Big Transfer (BiT): General Visual Representation Learning Supplementary Material

Alexander Kolesnikov[⋆], Lucas Beyer[⋆], Xiaohua Zhai[⋆],
Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby

Google Research, Brain Team
Zürich, Switzerland
{akolesnikov,lbeyer,xzhai}@google.com
{jpuigcerver,jessicayung,sylvaingelly,neilhoulsby}@google.com

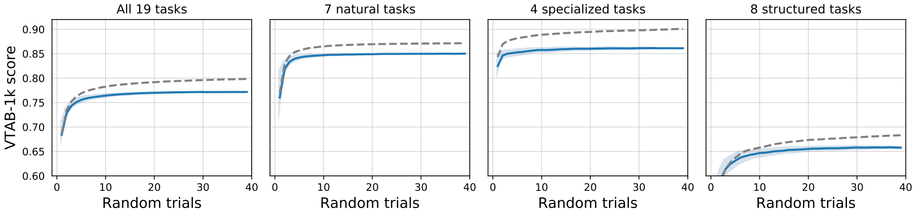

Fig. 1: Blue curves display VTAB-1k score (mean accuracy across tasks) depending on the total number of random hyperparameters tested. Reported VTAB-1k scores are averaged over 100 random hyperparameter orderings, the shaded blue area indicates the standard error. Dashed gray line displays the performance on the small hold-out validation split with 200 examples.

A Tuning hyperparameters for transfer

Throughout the paper we evaluate BiT using BiT-HyperRule. Here, we investigate whether BiT-L would benefit from additional computational budget for selecting fine-tuning hyperparameters.

For this investigation we use VTAB-1k as it contains a diverse set of 19 tasks. For each task we fine-tune BiT-L 40 times using 800 training images. Each trial uses randomly sampled hyperparameters as described below. We select the best model for each dataset using the validation set with 200 images. The results are shown in fig. 1. Overall, we observe that VTAB-1k score saturates roughly after 20 trials and that further tuning results in overfitting on the validation split.

[⋆] Equal contribution

This indicates that practitioners do not need to do very heavy tuning in order to find optimal parameters for their task.

After re-training BiT-L model with selected hyper-parameters using all union of training and validation splits (1000 images) we obtain the VTAB-1k score of 78.72%, an absolute improvement of 2.43% over 76.29% score obtained with BiT-HyperRule.

Our random search includes following hyperparameters with the following ranges and sampling strategies:

- Initial learning rate is sampled log-uniformly from the range $[10^{-1}, 10^{-4}]$.
- Total number of updates is sampled from the set $\{500, 1000, 2000, 4000, 8000, 16000\}$.
- Dropout rate for the penultimate layer is uniformly sampled from the range $[0.0, 0.7]$.
- Weight decay to the initial weight values is sampled log-uniformly from the range $[10^{-1}, 10^{-6}]$.
- MixUp α parameter is sampled from the set $\{\text{None}, 0.05, 0.1, 0.2, 0.4\}$.
- Input image resolution is sampled from the set $\{64, 128, 192, 256, 320, 384\}$.

Table 1: Performance of BiT-L on the original ("Full") and deduplicated ("Dedup") test data. The "Dups" column shows the total number of near-duplicates found.

	From JFT			From ImageNet21k			From ILSVRC-2012		
	Full	Dedup	Dups	Full	Dedup	Dups	Full	Dedup	Dups
ILSVRC-2012	87.8	87.9	6470	84.5	85.3	3834	80.3	81.3	879
CIFAR-10	99.4	99.3	435	98.5	98.4	687	97.2	97.2	82
CIFAR-100	93.6	93.4	491	91.2	90.7	890	85.3	85.2	136
Pets	96.8	96.4	600	94.6	94.5	80	93.7	93.6	58
Flowers	99.7	99.7	412	99.5	99.5	335	91.0	91.0	0

B Duplicates and near-duplicates

In order to make sure that our results are not inflated due to overlap between upstream training and downstream test data, we run extensive de-duplication experiments. For training our flagship model, BiT-L, we remove all images from JFT-300M dataset that are duplicates and near-duplicates of test images of all our downstream datasets. In total, we removed less than 50 k images from the JFT-300M dataset. Interestingly, we did not observe any drastic difference by doing de-duplication, evidenced by comparing the results (de-duplicated upstream) in the main paper and the first column of table 1 (full upstream).

In another realistic setting, eventual downstream tasks are not known in advance. To better understand this setting, we also investigate how duplicates affect performance by removing them from the downstream test data after the upstream model has already been trained. The results of this experiment are shown in table 1: "Full" is the accuracy on the original test set that contains near-duplicates, "Dedup" is the accuracy on the test set cleaned of near-duplicates, and "Dups" is the number of near-duplicates that have been removed from said test set. We observe that near-duplicates barely affect the results in all of our experiments. Note that near-duplicates between training and test sets have previously been reported by [8] for ILSVRC-2012, and by [1] for CIFAR.

In fig. 2, we present a few duplicates found between the ILSVRC-2012 training set and test splits of four standard downstream datasets.

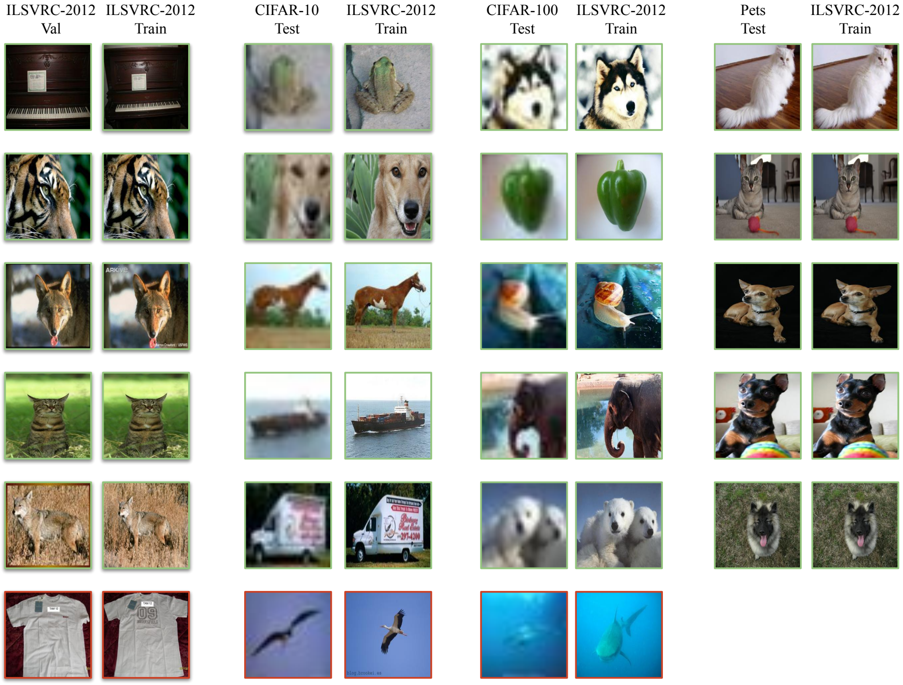

Fig. 2: Detected duplicates between the ILSVRC-2012 training set and test splits of various downstream datasets. Note that Flowers is not listed because there are no duplicates. Green borders mark true positives and red borders mark (rare) false positives.

C All of BiT-L's Mistakes

Here we take a closer look at the mistakes made by BiT-L[1]. Figure 3 and Figure 4 show *all* mistakes on the Pets and Flowers datasets, respectively. The first word always represents the model's prediction, while the second word represents the ground-truth label. The larger panels are best viewed on screen, where they can be magnified.

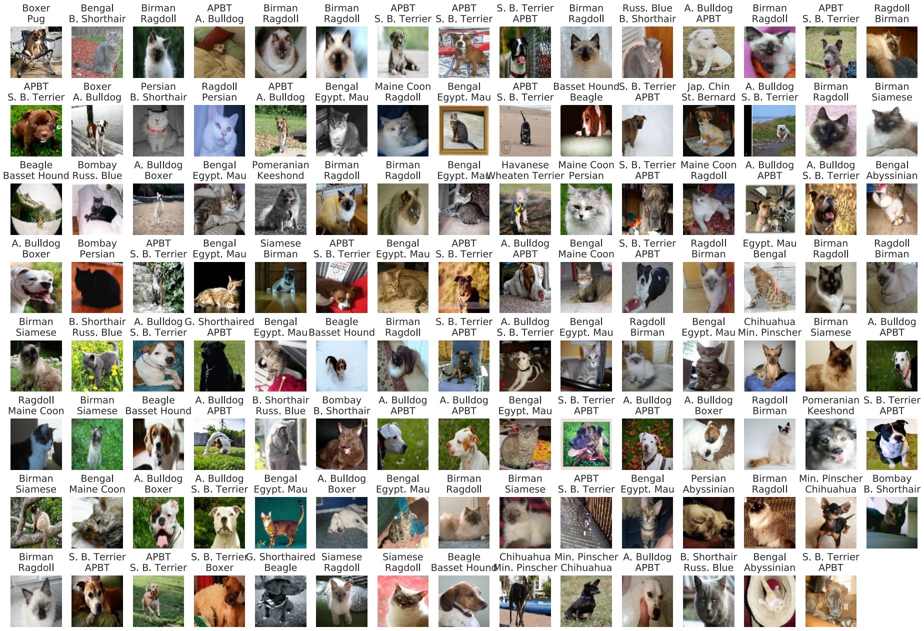

Fig. 3: All of BiT-L's mistakes on Oxford-IIIT-Pet.

[1] To be precise, the figures are obtained by an earlier version of our BiT-L model but which reaches almost the same accuracy. We did not re-run the figures and human evaluation with the latest model as they serve for illustration purposes and the models perform essentially the same, modulo a few flips.

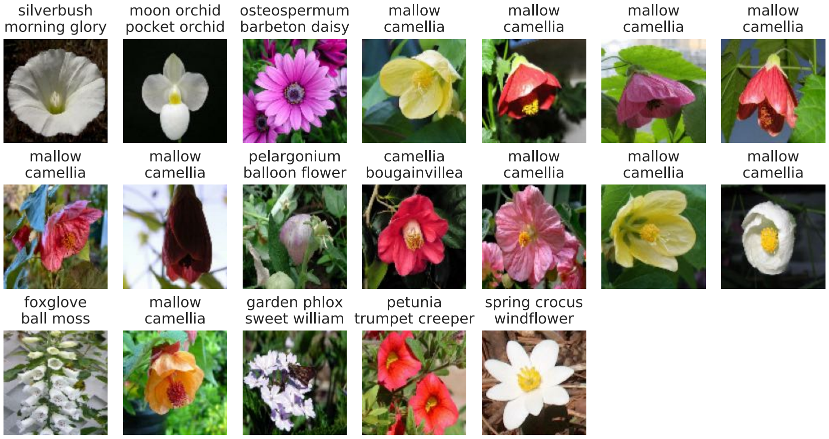

Fig. 4: All of BiT-L's mistakes on Oxford-Flowers102.

D Object detection experiments

As discussed in the main text, for object detection evaluation we use the RetinaNet model [4]. Our implementation is based on publicly available code[2] and uses standard hyper-parameters for training all detection models. We repeat training 5 times and report median performance.

Specifically, we train all of our models for 30 epochs using a batch size of 256 with stochastic gradient descent, 0.08 initial learning rate, 0.9 momentum and 10^{-4} weight decay. We decrease the initial learning rate by a factor of 10 at epochs number 16 and 22. We did try training for longer (60 epochs) and did not observe performance improvements. The input image resolution is 1024×1024. During training we use a data augmentation scheme as in [5]: random horizontal image flips and scale jittering. We set the classification loss parameters α to 0.25 and γ to 2.0, see [4] for the explanation of these parameters.

E Horizontal flipping and cropping for VTAB-1k tasks

When fine-tuning BiT models, we apply random horizontal flipping and cropping as image augmentations. However, these operations are not reasonable for certain VTAB tasks, where the semantic label (e.g. angle, location or object count) is not invariant to these operations.

Thus, we disable random horizontal flipping as preprocessing for dSprites-orientation, SmallNORB-azimuth and dSprites-location tasks. Random cropping preprocessing is disabled for Clevr-count, Clevr-distance, DMLab, KITTI-distance and dSprites-location tasks.

[2] https://github.com/tensorflow/tpu/tree/master/models/official/retinanet

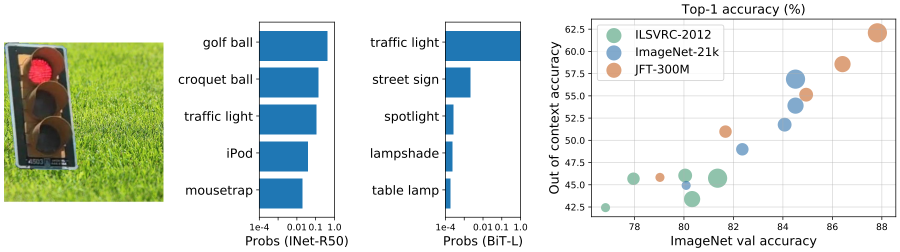

Fig. 5: **Left:** Top 5 predictions produced by an ILSVRC-2012 model (IN-R50) and BiT-L on an example out-of-context object. Bar lengths indicate predictive probability on a log scale. **Right:** Top-1 accuracy on the ILSVRC-2012 validation set plotted against top-1 accuracy on objects out-of-context. The legend indicates the pre-training data. All models are subsequently fine-tuned on ILSVRC-2012 with BiT-HyperRule. Larger markers size indicates larger architectures

F Robustness: Objects out-of-context

It has been shown that CNNs can lack robustness when classifying objects out-of-context [2,6,7]. We investigate whether BiT not only improves classification accuracy, but also out-of-context robustness. For this, we create a dataset of foreground objects corresponding to ILSVRC-2012 classes pasted onto miscellaneous backgrounds (fig. 5 left). We obtain images of foreground objects using OpenImages-v5 [3] segmentation masks. Figure 5 shows an example, and more are given in fig. 6. Sometime foreground objects are partially occluded, resulting in an additional challenge.

We transfer BiT models pre-trained on various datasets to ILSVRC-2012 and see how they perform on this out-of-context dataset. In fig. 5 we can see that the performance of models pre-trained on ILSVRC-2012 saturates on the out-of-context dataset, whereas by using more data during pre-training of larger models, better performance on ILSVRC-2012 *does* translate to better out-of-context performance.

More qualitatively, when we look at the predictions of the models on out-of-context data, we observe a tendency for BiT-L to confidently classify the foreground object regardless of the context, while ILSVRC-2012 models also predict objects absent from the image, but that could plausibly appear with the background. An example of this is shown in fig. 5 left.

F.1 Out of context dataset details

We generate this dataset by combining foreground objects extracted from OpenImages V5 [3] with backgrounds, licensed for reuse with modification, mined from search engine results.

Foreground objects In this study, we evaluate models that output predictions over ILSVRC-2012 classes. We therefore fine-tune BiT models on ILSVRC-2012 using BiT-HyperRule. We choose 20 classes from OpenImages that correspond to one such class or a subset thereof. These 20 classes cover a spectrum of different object types. We then extract foreground objects that belong to these classes from images in OpenImages using the provided segmentation masks. Note that this leads to some objects being partially occluded; however, humans can still easily recognize the objects, and we would like the same from our models.

Backgrounds We define a list of 41 backgrounds that cover a range of contexts such that (1) we have reasonable diversity, and (2) the objects we choose would not likely be seen in some of these backgrounds. We then collect a few examples of each background using a search engine, limiting to results licensed for reuse with modification. We take the largest square crop of the background from the top left corner.

We paste the extracted foreground objects onto the backgrounds. This results in a total of 3321 images in our dataset (81 foreground objects × 41 backgrounds). We fix the size of the objects such that the longest side corresponds to 80% of the width of the background; thus, the object is prominent in the image.

Figure 6 shows more examples of out-of-context images from our dataset, contrasting the predictions given by a standard ResNet50 trained on ILSVRC-2012 from scratch and the predictions of BiT-L fine-tuned on ILSVRC-2012.

F.2 Image Attributions

In this section we provide attributions for images used to generate the examples from the out-of-context dataset.
All images are licensed CC-BY-2.0 unless noted otherwise.

Foreground objects:

- Traffic light: U turn to Tophane by *Istanbul Photo Guide*.
- Sofa: Welcome by *woot*.
- Zebra: i like his tail in this one by *meg and rahul*.
- Starfish: Starfish by *Summer Skyes 11*.
- Limousine: Hummer limousine stopping at the door [nb: title translated] by *duncan_su*.

Backgrounds:

- Grass: Photo by *zoosnow*
  (Pexels license; Free to use, no attribution required).
- Wood: Tree Bark Texture 04 by *Jacob Gube, SixRevisions*.
- Street at night: City street calm buildings by *csr_ch*
  (Pixabay license; Free for commercial use, no attribution required).
- Underwater: Photo by *MaxX42*
  (Pixabay license; Free for commercial use, no attribution required).
- Kitchen: Interior of a modern modular home by *Riverview Homes, Inc.*
  (CC-BY-SA-3.0 Unported license).

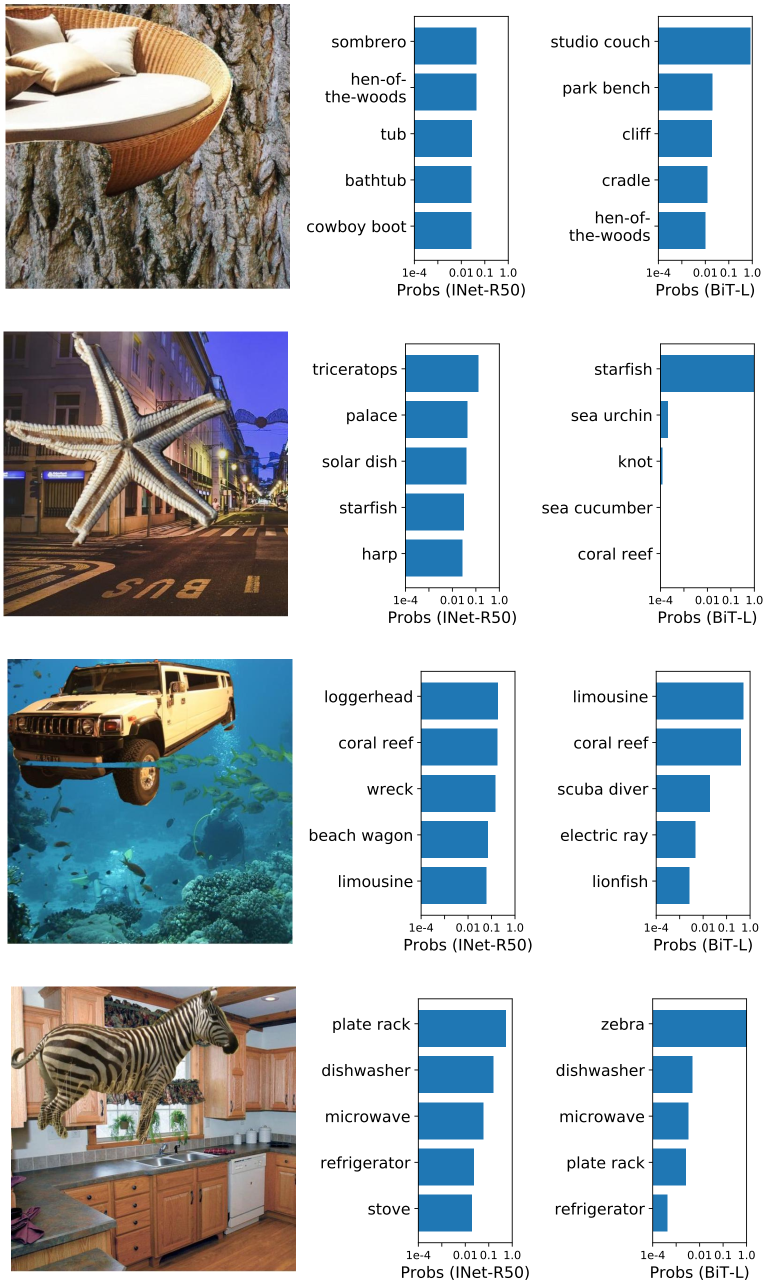

Fig. 6: Top 5 predictions produced by an ILSVRC-2012 model (INet-R50) and BiT-L on examples of out-of-context objects. Bar lengths indicate predicted probability on a log scale. We choose images that highlight the qualitative differences between INet-R50 and BiT-L predictions when the INet-R50 model makes mistakes.

References

1. Barz, B., Denzler, J.: Do we train on test data? purging CIFAR of near-duplicates. arXiv preprint arxiv:1902.00423 (2019)
2. Beery, S., Horn, G.V., Perona, P.: Recognition in terra incognita. arXiv preprint arXiv:1807.04975 (2018)
3. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., Ferrari, V.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. arXiv:1811.00982 (2018)
4. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV (2017)
5. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014)
6. Peyre, J., Laptev, I., Schmid, C., Sivic, J.: Weakly-supervised learning of visual relations. arXiv preprint arXiv:1707.09472 (2017)
7. Shetty, R., Schiele, B., Fritz, M.: Not using the car to see the sidewalk: Quantifying and controlling the effects of context in classification and segmentation. arXiv preprint arXiv:1812.06707 (2018)
8. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era. In: ICCV (2017)