Frustratingly Easy Transferability Estimation

Long-Kai Huang, Junzhou Huang, Yu Rong,
Qiang Yang, Ying Wei

Tencent AI Lab, City University of Hong Kong

Transferability

An Important Question in Transfer Learning

Which pre-trained model (source/architecture) and which layers of it should be transferred to benefit the target task the most?

Source Selection, Model Selection, Layer Selection

Transferability measure - Goal

To select a pre-trained model prior to training on a target task

Desired properties

1. Effectiveness
2. Computation-efficiency
   - Free of training on target tasks
   - Free of optimization
3. Widely applicable to
   - different per-training models
   - different layers
4. [Optional] Free of assessing source data

Summary of the existing transferability measures and ours

		Computation-efficiency		Wide Application	
Measures	Free of Assessing Source	Free of Training on Target	Free of Optimization	Applicable to Unsupervised Pre-trained Models	Applicable to Layer Selection
Taskonomy (Zamir et al., 2018)	×	×	√	√	√
Task2Vec (Achille et al., 2019)	×	×	×	√	×
RSA (Dwivedi & Roig, 2019)	√	×	√	√	√
DEPARA (Song et al., 2020)	√	×	√	√	√
$\mathcal{N}$LEEP (Li et al., 2021)	√	×	√	√	√
DS (Cui et al., 2018)	×	√	×	√	×
(Zhang et al., 2021) [1]	×	√	×	×	×
(Tong et al., 2021) [2]	×	√	×	×	×
NCE (Tran et al., 2019)	×	√	√	×	×
H-Score (Bao et al., 2019)	√	√	×	√	×
LogME (You et al., 2021)	√	√	×	√	×
LEEP (Nguyen et al., 2020)	√	√	√	×	×
TransRate	√	√	√	√	√

[1] Zhang, G., Zhao, H., Yu, Y., and Poupart, P. Quantifying and improving transferability in domain generalization. NeurIPS, 2021.

[2] Tong, X., Xu, X., Huang, S.-L., and Zheng, L. A mathemat- ical framework for quantifying transferability in multi- source transfer learning. NeurIPS, 2021.

Computation-Efficient Transferability Estimation: TransRate

Our Propose: TransRate

Mutual Information between the feature extracted by the pre-trained model and the labels.

$$TrR_{T_s \to T_t}(g) := h(Z) - h(Z|Y) \approx H\left(Z^{\Delta}\right) - H\left(Z^{\Delta}\middle|Y\right)$$

Applicable to ***layer selection****.*

Relation to transfer performance

Proposition 1. *Assume the target task has a uniform label distribution, i.e.* $p(Y = y^c) = \frac{1}{C}$ *holds for all* $c = 1,2,\ldots,C$. *We then have:*

$$TrR_{T_s \to T_t}(g) - H(Y) \gtrsim \mathcal{L}(g, w^*) \gtrsim TrR_{T_s \to T_t}(g) - H(Y) - H(Z^{\Delta})$$

Entropy, Rate Distortion (ε-Entropy) and Coding Rate

Difficulty of Entropy (Mutual Information) Estimation

bin-based	requires an extremely large memory capacity.
kernel density estimator	require sufficient number of sample
k-NN estimator	require exhaustive computation of nearest neighbors of all examples
NN-based (e.g. MINE)	require training a neural network

NOT Applicable

Entropy, Rate Distortion (ε-Entropy) and Coding Rate

Rate Distortion

$R(Z, \epsilon)$

$$R(Z, \epsilon) = \min_{p(\tilde{Z}|Z); d(\tilde{z}, z) \leq \epsilon} MI(Z, \tilde{Z})$$

where $\tilde{Z}$ is a reconstructed version of Z.

$\Delta = \sqrt{2\pi e \epsilon}$

Shannon Entropy

$\lim_{\Delta \to 0} H(Z^{\Delta})$

Coding Rate

$R(\hat{Z}, \epsilon)$

$$R(Z, \epsilon) = h(Z) + \frac{1}{2} \log \frac{1}{2\pi e \epsilon} + o(1)$$

Let $\Delta = \sqrt{2\pi e \epsilon}$ and let $\Delta \to 0$,

$$R(Z, \epsilon) = H(Z^{\Delta}) + o(1)$$

$$R(\hat{Z}, \epsilon) = \frac{1}{2} \log \det \left(I + \frac{1}{\epsilon} \frac{\hat{Z}\hat{Z}^T}{n} \right)$$

where $\hat{Z}$ is the features matrix

Coding Rate based TransRate

Our Propose: TransRate

$$TrR_{T_s \to T_t}(g) \approx H(Z^{\Delta}) - H(Z^{\Delta}|Y) \approx R(\hat{Z},\epsilon) - R(\hat{Z},\epsilon|\mathrm{Y})$$

We resort to coding rate $R(\hat{Z},\epsilon)$ as an approximation of $H(Z^{\Delta})$ with a small $\Delta = \sqrt{2\pi e \epsilon}$.

$$H(Z^{\Delta}) \approx R(\hat{Z},\epsilon) = \frac{1}{2}\log\det\left(I + \frac{1}{\epsilon}\frac{\hat{Z}\hat{Z}^T}{n}\right)$$

$$H(Z^{\Delta}|Y) \approx \sum_{c=1}^{C}\frac{n_c}{n}R(\hat{Z}^c,\epsilon) = \sum_{c=1}^{C}\frac{n_c}{2n}\log\det\left(I + \frac{1}{\epsilon}\frac{\hat{Z}^c\hat{Z}^{c\,T}}{n}\right) := R(\hat{Z},\epsilon|Y)$$

Computational Efficient !

Experiments

32 pre-trained models and 16 downstream tasks

Source Selection, Model Selction, Layer Selection

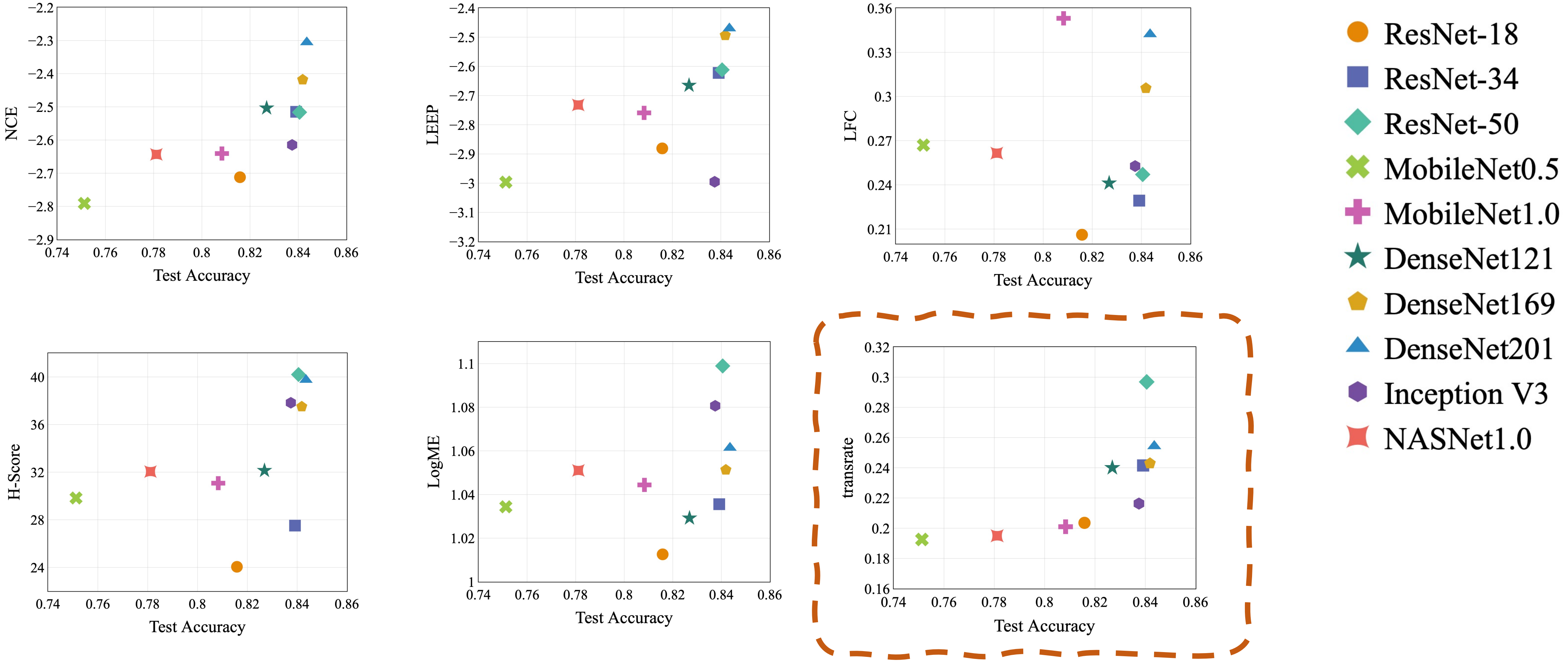
Supervised-trained models, Self-supervised trained models

Classification tasks, Regression tasks

Evaluation measure: correlation coefficient

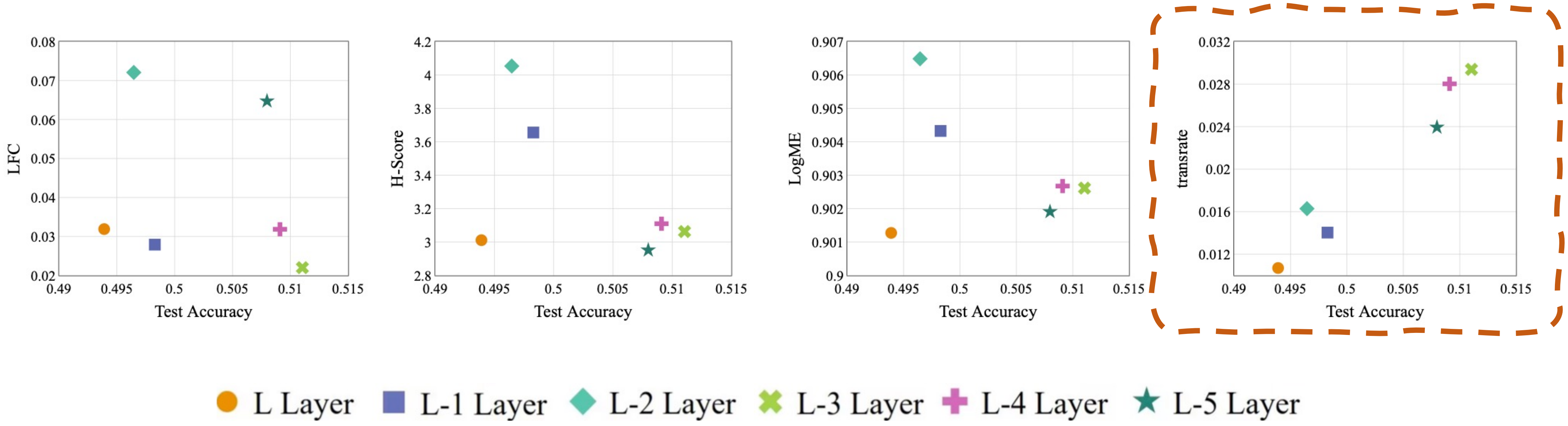
Pearson R_p, Kendall's τ_K, Weighted τ_w

Model Selection



Target Datasets	Measures	NCE	LEEP	LFC	H-Score	LogME	TransRate
CIFAR-100	R_p	0.7937	0.8506	-0.2159	0.5016	0.4965	**0.8780**
	τ_K	0.7436	0.7179	-0.0256	0.4872	0.4103	**0.9231**
	τ_ω	0.8315	0.8485	-0.0126	0.6058	0.5130	**0.8498**

Layer Selection



	Measures	LFC	H-Score	LogME	TransRate
Source: SVHN Model: ResNet-20	R_p	-0.1895	-0.5320	-0.3352	**0.9769**
	τ_K	-0.4667	-0.2000	-0.0667	**0.8667**
	τ_ω	-0.5497	-0.2993	-0.2340	**0.9265**

Comparison of the computational cost

	ResNet-18, Full Data		ResNet-18, Small Data		ResNet-50, Full Data	
	Wall-clock time (second)	Speedup	Wall-clock time (second)	Speedup	Wall-clock time (second)	Speedup
Fine-tune	8399.65	1×	882.33	1×	2.3×10^4	1×
Extract feature	30.1416		3.2986		72.787	
NCE	0.9126	9,204×	0.2119	4,164×	2.1220	10,839×
LEEP	0.7771	10,808×	0.1211	7,286×	1.9152	12,009×
LFC	30.1416	279×	0.7987	1,106×	149.3040	154×
H-Score	1.6285	5,158×	0.3998	2,207×	13.07	1,760×
LogME	9.2737	906×	2.0224	436×	50.1797	458×
TransRate	**1.3410**	**6,264**×	**0.2697**	**3,272**×	**10.6498**	**2,160**×

Summary

- A simple, efficient, and effective transferability measure named TransRate
    - Applicable to layer selection
- Coding Rate as an effective alternative to entropy in mutual information estimation
- Remarkably good performance in experiments in model selection, layer selection.