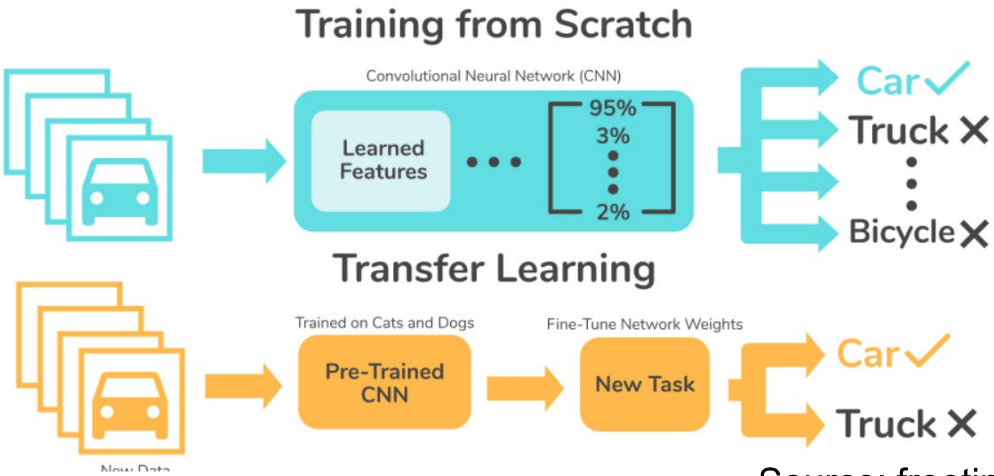
`TransferGraph`: Model Selection with Model Zoo via Graph Learning

Ziyu Li, Hilco van der Wilk, Danning Zhan
Megha Khosla, Alessandro Bozzon, Rihan Hai

TU Delft InfiniData

Transfer Learning

- Build a new model on top of a pre-trained model
- Re-train on limited data
- Work effectively for quickly training a model
- Enjoy tremendous success in both vision and language communities



Source: freetimelearning

Pre-trained models can be easily accessed nowadays

ONNX

Keras

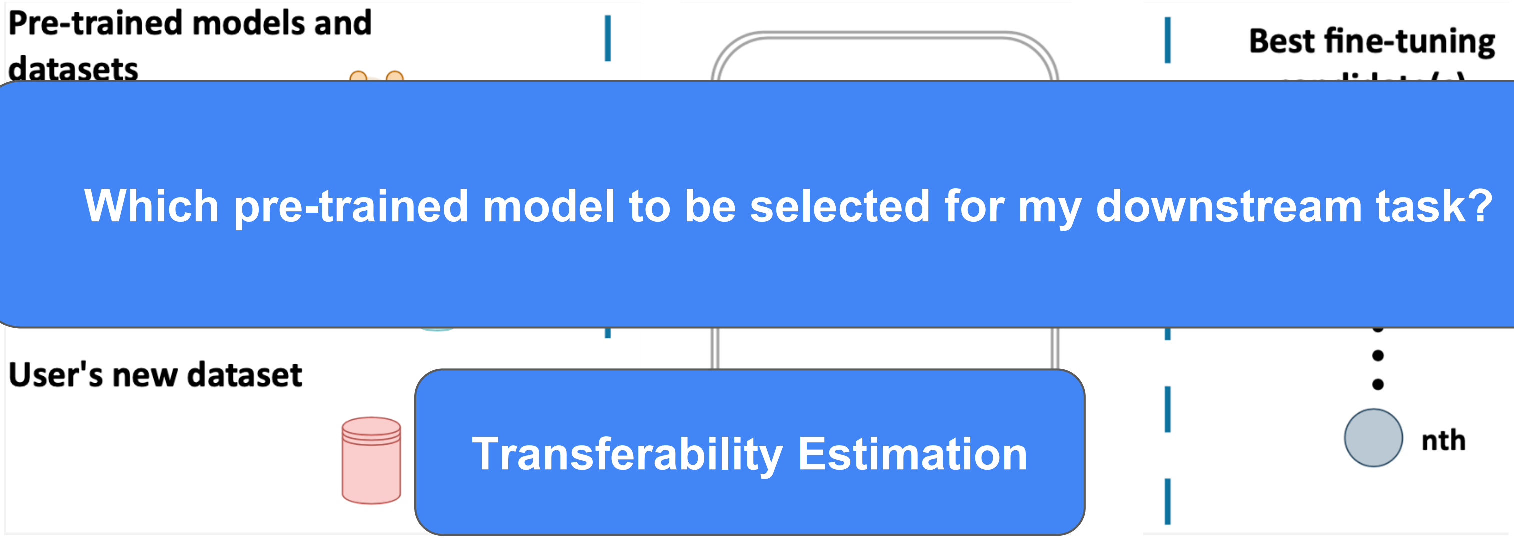
Hugging Face > 618K

kaggle > 3.4K

OpenVINO

PyTorch

Problem setting



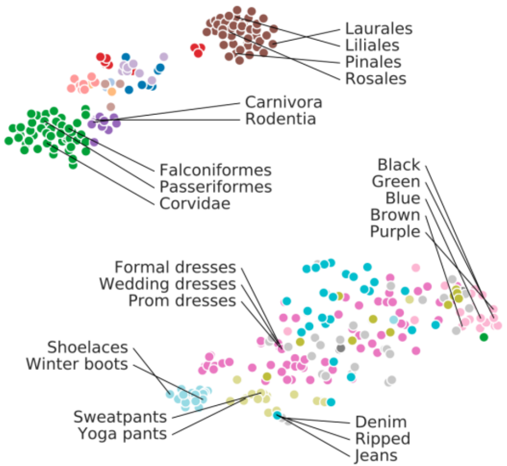
Which pre-trained model to be selected for my downstream task?

Transferability Estimation

Which pre-trained model to be selected for my downstream task?

Transferability Estimation

- **Dataset Similarity**
- Model Feature
- Learning from Fine-tuning History

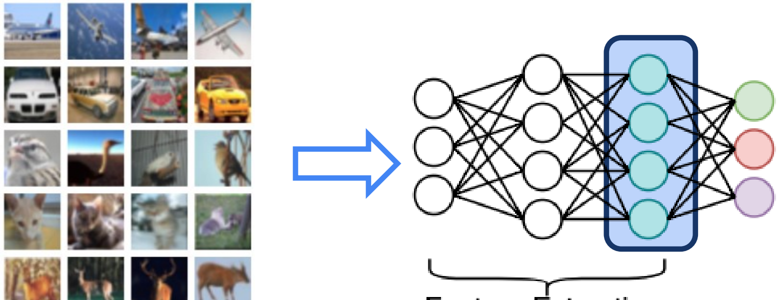


Find the most similar dataset

Which pre-trained model to be selected for my downstream task?

Transferability Estimation

- Dataset Similarity
- **Model Feature**
- Learning from Fine-tuning History



airplane
automobile
bird
cat
deer

LogME
LEEP
NCE
LFC [1]

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

[1] Deshpande, Aditya, et al. "A linearized framework and a new benchmark for model selection for fine-tuning." *arXiv preprint arXiv:2102.00084* (2021).

Which pre-trained model to be selected for my downstream task?

Transferability Estimation

- Dataset Similarity
- Model Feature
- **Learning from Fine-tuning 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

source: [2]

Amazon LR [2]

[2] Li, Hao, et al. "Guided recommendation for model fine-tuning." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

Model Selection Strategy

Transferability Estimation

Dataset Similarity

Model Feature

Learning from Fine-tuning History

LogME [3]

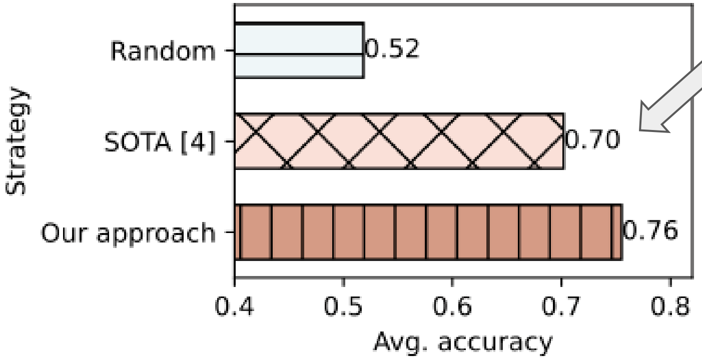
LEEP

NCE

LFC

Amazon LR

Ours



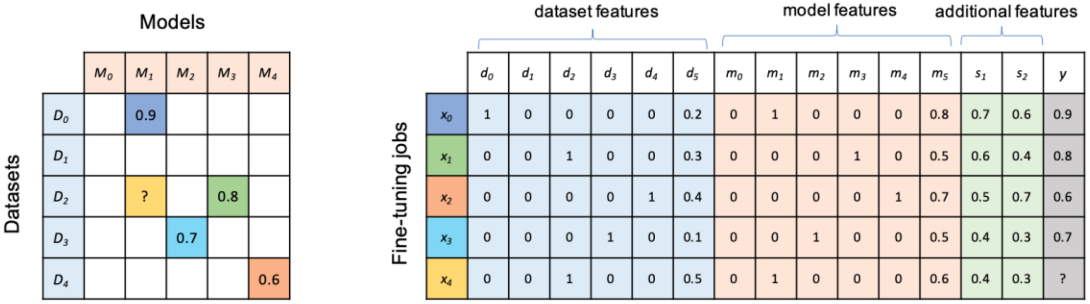
[3] You, Kaichao, et al. "Logme: Practical assessment of pre-trained models for transfer learning." *International Conference on Machine Learning*. PMLR, 2021.

Learning-based model selection strategy

- Features
  - features/metadata of dataset and model
- Label
  - Model performance on datasets

Amazon LR [2]

Ours



	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

	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	?

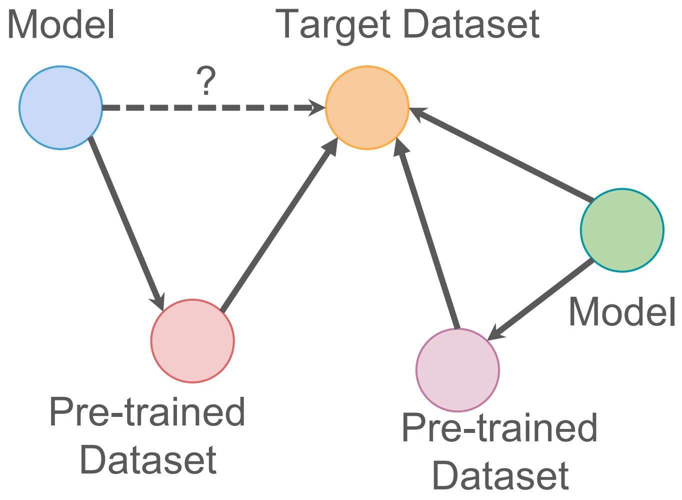
Embedding

source: [2]

Is there any other features than can be taken into account?

[2] Li, Hao, et al. "Guided recommendation for model fine-tuning." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

Model-dataset relationships as a graph

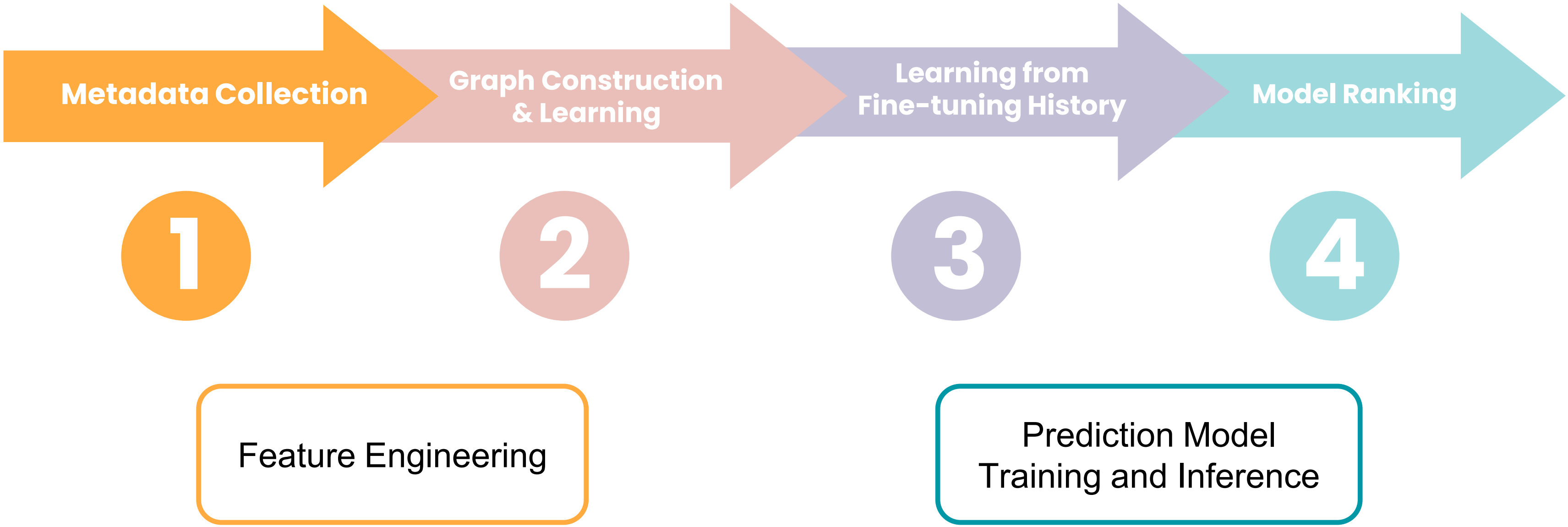


Can we learn from these inherent relationships between models and dataset from a graph?

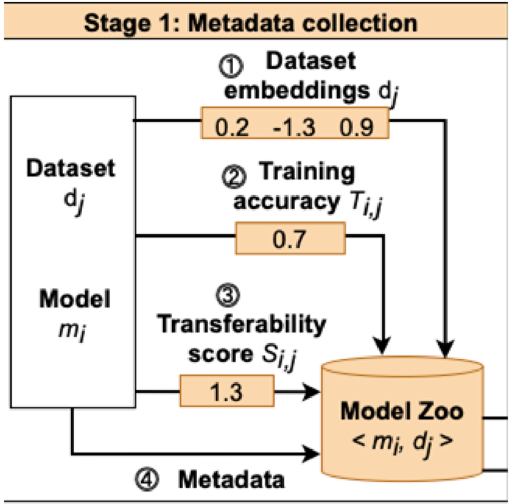
Framework overview

Mechanism:

- Learn from fine-tuning history
- **Metadata** and **node representations from graph** as features



Step 1 - Preparation



Dataset

- Number of classes
- Number of samples

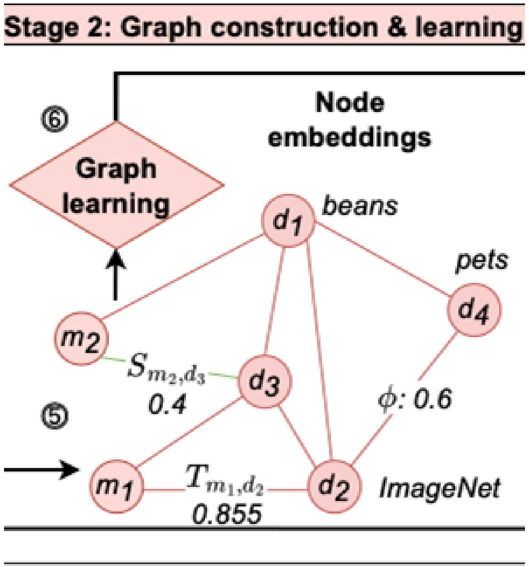
Model

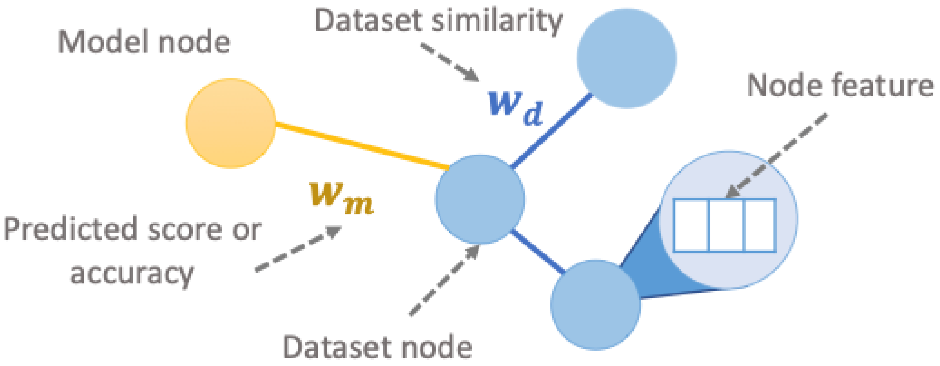
- Input shape
- Architecture
- Pre-trained dataset
- Number of parameters
- Memory consumption

Entity-wise feature

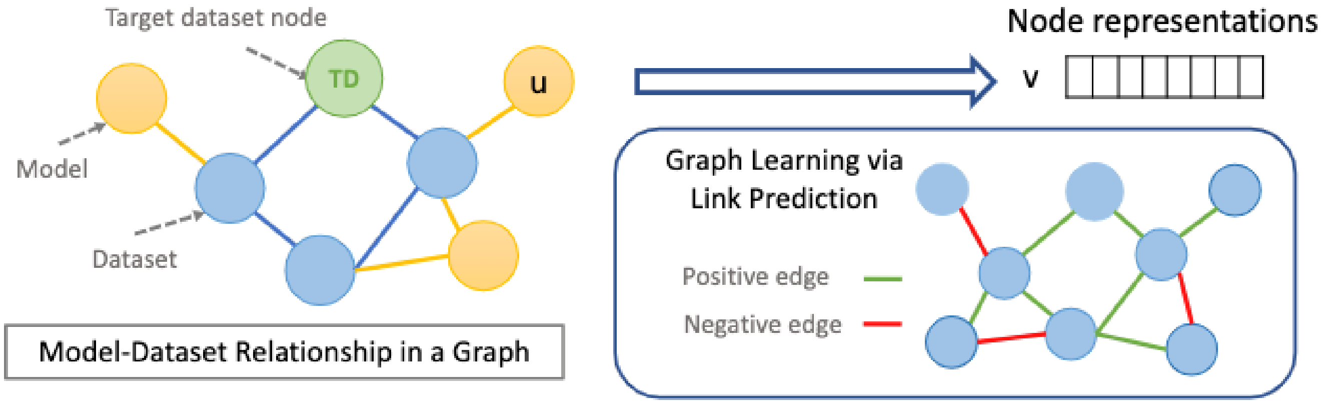
- Model performance
- Dataset representations
- Dataset similarity

Stage 2: Graph Construction & Learning





Graph construction for link prediction

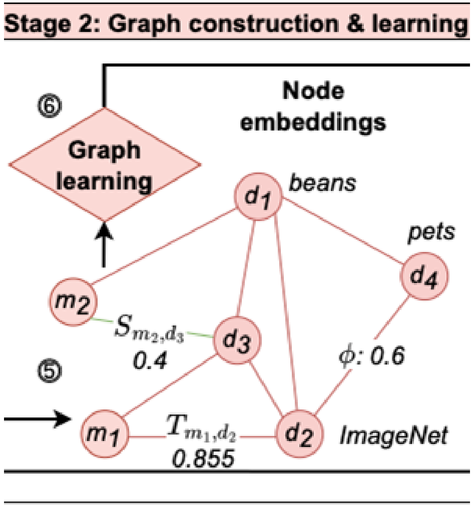


Graph learning algorithms

- Node2Vec(+)
- GraphSAGE
- GAT

Small sum up - graph construction & learning

- Exploit model-dataset, dataset-dataset relationships
- Learn inherent relationship via graph learning
- Objective: link prediction
  - Nodes with positive edges closer while further away within negative edges

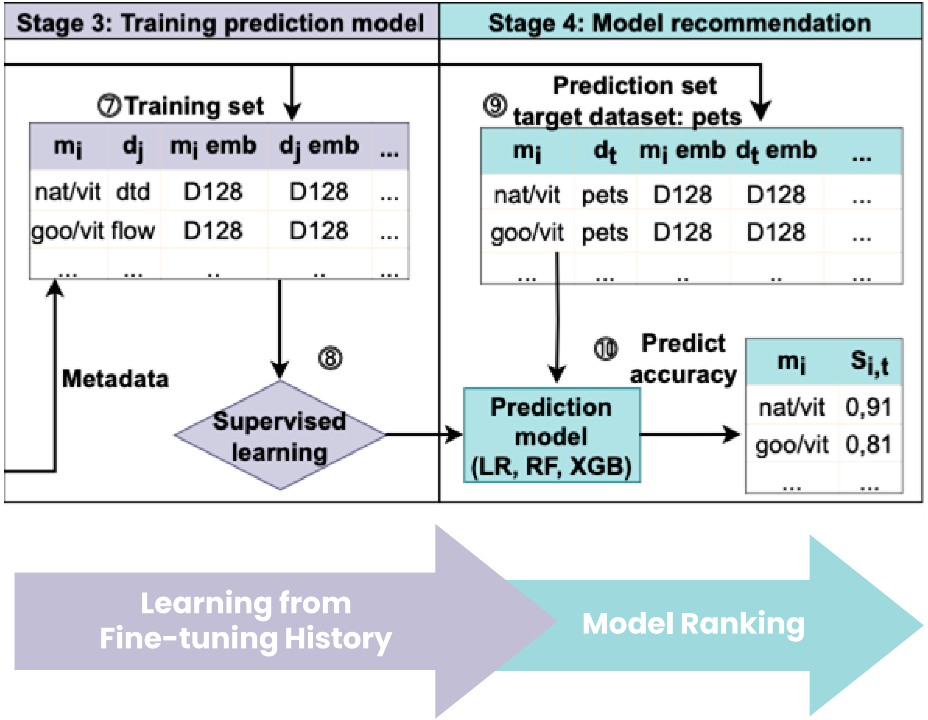


Step 3 - 4 training prediction model

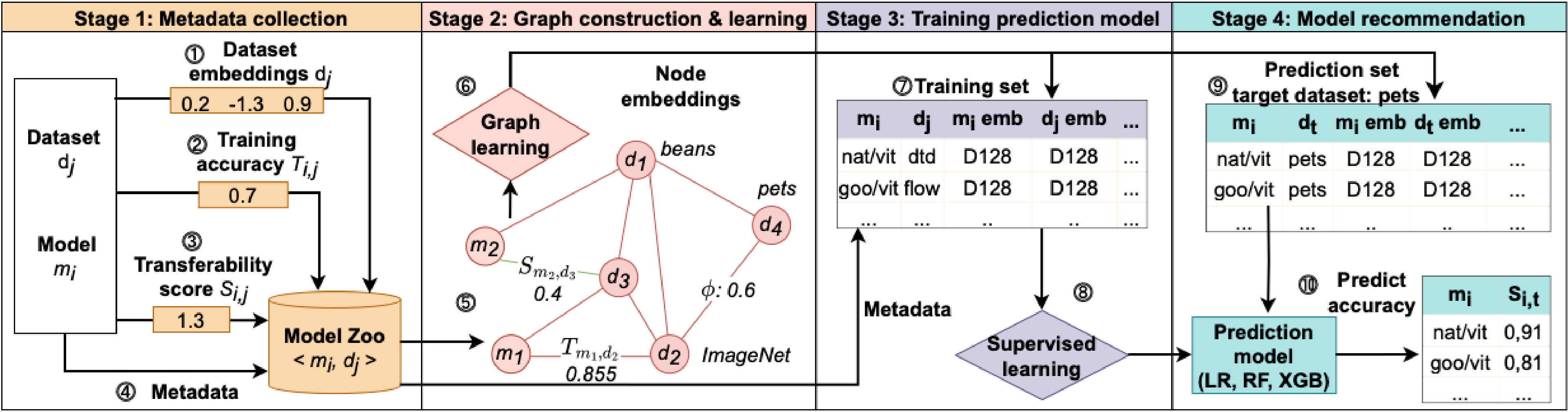
- Train regression model on past training history
- Combine metadata and graph features

Prediction models

- Linear regression (`LR`)
- Random forest (`RF`)
- XGBoost (`XGB`)



Framework overview



Experiments

Baselines

- `LogME*`
  - Mapping model features with target dataset labels
- `Amazon LR* - LR`
  - Training on fine-tuning history with basic metadata features
- Ours - `TG`
  - Training on fine-tuning history with graph extracted features (and metadata)
  - e.g., `TG:LR,N2V,all`

Our setting

- 186 image models, 164 text models
- 8 image datasets, 8 textual datasets
- 2800 fine-tuning trails

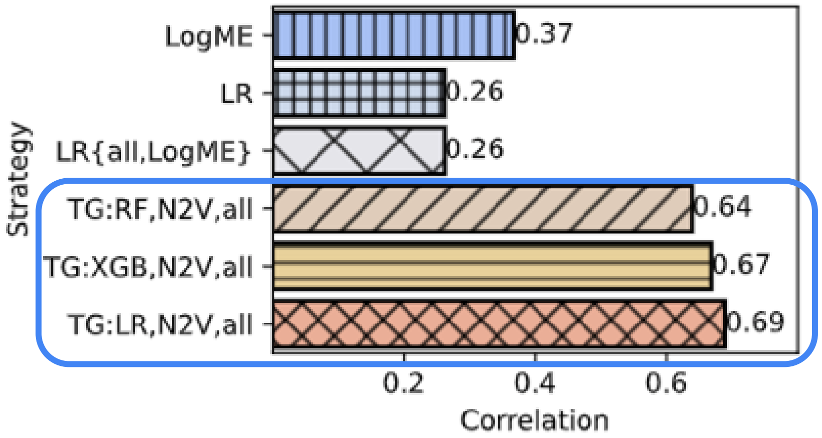
Prediction models

- Linear regression (`LR`)
- Random forest (`RF`)
- XGBoost (`XGB`)

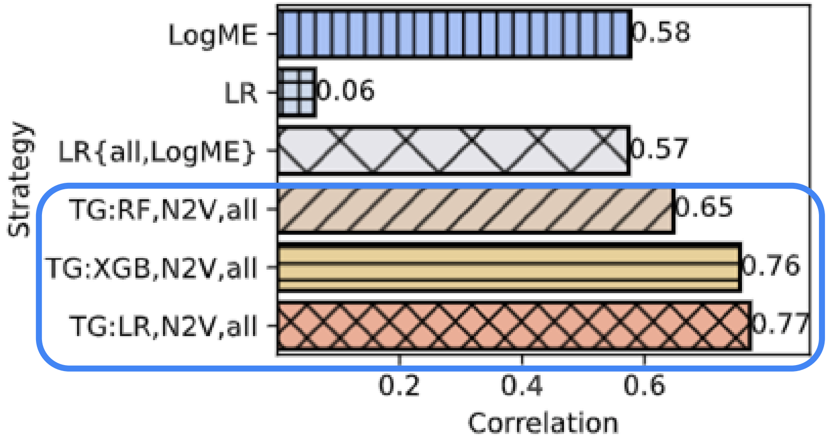
Graph learning algorithms

- Node2Vec(+) - `N2V(+)`
- GraphSAGE
- GAT

Results - Baseline comparison on model selection



(a) Image datasets



(b) Textual datasets

Up to a 32% improvement in correlation!

Results - Ablation Study

LR

- metadata

LR{all,LogME}

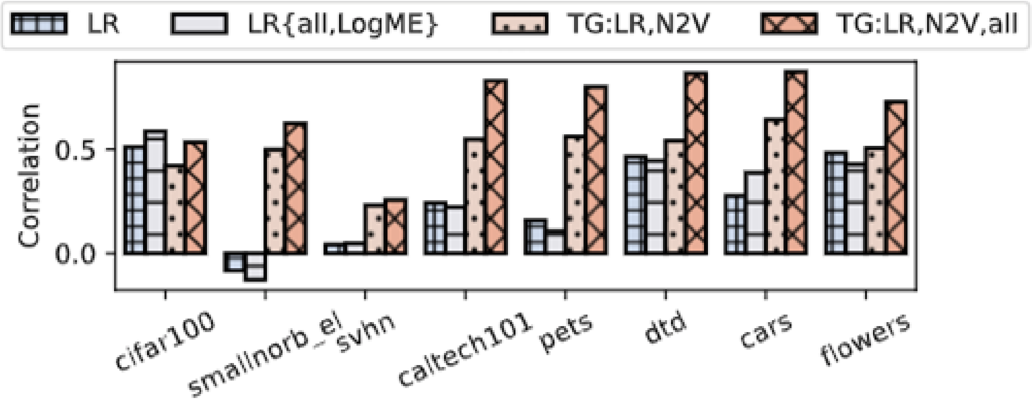
- metadata (with dataset similarity)
- LogME

TG:LR,N2V

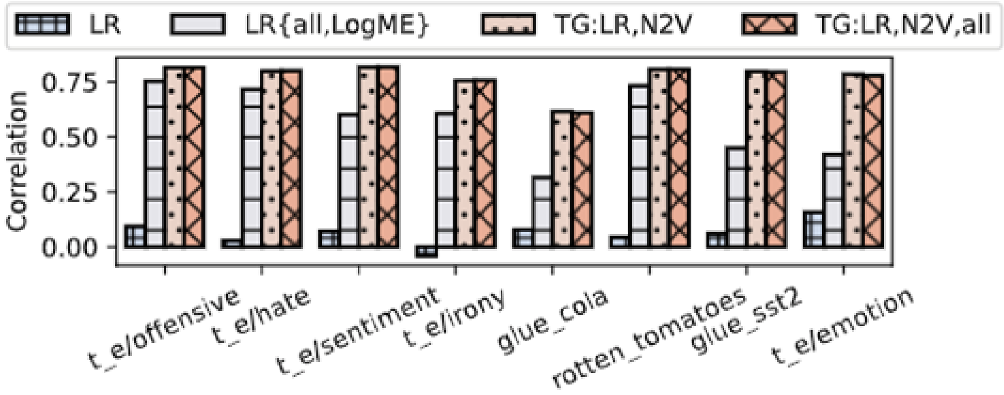
- without metadata
- only graph features

TG:LR,N2V,all

- metadata and graph features



(a) Image datasets



(b) Textual datasets

Discussion

- Cold-start problem
  - (Meta)Data preparation work
- Graph construction and learning
  - What information to be remained and removed
- Other metrics
  - Besides correlation

Take-away message (I'm open to work)

- We propose a **graph-learning-based** model selection strategy within the model zoo
- **Effectiveness** has been shown leveraging the intrinsic relationships between models and datasets for predicting the model performance
- Our model selection strategy **can continuously be improved** with more metadata and training history in the model zoo.

https://ziyuli.me/

Paper on arXiv

Reference

- [1] Deshpande, Aditya, et al. "A linearized framework and a new benchmark for model selection for fine-tuning." *arXiv preprint arXiv:2102.00084* (2021).
- [2] Li, Hao, et al. "Guided recommendation for model fine-tuning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
- [3] You, Kaichao, et al. "Logme: Practical assessment of pre-trained models for transfer learning." *International Conference on Machine Learning*. PMLR, 2021.

Properties of the constructed graph

Graph property		
Modality	image	text
graph type	homogenous	homogenous
Threshold on transferability score for edge pruning	0.5	0.5
Threshold on accuracy for edge pruning	0.5	0.5
Threshold of negative edge identification on accuracy	0.5	0.5
Number of nodes	265	188
Average node degree*	20.1	8.6
Number of dataset-dataset edge	5256	550
Number of model-dataset edge with accuracy weight*	1753	918
Number of model-datset edge with transferability weight*	916	419

More statistics

Image Classification Models		Text Classification Models	
Architecture	**Count**	**Architecture**	**Count**
vit	114	bert	80
swin	25	distilbert	37
convnext	24	roberta	27
beit	9	xlm-roberta	5
deit	5	electra	5
van	5	albert	3
resnet	1	fnet	2
data2vec-vision	1	camembert	2
		deberta-v2	1
		perceiver	1
		data2vec-text	1
Range of Parameters			
<100k	2		1
100k-1M	0		0
1M-10M	5		1
10M-100M	155		46
100M-1B	22		116