Active Transferability Estimation

Supplementary Material

In this appendix, we present the following details, which we could not include in the main paper due to space constraints.

- Theoretical discussion of our method, as presented in § S1
- Further investigation on the nature of samples in our informative subsets, as presented in § S2, and Figures S3, S4, and S5
- Further details on the experimental setup for all our experiments in Sec. 4, as presented in § S3
- Experimental results using additional baselines, alternate model architectures, correlation metrics, and experimental parameters and settings, as presented in § S4
- Additional ablation studies for other key components of ACT, as presented in § S5

S1. Theoretical Properties

Here, we show that ACT-augmented metrics inherit the theoretical properties of their baseline metric. Note that showing theoretical bounds for all transferability metrics is outside the scope of this work. Hence, we take one representative metric (LEEP) and show that ACT-LEEP retains its theoretical properties.

Preliminaries on Probability Estimations. Let $f_\theta^s(\mathbf{x}_i)$ denote the output softmax scores of the source model over the source dataset label space $\mathcal{Z}$ for an instance $\mathbf{x}_i$. We construct a "source label distribution" of the target dataset over the source label space $\mathcal{Z}$ by passing them through f_θ^s, and subsequently use it to build an empirical joint distribution over the source and target label spaces, *i.e.*, $\hat{P}(y,z) = \frac{1}{n}\sum_{i:y_i=y} f_\theta^s(\mathbf{x}_i)_z$, where $\mathcal{Y}$ denotes the target label space for dataset $\mathcal{D}_t$. Finally, the empirical marginal and conditional distributions can be computed using $\hat{P}(z)=\sum_{y\in\mathcal{Y}}\hat{P}(y,z)$ and $\hat{P}(y|z)=\frac{\hat{P}(y,z)}{\hat{P}(z)}$ respectively.

LEEP. Let source model f_θ^s predict the target label y by directly drawing from the label distribution $p(y|\mathbf{x}; f_\theta^s, \mathcal{D}_t^{\text{train}})=\sum_{z\in\mathcal{Z}}\hat{P}(y|z)f_\theta^s(\mathbf{x})_z$. The LEEP score is then defined as average log-likelihood:

$$\text{LEEP} = \mathcal{T}(f_\theta^s, \mathcal{D}_t^{\text{train}}) = \frac{1}{n}\sum_{i=1}^{n}\log\Big(\sum_{z\in\mathcal{Z}}\hat{P}(y|z)f_\theta^s(\mathbf{x})_z\Big) \tag{16}$$

Average Log-Likelihood. We fix the source model weights θ and re-train the classification model using maximum likelihood and $\mathcal{D}_t^{\text{train}}$ to obtain a new classifier f_θ^*, i.e.,

$$f_\theta^* = \arg\max_{k\in\mathcal{K}} \; l(\theta, k), \tag{17}$$

where $l(\theta,k)$ is the average likelihood for the weights θ and k on the target dataset $\mathcal{D}_t^{\text{train}}$, and k is selected from a space of classifiers $\mathcal{K}$.

Lemma 1. ACT-LEEP *is a lower bound of the optimal average log-likelihood for the informative subset.*

$$\mathcal{T}(f_\theta^s, \mathcal{D}_t^{inf}) \le l(w,k^*)^{inf} \le l(w,k^*) \tag{18}$$

Proof. This proof is true by definition as $\mathcal{D}_t^{\text{inf}}\subset\mathcal{D}_t^{\text{train}}$ represents the informative subset of the target dataset. Note that $l(w,k^*)$ is the maximal average log-likelihood over $k\in K$, and $\mathcal{T}(f_\theta^s,\mathcal{D}_t^{\text{train}})$ is the average log-likelihood in K. From [44] we know $\mathcal{T}(f_\theta^s,\mathcal{D}_t^{\text{train}}) \le l(w,k^*)$ and by definition of $\mathcal{D}_t^{\text{inf}}$, $\mathcal{T}(f_\theta^s,\mathcal{D}_t^{\text{inf}}) \le l(w,k^*)$. In addition, the model struggles to learn the samples in the informative subset, and hence $l(w,k^*)^{\text{inf}} \le l(w,k^*)$ □

Lemma 2. ACT-LEEP *is an upper bound of the NCE measure plus the average log-likelihood of the source label distribution, computed over the informative subset, i.e.,*

$$\mathcal{T}(f_\theta^s, \mathcal{D}_t^{inf}) \ge \text{ACT-}NCE(Y \mid Z) + \frac{1}{|\mathcal{D}_t^{inf}|}\Sigma_{i=1}^{|\mathcal{D}_t^{inf}|}\log f_\theta^s(\mathbf{x}_i)_{z_i}, \tag{19}$$

Proof Sketch. Let z_i be the dummy labels obtained when computing NCE and y_i be the true labels.

$$\begin{aligned}
\mathcal{T}(f_\theta^s, \mathcal{D}_t^{\text{inf}}) &= \frac{1}{|\mathcal{D}_t^{\text{inf}}|}\sum_{i=1}^{|\mathcal{D}_t^{\text{inf}}|}\log\left(\sum_{z\in\mathcal{Z}}\hat{P}(y_i|z)f_\theta^s(\mathbf{x}_i)_z\right) \\
&\ge \frac{1}{|\mathcal{D}_t^{\text{inf}}|}\Sigma_{i=1}^{|\mathcal{D}_t^{\text{inf}}|}\log\left(\hat{P}(y_i|z_i)f_\theta^s(\mathbf{x}_i)_{z_i}\right) \\
&= \frac{1}{|\mathcal{D}_t^{\text{inf}}|}\Sigma_{i=1}^{|\mathcal{D}_t^{\text{inf}}|}\log\hat{P}(y_i|z_i) + \frac{1}{|\mathcal{D}_t^{\text{inf}}|}\Sigma_{i=1}^{|\mathcal{D}_t^{\text{inf}}|}\log f_\theta^s(\mathbf{x}_i)_{z_i} \\
&= \text{ACT-NCE}(Y \mid Z) + \frac{1}{|\mathcal{D}_t^{\text{inf}}|}\Sigma_{i=1}^{|\mathcal{D}_t^{\text{inf}}|}\log f_\theta^s(\mathbf{x}_i)_{z_i}.
\end{aligned}$$

□

Empirical Analysis. We analytically evaluated the upper and lower bounds for ACT-LEEP by computing the RHS of Equations 18-19. In Figure S1, our results show ACT-LEEP and its corresponding theoretical upper and lower bounds, confirming that, across seven source model architectures, none of our theoretical bounds are violated. In addition, we empirically demonstrate that our bounds are tighter than LEEP.

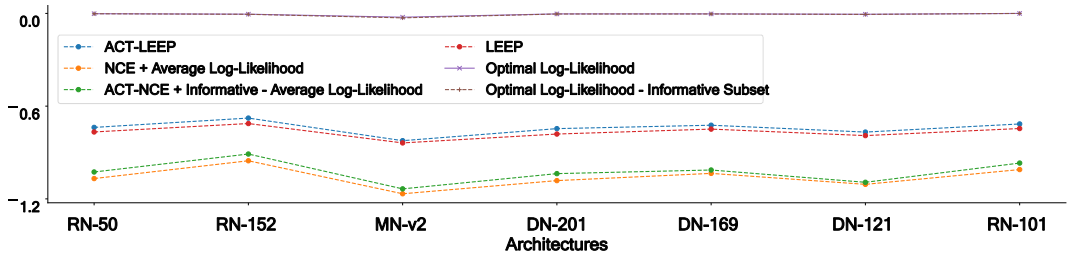

Figure S1. Empirically calculated ACT-LEEP (in blue), our theoretical upper (in purple) and lower (in green) bounds from Eqns. 18-19 across seven source model architectures trained on the ImageNet dataset, where RN=ResNet, MN=MobileNet, DN=DenseNet. Empirical results on the StanfordDogs target dataset show no violations of our theoretical bounds.

S2. Entropy of Samples in Informative Subsets

The improved performance of our proposed ACT augmentation lies in the usefulness of the samples in the informative subsets. The significant improvements observed in a variety of experimental settings, across multiple source architectures and datasets confirms this hypothesis. We further investigate this as follows: We sort the target dataset using either of our information scores defined in 3, and divide it into 5 uniform bins. For each bin, the average entropy of the samples in the bin. The entropy for a single target sample is calculated using the standard formula, utilizing the softmax output of the source model in the source dataset label space as the probability distribution for that sample. The results shown in Figure S2 show that the samples assigned higher information scores by our proposed method have a higher entropy on average than the samples with lower information scores. In addition, the samples in the most informative subset have a higher average entropy than the average entropy of the entire dataset.

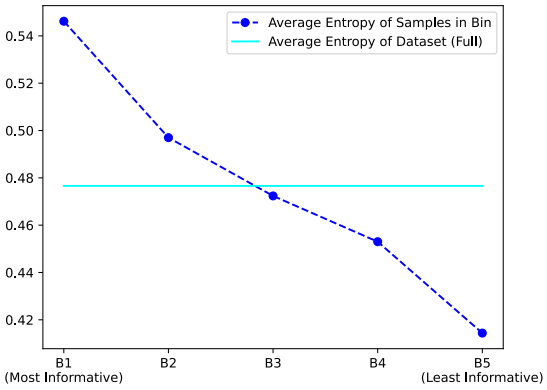

Figure S2. Average Entropy of samples in bins of decreasing information scores. More informative bins (subsets) have a higher average entropy, and vice-versa. The most informative bins also have a higher average entropy than the overall dataset. Results use an Imagenet Resnet50 source model, with StanfordDogs as the target dataset.

Algorithm 1 ACT

Require: Source model f_θ^s, Source dataset $\mathcal{D}_s$, Target dataset $\mathcal{D}_t$, Informative subset selection variant i_v, Transferability Metric $\mathcal{T}$

$k \leftarrow$ number of samples in the subset

if i_v='*CAW*' **then**

  Collect target dataset activations $\mathcal{E}_l(\mathbf{x}_j^t)$

  for class $c \in \mathcal{D}_t$ **do**

    Compute μ_c, Σ_c ▷ Eqns. 5, 6

  end for

  Compute $I(f_\theta^s, \mathbf{x}_j^t)_{CAW}$ ▷ Eqn. 7

else if i_v='*CAG*' **then**

  Collect source dataset activations $\mathcal{E}_l(\mathbf{x}_i^s)$

  Collect target dataset activations $\mathcal{E}_l(\mathbf{x}_j^t)$

  Compute similarity scores $(\mathbf{x}_i^s, \mathbf{x}_j^t)$ ▷ Eqn. 9

  Compute $I(f_\theta^s, \mathcal{D}_s, \mathbf{x}_j^t)_{\text{CAG}}$ ▷ Eqn. 10

end if

$\mathcal{D}_t^{\text{inf}} \leftarrow \{k$ most informative samples ordered by $I(\cdot)\}$

return $\mathcal{T}(f_\theta^s, \mathcal{D}_t^{\text{inf}})$

S3. Experimental Setup

Compute details. All experiments were run using the PyTorch library [52] with Nvidia A-100/V-100 GPUs.

Model Architectures. We use a variety of model architectures (VGG, ResNet, DenseNet), trained on different source datasets across our experiments. For each model architecture, we utilize embeddings from the pre-final layer for the class-aware method (Eq. 5). For the class-agnostic method, we utilize embeddings from intermediate layers for the similarity computation (Eq. 9). Particularly, we utilize the embeddings from the final layer of each block of convolutions (*e.g.*. output of each residual block in ResNet). We only include 3/4 layers for any architecture and do not consider embeddings from the first block.

Similarity Computation for Large Source Datasets. For the experiments with models pre-trained on ImageNet as the source, when using the class-agnostic method, it is in-

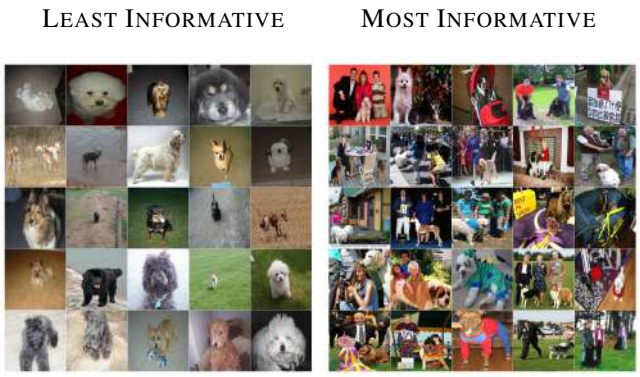

Figure S3. The 5×5 grid shows the top-25 images from the least informative (left) and most informative (right) subset of the target dataset using the class-agnostic technique for the ImageNet-StanfordDogs source-target pair. Images with *higher* information scores tend to feature cluttered images with atypical vantage points, whereas images with *lower* information scores mostly comprise dogs in an uncluttered background.

Figure S4. The 5×5 grid shows the top-25 images from the least (left) and most (right) informative subset of the target dataset using the class-agnostic technique for the ImageNet-OxfordIIIT Pets source-target pair. Images with *higher* information scores tend to feature cluttered images with atypical vantage points, whereas images with *lower* information scores mostly comprise dogs and cats in an uncluttered background.

feasible to use the entire ImageNet dataset for the similarity comparison to generate the similarity matrix (using Eqn. 9). Instead, we use a random 10% subset of the ImageNet dataset, uniformly sampled from each class, as the source dataset to compute the similarity matrix. We do not observe any performance drop due to this sub-sampling, and this can be extended to other datasets as well, for further computational speedup. Additionally, we do not re-do the similarity computation for each source architecture. Instead, we only compute the similarity matrix using the ResNet-50 model and use the informative subset obtained from this to compute ACT-augmented transferability metrics for all model architectures pre-trained on ImageNet. We repeat this in

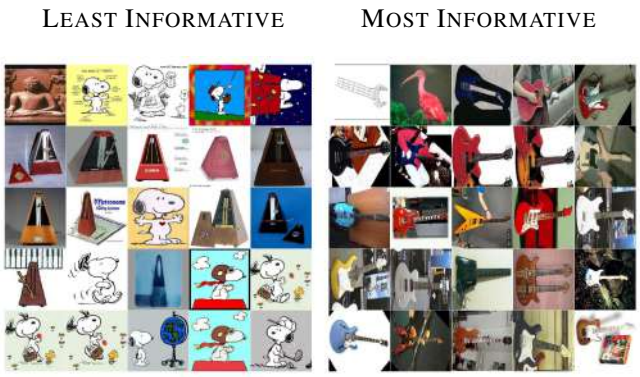

Figure S5. The 5×5 grid shows the top-25 images from the least (left) and most (right) informative subset of the target dataset using the class-aware technique for the ImageNet-Caltech101 source-target pair. Images with *higher* information scores tend to feature classes that are typically harder to classify (since they might have very less distinguishing features), whereas images with *lower* information scores mostly comprise classes that are easily distinguishable.

the model ensemble setting, utilizing a single similarity matrix for all model architectures trained on the same source dataset.

GBC Implementation. In all experiments, for computing GBC and ACT-GBC, we use a spherical covariance matrix, as we found this to yield better results for the base GBC score.

Size of Informative Subset. The size of the informative subset is a hyperparameter that can be tuned. Due to variations in dataset sizes, the exact value differs significantly. Instead of fixing a size, we set the size of the informative subset N_s to be a fraction of the size of the target dataset. In general, we found a value of 10%-25% to work well.

Models used in Ensemble Selection. We use the following pool of source models for ensemble selection experiment described in Section 4.4: i) DenseNet-201, ResNet-101, MobileNetv2 trained on on Imagenet, ii) DenseNet-201, ResNet-18, VGG-19 trained on Stanford Dogs, iii) DenseNet-201, ResNet-101 trained on Oxford IIIT Pets, iv) DenseNet-201, ResNet-101 trained on Flowers102, v) ResNet-18, VGG-19 trained on CUB200, and vi) ResNet-34 trained on Caltech101.

S4. Additional Results

We include the following additional results: i) Comparison with H-Score, TransRate, E-Train, SFDA, and PARC on Source Architecture Selection (Table S1), ii) Additional baselines on Target Dataset Selection (Table S2), iii) Language Models (Table S3); iv) Target Dataset Selection results with CUB20 source models (Table S7); v) Ensemble Selection results with $K = 3$ (Table S4), vi) Results

with Kendall Tau (Table S8) and Weighted Kendall Tau (Table S9), and vii) Results in a noisy dataset setting (Table S10).

S5. Additional Ablations

We include ablation studies on a few more key components of ACT: i) The architecture of the model used to compute information scores (Section S5.1), and ii) The selection of the optimal subset after computing information scores (Section S5.2), and iii) The size of the informative subset (Section S5.3)

S5.1. Architecture used to compute Information Score

ACT aims to achieve better transferability estimates irrespective of the source of the information scores, i.e., the source architecture we use to calculate informative subsets in Class-Agnostic way or Class-Aware way. We follow the experimental setup from the target task selection (Section 4.2) experiments. We calculate ACT-LEEP scores on informative subsets identified using i) ResNet18 and VGG19 trained on CUB200, and ii) ResNet50 trained on ImageNet. Results show that ACT-LEEP outperforms LEEP (baseline calculated using the entire dataset) across all three architectures (Table S5).

S5.2. Less Informative Samples added with Stochasticity

It is important to understand whether one can completely neglect the least informative samples for transferability estimation. We conduct an experiment where we add the least informative samples stochastically to our most informative subsets and then compute the respective metric correlation score. We follow the experiment setting of Table 1 and obtain results by iterating the addition of least informative samples 10 times, followed by taking the mean. We observe that addition of less informative samples do not improve the results. In fact, the results come out to be worse than when using only the most informative samples. Results for the same are shown in Table S6.

S5.3. Size of the informative subset (N_s)

The size of the informative subset is a key component of our propose ACT approach. We find an optimal subset size of $\approx$25% of the original target dataset works across all experiments. We observe that setting N_s too low does not leave enough samples for the transferability estimation, and setting N_s too high includes many samples, taking us away from informative subsets and closer to the baseline metrics themselves. This is observed in our results in Fig. S6, where we measure the performance of ACT-LEEP with different sizes of the informative subset.

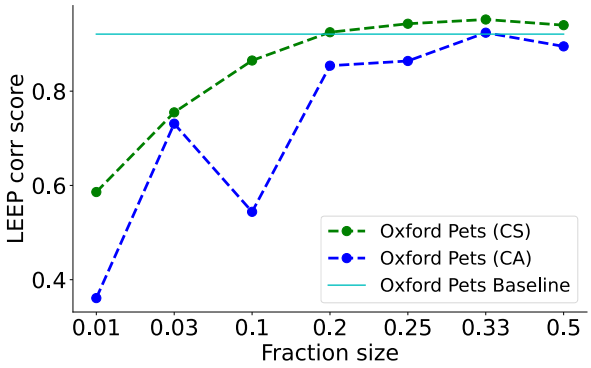

Figure S6. Correlation scores of ACT-LEEP on varying the size of the informative subset. Experiment was performed using a ResNet-50 model trained on ImageNet as the source model, and Oxford-IIIT Pets as the target dataset.

Table S1. Results on source architecture selection task with additional baselines. Shown are correlation scores (higher the better) computed across all source architectures trained on ImageNet. Results where ACT metrics perform better are in **bold**.

Target ($\mathcal{D}_t$)	H-Score	ACT-H-Score		TransRate	ACT-TransRate		PARC	ACT-PARC		SFDA	ACT-SFDA		E-Tran	ACT-E-Tran	
		CAG	CAW		CAG	CAW		CAG	CAW		CAG	CAW		CAG	CAW
CUB200	0.421	0.417	0.323	-0.367	**-0.254**	**-0.266**	0.692	**0.840**	**0.873**	0.47	**0.920**	**0.623**	-0.238	**0.821**	**0.764**
StanfordDogs	0.910	**0.924**	0.892	0.885	**0.892**	0.856	0.927	**0.945**	**0.937**	0.939	0.939	**0.940**	0.125	**0.944**	**0.881**
Flowers102	0.038	**0.238**	**0.192**	0.604	**0.662**	**0.641**	-0.399	**0.698**	**0.185**	0.606	0.583	0.581	0.490	**0.604**	0.409
OxfordPets	0.862	**0.940**	0.714	-0.025	**0.361**	**0.492**	0.773	**0.836**	**0.817**	0.753	**0.829**	**0.888**	-0.271	**0.924**	**0.427**
Imagenette	0.953	**0.954**	0.938	-0.898	**-0.880**	**-0.315**	0.782	**0.903**	**0.804**	0.919	**0.926**	**0.935**	0.724	**0.931**	**0.800**

Table S2. Results on target dataset selection task for Caltech101 source models with additional baselines. Shown are correlation scores (higher the better) computed across all target datasets. Results where ACT metrics perform better are in **bold**.

Target ($\mathcal{D}_t$)	H-Score	ACT-H-Score		TransRate	ACT-TransRate	
		CAG	CAW		CAG	CAW
CUB200	-0.45	**-0.24**	**0.300**	-0.662	**-0.58**	**0.197**
StanfordDogs	-0.131	**-0.112**	**0.388**	-0.49	**-0.385**	**0.321**
Flowers102	**-0.689**	**-0.584**	-0.341	-0.457	**-0.392**	**-0.13**
OxfordPets	-0.695	-0.707	**-0.291**	-0.429	**-0.372**	**-0.237**
PACS-Sketch	-0.543	-0.554	-0.55	-0.149	-0.179	**0.05**

Table S3. Results on target task selection for sentiment classification using additional baselines. Shown are correlation scores (higher the better) computed across all target candidates. Results where ACT metrics perform better than their counterparts are in **bold**.

Target ($\mathcal{D}_t$)	H-Score	ACT-H-Score		TransRate	ACT-TransRate	
		CAG	CAW		CAG	CAW
Emotion-IMDB	-0.296	**0.214**	**-0.040**	0.015	**0.210**	**0.145**
Emotion-Tweets	-0.138	**-0.046**	**0.358**	0.095	**0.227**	**0.602**
News-Tweets	-0.527	-0.606	**-0.227**	-0.638	**-0.071**	**0.381**

Table S4. Results on the ensemble model selection task for $K = 3$. Shown are correlation scores (higher the better) computed across all ensemble candidates. Results where ACT metrics outperform their baselines are **bolded**.

Target ($\mathcal{D}_t$)	MS-LEEP	ACT-MS-LEEP	E-LEEP	ACT-E-LEEP
Flowers102	-0.288	-0.376	-0.323	**-0.319**
Stanford Dogs	0.390	0.264	0.477	**0.494**
CUB200	0.345	**0.391**	0.405	0.405
Oxford-IIIT	0.115	**0.189**	0.253	**0.343**
Caltech101	0.430	**0.479**	0.480	0.478

Table S5. Results on target task selection task different source model architectures. Shown are correlation scores (higher the better) computed across the target dataset using ACT-LEEP. Irrespective of the architecture used for finding the most informative subset, ACT-LEEP outperforms the baseline LEEP score.

Information Score Source Model	Caltech101		Oxford-IIIT Pets	
	CAG	CAW	CAG	CAW
ResNet18	0.360	0.014	0.896	0.901
VGG19	0.200	0.267	0.881	0.894
ResNet50	0.630	0.196	0.896	0.898
Baseline (LEEP Score)	-0.03		0.863	

Table S6. Results using subsets obtained by stochastically adding the least informative samples to the most informative subset. Shown are correlation scores (higher the better) computed across all source architectures trained on ImageNet. Results where ACT-augmented metrics outperform their baselines are in **bold**.

Target ($\mathcal{D}_t$)	LEEP	ACT-LEEP		Stochasticity %			
		CAG	CAW	1%	3%	5%	10%
Caltech101	0.416	**0.439**	**0.475**	0.474	0.472	0.472	0.458
Flowers102	0.534	0.405	**0.626**	0.616	0.579	0.575	0.539
CUB200	0.504	**0.508**	**0.723**	0.719	0.714	0.728	0.679

Table S7. Results on target task selection using the fine-tuning method for CUB200 source models. Shown are correlation scores (higher the better) computed across all target datasets. Results where ACT metrics outperform are **bolded**.

Target ($\mathcal{D}_t$)	LEEP	ACT-LEEP		NCE	ACT-NCE		GBC	ACT-GBC	
		CAG	CAW		CAG	CAW		CAG	CAW
Caltech101	-0.035	**0.709**	**0.098**	0.081	**0.742**	**0.249**	0.507	**0.562**	**0.516**
Flowers102	0.612	**0.617**	**0.613**	0.593	0.579	**0.618**	0.535	0.526	**0.568**
StanfordDogs	0.929	**0.936**	0.929	0.928	0.927	**0.929**	0.909	**0.914**	**0.913**
Oxford-IIIT	0.863	**0.871**	0.860	0.812	**0.826**	**0.814**	0.859	**0.860**	**0.872**
PACS-Sketch	0.947	**0.965**	**0.960**	0.949	**0.958**	**0.950**	0.819	**0.909**	**0.883**

Table S8. Results on source architecture selection. Shown are Kendall Tau correlation scores (higher the better) computed across all source architectures trained on ImageNet. Results where ACT metrics outperform their baselines are in **bold**.

Target ($\mathcal{D}_T$)	LEEP	ACT-LEEP		NCE	ACT-NCE		GBC	ACT-GBC	
		CAG	CAW		CAG	CAW		CAG	CAW
CUB200	0.238	0.142	**0.714**	0.142	**0.238**	**0.619**	0.619	0.619	**0.714**
StanfordDogs	0.809	0.809	0.809	0.809	0.714	0.809	0.619	**0.904**	**0.714**
Flowers102	0.333	**0.619**	**0.523**	0.238	**0.428**	0.238	0.047	0.047	**0.238**
Oxford-IIIT	0.904	**1.000**	0.904	0.523	**0.619**	**0.714**	0.523	0.523	0.523
Caltech101	0.390	0.390	0.390	0.097	**0.292**	**0.195**	0.683	0.683	0.683
Imagenette	0.714	0.714	**1.000**	0.619	**0.714**	**0.683**	0.619	0.619	**0.714**
PACS-Sketch	0.000	**0.097**	**0.097**	0.000	**0.097**	**0.195**	0.487	0.487	**0.585**

Table S9. Results on source architecture selection task. Shown are Weighted Kendall Tau correlation scores (higher the better) computed across all source architectures trained on ImageNet. Results where ACT metrics outperform their baselines are in **bold**.

Target ($\mathcal{D}_T$)	LEEP	ACT-LEEP		NCE	ACT-NCE		GBC	ACT-GBC	
		CAG	CAW		CAG	CAW		CAG	CAW
CUB200	0.258	0.108	**0.638**	0.113	**0.247**	**0.659**	0.591	0.591	**0.805**
StanfordDogs	0.865	0.865	0.865	0.865	0.672	0.865	0.746	**0.952**	**0.805**
Flowers102	0.376	**0.705**	**0.644**	0.389	**0.611**	**0.400**	-0.119	-0.119	**0.031**
Oxford-IIIT	0.925	**1.000**	0.925	0.587	**0.678**	**0.721**	0.692	0.530	0.530
Caltech101	0.535	0.535	0.535	0.345	**0.482**	0.238	0.723	0.723	0.723
Imagenette	0.672	0.672	**1.000**	0.558	**0.758**	**0.693**	0.799	**0.808**	**0.851**
PACS-Sketch	-0.145	**0.026**	**0.095**	-0.063	**0.044**	**0.232**	0.567	0.406	**0.651**

LEEP	ACT-LEEP		NCE	ACT-NCE	
	20% Subset	25% Subset		20% Subset	25% Subset
0.91	0.89	**0.93**	0.91	**0.94**	**0.95**

Table S10. To test the performance of ACT in a noisy dataset setting, we test our approach using ImageNet source models, and Noisy ImageNette (target; w/ 20% Gaussian noised data) and find that ACT still improves the transferability estimation. This shows that ACT is able to correctly identify the correct samples to estimate transferability in this setting, even in the extreme case where the size of the informative subset is the same as the fraction of noised samples. The most informative subset may also include these noised samples, as they significantly affect the learning process, and thereby the finetuned model.