Test-time Assessment of a Model's Performance on Unseen Domains via Optimal Transport

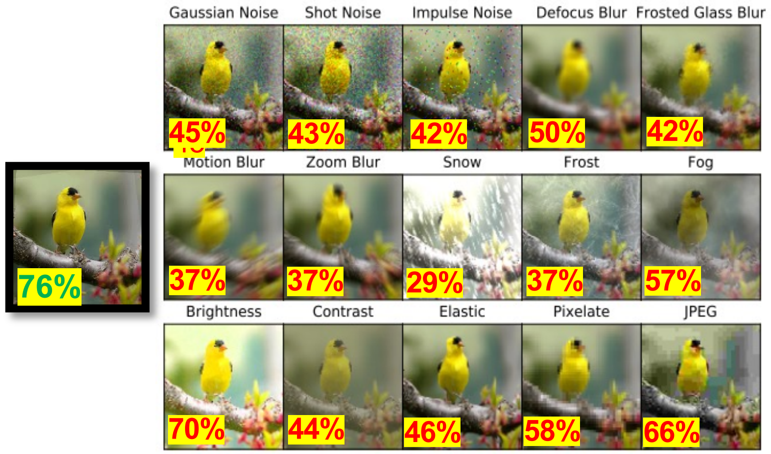
Akshay Mehra

Joint-work with **Yunbei Zhang** and **Jihun Hamm**.

Tulane University

Motivation

- Machine learning (ML) models often encounter data from domains unseen during training time.
- Performance of ML models suffers when faced with data from unseen domains.
- This makes the performance on in-distribution data is a poor indicator of their performance on unseen domains.
- Thus, metrics that can gauge the performance of ML models at test-time (a.k.a. ***transferability***) without access to labels are essential.

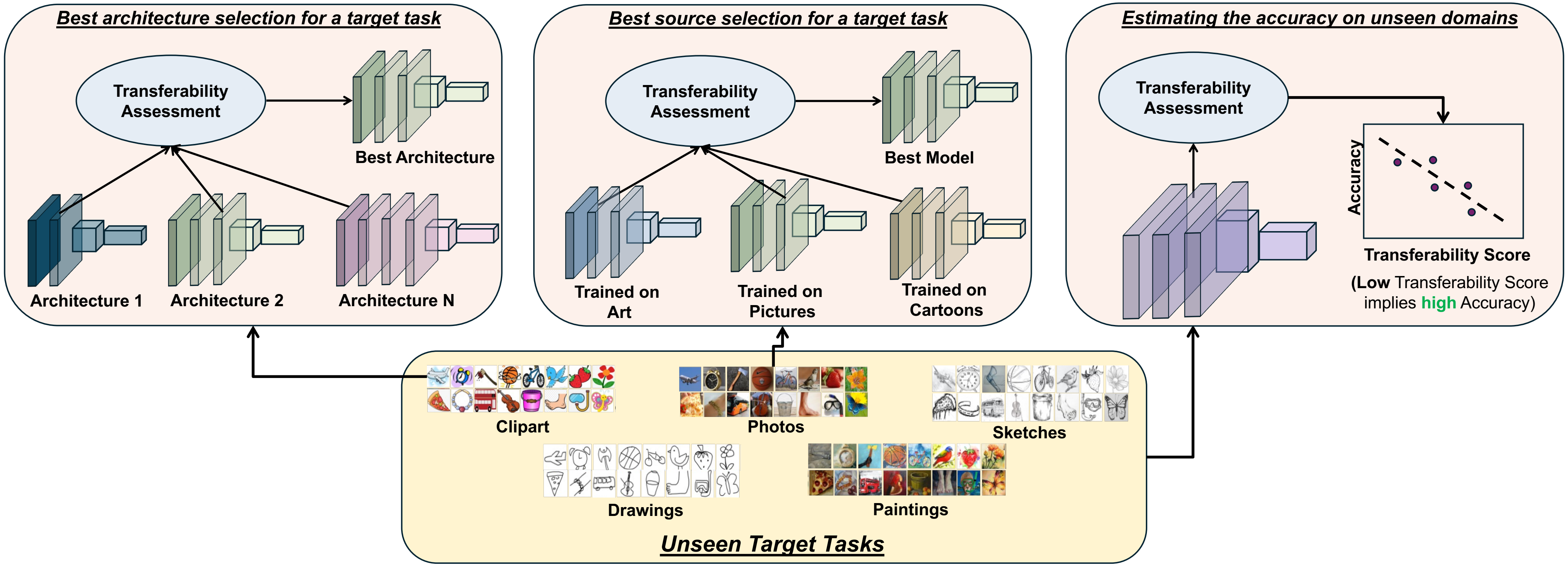


True label (top), predicted label (bottom).

Hendrycks, Dan, and Thomas Dietterich. “Benchmarking neural network robustness to common corruptions and perturbations.”
Nam, Hyeonseob, et al. "Reducing domain gap by reducing style bias." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

Motivation

- Such transferability estimation metrics can be useful for various practical applications.



Background on Optimal Transport

- Suppose every day you have move bread from M bakeries to N cafes
  - Bakery distribution: $P = \sum_{i=1}^{M} p_i \delta_i$
  - Cafe distribution: $Q = \sum_{j=1}^{N} q_j \delta'_j$
  - c_{ij}: (base) distance of the i^{th} bakery to j^{th} cafe
  - π_{ij}: amount of bread to be moved from i^{th} bakery to j^{th} cafe
  - $\mathbb{E}_{\pi}[c_{ij}] \coloneqq \sum_{ij} \pi_{ij} c_{ij}$ is the total cost

- Earth Movers Distance is defined as $\inf_{\pi \in \Pi} \mathbb{E}_{\pi}[c_{ij}]$.

- More generally, for two distributions (discrete/continuous) defined on a metric space $\mathbf{OT}(\mathbf{P}, \mathbf{Q}) \coloneqq \inf_{\pi \in \Pi(\mathrm{P},\mathrm{Q})} \mathbb{E}_{(x_1,x_2) \in \pi}[c(x_1, x_2)]$.

Test-time Estimation of Transferability via OT

- **Transferability** of a model trained on the source domain S to an unseen target domain T is defined as the model's accuracy on T i.e.,

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\in \boldsymbol{P_T}(\boldsymbol{x},\boldsymbol{y})}\left[\boldsymbol{accuracy}\left(\boldsymbol{h}\left(\boldsymbol{g}(\boldsymbol{x})\right),\boldsymbol{y}\right)\right],$$

where $g: \mathcal{X} \to \mathcal{Z}$ is the encoder and $h: \mathcal{Z} \to \mathcal{Y}$ is the classifier.

- The **base distance** c between two points is defined as

$$\boldsymbol{c}\left((\boldsymbol{x_S},\boldsymbol{y_S}),(\boldsymbol{x_T},\widehat{\boldsymbol{y}}_{\boldsymbol{T}})\right) \coloneqq \boldsymbol{c_{features}}(\boldsymbol{x_S},\boldsymbol{x_T}) + \boldsymbol{\lambda}\cdot \boldsymbol{c_{labels}}(\boldsymbol{y_S},\widehat{\boldsymbol{y}}_{\boldsymbol{T}}).$$

- We measure transferability of a model to T at test time using

$$\boldsymbol{TETOT} \coloneqq \boldsymbol{OT_c}(\boldsymbol{P_S},\boldsymbol{P_T}) \coloneqq \inf_{\boldsymbol{\pi}\in\Pi(\boldsymbol{P_S},\boldsymbol{P_T})} \mathbb{E}_{\boldsymbol{\pi}}[\boldsymbol{c}((\boldsymbol{x_S},\boldsymbol{y_S}),(\boldsymbol{x_T},\widehat{\boldsymbol{y}}_{\boldsymbol{T}}))]$$

Algorithm to compute TETOT

- We use labeled samples from S and unlabeled samples from T.
- We define the feature cost as $c_{features} \coloneqq \|g(x_S) - g(x_T)\|_2$ where $g(\cdot)$ denotes the features extracted from the encoder.
- We define the label cost as $c_{labels} \coloneqq \|y_S - h(g(x_T))\|_2$ where $h(g(\cdot))$ denotes the pseudo-labels from the model.

 # Select samples from S and T.
Randomly sample m samples, $(x_S^i, y_S^i) \sim \mathcal{D}_S$
Randomly sample n samples, $(x_T^j) \sim \mathcal{D}_T$

 # Compute pairwise cost.
for $i = 1, \dots, m$ and j $= 1, \dots, n$ **do**
$c_{features}^{ij} \coloneqq \|g(x_S^i) - g(x_T^j)\|_2$
$c_{labels}^{ij} \coloneqq \|y_S^i - h(g(x_T^j))\|_2$

$c \coloneqq c_{features} + \lambda \cdot c_{labels}$

$$\boldsymbol{TETOT} \coloneqq min_{\pi \in \Pi(P_S, P_T)} \sum_{i,j} \pi^{ij} \cdot c^{ij}$$

$$s.t. \sum_j \pi^{ij} = \frac{1}{m} \forall i, \sum_i \pi^{ij} = \frac{1}{n} \forall j$$

Empirical results

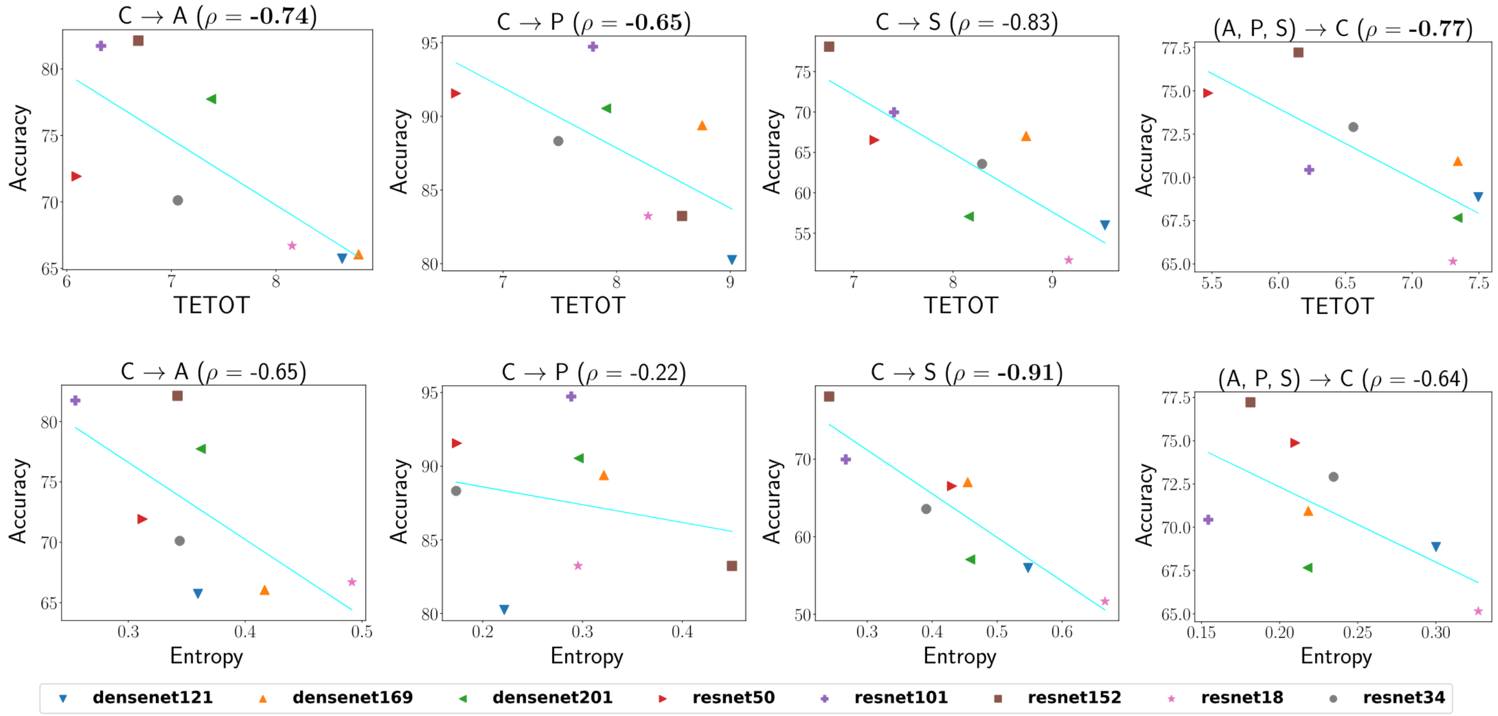
- We present evaluations on PACS and VLCS datasets and their variations in single and multiple source domain settings.

- We show the correlation of TETOT with transferability on
  - Best architecture selection for a given target task.
  - Best source selection for a given target task.
  - Estimating accuracy of unseen domains.

- We compare the correlation of TETOT with transferability with the popular entropy-based metric dependent only on the target domain data.

TETOT for model selection

- Given a target task, identify the best model architecture to use.



Dataset	Entropy	TETOT
PACS	-0.40	**-0.62**
VLCS	-0.29	**-0.40**
Average	-0.35	**-0.51**

Best source selection for a given target task

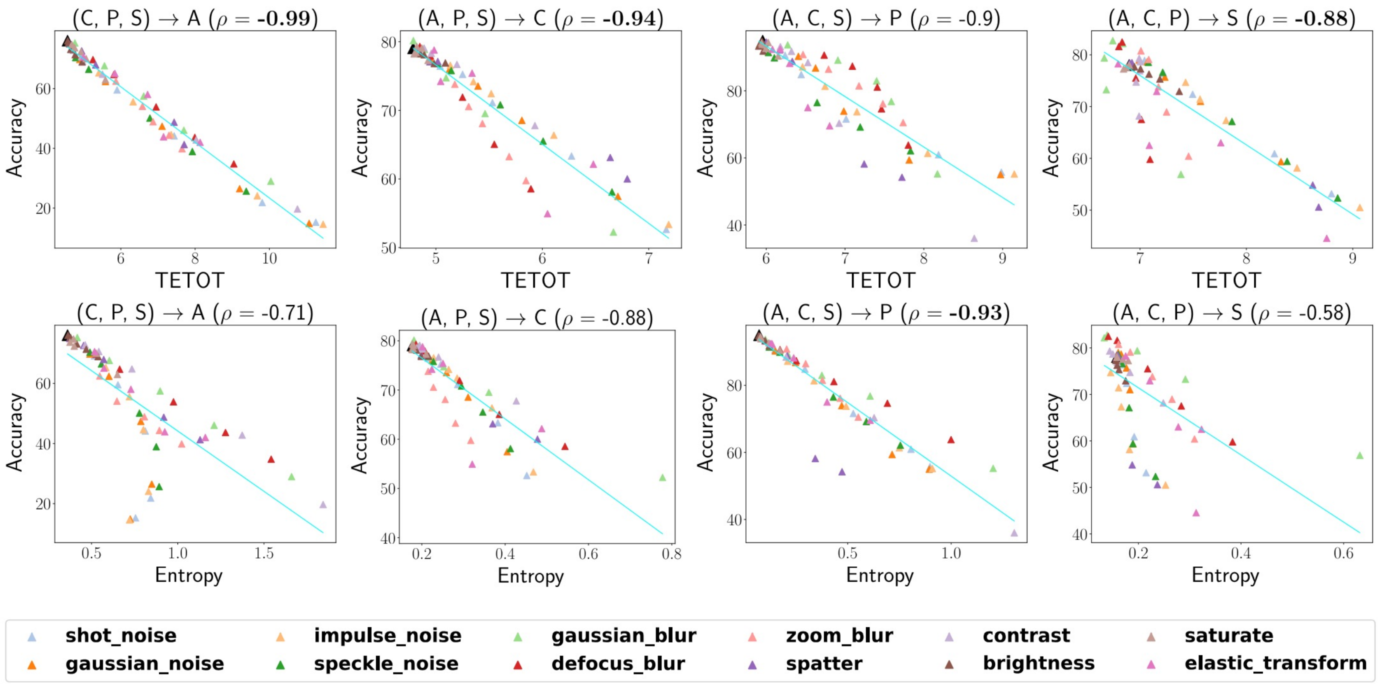
- Given a target task, select the model trained on the best source domain.

- We use the ResNet-50 model architecture trained in both single/multiple domain setting.

Dataset	Entropy	TETOT
PACS	-0.47	**-0.94**
VLCS	-0.58	**-0.92**
Average	-0.53	**-0.93**

Estimating transferability of unseen domains

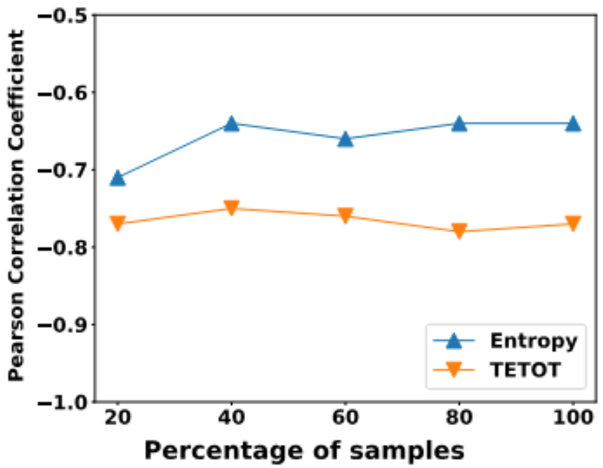
- Predict the transferability of a model on various unseen domains.



Dataset	Entropy	TETOT
PACS	-0.39	**-0.93**
VLCS	-0.34	**-0.80**
Average	-0.36	**-0.86**

Advantages of TETOT

- TETOT can be computed
  - Using a few samples from the two domains.
  - Using only the statistics from the two domains.
- Results on the architecture selection problem.



Metric	Pearson Corr. Coeff.
TETOT-approx	-0.60
TETOT	-0.75

Conclusion

- Estimating transferability at test time without access to labels of the target data is essential for various practical applications.
- We focused on proposing an efficiently computable metric (TETOT) to gauge transferability based on Optimal Transport distance between the source and the target domains.
- TETOT outperforms entropy and achieves a better correlation with transferability on the problems of model selection and predicting performance on unseen domains.

Paper