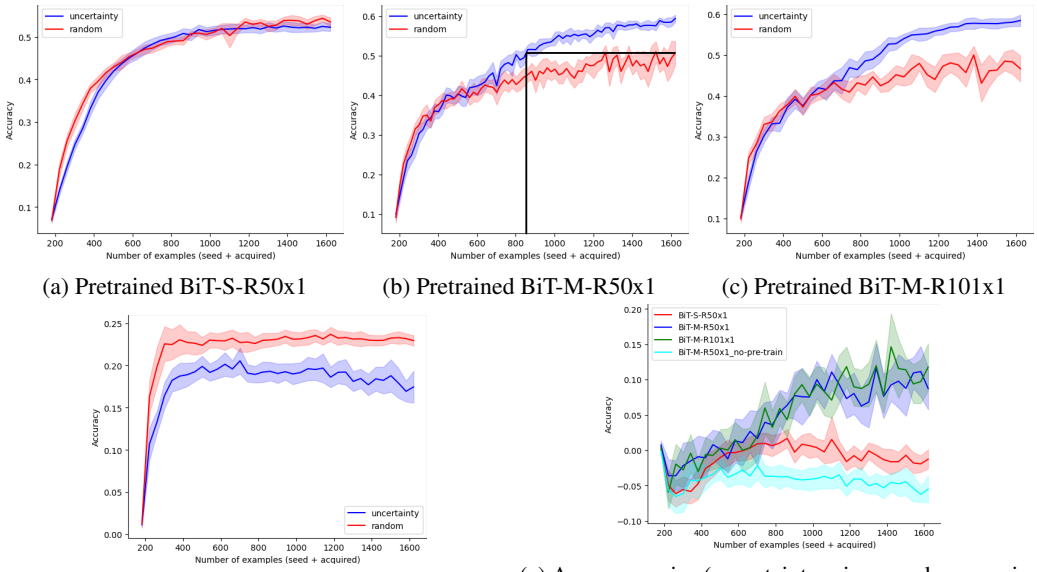
(e) Accuracy gains (uncertainty minus random acquisition) with different models

Figure 7: **Uncertainty sampling only provides gains over random sampling on Waterbirds when using pretrained models.** Shaded regions represent 95% CIs (Gaussian approx.).

success. D'Amour et al. (2020) describe the problem of *underspecification*, where high variance, instability, and poor model performance result from training overparameterized models on small amounts of data. Geirhos et al. (2020) describes how task ambiguity can arise when both desirable and undesirable features are predictive of the training labels, a problem which several works seek to better characterize and address (Nagarajan et al., 2021; Sagawa et al., 2019; 2020). Finally, Finn et al. (2018) address task ambiguity in few-shot settings via a probabilistic meta-learning algorithm, and perform an active learning experiment in a 1D regression setting. We build on these works by demonstrating that simple uncertainty sampling with pretrained models can be an effective approach to the task ambiguity problem across a wide variety of high-dimensional classification settings.

Uncertainty and distribution shift In the face of these challenges, several works have tried to quantify how much pretrained models know about problems or their own uncertainty about them. Rajpurkar et al. (2018) propose a question answering dataset with unanswerable questions, where a model must abstain rather than proceeding with an answer. Pretraining can also improve the calibration of model uncertainty (Hendrycks et al., 2019) and pretrained features can be used for out-of-distribution detection (Reiss et al., 2021; Wu & Goodman, 2020)—observations that align with our findings that uncertainty sampling can identify minority subgroups in datasets. Furthermore, our observation that upsampling latent minority groups results in better performance aligns well (Sagawa et al., 2020), which found that simply upweighting minority groups performed less well than increasing the relative fraction of minority group examples in the distribution. Importantly, however, our active learning setup does not require these groups to be known in advance.

Active learning and example selection Active learning (Lewis & Catlett, 1994; Settles & Craven, 2008; Settles, 2009; Houlsby et al., 2011; Aggarwal et al., 2020) is a well-studied field that investigates how machine learning algorithms might automatically select helpful additional data points to maximize their performance. Such strategies are especially helpful in imbalanced settings (Ertekin et al., 2007; Mussmann et al., 2020) and has been fruitfully applied to deep models (Gal et al., 2017; Beluch et al., 2018). Past work has also considered active learning for few-shot learning (Woodward & Finn, 2017). Our work extends these works into the regime of large pretrained models, showing how the application of these models to low-resource finetuning tasks is especially amenable to active learning strategies—perhaps due to the significance of each additional point required in a low-data setting. As we have shown, active learning strategies choose not only minority classes, but also



(a) Pretrained BiT-S-R50x1

(b) Pretrained BiT-M-R50x1

(c) Pretrained BiT-M-R101x1

(d) Non-pretrained ResNet R50x1

(e) Accuracy gains (uncertainty minus random acquisition) with different models

Figure 8: **Uncertainty sampling provides larger gains on iWildCam vs random sampling for pretrained models trained on more data.** Shaded regions represent 95% CIs (Gaussian approx.).

identify and oversample examples from *unlabeled* minority subgroups, as well as select examples which disambiguate causally-relevant features from spurious correlations.

Pretrained models and their emergent properties Our work contributes to a broader literature on how pretraining enables new kinds of model capabilities (Bommasani et al., 2021; Tamkin et al., 2021), especially when deployed in few-shot settings. For example, (Brown et al., 2020) identify the phenomenon of in-context learning, where tasks can be specified for models through a language modeling prompt, while Caron et al. (2021) discover that a self-supervised vision model implicitly learns high-quality segmentation maps visible through attention scores. Kaplan et al. (2020); Henighan et al. (2020) conduct scaling laws experiments which chart how capabilities emerge with scale. We identify a new model capability that emerges through pretraining: the capacity to actively learn and resolve task ambiguity in conceptually abstract hypothesis spaces.

6 Discussion and Limitations

We argue that pretrained models are good active learners, capable of identifying informative examples across a diverse range of settings where choosing examples can be challenging. Despite past work largely considering each separate problems, we find that active learning helps in cases where data is spuriously correlated, undergoes domain shift, or contains unlabeled subpopulations. These behaviors emerge most clearly as a result of large-scale pretraining, suggesting that active learning may be an underappreciated tool for increasing the reliability of systems in real-world settings. Importantly, pretraining appears necessary for these behaviors to manifest—and in some cases, models pretrained for longer on larger datasets appear to actively learn better.

Our method does suffer from a number of limitations. First, it requires a human in the loop for data acquisition, which significantly increases the time required to train a model compared to random sampling—a cost which must be weighed against potential benefits. Second, our method requires the labeling method to be relatively free of noise—this may be acceptable if annotators are domain experts or are well-trained, but may also increase the cost per acquired example.

Finally, we note the opportunity for much exciting future work, including deeper investigation of task ambiguity in real-world settings and better understanding how pretraining shapes active learning as models continue to scale.

REFERENCES

Umang Aggarwal, Adrian Popescu, and Céline Hudelot. Active learning for imbalanced datasets. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1428–1437, 2020.

Péter Bándi, Oscar G. F. Geessink, Quirine F Manson, Marcory Crf van Dijk, Maschenka C. A. Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe Thagaard, Anders Bjorholm Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Çetin, Eren Halıcı, Hunter Jackson, Richard Chen, Fabian Both, Jörg K.H. Franke, Heidi V. N. Küsters-Vandevelde, W. Vreuls, Peter Bult, Bram van Ginneken, Jeroen A. van der Laak, and Geert J. S. Litjens. From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 38:550–560, 2019.

Sara Beery, Elijah Cole, and Arvi Gjoka. The iwildcam 2020 competition dataset. *ArXiv*, abs/2004.10340, 2020.

William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9368–9377, 2018.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, D. Card, Rodrigo Castellon, Niladri S. Chatterji, Annie Chen, Kathleen Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, J.C. Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jackson K. Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *ArXiv*, abs/2108.07258, 2021.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *ArXiv*, abs/2104.14294, 2021.

Gordon A. Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6172–6180, 2018.

Alexander D'Amour, Katherine A. Heller, Dan I. Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yi-An Ma, Cory Y. McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin G. Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *ArXiv*, abs/2011.03395, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In *EMNLP*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, 2019.

Seyda Ertekin, Jian Huang, Leon Bottou, and Lee Giles. Learning on the border: active learning in imbalanced data classification. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 127–136, 2007.

Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. In *NeurIPS*, 2018.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pp. 1183–1192, 2017.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix Wichmann. Shortcut learning in deep neural networks. *ArXiv*, abs/2004.07780, 2020.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ICLR*, 2017.

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *ICML*, 2019.

T. J. Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling laws for autoregressive generative modeling. *ArXiv*, abs/2010.14701, 2020.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Jared Kaplan, Sam McCandlish, T. J. Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. Scaling laws for neural language models. *ArXiv*, abs/2001.08361, 2020.

Siddharth Karamcheti, Ranjay Krishna, Li Fei-Fei, and Christopher D. Manning. Mind your outliers! investigating the negative impact of outliers on active learning for visual question answering. *ArXiv*, abs/2107.02331, 2021.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Wei hua Hu, Michihiro Yasunaga, Richard L. Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *ICML*, 2021.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *ECCV*, 2020.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2016.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

David D Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pp. 148–156. Elsevier, 1994.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

David Lowell, Zachary Chase Lipton, and Byron C. Wallace. Practical obstacles to deploying active learning. In *EMNLP*, 2019.

Stephen Mussmann, Robin Jia, and Percy Liang. On the importance of adaptive data collection for extremely imbalanced pairwise tasks. In *FINDINGS*, 2020.

Vaishnavh Nagarajan, Anders Johan Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *ArXiv*, abs/2010.15775, 2021.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *EMNLP*, 2019.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True few-shot learning with language models. *ArXiv*, abs/2105.11447, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *ACL*, 2018.

Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation. In *CVPR*, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *ArXiv*, abs/1911.08731, 2019.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. *ArXiv*, abs/2005.04345, 2020.

Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 1070–1079, 2008.

Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. Understanding the capabilities, limitations, and societal impact of large language models. *ArXiv*, abs/2102.02503, 2021.

Jessica Taylor, Eliezer Yudkowsky, Patrick LaVictoire, and Andrew Critch. Alignment for advanced machine learning systems. 2020.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2018.

Mark P. Woodward and Chelsea Finn. Active one-shot learning. *ArXiv*, abs/1702.06559, 2017.

Mike Wu and Noah D. Goodman. A simple framework for uncertainty in contrastive learning. *ArXiv*, abs/2010.02038, 2020.

A CODE RELEASE

Our code and training scripts will be released at `https://redacted`.

B ADDITIONAL EXPERIMENTAL DETAILS

Here we provide some additional experimental details.

All BiT runs use the same default settings specified in the BiT paper (Kolesnikov et al., 2020) and the BiT GitHub repo: `https://github.com/google-research/big_transfer`

Settings that are constant across all BiT runs:

1. **Optimizer** - SGD with momentum 0.9
2. **Learning Rate** - Base learning rate 0.003. Linear warm up to this rate, then staircase decay. Exact schedule depends on dataset size, but for our few shot setting, this means: (1) linear warm up in the first 100 steps to 0.003, then (2) decay 10 fold every 100 steps; (3) after 500 steps, stop and move on to the next acquisition round.
3. **Data Augmentation** - Random cropping and flipping. See our `repo/utils/datasets/load` for details. Also available in BiT repo.
4. **Batch Size** - 32 when training, split into gradient accumulation microbatches of size 8.
5. **Early stopping condition** - When the training loss reaches below 0.001 times the original training loss.

Settings that varied between image datasets:

1. **Size of Initial Seed Set** - 40 for Waterbirds, 40 for Treeperson, 182 for iWildCam, 20 for CIFAR10, 8 for Camelyon17
2. **Size of Training Pool** - Entire training set for Waterbirds, Treeperson, Camelyon17. A random subset of 5000 examples and 12000 examples (re-drawn for each acquisition) from the entire training set of CIFAR10 and iWildCam, respectively.
3. **Number of Acquired Examples** - 320 for Waterbirds, 320 for Treeperson, 1456 for iWildCam, 180 for CIFAR10, 128 for Camelyon17
4. **Number of Examples Acquired Each Acquisition** - 5 for Waterbirds, 5 for Treeperson, 20 for iWildCam, 2 for CIFAR10, 1 for Camelyon17.

Settings used for Amazon + RoBERTa-Large runs:

1. **Optimizer** - AdamW with default hyperparameters ($\beta_1 = 0.99, \beta_2 = 0.999$, weight decay = 0.1).
2. **Learning Rate** - LR = 1e-6.
3. **Batch Size** - 2 (due to memory considerations).

   Early Stopping Condition When the training loss reaches below 0.001 times the original training loss or when 4000 gradient steps have been taken (whichever comes first).
4. **Size of Initial Seed Set** - 10.
5. **Size of Training Pool** - A random subset of 2000 examples (re-drawn for each acquisition) from the entire training set.
6. **Number of Acquired Examples** - 90.
7. **Number of Examples Acquired Each Acquisition** - 2.

C TREEPERSON DATASET

The Treeperson dataset is composed of images from Visual Genome (Krishna et al., 2016) with different compositions of detected objects.

Training set composition by subclass:

- Person and Building: 3700
- Person and Tree: 370
- No Person and Building: 370
- No Person and Tree: 3700

The validation set contains 498 examples of each subclass.

The following annotated objects were used to form the different subclasses:

- Person: person, people, man, men, woman, women
- Building: building, buildings
- Tree: tree, trees, leaf, leaves, grass

The training set and validation set were drawn randomly from qualifying images in Visual Genome's training set and validation set, respectively.

D EARLY STOPPING CONDITION

At the outset of this work, we explored if we could identify a heuristic for stopping training when there was no validation set present. We compared how the BiT model would perform if it stopped the finetuning step based off of when the validation accuracy plateaued versus when the training loss decayed to be 0.001 of its original value.

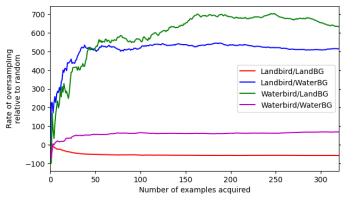
We ran a smaller experiment than the standard Waterbirds parameters we described in Appendix B. Namely, our seed set was of size 32, we acquired 64 examples on top of that, and we acquired one example at a time. We also defined the validation accuracy as having plateaued the fifth time it did not increase.

We ran twelve paired experiments where for twelve different randomized seed sets, we performed the Waterbirds experiment four times - with random sampling and stopping when validation accuracy plateaued, with random sampling and stopping when training loss decayed, with uncertainty sampling and stopping when validation accuracy plateaued, and with uncertainty sampling and stopping when training loss decayed.

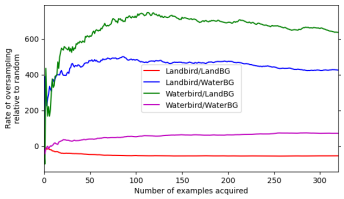
We found that these experiments achieved:

1. Random sampling + Stop from validation accuracy: average accuracy = 69.11%, average duration = 3.5 hours
2. Random sampling + Stop from training loss: average accuracy = 69.93%, average duration = 38 minutes
3. Uncertainty sampling + Stop from validation accuracy: average accuracy = 83.64%, average duration = 4.5 hours
4. Uncertainty sampling + Stop from training loss: average accuracy = 85.62%, average duration = 74 minutes

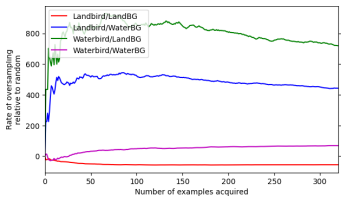
Thus, we concluded that the by stopping our finetuning step just by waiting for the training loss to decay to 0.001 of its original values, we could achieve comparable if not better accuracies, spend a fraction of the time, and remove the need for a labeled validation set.



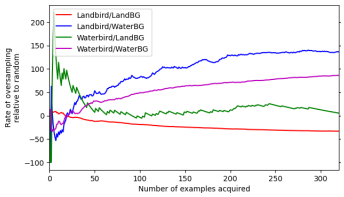
(a) Pretrained BiT-S-R50x1



(b) Pretrained BiT-M-R50x1



(c) Pretrained BiT-M-R101x1



(d) Non-pretrained ResNet R50x1

Figure 9: **All pretrained models acquire disambiguating subgroups much more efficiently than their non-pretrained counterparts.** The pretrained models do not simply oversample based on the the bird or the background; instead they oversample disambiguating examples which have mismatched backgrounds. By contrast, the non-pretrained model only oversamples images with a water background, and accordingly is less able to perform well on the balanced validation set.

E EFFECT OF MODEL SCALING AND PRETRAINING ON ACQUISITION

For the Waterbirds model scaling experiment, we track the examples each model acquires. These are presented in Figure 9. The acquisition patterns of all the pretrained models look fairly similar—they upsample both minority (landbird/water-background, waterbird/land-background) subclasses. However, the non-pretrained model is unable to capture that distinction, and is only able to upsample images with a water background, resulting in worse performance. A summary of final class acquisition ratios is available in Figure 10.

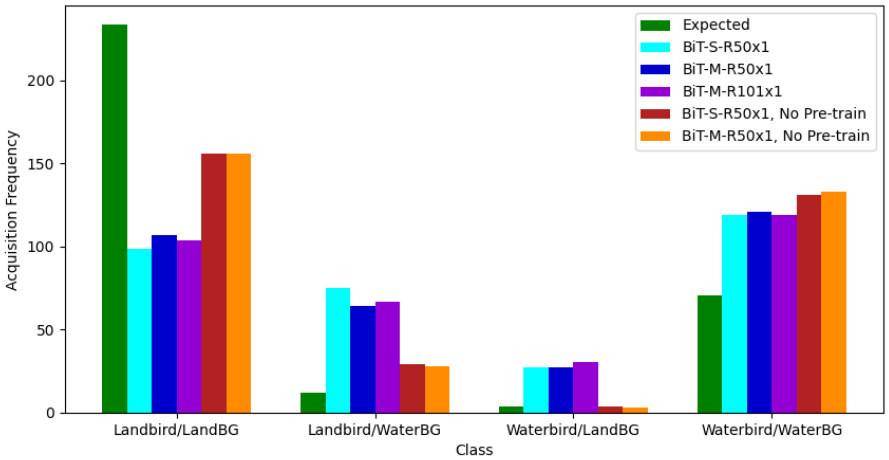
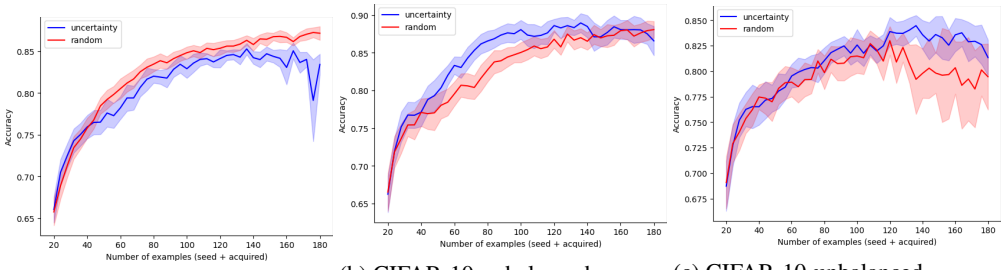


Figure 10: **Pretrained models ask for labels of images with mismatched backgrounds, while non-pretrained models do not.**

F ROLE OF LABEL IMBALANCE

We attempt to isolate the role of label imbalance by investigating CIFAR-10 (Krizhevsky, 2009), a widely-used balanced dataset where task ambiguity is not known to be a common problem. We train uncertainty- and random-sampling models on both the original CIFAR-10 dataset, as well as an imbalanceed variant where half of the classes have 90% of their examples removed. As shown in Figure 11, random sampling slightly outperforms uncertainty sampling on the original CIFAR-10 dataset, but uncertainty sampling performs better in unbalanced settings, both when the validation dataset is imbalanced as well as when it is balanced.



(a) CIFAR-10

(b) CIFAR-10-unbalanced (unbalanced validation)

(c) CIFAR-10-unbalanced (balanced validation)

Figure 11: Accuracy on CIFAR10 as more samples are acquired with uncertainty vs random acquisition. Shaded regions represent 95% CIs (Gaussian approx.). All runs are with the pretrained model BiT-M-R50x1. (a): The usual CIFAR10. (b) and (c): CIFAR10 where the training set has 5 out of 10 classes from which 90% samples are removed. (b) is accuracy on validation split with the same distribution as train. (c) is on balanced validation split.