Guided Recommendation for Model Fine-tuning

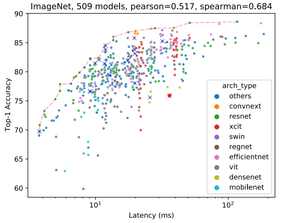
Hao Li, Charless Fowlkes, Hao Yang, Onkar Dabeer, Zhuowen Tu, Stefano Soatto

AWS AI Labs

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Model Selection

- Given a task and a large zoo of pre-trained models, Model Selection (MS) selects the top few models for the best fine-tuning performance, avoiding fine-tuning of all models.



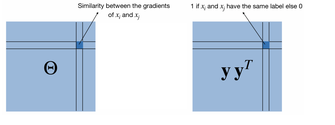
Feature-based Model Selection

- Linearization Assumption

$$f_w(x) = f_{w_0}(x) + \nabla_w f_{w_0}(x)(w - w_0)$$

- Label-Feature Correlation (**LFC**)[1]

$$S_{\mathbf{LFC}}(\mathbf{x}, \mathbf{y}) = (f_w(\mathbf{x}) f_w(\mathbf{x})^T) \cdot \mathbf{y}\mathbf{y}^T$$

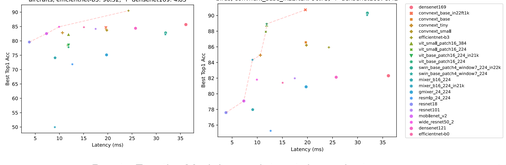


- **PARC**[2] introduces heuristics of layer depth

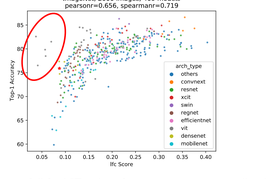
$$S'_{\mathrm{PARC}} = \frac{S_{\mathrm{PARC}} - \mu^t}{\sigma^t} + \frac{\ell_s}{\ell_{\max}}$$

Limitations

- Difficulty with heterogeneous models

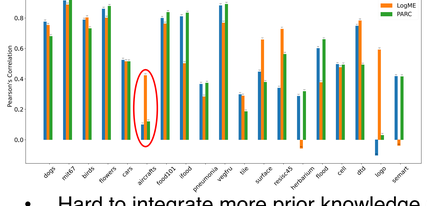


Pareto Frontier Models are dataset dependent.



MS can fail for ViTs when features are not normalized

- Weights may change a lot for hard/dissimilar datasets



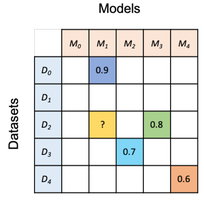
Stanford Dogs is very close to ImageNet, its MS score is more accurate.

However, dataset like Aircrafts is much different from ImageNet, so its MS scores are much lower.

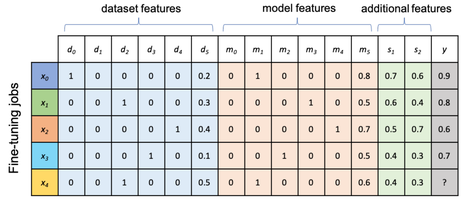
- Hard to integrate more prior knowledge (e.g., capacity, dataset size)
  The scale is ad-hoc and it is not easy to integrate more signals or prior knowledge.

Learning to Recommend Models

- We convert model selection as a model **recommendation** problem, which **learns** the model selection criteria from the past **fine-tuning history**.
- The goal is to predict performance on the target dataset for a given model.

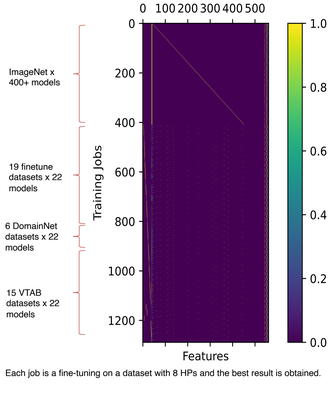


Training History

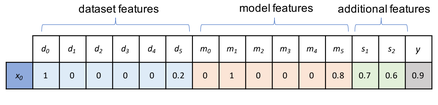


Embedding

Training History and Meta Features



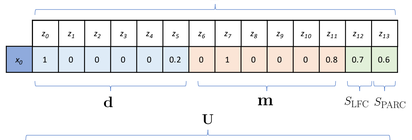
Each job is a fine-tuning on a dataset with 8 HPs and the best result is obtained.



- **task difficulty**: If a task can be solved with a simple model, then the task is relatively easy in comparison with other dataset.
- **number of samples**: a few-shot task is generally harder and often requires a strong model than a larger dataset size.
- **number of classes:** the task difficulty usually increase as the number of classes when the total images are fixed.
- **architecture family**: architectures of the same family usually have similar inductive biases as they consist of similar modules.
- **input size**: architectures with higher resolution usually helps for downstream tasks.
- **model capacity**: a model with high capacity usually generalizes better with more data.
- **model complexity**: GMACs
- **pre-trained domain**: the pre-trained domain matters for the downstream task performance.
- **MS score:** it considers the feasibility of the model's initial features.
- **semantic distance:** semantic embedding of labels of the target task and the source task
- any features that are relevant for performance prediction

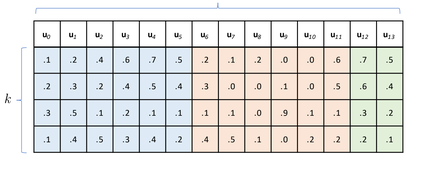
Recommendation Models

- **Linear Regression (LR)** $S_{\mathrm{LR}}(\mathbf{z}) = w_0 + \sum_{i=1}^{|z|} w_i z_i$

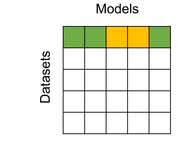


- **Factorization Machines (FM)**

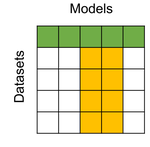
$$S_{\mathrm{FM}}(\mathbf{z}) = w_0 + \underbrace{\sum_{i=1}^{|z|} w_i z_i}_{\text{linear regression}} + \underbrace{\sum_{i=1}^{|z|-1} \sum_{j=i+1}^{|z|} \langle \mathbf{u}_i, \mathbf{u}_j \rangle z_i z_j}_{\text{feature interactions}}$$



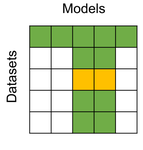
Experiments: More Training Data Helps!



- Learning from the history of **single** dataset with **a subset** of models.
- Evaluating **unseen models** on the **same** dataset.



- Learning from the history of **single** dataset with **all** models.
- Evaluating **known models** on **unseen** datasets



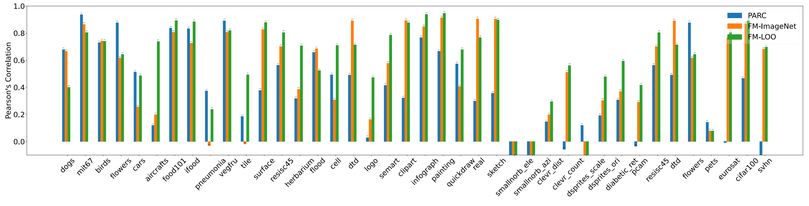
- Learning from the history of leave-one-out **datasets**.
- Evaluating **known models** on **unseen tasks**.

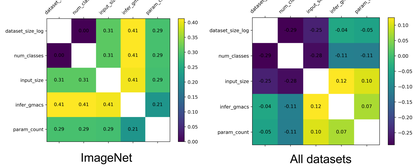
Setting 1. MS learned with **only ImageNet** training history. 80% of the 409 models as sampled as training set. Our methods still get reasonable corr. scores even when models are ***random initialized***.

Methods	Features	ImageNet Pre-trained	ImageNet Random Init.
feature-based	S_{LFC}	0.65 ± 0.07	0.03 ± 0.10
	S_{LogME}	0.35 ± 0.09	0.04 ± 0.08
	S_{PARC}	$\mathbf{0.83} \pm 0.04$	0.08 ± 0.09
LR (ours)	$\mathbf{d}, \mathbf{m}$	0.53 ± 0.07	$\mathbf{0.57} \pm 0.10$
	$\mathbf{d}, \mathbf{m}, S_{\mathrm{LFC}}$	0.73 ± 0.06	0.56 ± 0.10
	$\mathbf{d}, \mathbf{m}, S_{\mathrm{LogME}}$	0.55 ± 0.08	0.56 ± 0.09
	$\mathbf{d}, \mathbf{m}, S_{\mathrm{PARC}}$	$\mathbf{0.85} \pm 0.04$	0.57 ± 0.11
FM (ours)	$\mathbf{d}, \mathbf{m}$	0.54 ± 0.06	$\mathbf{0.57} \pm 0.10$
	$\mathbf{d}, \mathbf{m}, S_{\mathrm{LFC}}$	0.70 ± 0.12	0.56 ± 0.10
	$\mathbf{d}, \mathbf{m}, S_{\mathrm{LogME}}$	0.55 ± 0.09	0.56 ± 0.10
	$\mathbf{d}, \mathbf{m}, S_{\mathrm{PARC}}$	$\mathbf{0.84} \pm 0.05$	0.57 ± 0.11

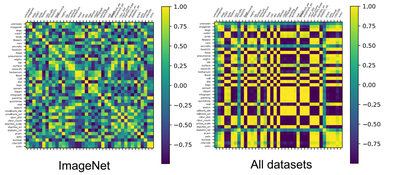
Setting 2 & 3. Average Pearson Correlation of predicted performance and the ground-truth performance of 22 models. The **ImageNet** column is trained on 409 ImageNet training jobs. The **LOO** column denotes MS learned with the training history combining ImageNet and all other downstream jobs. MS trained with LOO outperforms over ImageNet only.

Methods	Features	19 fine-grained		6 DomainNet		15 VTAB	
feature-based MS	S_{LFC}	0.55		0.63		0.14	
	S_{LogME}	0.54		0.52		0.20	
	S_{PARC}	0.54		0.50		0.13	
		ImageNet	LOO	ImageNet	LOO	ImageNet	LOO
LR (ours)	$\mathbf{d}, \mathbf{m}$	0.53	0.66	0.80	0.82	0.29	0.37
	$\mathbf{d}, \mathbf{m}, S_{\mathrm{LFC}}$	**0.67**	**0.74**	**0.84**	**0.85**	**0.38**	**0.41**
	$\mathbf{d}, \mathbf{m}, S_{\mathrm{LogME}}$	0.54	0.65	0.81	0.84	0.30	0.36
	$\mathbf{d}, \mathbf{m}, S_{\mathrm{PARC}}$	0.54	0.66	0.81	0.85	0.30	0.40
FM (ours)	$\mathbf{d}, \mathbf{m}$	0.53	0.65	0.81	0.85	0.35	0.39
	$\mathbf{d}, \mathbf{m}, S_{\mathrm{LFC}}$	**0.64**	**0.74**	0.82	**0.87**	**0.39**	0.41
	$\mathbf{d}, \mathbf{m}, S_{\mathrm{LogME}}$	0.60	0.67	0.82	0.86	0.31	0.40
	$\mathbf{d}, \mathbf{m}, S_{\mathrm{PARC}}$	0.56	0.69	**0.86**	0.86	0.30	**0.43**





Pearson Corr. between scalar feature $\mathbf{z}_i \mathbf{z}_j$ and $\mathbf{y}$.



Corr. between Latent Features of Dataset IDs

Reference

[1] A linearized framework and a new benchmark for model selection for fine-tuning, Deshpande et al, arXiv 2021
[2] Scalable Diverse Model Selection for Accessible Transfer Learning, Bolya et al, NeurIPS 2021
[3] LogME: Practical Assessment of Pre-trained Models for Transfer Learning, You et al, ICML 2021