Supplementary Material
How Stable are Transferability Metrics evaluations?

Andrea Agostinelli, Michal Pándy*, Jasper Uijlings, Thomas Mensink, and Vittorio Ferrari

Google Research
{agostinelli, michalpandy, jrru, mensink, vittoferrari}@google.com

Abstract. In the supplementary material we study a third, somewhat separate, scenario of correlating transferability metrics of subsampled target datasets with the accuracy of fine-tuned models.

A Bonus Scenario: target dataset transferability on image classification

As a somewhat separate investigation, we take a closer look at a commonly used scenario where the source model is fixed and the target task is constructed by sub-sampling classes from a large target dataset [5, 3, 6]. More formally, in this scenario an experiment consists of the following three components: (1) A source model S. (2) A target dataset from which we create a *target dataset pool* $\mathcal{T}$. (3) An evaluation measure E. In this scenario, we investigate the effect of how to use a single target dataset to create the target dataset pool $\mathcal{T}$.

Experimental setup and creating $\mathcal{T}$. It is common practice to construct $\mathcal{T}$ by sampling uniformly between 2% and a 100% of the target classes [5, 3]. This results in target datasets which vary in the number of classes. In this paper we compare this approach with sampling uniform 50% of the target classes, resulting in target datasets with an *equal* number of classes. In both cases we use all images for the selected classes both for training and testing.

For our setup, we use as source model S a ResNet50 pre-trained on ImageNet. We consider four target datasets: CIFAR100 [2], Stanford Dogs [1], Sun397 [7], and Oxford Flowers 102 [4]. For every target dataset we construct 100 datasets. The transferability metrics are evaluated as described in Sec. 3.1 (main paper).

New transferability metric: NumClasses (#C). For the purpose of our investigation, we define a new transferability metric. Intuitively, target datasets that contain more classes are more complex than target datasets with fewer classes. Therefore, to determine to what extent is transferability is *trivially* explained by the number of target classes, our *NumClasses* metric is simply defined as the number of classes of a given target dataset in $\mathcal{T}$.

[*] Currently at Waymo.

	GBC	**LEEP**	**$\mathcal{N}$LEEP**	**LogME**	**H-score**	**NumC**
CIFAR100	**0.935**	0.909	0.933	-0.787	0.686	0.899
Dogs	0.948	**0.949**	0.936	-0.789	0.200	0.913
SUN	**0.960**	0.947	0.950	-0.920	0.473	0.953
Flowers	**0.748**	0.699	0.712	-0.628	-0.642	0.697
Average	**0.898**	0.876	0.883	-0.781	0.179	0.866

(a) Sampling between 2% and 100% classes

	GBC	**LEEP**	**$\mathcal{N}$LEEP**	**LogME**	**H-score**	**NumC**
CIFAR100	0.580	0.304	0.470	**0.675**	0.386	-
Dogs	**0.740**	0.703	0.611	0.521	0.191	-
SUN	0.509	0.421	**0.681**	0.263	0.256	-
Flowers	0.344	0.266	0.341	**0.388**	0.167	-
Average	**0.543**	0.424	0.526	0.462	0.250	-

(b) Sampling always 50% classes

Table 1: τ_w performance of transferability metrics in ranking 100 randomly subsampled datasets out of a single large target dataset. We compare two different sampling strategies: uniformly sampling $2-100\%$ of the target classes (1a) or always uniformly sampling 50% of the target classes (1b). We repeat the experiment for each of 4 target datasets in turn (rows)

Results. The results presented in Tab. 1a show that the trivial *NumClasses* performs on par with the top transferability metrics (LEEP, GBC, and $\mathcal{N}$LEEP) on all datasets in terms of τ_w and outperforms two metrics (LogMe and H-score) on average. We also note that while LogMe is the best performing method in Sec. 5 (main paper), it now is the worst method and has even negative rank correlations (Tab. 1a). In contrast, if we fix the number of target classes across all target datasets, suddenly LogME has decent rank correlations and outperforms LEEP and H-score. The trivial method *NumClasses* becomes unusable. To conclude, this suggests that a scenario where the target dataset pool $\mathcal{T}$ is created by sampling a variable number of classes is not suitable for evaluating transferability metrics. Instead, it is preferable to sample a fixed number of classes for the whole target pool $\mathcal{T}$.

If we look at the overall winning transferability metric, we find that GBC works best in the current scenario. Interestingly, this is again different from the winning metric in Sec. 5 (LogME) and Sec. 6 ($\mathcal{N}$LEEP) of the main paper.

References

1. Khosla, A., Jayadevaprakash, N., Yao, B., Fei-Fei, L.: Novel dataset for fine-grained image categorization. In: CVPR Workshops (2011)
2. Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech. rep., University of Toronto (2009)
3. Nguyen, C., Hassner, T., Seeger, M., Archambeau, C.: LEEP: A new measure to evaluate transferability of learned representations. In: ICML (2020)
4. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Indian Conf. on CVGIP (2008)
5. Pándy, M., Agostinelli, A., Uijlings, J., Ferrari, V., Mensink, T.: Transferability estimation using bhattacharyya class separability. In: CVPR (2022)
6. Tan, Y., Li, Y., Huang, S.L.: OTCE: A transferability metric for cross-domain cross-task representations. In: CVPR (2021)
7. Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A.: SUN database: Large-scale scene recognition from Abbey to Zoo. In: CVPR (2010)