How NOT to Benchmark your SITE Metric

Beyond Static Leaderboards and Towards Realistic Evaluation.

Prabhant Singh, Sibyl Hess and Joaquin Vanschoren

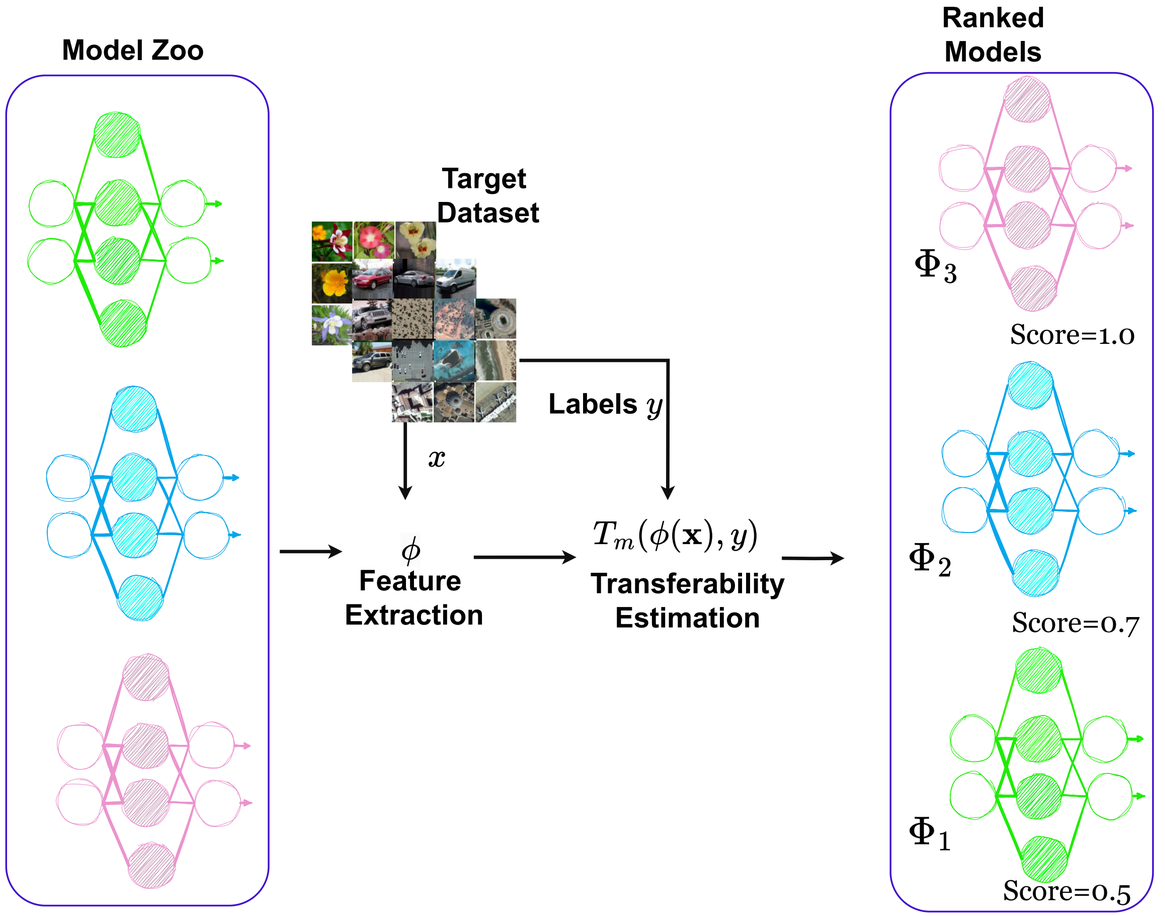
AMOR/e
Advanced Models by
Open Research & Engineering

TU/e
EINDHOVEN
UNIVERSITY OF
TECHNOLOGY

What's SITE

Source Independent Transferability Estimation

- How to find the best model for a given task without finetuning the models on the tasks.
- Recently attracted a lot of attention with works like LogME, H-Score, LEEP etc



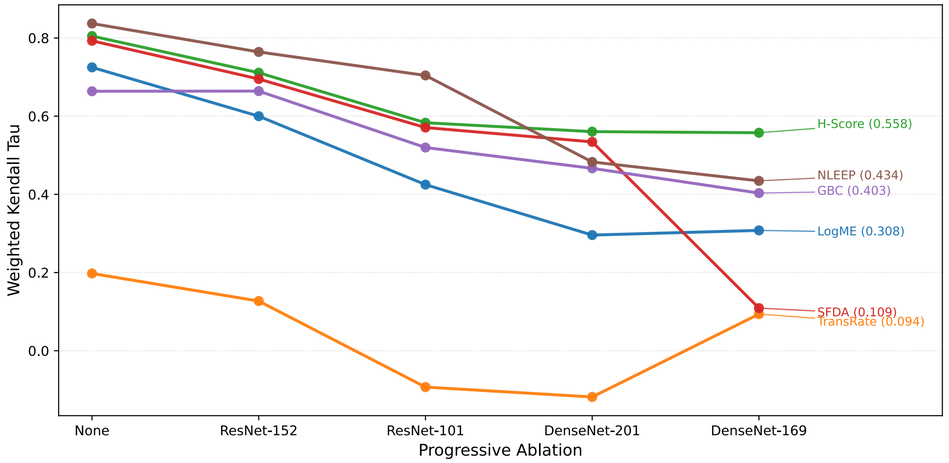
Standard Setup

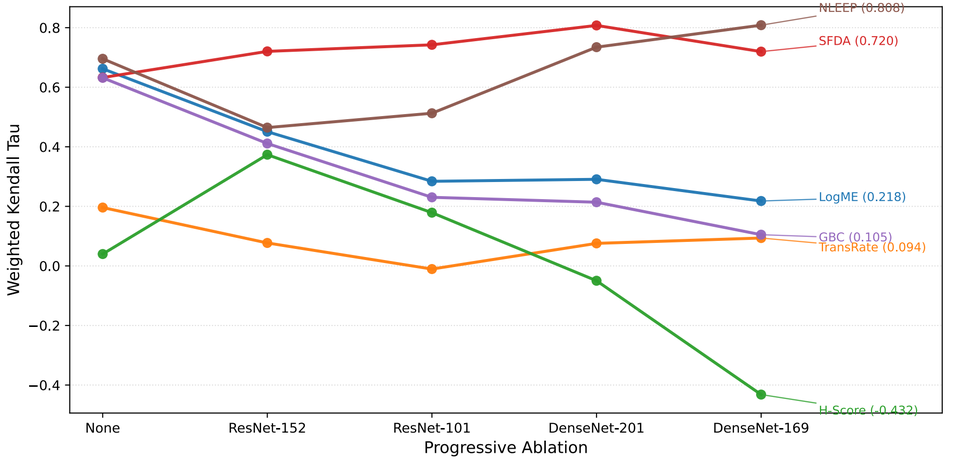
Search space, Hyperparameters, tasks

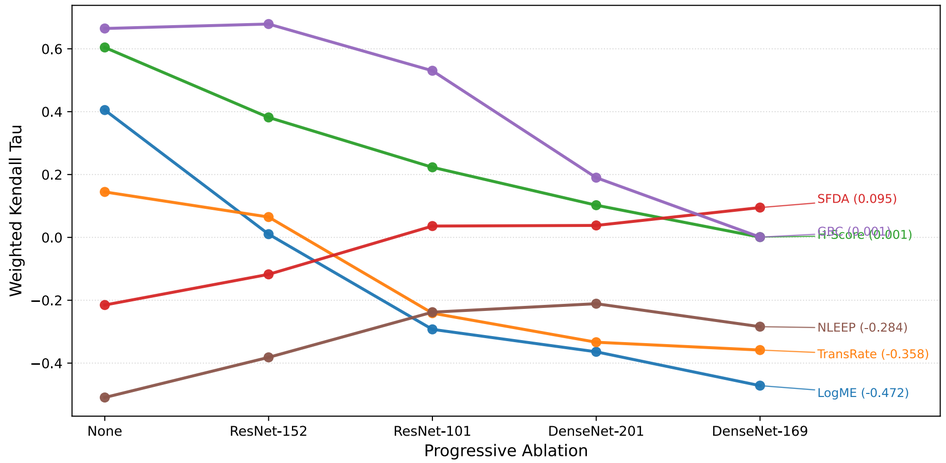
- Search space only consists of ResNets, DenseNets and MobileNet and inception models of different sizes
- Standard finetuning of all models on test train set of 8 popular datasets

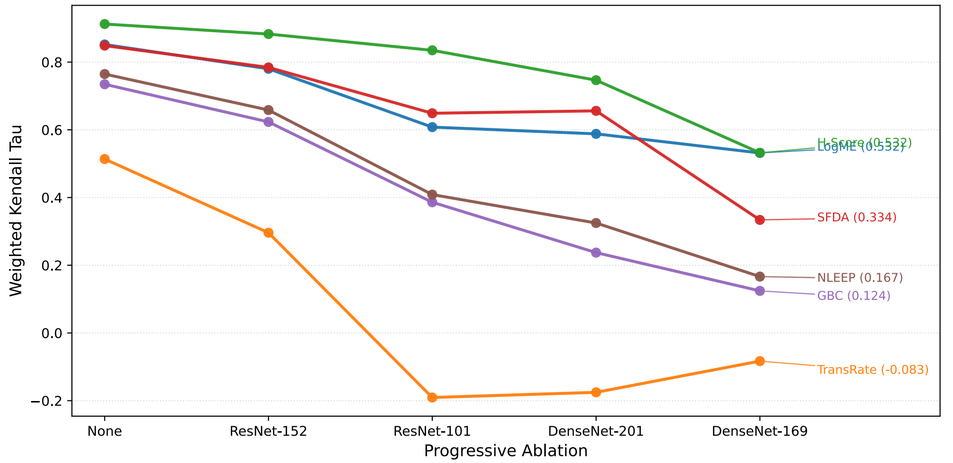
Problem 1: THE MODEL SEARCH SPACE IS UNREALISTIC

Verification by ablation



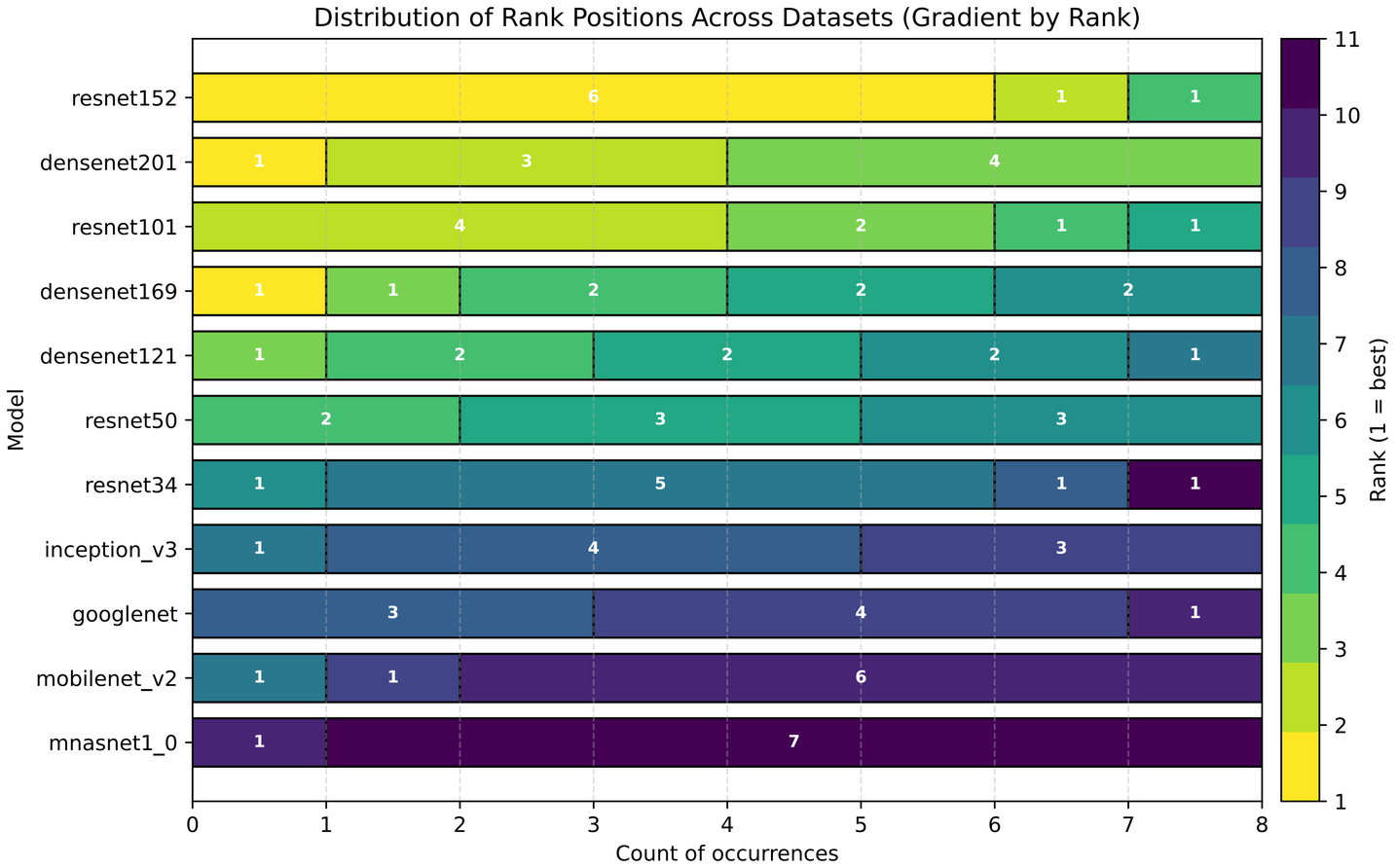






Problem 2: Static leaderboard

Same models are always #1



Problem 2: Static Leaderboards

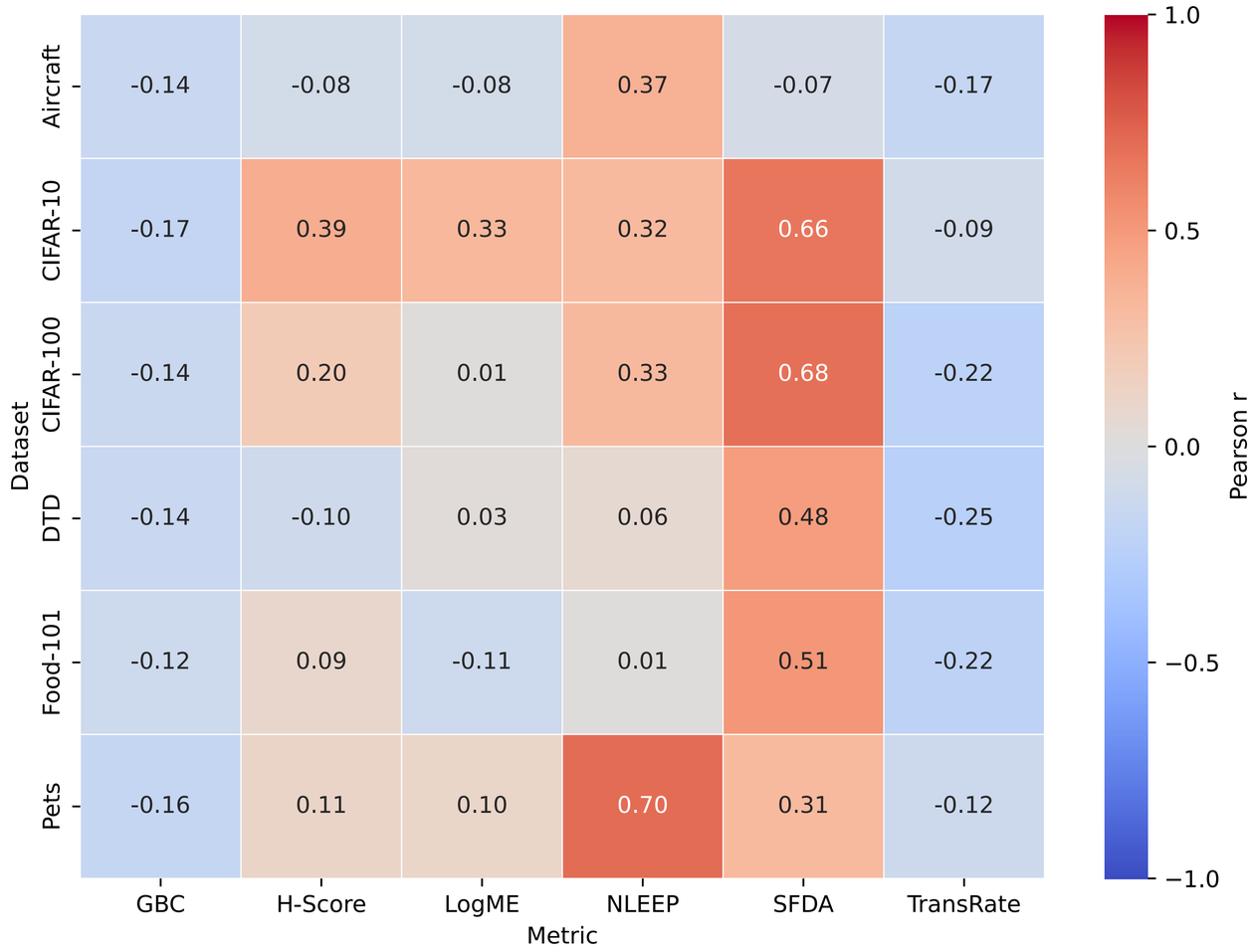
Verification by Static Ranker

Metric	Aircraft	CIFAR10	CIFAR100	DTD	Food	Pets	Average
GBC	-0.12	-0.02	0.09	0.14	0.10	-0.15	0.7
TransRate	0.14	0.51	0.20	0.20	-0.05	0.17	0.19
SFDA	-0.22	0.85	0.79	0.63	0.30	0.34	0.44
H-Score	0.60	0.91	0.80	0.04	0.59	0.37	0.55
NLEEP	-0.51	0.76	0.84	0.70	0.69	0.84	0.55
LogME	0.41	0.85	0.72	0.66	0.39	0.41	0.57
Static Ranking	**0.84**	**0.91**	**0.98**	**0.99**	**0.80**	**0.94**	**0.91**

Problem 3: No Fiedelity

Verification via correlation

- SITE Metrics do not capture the differences in performance and give uneven score based on model and performance



Solutions

How can we mitigate falling into this benchmarking trap

- Checklist with best practices
- Example benchmark implementing these practices

Dataset	TransRate	LogME	NLEEP	SFDA	HScore	GBC	**Static**
Sports	0.39	0.25	0.30	**0.70**	-0.08	0.38	0.46
PlantVillage	0.18	**0.61**	**0.61**	-0.05	0.30	0.14	-0.30
RESISC	0.24	0.11	0.14	**0.76**	0.23	0.36	0.55
Stanford Actions	-0.16	-0.37	-0.28	-0.07	0.01	0.03	**0.27**
Insects	0.72	0.57	0.84	0.53	0.52	**0.87**	0.56
DTD	-0.53	-0.37	-0.37	-0.48	-0.33	-0.42	**0.01**
PanNuke	0.14	-0.06	0.24	**0.68**	0.13	0.40	0.00
Dogs	-0.71	-0.41	-0.62	-0.32	-0.30	-0.59	**-0.15**
MPII Human	0.34	0.27	0.25	0.18	0.24	0.23	**0.50**
Fungi	0.44	0.70	-0.22	**0.77**	0.40	0.34	**0.77**
Plant Doc	0.54	0.30	0.30	0.10	0.10	0.48	**0.58**
SPIPOLL	-0.15	0.10	-0.15	-0.28	-0.18	-0.15	**0.32**
RSD	0.02	-0.20	-0.34	-0.07	-0.06	-0.04	**0.60**
PRTA	0.28	-0.09	-0.44	0.14	**0.38**	0.29	0.37
Boats	0.20	-0.49	**0.38**	-0.32	-0.33	0.27	0.09
Average	0.13	0.061	0.04	0.15	0.06	0.17	**0.31**

Thank You!