JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Guided Recommendation for Model Fine-Tuning






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

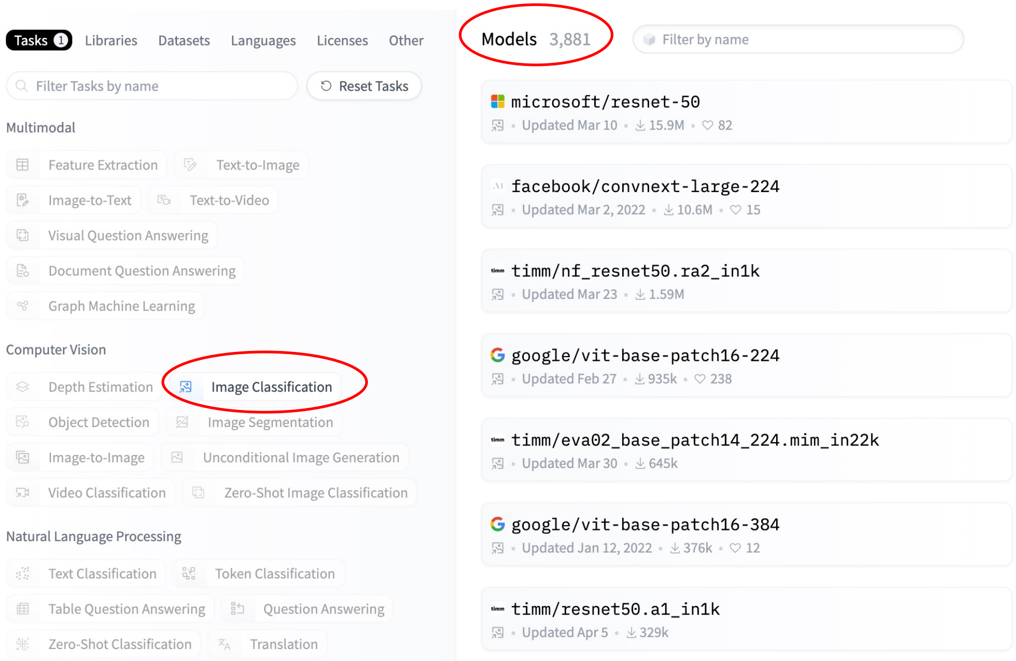
AWS AI Labs

Tue-AM-346

Overview

Fine-tuning Pre-trained Models

- A large-scale pre-trained model zoo are important for broad domain coverage.



Model Selection

- Given a task and a bank of pre-trained models, MS selects the top few models for the best fine-tuning performance, avoiding the brute-forth fine-tuning.

Issues

- Improper assumption: fixed backbones.
- Hard to integrate prior knowledge (e.g., model capacity, dataset size).

Overview

Learning to Recommend Models

- We convert model selection as a model **recommendation** problem, which learns the model selection criteria from the past **training history**.

Models

Datasets

	M_0	M_1	M_2	M_3	M_4
D_0		0.9			
D_1					
D_2		?		0.8	
D_3			0.7		
D_4					0.6

Training History

Fine-tuning jobs

	dataset features						model features						additional features		
	d_0	d_1	d_2	d_3	d_4	d_5	m_0	m_1	m_2	m_3	m_4	m_5	s_1	s_2	y
x_0	1	0	0	0	0	0.2	0	1	0	0	0	0.8	0.7	0.6	0.9
x_1	0	0	1	0	0	0.3	0	0	0	1	0	0.5	0.6	0.4	0.8
x_2	0	0	0	0	1	0.4	0	0	0	0	1	0.7	0.5	0.7	0.6
x_3	0	0	0	1	0	0.1	0	0	1	0	0	0.5	0.4	0.3	0.7
x_4	0	0	1	0	0	0.5	0	1	0	0	0	0.6	0.4	0.3	?

Embedding

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

Recommendation Model

Feature-Based MS: Basic Assumption

- **Linearization Assumption**
  - It is assumed that the model weights do not change much during fine-tuning, i.e., the final fine-tuned solution can be a linearized approximation

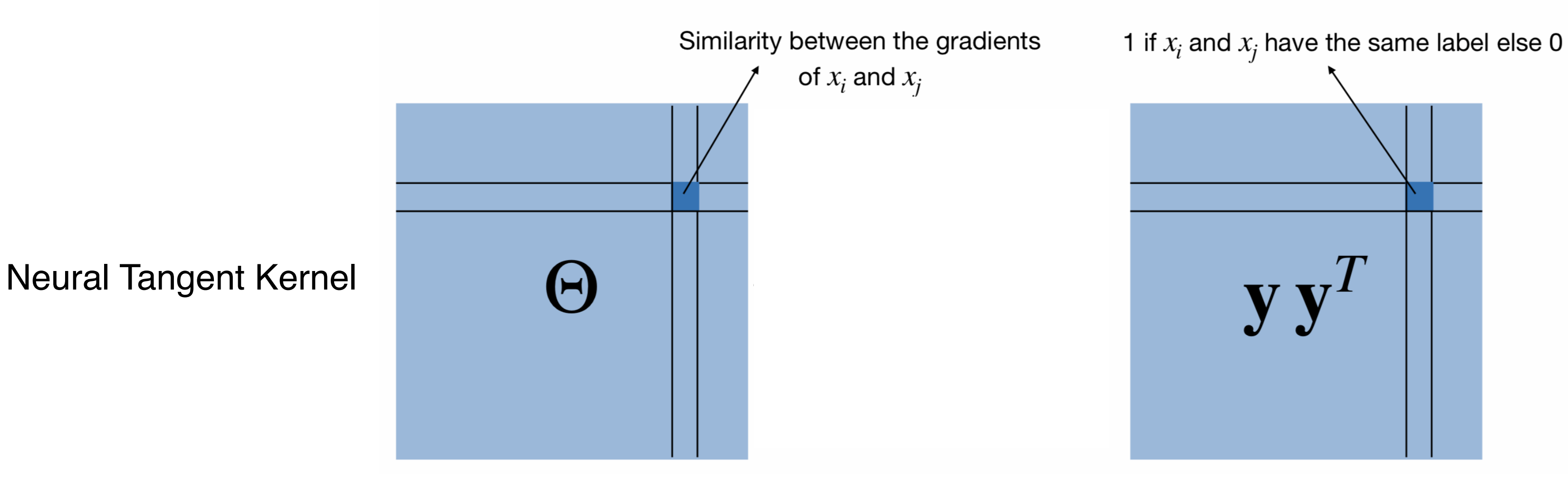
$$f_w(x) = f_{w_0}(x) + \nabla_w f_{w_0}(x)(w - w_0)$$

  - The backbone can extract features on the target training set and the generalization ability is estimated based on these features with algorithms like LFC, PARC and LogME.

Feature-Based MS: LFC

- **Label-Gradient/Feature Correlation (LGC/LFC)** [Deshpande et al, 2021]

$$\mathbf{y}^T \Theta \mathbf{y} = (\nabla f_w(\mathbf{x}) \nabla f_w(\mathbf{x})^T) \cdot \mathbf{y}\mathbf{y}^T$$



Θ can be approximated using features instead of gradients $\Theta_F = f_w(\mathbf{x}) f_w(\mathbf{x})^T$

Deshpande et al, A linearized framework and a new benchmark for model selection for fine-tuning, arXiv 2021

Feature-Based MS: PARC

- **PARC** [Bolya et al, 2021]
  - Similar to LFC, it calculates the Spearman's Rank Correlation between the two distance matrices for all pair of images.

$$D_\theta = 1 - \text{corrcoef}(f_\theta(x)) \qquad D_y = 1 - \text{corrcoef}(g(y))$$

$$\text{PARC}(\theta, \mathcal{P}_n) = \text{spearmanr}(\{D_\theta[i,j] : i < j\}, \{D_y[i,j] : i < j\})$$

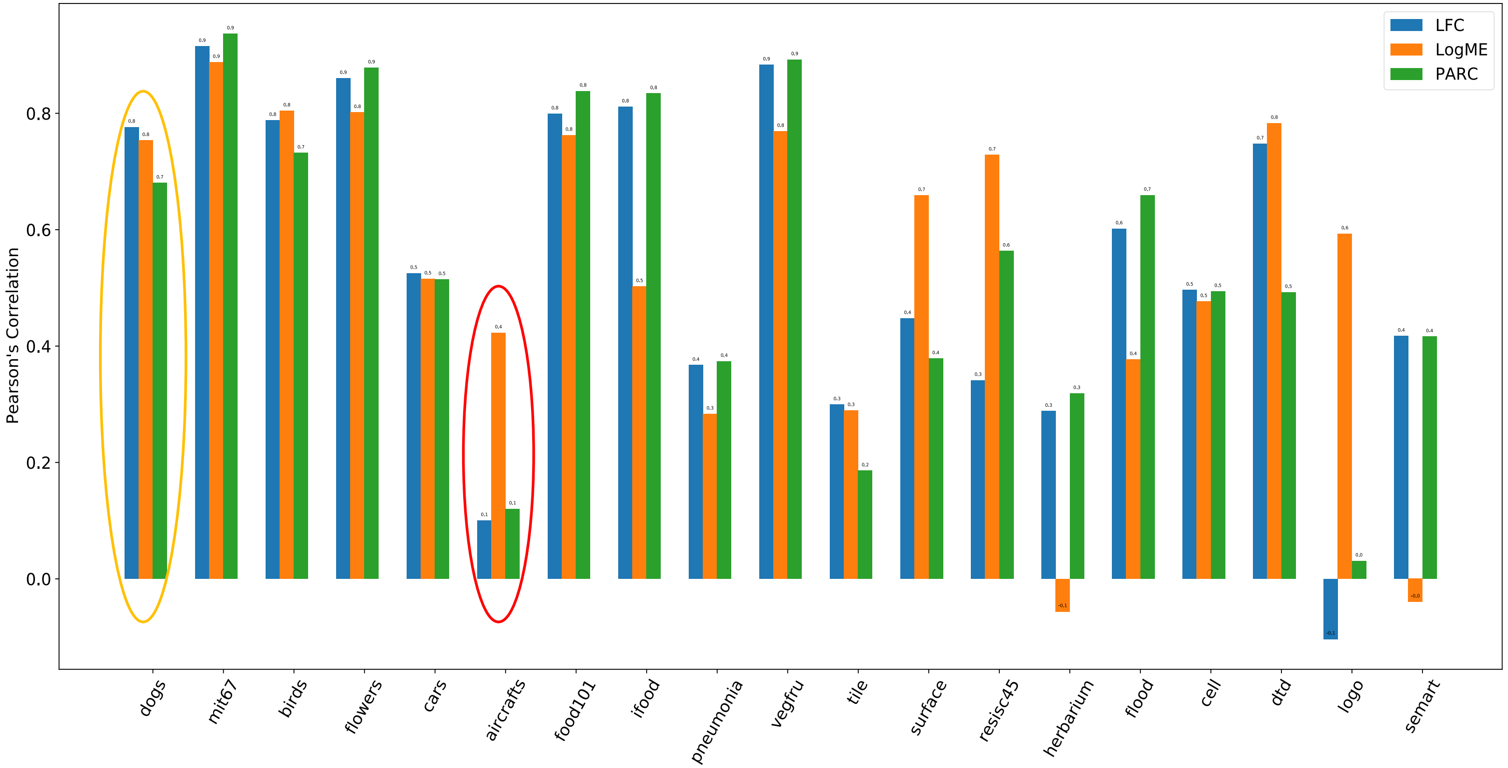
  - Add heuristic of model depth with ad-hoc scaling.

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

Bolya et al, Scalable Diverse Model Selection for Accessible Transfer Learning, NeurIPS 2021

MS Issues I: Linearization Assumption

- **The linearization assumption could fail**
    - When the target data is much different from to the source data or the training dataset size is large.
    - The MS score then becomes less accurate and the effect of model initialization diminishes.



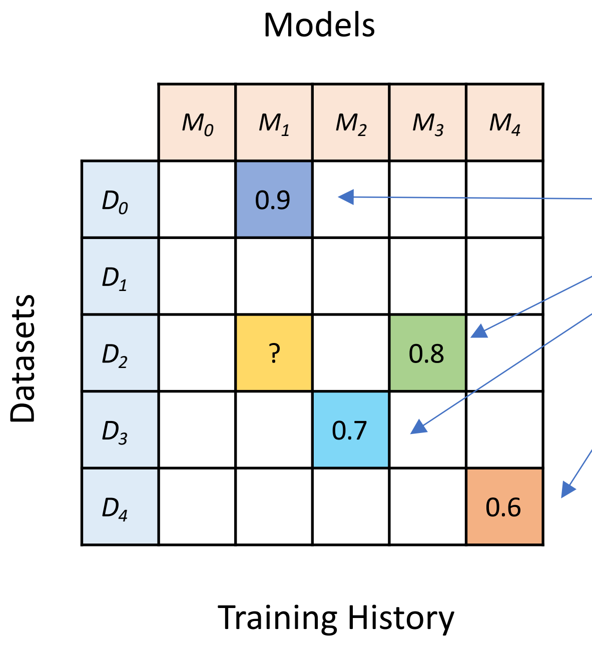
MS Issues II: Integrating Additional Knowledge

- **Ad-hoc scaling for additional heuristic scores**
  - The heuristics (e.g. model depth/layers) may not apply for different architectures, such as ViTs.
  - The scale of heuristics requires ad-hoc tuning.

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

- **Missing meta feature and feature correlations**
  - The effect of model's inductive bias is correlated with dataset characteristics, e.g., *"a **random initialized large** model could generalize better than a **small pre-trained** model on a **large** dataset"*.
    - This correlation between model and dataset is not considered.

Learning To Recommend Models



the training history of same or similar datasets/models will help the prediction, and the performance can be continuously improved with more data.

Learning To Recommend Models

Models

Datasets

	M_0	M_1	M_2	M_3	M_4
D_0		0.9			
D_1					
D_2		?		0.8	
D_3			0.7		
D_4					0.6

Training History

Fine-tuning jobs

	dataset features						model features						additional features		
	d_0	d_1	d_2	d_3	d_4	d_5	m_0	m_1	m_2	m_3	m_4	m_5	s_1	s_2	y
x_0	1	0	0	0	0	0.2	0	1	0	0	0	0.8	0.7	0.6	0.9
x_1	0	0	1	0	0	0.3	0	0	0	1	0	0.5	0.6	0.4	0.8
x_2	0	0	0	0	1	0.4	0	0	0	0	1	0.7	0.5	0.7	0.6
x_3	0	0	0	1	0	0.1	0	0	1	0	0	0.5	0.4	0.3	0.7
x_4	0	0	1	0	0	0.5	0	1	0	0	0	0.6	0.4	0.3	?

Embedding

Dataset and Model Representation

	dataset features						model features						additional features		
Fine-tuning jobs	d_0	d_1	d_2	d_3	d_4	d_5	m_0	m_1	m_2	m_3	m_4	m_5	s_1	s_2	y
x_0	1	0	0	0	0	0.2	0	1	0	0	0	0.8	0.7	0.6	0.9
x_1	0	0	1	0	0	0.3	0	0	0	1	0	0.5	0.6	0.4	0.8
x_2	0	0	0	0	1	0.4	0	0	0	0	1	0.7	0.5	0.7	0.6
x_3	0	0	0	1	0	0.1	0	0	1	0	0	0.5	0.4	0.3	0.7
x_4	0	0	1	0	0	0.5	0	1	0	0	0	0.6	0.4	0.3	?

Dataset and Model Representation

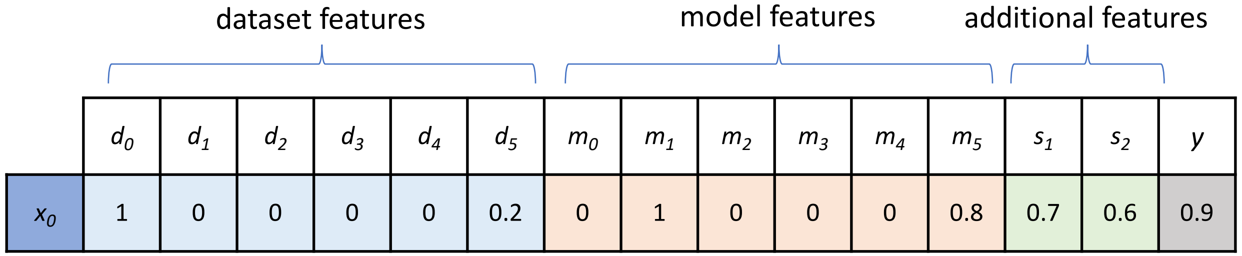
	dataset features						model features						additional features		
	d_0	d_1	d_2	d_3	d_4	d_5	m_0	m_1	m_2	m_3	m_4	m_5	s_1	s_2	y
x_0	1	0	0	0	0	0.2	0	1	0	0	0	0.8	0.7	0.6	0.9

- **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**: archs with higher resolution usually helps for downstream tasks.
- **model capacity**: a model with high capacity usually generalizes better with more data.
- **model complexity**: the calculation cost (GMACs) can represent the complexities.
- **pre-trained domain**: he 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

Dataset and Model Representation



Field idx	Field Name	Feature Name	Type	One-hot	Log	Dimension	Min	Max
1	dataset	dataset id	category	Yes	No	41	0	40
1	dataset	dataset size	scalar	No	Yes	1	1008	1200000
1	dataset	number of classes	scalar	No	Yes	1	2	1000
2	model	architecture id	category	Yes	No	405	0	404
2	model	architecture family id	category	Yes	No	10	0	1
2	model	pre-trained dataset id	category	Yes	No	3	0	2
2	model	input size	scalar	No	Yes	1	106	448
2	model	GMACs (G)	scalar	No	Yes	1	0.03	46.95
2	model	#Parameters (G)	scalar	No	Yes	1	1.88	88.59
3	MS score	LFC	scalar	No	No	1	0.002	0.792
3	MS score	LogME	scalar	No	No	1	-0.905	2.209
3	MS score	PARC	scalar	No	No	1	0.085	80.358

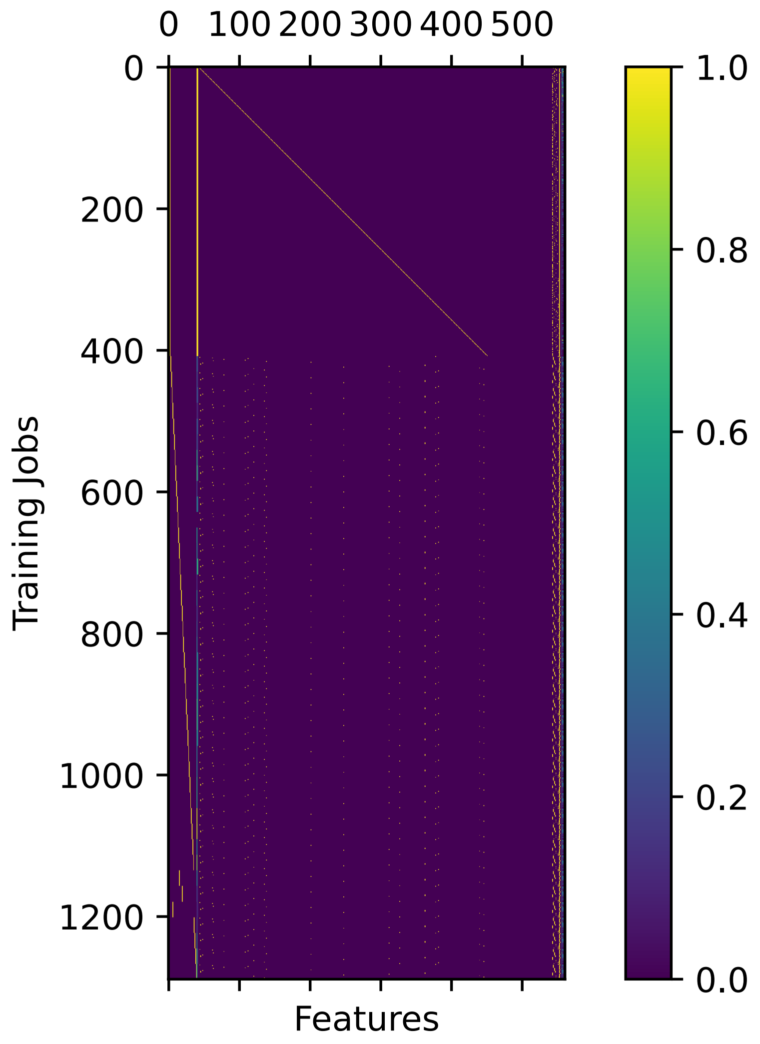
Embedding the Training History

ImageNet x 400+ models

19 finetune datasets x 22 models

6 DomainNet datasets x 22 models

15 VTAB datasets x 22 models



Recommendation Models

$\mathbf{z}$

	z_0	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9	z_{10}	z_{11}	z_{12}	z_{13}
x_0	1	0	0	0	0	0.2	0	1	0	0	0	0.8	0.7	0.6

$\mathbf{d}$ $\mathbf{m}$ S_{LFC} S_{PARC}

$\mathbf{U}$

$\mathbf{u}_0$	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$	$\mathbf{u}_4$	$\mathbf{u}_5$	$\mathbf{u}_6$	$\mathbf{u}_7$	$\mathbf{u}_8$	$\mathbf{u}_9$	$\mathbf{u}_{10}$	$\mathbf{u}_{11}$	$\mathbf{u}_{12}$	$\mathbf{u}_{13}$
.1	.2	.4	.6	.7	.5	.2	.1	.2	.0	.0	.6	.7	.5
.2	.3	.2	.4	.5	.4	.3	.0	.0	.1	.0	.5	.6	.4
.3	.5	.1	.2	.1	.1	.1	.1	.0	.9	.1	.1	.3	.2
.1	.4	.5	.3	.4	.2	.4	.5	.1	.0	.2	.2	.2	.1

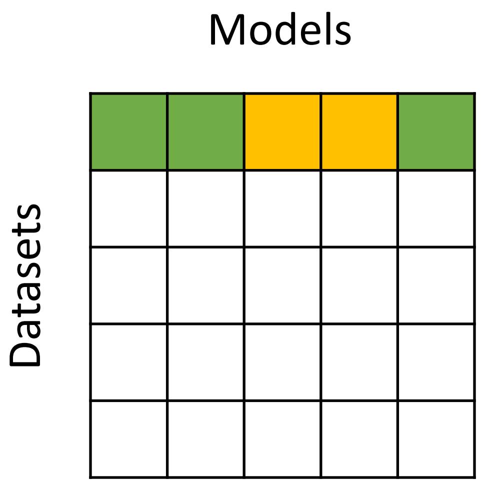
Linear Regression (LR)

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

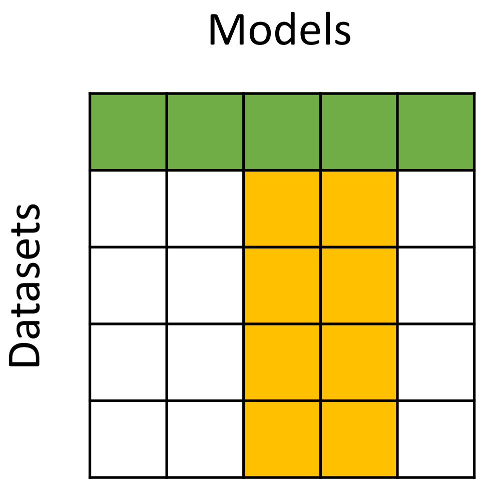
Factorization Machines (FM)

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

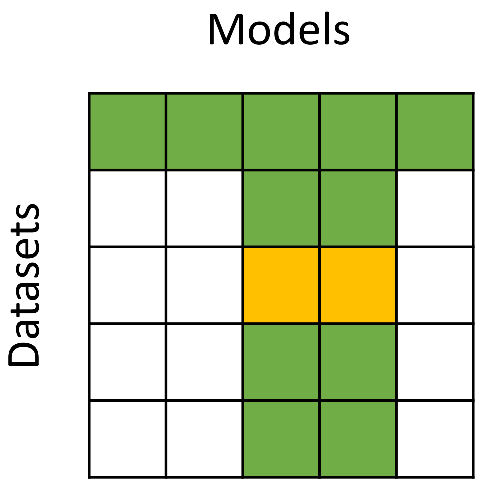
Experimental Settings



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

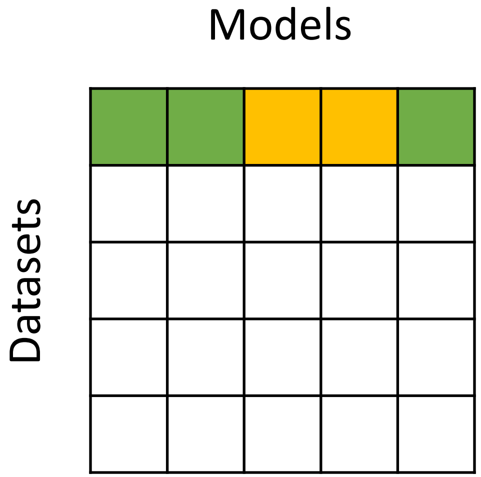


- Learning from the history of **single** dataset will **all** models.
- Evaluating **known models** on **unseen** datasets



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

Learning from ImageNet and Predict for New Models



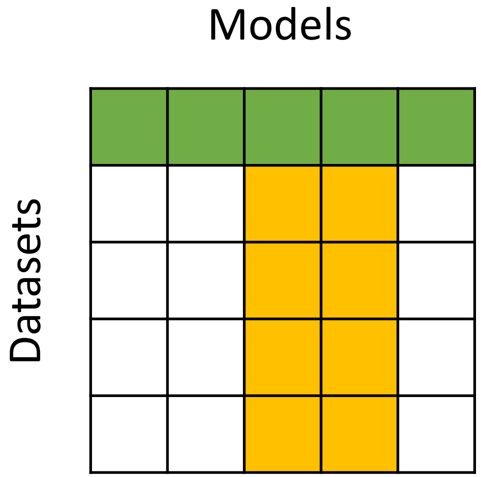
- Learning from the history of **single** dataset with a **subset** of models.
- Evaluating **unseen models** on the **same** dataset.

80% of the 400+ models are used for training and the rest 20% models are used for evaluation.

Methods	Features	ImageNet Pre-trained
feature-based	S_{LFC} [10]	0.65 ± 0.07
	S_{LogME} [60]	0.35 ± 0.09
	S_{PARC} [5]	$\mathbf{0.83} \pm 0.04$
LR (ours)	$\mathbf{d}, \mathbf{m}$	0.53 ± 0.07
	$\mathbf{d}, \mathbf{m}, S_{\text{LFC}}$	0.73 ± 0.06
	$\mathbf{d}, \mathbf{m}, S_{\text{LogME}}$	0.55 ± 0.08
	$\mathbf{d}, \mathbf{m}, S_{\text{PARC}}$	$\mathbf{0.85} \pm 0.04$
FM (ours)	$\mathbf{d}, \mathbf{m}$	0.54 ± 0.06
	$\mathbf{d}, \mathbf{m}, S_{\text{LFC}}$	0.70 ± 0.12
	$\mathbf{d}, \mathbf{m}, S_{\text{LogME}}$	0.55 ± 0.09
	$\mathbf{d}, \mathbf{m}, S_{\text{PARC}}$	$\mathbf{0.84} \pm 0.05$

Feature-based MS scores completely fail with random init, while learning-based MS can still get reasonable scores

Learning from ImageNet and Predict for New Datasets

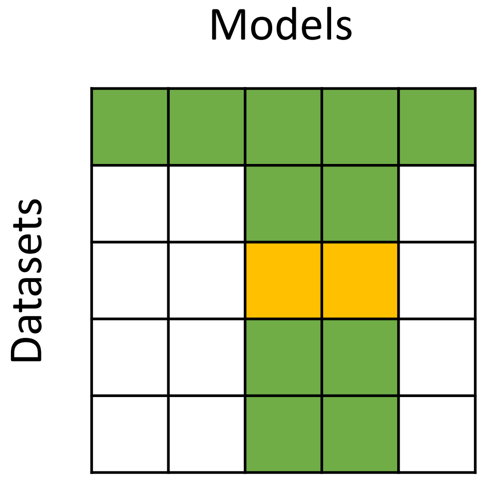


- Learning from the history of **single** dataset will **all** models.
- Evaluating **known models** on **unseen** datasets

The ImageNet column is the MS learned with all 409 ImageNet training jobs.

Methods	Features	**19 fine-grained**	**6 DomainNet**	**15 VTAB**
feature-based MS	S_{LFC} [10]	0.55	0.63	0.14
	S_{LogME} [60]	0.54	0.52	0.20
	S_{PARC} [5]	0.54	0.50	0.13
		ImageNet	ImageNet	ImageNet
LR (ours)	$\mathbf{d}, \mathbf{m}$	0.53	0.80	0.29
	$\mathbf{d}, \mathbf{m}, S_{\text{LFC}}$	**0.67**	**0.84**	**0.38**
	$\mathbf{d}, \mathbf{m}, S_{\text{LogME}}$	0.54	0.81	0.30
	$\mathbf{d}, \mathbf{m}, S_{\text{PARC}}$	0.54	0.81	0.30
FM (ours)	$\mathbf{d}, \mathbf{m}$	0.53	0.81	0.35
	$\mathbf{d}, \mathbf{m}, S_{\text{LFC}}$	**0.64**	0.82	**0.39**
	$\mathbf{d}, \mathbf{m}, S_{\text{LogME}}$	0.60	0.82	0.31
	$\mathbf{d}, \mathbf{m}, S_{\text{PARC}}$	0.56	**0.86**	0.30

Learning from All History and Predict for New Datasets



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

The column of LOO (leave-one-out) denotes MS learned with combined training history of ImageNet jobs and all downstream jobs except jobs on the test dataset

Methods	Features	**19 fine-grained**		**6 DomainNet**		**15 VTAB**	
feature-based MS	S_{LFC} [10]	0.55		0.63		0.14	
	S_{LogME} [60]	0.54		0.52		0.20	
	S_{PARC} [5]	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
	$\mathbf{d}, \mathbf{m}, S_{\text{LFC}}$	**0.67**	**0.74**	**0.84**	**0.85**	**0.38**	**0.41**
	$\mathbf{d}, \mathbf{m}, S_{\text{LogME}}$	0.54	0.65	0.81	0.84	0.30	0.36
	$\mathbf{d}, \mathbf{m}, S_{\text{PARC}}$	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
	$\mathbf{d}, \mathbf{m}, S_{\text{LFC}}$	**0.64**	**0.74**	0.82	**0.87**	**0.39**	0.41
	$\mathbf{d}, \mathbf{m}, S_{\text{LogME}}$	0.60	0.67	0.82	0.86	0.31	0.40
	$\mathbf{d}, \mathbf{m}, S_{\text{PARC}}$	0.56	0.69	**0.86**	0.86	0.30	**0.43**

Continuously Improved Model Recommendation