LEEP: A New Measure to Evaluate Transferability of Learned Representations — Supplementary Material —

Cuong V. Nguyen [1] **Tal Hassner** [2] **Matthias Seeger** [1] **Cedric Archambeau** [1]

We present proofs of theoretical claims and the full experimental results discussed in the main paper.

1. Proofs

1.1. Proof of Property 1

This proof is straight-forward because $l(w, k^*)$ is the maximal average log-likelihood over $k \in \mathcal{K}$, $T(\theta, \mathcal{D})$ is the average log-likelihood of the EEP, and the EEP is in $\mathcal{K}$. Thus, $T(\theta, \mathcal{D}) \leq l(w, k^*)$.

1.2. Proof of Property 2

Let $Z = (z_1, z_2, \ldots, z_n)$ be the dummy labels of $(x_1, x_2, \ldots, x_n)$ obtained when computing the NCE, and let $Y = (y_1, y_2, \ldots, y_n)$ be the true label set. We have:

$$T(\theta, \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\sum_{z\in\mathcal{Z}} \hat{P}(y_i|z)\ \theta(x_i)_z\right) \quad \text{(by definition)}$$

$$\geq \frac{1}{n}\sum_{i=1}^{n} \log\left(\hat{P}(y_i|z_i)\ \theta(x_i)_{z_i}\right) \quad \text{(monotonicity of log)}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \log \hat{P}(y_i|z_i) + \frac{1}{n}\sum_{i=1}^{n} \log \theta(x_i)_{z_i}.$$

According to the proof of Theorem 1 of Tran et al. (2019), we have:

$$\mathrm{NCE}(Y|Z) = \frac{1}{n}\sum_{i=1}^{n} \log \hat{P}(y_i|z_i).$$

Thus, we have:

$$T(\theta, \mathcal{D}) \geq \mathrm{NCE}(Y|Z) + \frac{1}{n}\sum_{i=1}^{n} \log \theta(x_i)_{z_i}.$$

[1]Amazon Web Services [2]Facebook AI (Work done before joining Facebook). Correspondence to: Cuong V. Nguyen <nguycuo@amazon.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

2. Full Experimental Results

Fig. 1 shows the results for all experimental settings with small balanced target data sets.

Fig. 2 shows the results for all experimental settings with small imbalanced target data sets.

Fig. 3 shows the results for all experimental settings with the convergence speed of fine-tuned models. For a clearer comparison, we only consider two LEEP transferability levels for target tasks constructed from FashionMNIST.

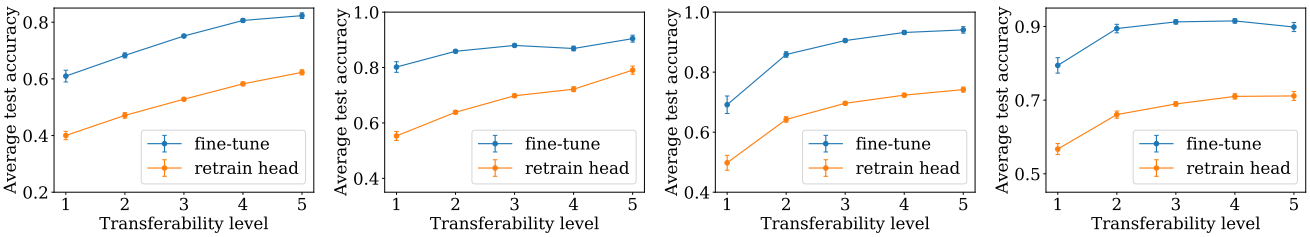
Fig. 4 shows the results for all experimental settings in the source model selection problem.

References

Bao, Y., Li, Y., Huang, S.-L., Zhang, L., Zheng, L., Zamir, A., and Guibas, L. An information-theoretic approach to transferability in task transfer learning. In *IEEE International Conference on Image Processing*, pp. 2309–2313, 2019.

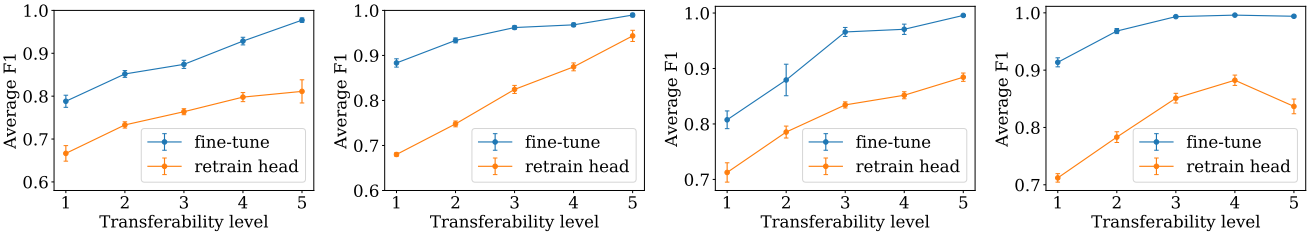
Kornblith, S., Shlens, J., and Le, Q. V. Do better imagenet models transfer better? In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2661–2671, 2019.

Tran, A. T., Nguyen, C. V., and Hassner, T. Transferability and hardness of supervised classification tasks. In *International Conference on Computer Vision*, pp. 1395–1405, 2019.



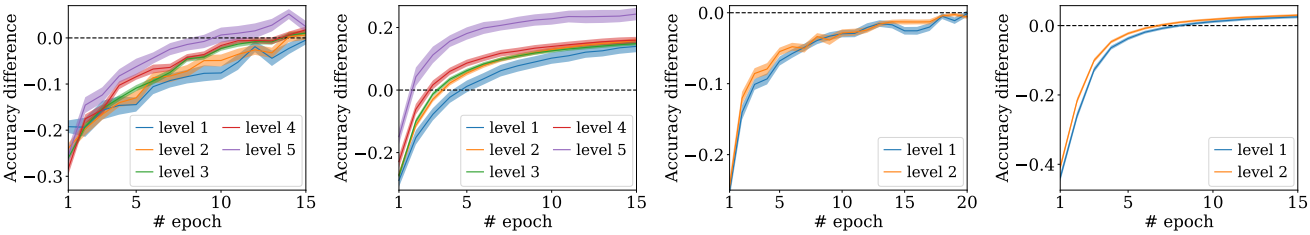
(a) ImageNet → CIFAR100 (b) CIFAR10 → CIFAR100 (c) ImageNet → FashionMNIST (d) CIFAR10 → FashionMNIST

Figure 1. **Average test accuracy of transferred models on small, balanced target data sets in five transferability levels** obtained from LEEP scores. The higher the level, the easier the transfer. A → B in the subcaptions indicate that the source model is trained on A and the target datasets are constructed from B. The source models are ResNet18 for ImageNet (a,c) and ResNet20 for CIFAR10 (b,d).



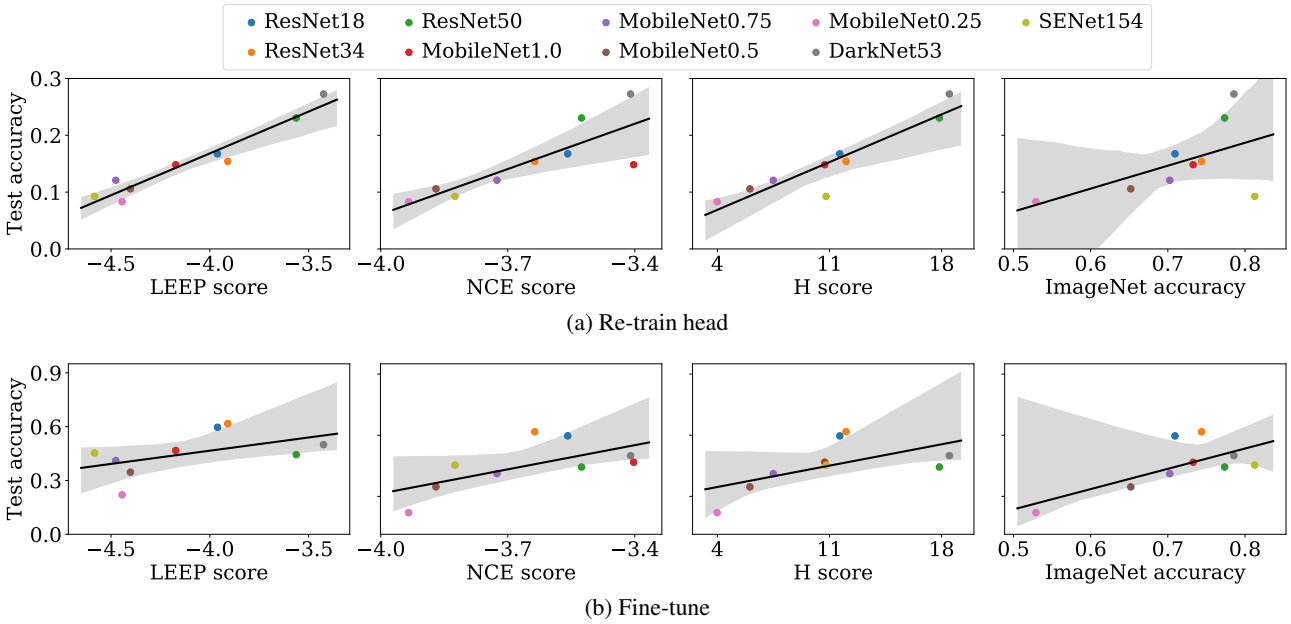
(a) ImageNet → CIFAR100 (b) CIFAR10 → CIFAR100 (c) ImageNet → FashionMNIST (d) CIFAR10 → FashionMNIST

Figure 2. **Average test F1 score of transferred models on small, imbalanced target data sets in five transferability levels** obtained from LEEP scores. The higher the level, the easier the transfer. A → B in the subcaptions indicate that the source model is trained on A and the target datasets are constructed from B. The source models are ResNet18 for ImageNet (a,c) and ResNet20 for CIFAR10 (b,d).



(a) ImageNet → CIFAR100 (b) CIFAR10 → CIFAR100 (c) ImageNet → FashionMNIST (d) CIFAR10 → FashionMNIST

Figure 3. **Convergence of accuracy for fine-tuned models** to the accuracy of a *reference* model trained from scratch using only the target dataset. The convergence is represented by the accuracy difference between the fine-tune model and the reference model. Each curve is the average of the accuracy difference curves over tasks within the same transferability level. The zero lines indicate where the fine-tuned models match the accuracy of the reference model.



(a) Re-train head

(b) Fine-tune

Figure 4. **Test accuracy vs. transferability** according to LEEP score, NCE score (Tran et al., 2019), H score (Bao et al., 2019), and ImageNet accuracy (Kornblith et al., 2019) for 9 candidate source models (see the legend) pre-trained on ImageNet. The transferred models are obtained by (a) re-training the head classifier, and (b) fine-tuning the source model.