Active Transferability Estimation

Tarun Ram Menta [*] Surgan Jandial[*] Akash Patil[†] Saketh Bachu[‡] Vimal K.B[‡]

Balaji Krishnamurthy[*] Vineeth N. Balasubramanian[‡] Mausoom Sarkar[*] Chirag Agarwal[§ ¶]

Abstract

As transfer learning techniques are increasingly used to transfer knowledge from the source model to the target task, it becomes important to quantify which source models are suitable for a given target task without performing computationally expensive fine-tuning. Inspired by active learning techniques, we propose ACT *(ACtive Transferability), a new strategy to improve the performance of transferability estimation methods, by leveraging an informative subset of the target data. By leveraging the model's internal and output representations, we introduce two techniques – class-agnostic and class-aware – to identify informative subsets and show that* ACT *can be applied to any existing transferability metric to improve their performance and reliability. Our experimental results across multiple source model architectures, target datasets, and transfer learning tasks show that* ACT *metrics are consistently better or on par with the state-of-the-art transferability metrics.*

1. Introduction

Transfer learning (TL) [49, 69, 74] aims to improve the performance of pre-trained models on target tasks by utilizing the knowledge from source tasks. With the increasing development of large-scale pre-trained models [12, 13, 16, 55] and the availability of multiple model choices (*e.g.*, model hubs of Pytorch, Tensorflow, and HuggingFace) for TL, it is critical to estimate their transferability without training on the target task and determine how effectively TL algorithms will transfer knowledge from the source to the target task. To this end, transferability estimation metrics have been recently proposed to estimate the ease of transferring the knowledge learned from these models to any given target dataset. They work with little to no training on the target dataset, avoiding the infeasible fine-tuning techniques to find the best source-target combination.

Recent years have seen a surge of transferability estimation techniques [1, 44, 50, 71, 81] for a given pair of source models and target tasks. Transferability estimation techniques find application in a wide range of tasks (explored and evaluated in Sec. 4). However, these methods possess some limitations, and the stability and generalizability of these metrics across various settings is a major area of concern. For instance, Agostinelli et al. [4] shows that existing transferability metrics do not work consistently across all settings, with different metrics showing superiority in different settings. Most transferability metrics achieve lower performance in experimental settings [78], when the source and target datasets possess large domain differences [44], or when the experimental parameters are slightly modified [4].

In this work, we aim to alleviate the above drawbacks of transferability metrics and improve their performance across a wide range of tasks. While investigating these shortcomings, we found a small performance gap when a set of good source models is transferred to a target dataset, where this gap primarily stems from a *subset of the dataset*, as we later show in Sec. 3. In addition, Agostinelli et al. [4] argues that transferability metrics are highly sensitive to changes in the target dataset. Hence, we hypothesize that a refined subset of the target dataset can provide the most *new information* to the source model, and will help boost the performance of existing transferability metrics. An analogous of the above phenomenon is also observed in active learning methods [56], where they attempt to maintain model performance using a lower annotation cost by selecting a small subset of samples. This informative subset forms a representative sample of the entire dataset, providing information on the data geometry and class boundaries with far fewer examples [3]. Similar findings have been shown in other contexts [2, 17, 35, 35, 66, 83], where a small subset of representative samples can be beneficial to the learning process, and that certain samples are redundant [37], with minimal impact to model performance.

Present work. Building on the above observations, in this

[*]Adobe Systems
[†]Indian Institute of Technology, Madras
[‡]Indian Institute of Technology, Hyderabad
[§]Harvard University
[¶]Agyeya Foundation

work, we follow an information-theoretic approach to improve the performance of existing transferability methods by identifying a small subset of the target dataset that provides the most *new information* to the source model. We show that using this subset in conjunction with existing transferability metrics helps boost their performance. Our identified subsets of maximal new information follow the observations in [3] and are found to lie closer to the decision boundary (Fig. 3), thus providing critical information on the usefulness of samples from one domain for another. To this end, we propose a simple framework, ACT, that can be applied to any existing transferability metric to estimate transferability using a carefully selected subset of the target dataset. More specifically, we introduce two independent techniques — *class-agnostic* and *class-aware* — to identify the subsets from a target dataset that provide the most *new information* to the source model, using the model's internal (class-agnostic) and output (class-aware) representations (Sec. 3). We utilize these subsets to improve the performance and reliability of existing transferability metrics. Our empirical analysis across a range of transfer learning tasks like source architecture selection (Sec. 4.1), target dataset selection (Sec. 4.2), ensemble model selection (Sec. 4.4), semantic segmentation (Sec. 4.3) and language models (Sec. 4.5) show that ACT scores better correlate with the transfer accuracy than their counterparts.

2. Related Work

Our work lies at the intersection of transfer learning, transferability estimation metrics, and active learning.

Transfer Learning. Such methods can be broadly organized into three categories: (i) *Inductive Transfer* [19, 77] methods, which leverage inductive bias; (ii) *Transductive Transfer* methods that include domain adaptation methods [73, 76]; and (iii) *Task Transfer* [48, 80] methods, which transfer between different tasks instead of models. Amongst these, the most common form of transfer learning is fine-tuning a pre-trained source model on a given target dataset. Recent works show the use of large-scale pre-trained models [15, 54] for learning representations for different source tasks. We request the readers to refer to [47, 86] for further details and strategies used in transfer learning.

Transferability Metrics. Despite the availability of large numbers of source models, achieving an optimal transfer for a given target task is still an open research area as it is non-trivial to identify the best source model or dataset for efficient transfer learning. Transferability metrics are used as proxy scores to estimate the transfer accuracy from a source model to a target task. Recent years have witnessed the development of such metrics. For instance, NCE [70] and LEEP [44] utilize the labels in the source and target task domains to estimate transferability, whereas metrics like H-Score [7], GBC [50], TransRate [30], PARC [10], SFDA [62], and E-Tran [23] use the embeddings from the source model to estimate transferability. In contrast to the above metrics that focus on a single source model, Agostinelli et al. explored metrics to estimate the transferability of an ensemble of models and introduced two metrics [5] – MS-LEEP and E-LEEP – for identifying a subset of model ensembles from the pool of available source models. None of these existing efforts however considered the use of an informative subset in the target dataset to estimate transferability, which is the focus of this work.

Task Transferability. Our work focuses on the problem of estimating the transferability of a source model to the target dataset in two settings: (i) identifying the most suitable model from a pool of pre-trained source models to perform transfer learning on a given target dataset, and (ii) for a given pre-trained source model, finding the most suitable target dataset from a collection of datasets to perform transfer learning. Some other recent transferability works [18, 64, 65, 81] consider models that are pre-trained on one or more tasks, and study transfer of such models to another task, often requiring an expensive fine-tuning process. These works only discuss task transferability, *i.e.*, they have a different objective and establish task relatedness. In addition, these works either perform fine-tuning from scratch or have computational costs similar to fine-tuning, and do not aim to study or propose transferability metrics, which is the focus of our work. In a related area, [82] quantified transferability for the task of Domain Generalization, [68] discussed transferability for multi-source transfer, both of which operate in a setting different from ours.

Active Learning. Given a large, unlabeled dataset, active learning methods aim to select the best possible samples to annotate, assuming the availability of a labeling oracle. Traditional active learning methods can be broadly categorized into (i) *Uncertainty-based* approaches [9, 26, 33, 61] which select samples that the model is most uncertain about and (ii) *Diversity-based* methods [22, 24, 45], which encourage the model to learn more generalized representations. The rise in popularity of deep learning methods has brought attention to deep active learning [3, 56] methods that bring down annotation costs. Existing methods for deep active learning [21, 27, 32, 60] largely build upon traditional active learning methods, or a hybrid of a few methods to achieve their objective. Our objective in this work, however, is different from mainstream active learning and is rather focused on subset selection for transferability estimation.

3. Methodology

3.1. Notations and Preliminaries

Problem Statement. Given a pre-trained model f_θ^s trained on a source dataset $\mathcal{D}_s$, and a target dataset $\mathcal{D}_t$, a transferability metric aims to produce a score $\hat{\mathcal{A}}$ which estimates

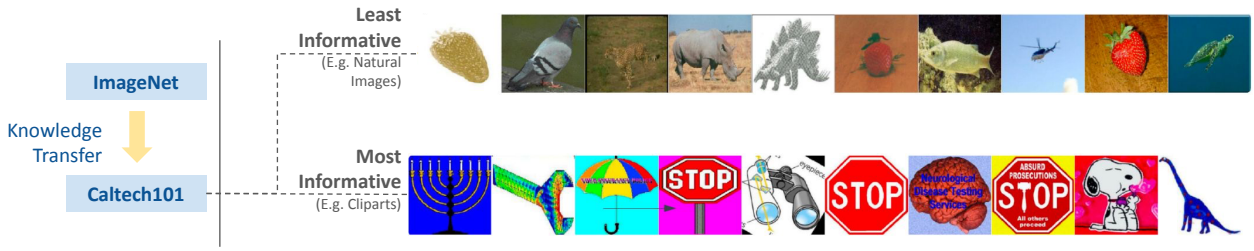


Figure 1. Top-10 images from subsets from Caltech101 containing *most* and *least* new information as estimated w.r.t. ImageNet source dataset. These show that the subsets with the most informative images (cliparts in this case) that are out-of-distribution when compared to the source images. See the Appendix for more qualitative results.

how effectively transfer learning algorithms can transfer knowledge from the source task to the target task, *without fine-tuning on the target task*. In contrast to existing transferability metrics [5, 44, 50, 70, 78], our objective is to improve the transferability estimation performance of existing metrics by focusing on a carefully selected subset of examples from the target dataset. Note that this subset selection is only carried out for transferability estimation, and has no bearing on the underlying transfer learning process itself.

Notations. Let f_θ^s be a pre-trained source model trained on a source dataset $\mathcal{D}_s=\{\mathcal{D}_s^{\text{train}}, \mathcal{D}_s^{\text{test}}\}$, and a target dataset $\mathcal{D}_t=\{\mathcal{D}_t^{\text{train}}, \mathcal{D}_t^{\text{test}}\}$. In transfer learning, we obtain a target model $f_\theta^{s\to t}$ initialized using the source model weights and fine-tuned on the target dataset $\mathcal{D}_t^{\text{train}}$. The performance of the target model $f_\theta^{s\to t}$ is quantified using the target model accuracy $\mathcal{A}^{s\to t}$ when evaluated on the unseen target test data $\mathcal{D}_t^{\text{test}}$. Let $\mathcal{T}^{s\to t}$ be a transferability metric that estimates the ease of transferring knowledge from the source model to the target dataset, where the metric only has access to the pre-trained source model f_θ^s, and dataset $\mathcal{D}_t^{\text{train}}$, and produce their estimates *without expensive fine-tuning on the target dataset*. Hence, the key task of transferability estimation is to define a metric that produces a score $\hat{\mathcal{A}}^{s\to t}$, that reliably estimates the target test accuracy $\mathcal{A}^{s\to t}$:

$$\mathcal{T}^{s\to t}(f_\theta^s, \mathcal{D}_t^{\text{train}}) \mapsto \hat{\mathcal{A}}^{s\to t} \tag{1}$$

Evaluation. A good transferability metric $\mathcal{T}^{s\to t}$ should reliably estimate the target model accuracy, $\hat{\mathcal{A}}^{s\to t}$. Hence, the performance of a transferability metric is evaluated as the correlation between $\hat{\mathcal{A}}^{s\to t}$ and $\mathcal{A}^{s\to t}$ across multiple combinations of source model and target dataset. Following earlier work [44], we employ the Pearson Correlation Coefficient (PCC) and Kendall Tau metrics for this purpose.

3.2. Informative Subsets for Transferability Estimation

To understand the limitations of existing transferability metrics, we begin by investigating the behavior of models after fine-tuning. As shown in Fig. 2, we observe that across multiple source models, certain subsets of the target dataset

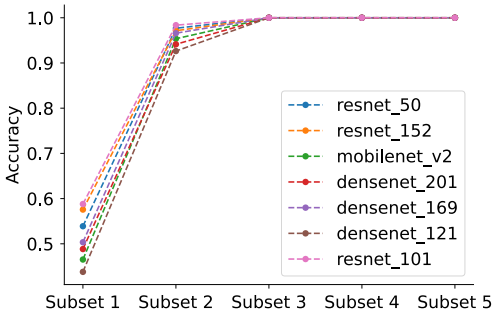


Figure 2. Transfer learning accuracies on different subsets of the Oxford-IIIT target dataset, after fine-tuning multiple source models trained on ImageNet. Evidently, across all models, we note that some subsets demonstrate low accuracy when compared to others, motivating us to identify such subsets from the target dataset.

have very uncertain predictions from the model, leading to a lower accuracy on this subset in relation to other subsets. The samples of such subsets of the target dataset bear higher uncertainty, and hence provide more *new information* about the target dataset to the source model. Besides being indicative of the performance after finetuning, these samples also most strongly influence the performance gaps between different models, as can be seen in the figure.

We posit that one can view the transferability estimation as how well a source model captures information about a target domain. In particular, we hypothesize that carefully selecting a subset of samples from the target dataset that provides the most *new information* w.r.t. the source model will help better estimate transferability. This is similar to Support Vector Machines, where the samples with most uncertainty provide the most information, or active learning methods, where information-theoretic methods are used to identify a subset of instances from a given pool to label, so as to achieve the most effective training. We define ACtive Transferability (ACT) estimation as a technique which utilizes these informative subsets to improve the quality of existing transferability metrics:

$$\mathcal{T}_{\text{ACT}} = \mathcal{T}(f_\theta^s, \mathcal{D}_t^{\text{inf}}), \tag{2}$$

where $\mathcal{T}$ is any existing transferability metric, and $\mathcal{D}_t^{\text{inf}}$ is the identified subset of most informative samples. For instance, in Fig. 2, the first subset may be the most informative one to choose for estimating transferability, but we do not have access to the fine-tuned model during transferability estimation. Hence, we need to identify such a subset using an estimate of information using the target dataset samples.

Taking inspiration from active learning methods [3, 56], we aim to approach the problem of subset selection for transferability estimation using an information-theoretic approach. We use estimates of *mutual information* between the source and target domains to identify the most informative samples. More precisely, we aim to select a subset of samples $\mathcal{D}_t^{\text{inf}}$ from the target dataset, which has the *lowest* mutual information w.r.t. the source domain.

$$\mathcal{D}_t^{\text{inf}} = \underset{\mathcal{D} \subset \mathcal{D}_t^{\text{train}}}{\arg\min}\, I(\mathcal{D}_s^{\text{train}}; \mathcal{D}), \tag{3}$$

where I is the mutual information. Intuitively, a subset of target domain samples with the least mutual information w.r.t. the source domain is the source of most *new information* for the source model. Further, this subset provides an estimate of the difficulty of fine-tuning on the target domain. Note that the subset is from the training set of the target domain, but we evaluate the transferability performance on a held-out test set from the target domain. We propose two methods to identify these samples; Fig. 3 shows that the subset of samples identified by our methods indeed have higher uncertainties than other subsets, lie closer to the decision boundary, and thus help provide useful information about performance on the target dataset. For convenience and ease of reading, we refer to these subsets of minimal mutual information as *informative* subsets, going forward.

3.3. Identifying Informative Subsets

We propose two methods – *class-aware* and *class-agnostic* – to estimate the mutual information between the source and target domains, where we make no assumptions about the source model or source dataset/task. While the *class-aware* method utilizes the label information in the target dataset, the *class-agnostic* method does not consider the label information, and only utilizes the distance between source and target datasets to estimate the mutual information. We use this estimate to model the *informativeness* of the target dataset samples and choose the most informative subset.

Class-Aware Method. When the target task is classification, we leverage the fact that mutual information I between two random variables X and Y can be written as the difference between entropy $H(X)$ and conditional entropy $H(X|Y)$. The conditional entropy can then be written in terms of the negative expectation of the conditional probability $P(X|Y)$, *i.e.*:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ \implies I(X;Y) &= H(X) + \mathbb{E}[\ \log\ P(X|Y)\ ] \end{aligned} \tag{4}$$

Hence, a low value of the conditional probability corresponds to a low value of mutual information between the two domains. We use the source model as a proxy for the information in the source domain and compute the conditional probability between *embeddings* of the target samples in the source model space $f_\theta^s(x_t)$ and target dataset labels y_t. Following [50], we model the conditional distribution $P(f_\theta^s(x_t)|y_t)$ as a Gaussian in the source model embedding space for each class in the target dataset. The Gaussian is parametrized by the mean and variance as given below:

$$\mu_c = \frac{1}{N_c} \sum_{j:y_j^t=c} f_\theta^s(\mathbf{x}_j^t) \tag{5}$$

$$\Sigma_c = \frac{1}{N_c} \sum_{j:y_j^t=c} (f_\theta^s(\mathbf{x}_j) - \mu_c)(f_\theta^s(\mathbf{x}_j) - \mu_c)^\top \tag{6}$$

where $y_j^t{=}c$, and N_c is the number of samples in class c.

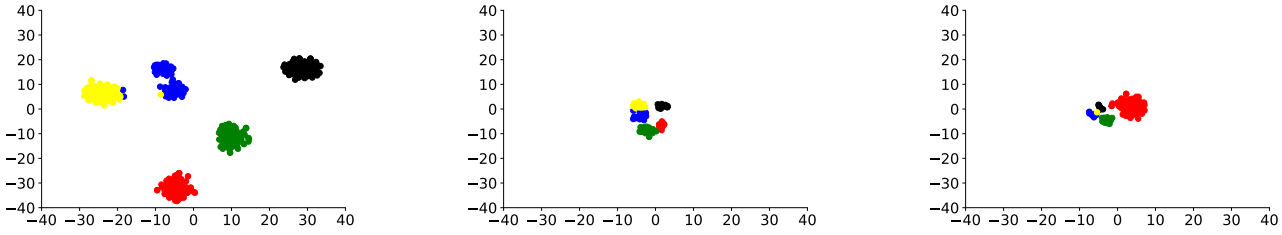
To find the target dataset samples that contain the least mutual information with the source dataset, we identify the samples that minimize the conditional probability. Hence, the *informativeness* of a target dataset sample $\mathbf{x}_t$, with label y_t with respect to the source model f_θ^s can be computed as the distance from the corresponding class mean, *i.e*,

$$I(f_\theta^s, \mathbf{x}^t)_{\text{CAW}} = \sqrt{(f_\theta^s(\mathbf{x}^t) - \mu_c)^\top \Sigma_c^{-1} (f_\theta^s(\mathbf{x}^t) - \mu_c)} \tag{7}$$

Class-Agnostic Method. The *class-aware* method is efficient to compute, and as seen in Sec. 4, helps improve the performance of a wide range of existing transferability metrics across multiple settings, but is limited to the classification setting. To ameliorate this shortcoming, we propose the *class-agnostic* method for identifying informative samples. The mutual information I can be written in terms of the joint distribution, *i.e.,*

$$I(X;Y) = \mathbb{E}\left[\frac{P(X,Y)}{P(X)P(Y)}\right] \tag{8}$$

To identify subsets of the target dataset that have low mutual information with the source dataset, we simply identify a subset that lies in a region of low joint probability density. We take inspiration from [6, 67], which estimate the distance betweeen datasets using optimal transport. Given two discrete empirical distributions $\alpha = (a_1, \ldots, a_m)$ and $\beta = (b1, \ldots, b_n)$ and a cost function $\mathcal{C}$, Optimal Transport [34, 53] estimates an optimal joint distribution (pairing matrix) π. For any pair of points, the optimal joint distribution is inversely proportional to the cost function for that pair of points, *i.e.,* $\pi_{ij} \propto \frac{1}{\mathcal{C}(a_i, b_j)}$. In particular, we follow [67] and model the optimal transport problem for



(a) Dogs - Full Dataset (b) Informative Subset - CAG (c) Informative Subset - CAW

Figure 3. t-SNE embeddings of the whole target dataset and its most informative subset using a ResNet-50 source model trained on ImageNet. We show sample embeddings from five random classes from the StanfordDogs target dataset, using class-agnostic (b) and class-aware (c) methods. We observe that embeddings from informative subsets are more entangled than the entire dataset.

the distributions of the source $\mathcal{D}_s$ and target $\mathcal{D}_t$ datasets by utilizing the dot product between the source model embeddings of the source and target samples as a measure of cost/similarity. In addition, we average the ℓ_2 distances over multiple layers of the model, *i.e.,*

$$\mathcal{C}(\mathbf{x}_i^s, \mathbf{x}_j^t) = \frac{1}{L}\sum_{l=1}^{L} ||\mathcal{E}_l(\mathbf{x}_i^s) - \mathcal{E}_l(\mathbf{x}_j^t)||_2, \tag{9}$$

where $\mathcal{E}_l(\cdot)$ is the intermediate representation from the l-th layer of f_θ^s and L is the total number of layers in f_θ^s. A similar approach of averaging the L2 distance over multiple layers has also been used for other settings as in [84].

Following from Eqn. 9, the target dataset samples lie far away from the source dataset, as these points lie in regions of low probability density, and hence are those that possess the lowest mutual information with the source domain. Hence, we score the information of target samples by the average distance from the source domain, *i.e.*:

$$I(f_\theta^s, \mathcal{D}_s, \mathbf{x}_j^t)_{\text{CAG}} = \frac{1}{M}\sum_{i=1}^{M} \mathcal{C}(\mathbf{x}_i^s, \mathbf{x}_j^t), \text{ where } M = |\mathcal{D}_s| \tag{10}$$

3.4. ACT (ACtive Transferability) Estimation

Given the above methods to estimate informative subsets, we define our overall proposed technique, ACT, as a means to improve existing transferability metrics. Building on our observations in Sec. 3.2, we first select the most informative subset of the target dataset $\mathcal{D}_t^{\text{inf}}$ using either of the information scores described in Sec. 3.3, *i.e.*:

$$\mathcal{D}_t^{\text{inf}} = \{(\mathbf{x}_{q_1}^t, y_{q_1}^t), \ldots, (\mathbf{x}_{q_{N_s}}^t, y_{q_{N_s}}^t)\} \tag{11}$$

where $\{q_1, q_2, \ldots, q_N\}$, $N_s \leq N$ denote the indices of the sorted samples, and the information score of each sample follows $I(\mathbf{x}_{q_1}^t) \geq I(\mathbf{x}_{q_2}^t) \geq \ldots \geq I(\mathbf{x}_{q_N}^t)$. This informative subset is then passed to existing transferability metrics, *i.e*:

$$\mathcal{T}_{\text{ACT}} = \mathcal{T}(f_\theta^s, \mathcal{D}_t^{\text{inf}}) \tag{12}$$

3.5. Extending Active Transferability to Ensembles

We also extend our proposed metrics to the setting of selecting source model *ensembles*. MS-LEEP and E-LEEP [5] are state-of-the-art transferability metrics for selecting source model *ensembles*. In this setting, our objective becomes to measure how well an ensemble of source models can capture information about the target domain. Formally, this setting involves N_e source models, *i.e.,* $\{f_\theta^{s_1}, \ldots, f_\theta^{s_{N_e}}\}$, trained on N_e source datasets, $\{\mathcal{D}_{s_1}, \ldots, \mathcal{D}_{s_{N_e}}\}$ (which may repeat when different source models are trained on the same source dataset), and a target dataset $\mathcal{D}_t$. For a given set of source models, the number of possible ensembles is 2^{N_e}. While applying transferability metrics on each candidate ensemble is possible, evaluating the ground truth $\mathcal{A}^{s \to t}$ for each candidate is computationally expensive. Hence, following [5], we set the size of the candidate ensembles to k, and evaluate $\mathcal{T}_{\text{ACT}}$, and $\mathcal{A}^{s \to t}$ for each of the the $\binom{N}{k}$ possible candidate ensembles. Given a candidate ensemble of k models, the most informative subset obtained via our ACTframework is different for each source model, leading to k most informative subsets, $\{\mathcal{D}_1^{\text{inf}}, \ldots, \mathcal{D}_k^{\text{inf}}\}$, corresponding to each model respectively. Then, we treat our active variants of MS-LEEP and E-LEEP differently. In case of MS-LEEP, we note that it is simply the sum of LEEP [44] scores over each source model in the candidate ensemble. Hence, our ACT counterpart of MS-LEEP follows as the sum of ACT-LEEP scores over the candidate set:

$$\text{MS-LEEP}_{\text{ACT}} = \sum_{i=1}^{k} \text{LEEP}(f_\theta^{s_i}, \mathcal{D}_i^{\text{inf}}) \tag{13}$$

E-LEEP considers the ensembled predictor $p_{\text{ens}}(y_i|x_i)$ by evaluating the prediction probability for each sample in the target dataset, across all models in the candidate ensemble:

$$\text{E-LEEP} = \frac{1}{|\mathcal{D}_t|} \sum_{(x^t, y^t) \in \mathcal{D}_t} log\ p_{\text{ens}}(y^t|x^t) \tag{14}$$

This requires a common target dataset to be considered across all models in the ensemble. To apply ACT for E-LEEP, we take the union of the most informative subsets

obtained for each source model and use this for the E-LEEP calculation, *i.e.*:

$$\text{E-LEEP}_{\text{ACT}} = \frac{1}{|\mathcal{D}_t^{\text{ens}}|} \sum_{(x^t, y^t) \in \mathcal{D}_t^{\text{ens}}} \log\ p_{\text{ens}}(y^t | x^t), \quad (15)$$

where $\mathcal{D}_t^{\text{ens}} = \bigcup_{i=1}^{k} \mathcal{D}_i^{\text{inf}}$.

4. Experiments

Next, we present experimental results to show the effectiveness of ACT metrics for different transfer learning tasks, including source architecture selection (Sec. 4.1), target dataset selection (Sec. 4.2), semantic segmentation task (Sec. 4.3), ensemble model selection (Sec. 4.4), and language domain transferability (Sec. 4.5).

Evaluation Metrics and Baselines. We use the Pearson Correlation Coefficient (PCC) for correlation between $\mathcal{T}^{s \to t}$ and $\mathcal{A}^{s \to t}$. For baselines, we use LEEP, NCE, LogME, H-Score, TransRate, GBC, E-Tran, SFDA, and PARC for single model transferability tasks, and MS-LEEP and E-LEEP for ensemble model selection. See Appendix for more details (Sec. S3), and results using other correlation coefficients (Sec. S4).

4.1. Source Architecture Selection

Experimental setup. We follow the experimental setup from [50] where the target dataset is fixed and $\mathcal{T}^{s \to t}$ is computed over multiple source architectures. The correlation scores are computed between $\mathcal{T}^{s \to t}$ and the transfer accuracies $\mathcal{A}^{s \to t}$. We consider seven target datasets for our these experiments: i) Caltech101 [20], ii) CUB200 [75], iii) Oxford-IIIT Pets [51], iv) Flowers102 [46], v) Stanford Dogs [36], vi) Imagenette [28], and vii) PACS-Sketch [38].

Model architectures and Training. We consider seven source architectures pre-trained on ImageNet [57] dataset, including ResNet-50, ResNet-101, ResNet-152 [25], DenseNet-121, DenseNet-169, DenseNet-201 [29], and MobileNetV2 [58]. All models were set using the publicly available pre-trained weights from the Torchvision library [43]. Following Pandy et al. [50], we calculate the target accuracy $\mathcal{A}^{s \to t}$ by *fine-tuning* the source model on each target dataset. We fine-tune the source model for 100 epochs using an SGD optimizer with a momentum of 0.9, a learning rate of 10^{-4}, and a batch size of 64.

Results. On average, across seven target datasets, ACT metrics show an improvement in correlation scores of **+129.74%** for LEEP, **+29.38%** for NCE, **+120.53%** for LogME, and **-0.07%** for GBC (Table 1). Interestingly, for most target datasets, both Class-Agnostic and Class-Aware variants of the ACT metrics outperform the baseline scores. We also observe an improvement of **+1.13%** for H-Score, **+3.21%** for TransRate, **50.40%** for PARC, **15.10%** for SFDA, **287.03%** for E-Tran (See App. S4 for more results).

4.2. Target Dataset Selection

Experimental setup. Here, the source model is fixed and the transferability metric is computed over multiple target datasets [44]. We construct 50 target datasets by randomly selecting a subset containing 40% to 100% of the total classes from the original target dataset. For each class, all train and test images are included in the respective train and test subsets. The PCC is computed between $\mathcal{T}^{s \to t}$ and $\mathcal{A}^{s \to t}$ across all 50 target tasks. We consider six target datasets including Caltech101, CUB200, Oxford-IIIT Pets, Flowers102, Stanford Dogs, and PACS-Sketch.

Model architectures and Training. We consider two source models: ResNet-18 pre-trained on CUB200 and ResNet-34 pre-trained on Caltech101. We train the transferred models for 100 epochs using SGD with a momentum of 0.9, a learning rate of 10^{-3}, and a batch size of 64.

Results. Across multiple source and target datasets, ACT-LEEP achieves the highest correlation for the target selection task and outperforms their respective baseline methods (Table 2). In particular, we observe an improvement in correlation scores of **236.16%** for LogME, **+0.99%** for LEEP, **+1.15%** for NCE, and **+5.11%** for GBC. We also observe an improvement of **+79.2%** for H-Score, and **+58.6%** for TransRate (See Appendix S4 for detailed results).

4.3. Semantic Segmentation

Experimental setup. We follow the fixed target setting described in [50] and report the correlation between meanIoU and $\mathcal{T}^{s \to t}$ for each target dataset. We consider a Fully Connected Network (FCN) [41] with a ResNet-50 backbone pre-trained on a subset of COCO-2017 [40]. We consider CityScapes [14], CamVid [11], BDD100k [79], IDD [72], PascalVOC [39] and SUIM [31] datasets. Among them, we consider the target datasets CityScapes, CamVid, BDD100k, and SUIM. Note that we use the Class-Agnostic variant of ACT for semantic segmentation as segmentation does not have class labels.

Model architectures and Training. We train an FCN Resnet50 [41] model for each source dataset and individually fine-tune them on $\mathcal{D}_t^{\text{train}}$. We train the individual models independently for 100 epochs using SGD with a momentum of 0.9, weight decay of 10^{-4}, a batch size of 16, a learning rate of 10^{-3}, and reduce it on plateau by a factor of 0.5.

Results. On average across four target datasets, ACT metrics outperform their baseline methods (Table 3). We observe an improvement in correlation scores of **+182.23%** for LEEP, **+33.34%** for NCE, and **+149.30%** for GBC.

4.4. Ensemble Model Selection

Experimental setup. Given a pool with P number of source models, the ensemble model selection task aims to select the subset of models whose ensemble yields the best

Table 1. Results on source architecture selection task. 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 ($\mathcal{D}_t$)	LEEP	ACT-LEEP CAG	ACT-LEEP CAW	GBC	ACT-GBC CAG	ACT-GBC CAW	LogMe	ACT-LogMe CAG	ACT-LogMe CAW	NCE	ACT-NCE CAG	ACT-NCE CAW
CUB200	0.534	0.405	**0.667**	0.790	**0.811**	0.785	-0.310	**0.082**	**0.365**	0.330	0.040	**0.500**
StanfordDogs	0.926	**0.943**	**0.931**	0.784	**0.944**	**0.834**	0.921	**0.953**	**0.943**	0.930	0.924	**0.955**
Flowers102	0.504	**0.508**	**0.723**	-0.012	-0.013	-0.02	-0.210	**0.483**	**0.614**	0.382	**0.390**	**0.388**
Oxford-IIIT	0.921	**0.952**	**0.927**	0.668	**0.867**	**0.745**	0.940	**0.973**	0.930	0.846	**0.851**	**0.916**
Caltech101	0.416	**0.439**	**0.458**	0.810	0.793	**0.821**	0.358	**0.712**	**0.792**	0.204	**0.461**	**0.504**
Imagenette	0.950	**0.950**	**0.962**	0.709	**0.723**	**0.711**	0.928	**0.930**	**0.971**	0.927	**0.940**	0.889
PACS-Sketch	-0.029	**0.196**	**0.253**	0.612	**0.637**	0.601	-0.423	**0.677**	**0.117**	-0.129	**0.160**	-0.208

Table 2. Results on target task selection using the fine-tuning method for Caltech101 source models. Shown are correlation scores (higher the better) computed across all target datasets. Results where ACT metrics perform better than their counterparts are in **bold**. See the Appendix for results on CUB200 source models.

Target ($\mathcal{D}_t$)	LEEP	ACT-LEEP CAG	ACT-LEEP CAW	GBC	ACT-GBC CAG	ACT-GBC CAW	LogMe	ACT-LogMe CAG	ACT-LogMe CAW	NCE	ACT-NCE CAG	ACT-NCE CAW
CUB200	0.948	**0.950**	0.948	0.916	**0.917**	0.916	-0.951	**0.945**	**0.943**	0.944	**0.948**	0.944
Flowers102	0.769	**0.820**	0.761	0.743	0.742	0.727	-0.759	**0.795**	**0.723**	0.762	**0.823**	0.758
StanfordDogs	0.884	**0.901**	0.884	0.873	**0.876**	0.856	-0.887	**0.847**	**0.842**	0.885	**0.899**	**0.886**
Oxford-IIIT	0.899	**0.907**	**0.905**	0.845	**0.854**	**0.858**	-0.899	**0.852**	**0.476**	0.899	**0.905**	**0.908**
PACS-Sketch	0.940	**0.943**	**0.944**	0.692	**0.852**	**0.894**	0.044	0.035	**0.416**	0.939	**0.940**	**0.941**

Table 3. Results on the semantic segmentation source architecture selection task. Shown are correlation scores (↑) computed across all source architectures. Results where ACT metrics (denoted by 'A') perform better than their counterparts are in **bold**.

Target ($\mathcal{D}_t$)	LEEP	A-LEEP	NCE	A-NCE	GBC	A-GBC
BDD100k	0.147	**0.197**	0.731	**0.743**	0.645	**0.660**
CamVid	0.063	**0.374**	0.573	**0.583**	0.334	**0.796**
SUIM	0.823	**0.980**	0.204	**0.461**	-0.218	**0.784**
CityScapes	0.045	**0.127**	0.524	**0.545**	0.952	0.923

Table 4. Results on the ensemble model selection task. Shown are correlation scores (higher the better) computed across all ensemble candidates. Results where ACT metrics (denoted by 'A') perform better than their counterparts are in **bold**.

Target ($\mathcal{D}_t$)	MS-LEEP	A-MS-LEEP CAG	A-MS-LEEP CAW	E-LEEP	A-E-LEEP CAG	A-E-LEEP CAW
Flowers102	0.230	**0.368**	**0.251**	0.271	**0.314**	0.244
Stanford Dogs	0.400	0.378	0.400	0.503	**0.522**	**0.506**
CUB200	0.334	**0.411**	0.324	0.402	**0.403**	**0.434**
OxfordPets	0.112	**0.148**	**0.133**	0.276	**0.338**	**0.281**
Caltech101	0.462	**0.502**	**0.467**	**0.520**	0.513	0.518

performance on a fixed target dataset [5]. Since evaluating every ensemble combination of the P source models is computationally expensive, the ensemble size K (*i.e.*, number of models per ensemble) is fixed, which yields $\binom{P}{K}$ candidate ensembles. The correlation is then computed between $\mathcal{T}^{s\to t}$ and $\mathcal{A}^{s\to t}$ across all candidate ensemble. We use K=4 and P=11 in our experiments and consider Caltech101, CUB200, Oxford-IIIT Pets, Flowers102, and StanfordDogs datasets as our target datasets.

Model architectures and Training. We include source models pre-trained on the above datasets as well as ImageNet. Each ensemble of model architectures consists of one or more models from the pool of ResNet-101, VGG-19 [63], and DenseNet-201, pre-trained on the mentioned datasets. For a given candidate ensemble, each member model is fine-tuned individually on the target train dataset $\mathcal{D}_t^{\text{train}}$, and the ensemble prediction is calculated as the mean of all individual predictions. Each model is fine-tuned on the target dataset independently, using SGD with a momentum of 0.9, a learning rate of 10^{-4}, and a batch size of 64.

Results. Our empirical analysis in Table 4 shows that, on average across five datasets, ACT metrics outperform the baselines. We observe an improvement in correlation scores of **+14.43%** for MS-LEEP and **+4.10%** for E-LEEP.

4.5. Additional Results on Language Models

Experimental setup. To show the generalizability of our proposed approach, we evaluate the performance of ACT on a language task – sentiment classification and show results in the target dataset selection setting. We consider three target datasets, including TweetEval [8], IMDB Reviews [42], and CARER [59], for our language experiments.

Model architecture and Training. We include source models trained using a classification head on a pre-trained BERT [16] model on CARER [59] and AG-News [85] datasets. We fine-tune the entire source model, including the BERT layers for 3 epochs using the Adam optimizer, with a learning rate of 5×10^{-5}, and a batch size of 8.

Results. Table 5 show that ACT metrics outperform their baseline counterparts. On average, across four source-target pairs and two techniques, we observe an improvement of **+38.13%** for LEEP, **+33.40%** for NCE, and **+57.24%** for GBC using ACT metrics. We also observe an improvement of **+121.11%** for H-Score, and **+52.51%** for TransRate (See Appendix S4 for detailed results).

4.6. Ablation Studies

We conduct ablations on two key components of ACT: i) choice of information scoring method and ii) correlation

Table 5. Results on target task selection for sentiment classification. Shown are correlation scores (higher the better) computed across all target candidates. Results where ACT metrics perform better than their counterparts are in **bold**.

Source - Target Pair	LEEP	ACT-LEEP		NCE	ACT-NCE		GBC	ACT-GBC	
		CAG	CAW		CAG	CAW		CAG	CAW
Emotion - IMDB	-0.172	**0.115**	**0.06**	-0.192	**0.073**	**0.050**	-0.097	**0.141**	**0.109**
Emotion - TweetEval	0.884	**0.892**	**0.885**	0.884	**0.892**	**0.885**	0.828	**0.834**	0.824
AGNews - Emotion	0.939	**0.943**	**0.944**	0.940	**0.947**	**0.944**	0.808	0.808	0.808
AGNews - TweetEval	0.776	**0.779**	**0.784**	0.884	**0.892**	**0.885**	0.549	0.549	0.549

Table 6. Results for ACT transferability metrics compared against other information scoring methods. ACT transferability metrics outperform their base counterparts (shown in **bold**).

Transferability Metric	CAG	CAW	Entropy	Random Subset	Base
LEEP	**0.943**	**0.931**	0.924	0.924	0.926
NCE	**0.953**	**0.943**	0.920	0.919	0.921

score of informative subsets. We also study the impact of different model architectures, size of the informative subset, and less informative samples on the performance of ACT metrics (Appendix S5)

Choice of Information Score. The key component of ACT is the process of identifying informative subsets, as the quality of these identified subsets determines the performance of ACT when applied on existing transferability metrics. We use the source architecture selection setting, and study the effect of using alternative measures to select subsets for the ACT estimation. In particular, we compare the performance of four methods (using PCC scores) – our proposed class-agnostic and class-aware methods, entropy of the target samples in the source model output space, and random selection – on LEEP, and NCE. (Table 6). We also include a comparison with the respective base metrics. Our proposed class-agnostic and class-aware methods consistently pick more informative subsets, leading to better performance over the base transferability metrics.

Transferability for different Informative Buckets. A key question in ACT is to explore the effect of the informativeness of a subset on estimating transferability. We follow the source architecture selection experimental setup (Sec. 4.1), calculate ACT-LEEP using different buckets, and compare it with the baseline LEEP score (full dataset). Results show that transferability estimates are the best (highest correlation scores) for the most informative subsets and gradually degrade while moving towards less informative subsets (Fig. 4), confirming the core hypothesis of ACT.

4.7. Computational Cost

The computational cost associated with transferability estimation methods is crucial as it aims to replace the expensive finetuning-based trial-and-error method. While ACT significantly improves the performance of the base metrics, it adds computational overhead over these methods, when identifying the informative subsets. However, our results in Table 7 show that this overhead is still far lower than the computational cost associated with model fine-tuning, thereby showing that ACT transferability estimators are an *effective*, and *efficient* method for transferability estimation.

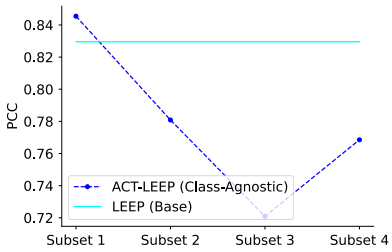


Figure 4. Correlation scores for LEEP (y-axis) for most-to-least informative buckets (x-axis) show that the correlation scores are the highest for the most informative subset.

Table 7. Runtimes for base and ACT transferability metrics using a Resnet-18 trained on CUB200 as the source model. The timings shown for CAG and CAW variants of ACT are an overhead cost upon the base metrics using the experimental setup from Sec. 4. We found a minimum of 50 finetuning epochs necessary for convergence. All values are in seconds.

Dataset	LEEP	NCE	GBC	CAG	CAW	Finetuning
Oxford-IIIT	7.2s	7.3s	8.3s	24.9s	0.6s	$\approx$ 1250s
StanfordDogs	21.1s	21.0s	29.6s	41.8s	1.1s	$\approx$ 3660s

5. Conclusion

We propose and address the problem of estimating transferability from a source model to a target task using examples from an informative subset of the target dataset. To this end, we introduce ACT which leverages class-agnostic and class-aware strategies to identify informative subsets from a target dataset and can be used with any existing transferability metric. We show that ACT metrics outperform their counterparts across different transfer learning tasks, data modalities, models, and datasets. In contrast to the findings in [5] (*i.e.*, one metric doesn't work for all transfer learning tasks), we show that ACT metrics achieve favorable results across diverse transfer learning settings (Sec. 4). We anticipate that using ACT could open new frontiers in estimating transferability and pave the way for several exciting future directions, like developing new techniques to identify informative subsets and extending ACT analysis to other transfer learning tasks.

References

[1] Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6430–6439, 2019. 1

[2] Chirag Agarwal, Daniel D'souza, and Sara Hooker. Estimating example difficulty using variance of gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10368–10378, 2022. 1

[3] Charu C Aggarwal, Xiangnan Kong, Quanquan Gu, Jiawei Han, and S Yu Philip. Active learning: A survey. In *Data classification*, pages 599–634. Chapman and Hall/CRC, 2014. 1, 2, 4

[4] Andrea Agostinelli, Michal Pándy, Jasper Uijlings, Thomas Mensink, and Vittorio Ferrari. How stable are transferability metrics evaluations? *arXiv preprint arXiv:2204.01403*, 2022. 1

[5] Andrea Agostinelli, Jasper Uijlings, Thomas Mensink, and Vittorio Ferrari. Transferability metrics for selecting source model ensembles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7936–7946, 2022. 2, 3, 5, 7, 8

[6] David Alvarez-Melis and Nicolò Fusi. Geometric dataset distances via optimal transport. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. 4

[7] Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas Guibas. An information-theoretic approach to transferability in task transfer learning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2309–2313, 2019. 2

[8] Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa-Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. Semeval 2018 task 2: Multilingual emoji prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 24–33, 2018. 7

[9] William H. Beluch, Tim Genewein, Andreas Nurnberger, and Jan M. Kohler. The power of ensembles for active learning in image classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9368–9377, 2018. 2

[10] Daniel Bolya, Rohit Mittapalli, and Judy Hoffman. Scalable diverse model selection for accessible transfer learning. *Advances in Neural Information Processing Systems*, 34:19301–19312, 2021. 2

[11] Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognit. Lett.*, 30:88–97, 2009. 6

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 1

[13] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 1

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[15] Karan Desai and Justin Johnson. VirTex: Learning Visual Representations from Textual Annotations. In *CVPR*, 2021. 2

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 1, 7

[17] Daniel D'souza, Zach Nussbaum, Chirag Agarwal, and Sara Hooker. A tale of two long tails. *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2021. 1

[18] Kshitij Dwivedi and Gemma Roig. Representation similarity analysis for efficient task taxonomy & transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12387–12396, 2019. 2

[19] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 201–208. JMLR Workshop and Conference Proceedings, 2010. 2

[20] Li Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178, 2004. 6

[21] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 2

[22] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR, 2017. 2

[23] Mohsen Gholami, Mohammad Akbari, Xinglu Wang, Behnam Kamranian, and Yong Zhang. Etran: Energy-based transferability estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18613–18622, 2023. 2

[24] Yuhong Guo. Active instance sampling via matrix partition. In *Proceedings of the 23rd International Conference*

on Neural Information Processing Systems - Volume 1, page 802–810, Red Hook, NY, USA, 2010. Curran Associates Inc. 2

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6

[26] Patrick Hemmer, Niklas Kühl, and Jakob Schöffer. Deal: Deep evidential active learning for image classification. *Deep Learning Applications, Volume 3*, pages 171–192, 2022. 2

[27] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011. 2

[28] Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from imagenet, 2019. 6

[29] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. 6

[30] Long-Kai Huang, Junzhou Huang, Yu Rong, Qiang Yang, and Ying Wei. Frustratingly easy transferability estimation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 9201–9225. PMLR, 2022. 2

[31] Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic Segmentation of Underwater Imagery: Dataset and Benchmark. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE/RSJ, 2020. 6

[32] David Janz, Jos van der Westhuizen, and José Miguel Hernández-Lobato. Actively learning what makes a discrete sequence valid. *arXiv preprint arXiv:1708.04465*, 2017. 2

[33] Ajay J. Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2372–2379, 2009. 2

[34] L. Kantorovich. On the translocation of masses. *Journal of Mathematical Sciences*, 133, 2006. 4

[35] Salman H. Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A. Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8):3573–3587, 2018. 1

[36] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, 2011. 6

[37] Ikki Kishida and Hideki Nakayama. Empirical study of easy and hard examples in cnn training. In *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part IV 26*, pages 179–188. Springer, 2019. 1

[38] D. Li, Y. Yang, Y. Song, and T. M. Hospedales. Deeper, broader and artier domain generalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5543–5551, Los Alamitos, CA, USA, 2017. IEEE Computer Society. 6

[39] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. FSS-1000: A 1000-class dataset for few-shot segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020. 6

[40] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 6

[41] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038, 2014. 6

[42] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, 2011. Association for Computational Linguistics. 7

[43] Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM International Conference on Multimedia*, page 1485–1488, New York, NY, USA, 2010. Association for Computing Machinery. 6

[44] Cuong V. Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. Leep: A new measure to evaluate transferability of learned representations. In *Proceedings of the 37th International Conference on Machine Learning*. JMLR.org, 2020. 1, 2, 3, 5, 6

[45] Hieu T. Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the Twenty-First International Conference on Machine Learning*, page 79, New York, NY, USA, 2004. Association for Computing Machinery. 2

[46] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. 6

[47] Shuteng Niu, Yongxin Liu, Jian Wang, and Houbing Song. A decade survey of transfer learning (2010–2020). *IEEE Transactions on Artificial Intelligence*, 1(2):151–166, 2020. 2

[48] Arghya Pal and Vineeth N Balasubramanian. Zero-shot task transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[49] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009. 1

[50] Michal Pándy, Andrea Agostinelli, Jasper Uijlings, Vittorio Ferrari, and Thomas Mensink. Transferability estimation using bhattacharyya class separability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9172–9182, 2022. 1, 2, 3, 4, 6

[51] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 6

[52] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 2

[53] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. 4

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[56] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Comput. Surv.*, 54(9), 2021. 1, 2, 4

[57] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. 6

[58] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 6

[59] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium, 2018. Association for Computational Linguistics. 7

[60] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. 2

[61] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, page 287–294, New York, NY, USA, 1992. Association for Computing Machinery. 2

[62] Wenqi Shao, Xun Zhao, Yixiao Ge, Zhaoyang Zhang, Lei Yang, Xiaogang Wang, Ying Shan, and Ping Luo. Not all models are equal: Predicting model transferability in a self-challenging fisher space. In *European Conference on Computer Vision*, pages 286–302. Springer, 2022. 2

[63] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 7

[64] Jie Song, Yixin Chen, Xinchao Wang, Chengchao Shen, and Mingli Song. Deep model transferability from attribution maps. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[65] Jie Song, Yixin Chen, Jingwen Ye, Xinchao Wang, Chengchao Shen, Feng Mao, and Mingli Song. Depara: Deep attribution graph for deep knowledge transferability. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3921–3929, 2020. 2

[66] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *International Journal on Computer Vision*, 130, 2022. 1

[67] Yang Tan, Yang Li, and Shao-Lun Huang. Otce: A transferability metric for cross-domain cross-task representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15779–15788, 2021. 4

[68] Xinyi Tong, Xiangxiang Xu, Shao-Lun Huang, and Lizhong Zheng. A mathematical framework for quantifying transferability in multi-source transfer learning. In *Advances in Neural Information Processing Systems*, pages 26103–26116. Curran Associates, Inc., 2021. 2

[69] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010. 1

[70] Anh Tran, Cuong Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1395–1405, 2019. 2, 3

[71] Anh T Tran, Cuong V Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1395–1405, 2019. 1

[72] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and C.V. Jawahar. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. pages 1743–1751, 2019. 6

[73] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. 2

[74] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016. 1

[75] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. Technical Report CNS-TR-201, Caltech, 2010. 6

[76] Garrett Wilson and Diane J. Cook. A survey of unsupervised deep domain adaptation. *ACM Trans. Intell. Syst. Technol.*, 11(5), 2020. 2

[77] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2014. 2

[78] Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. Logme: Practical assessment of pre-trained models for transfer learning. In *ICML*, 2021. 1, 3

[79] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving video database with scalable annotation tooling. *CoRR*, abs/1805.04687, 2018. 6

[80] Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[81] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018. 1, 2

[82] Guojun Zhang, Han Zhao, Yaoliang Yu, and Pascal Poupart. Quantifying and improving transferability in domain generalization. *Advances in Neural Information Processing Systems*, 34:10957–10970, 2021. 2

[83] Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan Kankanhalli. Geometry-aware instance-reweighted adversarial training. In *International Conference on Learning Representations*, 2021. 1

[84] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[85] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NIPS*, 2015. 7

[86] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020. 2