Supplementary Material
Transferability Metrics for Selecting Source Model Ensembles

A. Limitations

Here we discuss some limitations of our proposed method and of the general transfer learning field.

Diverse Pool of Source Models. As discussed in Sec.2, ensembling machine learning models is a classical method for increasing accuracy, where having diverse models is typically important [2–7]. Hence, the benefits of selecting an ensemble of source models over a single source model could decrease if the pool of source models is not diverse enough.

Focus on Limited Training Data. The general field of transfer learning focuses on the scenario of having a limited target training data, which is not sufficient for the model to achieve good performance and generalize to the target test data. Hence, by exploiting the knowledge of one or more source models, performance can be improved. For this reason, advances on this line of work may not be beneficial to scenarios where the target training data is enough for solving the final target test task.

General risks of transfer learning. As described in detail in [1], transfer learning demands caution, especially when using source models trained on broad data. In particular, problems related to the source models, such as intrinsic biases, can easily propagate to the target task.

B. Training source models.

We describe here implementation details on the training procedure of our source models. We train using a pixel-wise cross-entropy loss, optimized by SGD with momentum [8]. We use 8 Google Cloud TPUs v3 with synchronized batch norm and batch size 32. We decrease the learning rate by 10× after $2/3$ of the total training steps. For each model, we tune the initial learning rate and number of training steps on the source dataset.

C. Ablation study: number of ensemble members

We perform an ablation study to understand to which extend an ensemble can benefit from combining more source models together, given our pool of source models. We use ensembles composed of up to 5 members. We measure the actual performance of each ensemble using the target datasets and source models from Sec. 4. For example, when evaluating ensembles of 5 members out of a pool of 15 source models, we consider a total of $\binom{15}{5} = 3003$ combinations. To limit computation, the source models are aggregated using an unweighted average of their predictions.

We are interested at the best performing ensemble, which is the one a transferability metric aims at selecting. Hence, we show in Fig. 1 the performance of the best ensemble across all combinations for a given number of ensemble members. In detail, we show the relative performance gain compared to the best single source model (ensemble of 1 member). Most performance gain comes from combining 3 models. This validates our choice to use 3 source models as compromise between performance and total capacity, as done in Sec.4 and Sec.5.

References

[1] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1

[2] Leo Breiman. Bagging predictors. *Machine learning*, 1996. 1

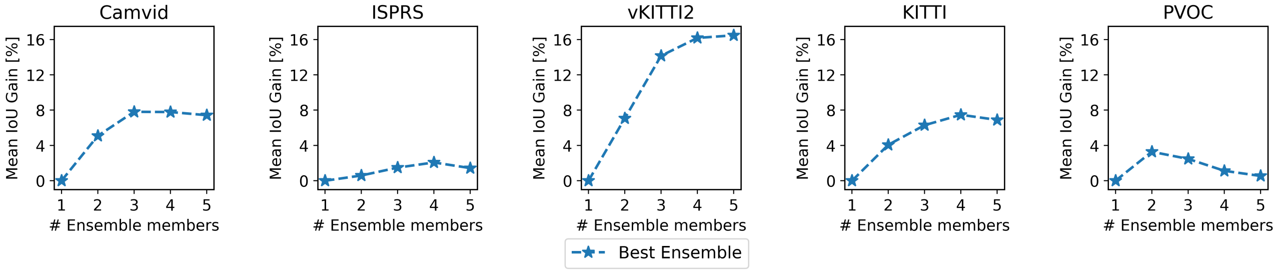

Figure 1. Relative performance gain (Mean IoU) over the best single source model, of the best ensemble across $\binom{S}{15}$ possible combinations of S ensemble members. We use a pool of 15 source models for each target dataset. We consider S ranging from 1 to 5. Most performance gain comes from combining 3 source models together.

[3] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000. 1

[4] Yoav Freund and Robert Schapire. Experiments with a new boosting algorithm. In *ICML*, 1996. 1

[5] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001, 1990. 1

[6] Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *NeurIPS*, 1995. 1

[7] Stefan Lee, Senthil Purushwalkam, Michael Cogswell, David Crandall, and Dhruv Batra. Why m heads are better than one: Training a diverse ensemble of deep networks. In *arXiv*, 2015. 1

[8] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145–151, 1999. 1