TU Delft

How Does the Downstream Accuracy of Barlow Twins Scale with Pre-training Dataset Size?

A small-compute characterization with a ViT-Tiny on Tiny-ImageNet subsets · Yan Olerinskiy

Supervisors: Jan van Gemert, Alex Manolache, Petter Reijalt · EEMCS, Delft University of Technology

At this small scale, the amount of data is not the only thing that matters: the kind of downstream task and the checkpoint we keep matter just as much.

1. Why it matters

- Vision models are often pre-trained once and reused for many new tasks.
- Labels are expensive, so it's common to use **self-supervised learning (SSL)**, which learns from unlabeled images.
- SSL is usually used with *millions* of images, does it work with less?

2. Why Barlow Twins

- A well-established self-supervised method.
- No negative pairs, no momentum encoder, no large batches, simple and cheap.
- Fits the small-data setup, but was never measured at this scale.

3. Research questions

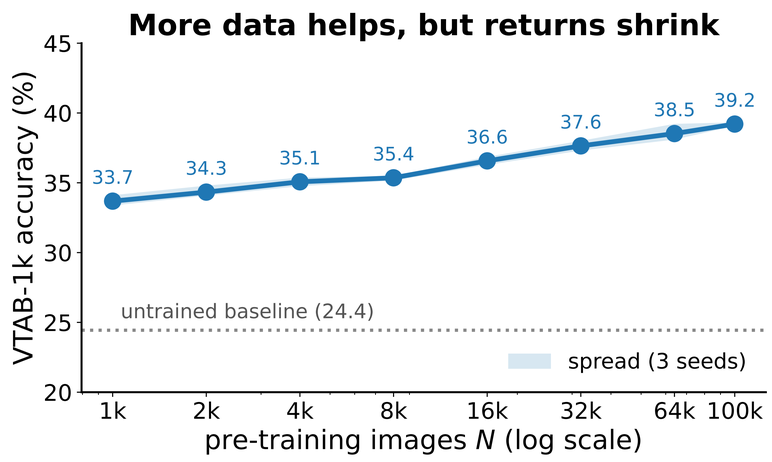
For Barlow Twins on a small vision transformer:

- How does accuracy on new tasks change with the number of unlabeled pre-training images?
- How does this depend on the *kind* of downstream task?
- How does the representation behave at the *smallest* dataset sizes?

4. Setup

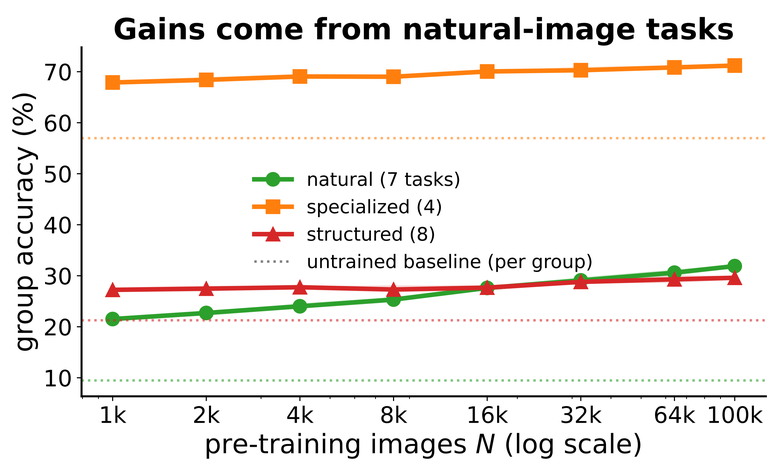
- ViT-Tiny/8 (5.4M params), pre-trained on Tiny-ImageNet subsets of 1k to 100k images.
- Same 1000 epochs every run, only the data size changes.
- Same schedule for every size, keep the best-validation checkpoint.
- Freeze each model, measure transfer to the 19 **VTAB-1k** tasks (natural / specialized / structured).
- Report mean of 3 seeds and min-max band.

5. Q: Does more data help? A: Yes, with diminishing returns



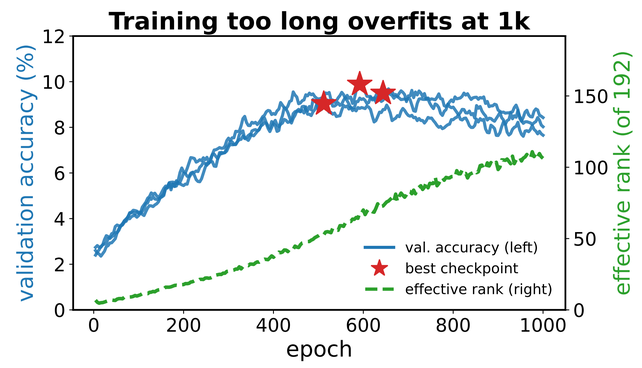
- Beats the untrained baseline at every size.
- Diminishing returns: $4\times$ the data (1k→4k) adds only $+1.4$ points, but the curve is still rising at 100k.

6. Q: Does the average tell the whole story? A: No



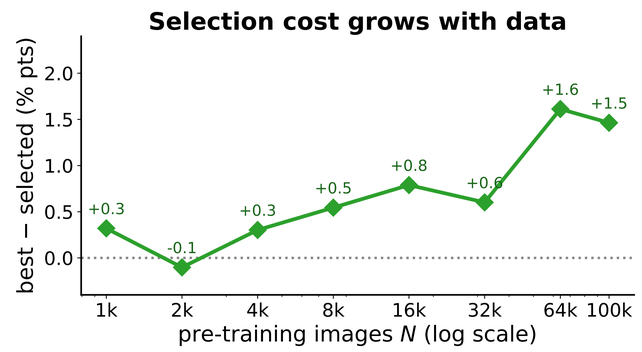
- Almost all the gain is on natural-image tasks.
- Specialized (medical, satellite) improves slightly, and structured (counting, geometry) is flat across.

7. Q: Does training longer help at 1k? A: No, the accuracy degrades



- Validation accuracy peaks near epoch 580, then falls ∼15%.
- Effective rank stays high, so representations don't collapse. We interpret this decline as overfitting.

8. Q: Does the validation selection pick the best checkpoint? A: Less and less as data grows



- The best-validation checkpoint drifts from the best-for-transfer one.
- Gap ir near zero at 1k and grows to ∼1.5 points at 64k–100k.

9. Limitations

- One backbone, one pre-training dataset (ViT-Tiny/8 on Tiny-ImageNet).
- The probe reports its best epoch, so absolute numbers are optimistic.
- No fine-tuning evaluation.