Building a Winning Team: Selecting Source Model Ensembles using a Submodular Transferability Estimation Approach

Vimal K B*[1], Saketh Bachu*[1,3], Tanmay Garg[1], Niveditha Lakshmi Narasimhan[2], Raghavan Konuru[2], and Vineeth N Balasubramanian[1]

[1]Indian Institute of Technology, Hyderabad [2]KLA [3]University of California, Riverside
* Equal Contribution. Corresponding author: vimalkb96@gmail.com

Abstract

Estimating the transferability of publicly available pre-trained models to a target task has assumed an important place for transfer learning tasks in recent years. Existing efforts propose metrics that allow a user to choose one model from a pool of pre-trained models without having to fine-tune each model individually and identify one explicitly. With the growth in the number of available pre-trained models and the popularity of model ensembles, it also becomes essential to study the transferability of multiple-source models for a given target task. The few existing efforts study transferability in such multi-source ensemble settings using just the outputs of the classification layer and neglect possible domain or task mismatch. Moreover, they overlook the most important factor while selecting the source models, viz., the cohesiveness factor between them, which can impact the performance and confidence in the prediction of the ensemble. To address these gaps, we propose a novel Optimal tranSport-based suBmOdular tRaNsferability metric (OSBORN) to estimate the transferability of an ensemble of models to a downstream task. OSBORN collectively accounts for image domain difference, task difference, and cohesiveness of models in the ensemble to provide reliable estimates of transferability. We gauge the performance of OSBORN on both image classification and semantic segmentation tasks. Our setup includes 28 source datasets, 11 target datasets, 5 model architectures, and 2 pre-training methods. We benchmark our method against current state-of-the-art metrics MS-LEEP and E-LEEP, and outperform them consistently using the proposed approach.

1. Introduction

In computer vision, transfer learning is a go-to strategy to train Deep Neural Networks (DNNs) on newer domains and datasets across tasks such as image classification [30, 21], image segmentation [44, 60] and object detection [16, 42]. This widespread usage is due to the easy availability of a large pool of open-sourced pre-trained models (trained on large-scale datasets such as ImageNet [31, 3]), which, when fine-tuned, achieve faster convergence and better performance than training from scratch. However, every time a user wants to employ transfer learning, the question that has increasingly grown relevant with an increased number of source models is: "Which combination of dataset and architecture should I pick to fine-tune to achieve the best performance on my target dataset?". To solve this, we need a tool that helps us choose a source or set of source models, which require minimal fine-tuning and achieves maximal performance.

Transferability estimation (TE) metrics have been proposed in recent years to tackle this problem [48, 36, 58, 47, 39]. With these metrics, a particular source model can be selected without conducting expensive fine-tuning of all available source models on the target training set. Most efforts in this direction are, however limited by their capability of selecting only a single source model, thus restricting their use in an ensemble learning setting. There has been only one work so far [1] which extends an existing single-source transferability estimation method [36] to an ensemble setting. While this work showed promising results, it did not consider the similarity between source and target datasets in the latent representation space, or account for the relationships between individual models in the ensemble. This problem space remains nascent at this time, necessitating more efforts to estimate transferability reliably in different conditions.

Ensemble models have been popular for a few decades now in machine learning [14, 7, 51]. Ensemble models are known to increase task accuracy, decrease overall predictive variance and increase robustness against out-of-distribution

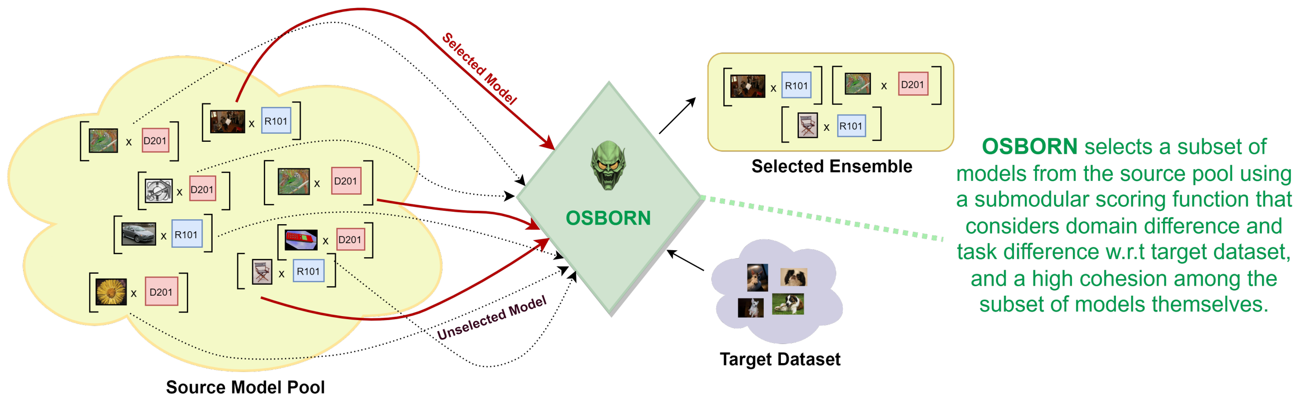

Figure 1. Illustration of the objective and problem setting of our proposed metric. (*Trivia:* OSBORN is also the main antagonist in the Spider-Man movie (2002), hence the emoji.)

data samples [15]. Recent efforts have shown the usefulness of ensembles of pre-trained models [52], especially considering the widespread availability of pre-trained models in the community [41]. The problem of estimating transferability for a model ensemble from a large source model pool becomes even more relevant in this context.

In this work, we introduce a novel transferability estimation metric specifically designed for ensemble selection called Optimal Transport-based Submodular Transferability metric (OSBORN). As stated earlier, a recent effort in this direction [1] showed promising results for such a score, but focused on individual model's performance (via the classifier's outputs) and did not consider the feature (latent representation) space mismatch, or how these models interact with each other in the ensemble. To address this, OSBORN measures the latent space mismatch between the source and the target datasets (domain difference) in addition to the mismatch in the classifier's outputs (task difference). Also, to account for the interaction between models in the ensemble, we introduce a novel model cohesion term, which captures the mutual cooperation between models towards forming an ensemble. Cohesion is required to ensure that individual models in an ensemble are in agreement with each other in terms of predictions (and not voting out each other). Thus, in this work, we propose a domain, task and cohesion-aware transferability estimator for ensemble selection from a source pool of multiple models.

Beyond bringing the abovementioned factors into transferability estimation for ensembles, we show that the proposed score can be viewed as a submodular set function [4]. This allows us to follow a greedy maximization strategy, which is known to provide a high-quality solution for the problem based on well-known theoretical guarantees [34]. We thus select cohesive and closely related models for a particular target dataset. To evaluate our metric, we conduct extensive experiments using 28 source datasets, 11 target datasets, and 5 model architectures. In downstream tasks, we consider fully-supervised pre-training-based image classification, self-supervised pre-training-based image classification, semantic segmentation as well as domain adaptation. Table 1 presents an overview of our experiment breadth, as compared to other recent efforts on this problem. In particular, to the best of our knowledge, we are the first to perform transferability estimation of ensembles for image classification and domain adaptation tasks.

To summarize, we make the following contributions: (1) We introduce a novel transferability estimation metric for ensemble selection that considers domain similarity, task similarity and inter-model cohesion in its design; (2) We show that viewing the proposed metric as a submodular set function allows us to use a simple greedy maximization strategy to select a source model ensemble for a given target dataset; (3) We study the performance of our metric across a wide range of downstream tasks and model pools;(4) We evaluate the reliability of our metric using different correlation metrics in our studies, and also carry out additional analysis and ablation studies to study its usefulness. We outperform earlier methods by a margin of 58.62%, 66.06%, and 96.36% in terms of Pearson Correlation Coefficient (PCC), Kendall τ (KT) [25] and Weighted Kendall τ (WKT) [50] for the image classification task. [1]

2. Related Work

Transfer Learning: Over the years, transfer learning has been applied and explored across various fields [10, 33, 2, 5], as well as across datasets, model architectures, and pre-training strategies [32, 12, 20]. These efforts have included the study of interesting and practical questions such as which particular layers are more transferable [57] or estimating the correlation between pre-training and fine-tuning performance [27]. Beyond finetuning of source models to target datasets, task transfer methods [59, 11] have also

[1] Project page: https://vimalkb007.github.io/OSBORN/

	Single Source TE		
	Classification	Segmentation	DA Classification
# LEEP [36]	✓	×	×
# LogME [58]	✓	×	×
# OTCE [47]	×	×	✓
	Multi Source TE		
# MS-LEEP [1]	×	✓	×
# Ours	✓	✓	✓

Table 1. Experimental settings studied by different methods in single-source TE and multi-source TE settings (DA: Domain Adaptation). We note the wide range of our experimental settings when compared to earlier work.

studied relationships between visual tasks such as semantic segmentation, depth prediction and vanishing point prediction, or used attribution maps to relate such tasks [45, 46]. In contrast to the aforementioned methods, the objective of our work is dataset transferability estimation.

Transferability Estimation Metrics (Single Source): As stated earlier, gauging transferability reduces the effort in finding an optimal source model for a particular target dataset because it averts the expensive fine-tuning process. In recent years, significant efforts have been made in this problem space, considering the relevance of this problem to practitioners. The H-Score was proposed [6] to measure the usefulness (in terms of discriminativeness) of pre-trained source models for the target task. While this method shows promising results as a pioneer work in this field, it misses considering the scenarios where the source and target data have different distributions. Subsequently, NCE [48], and LEEP [36] developed methods that used the classifier outputs of pre-trained source models when the target dataset is forward-propagated through the model to estimate the log-likelihood of the target dataset. NCE largely focused on estimating transferability in scenarios where the source and target tasks share the same input data (e.g., face recognition and facial attribute classification). Subsequent methods such as LogME [58] also showed that likelihood methods might be prone to over-fitting. To tackle this, LogME [58] estimated the maximum value of label evidence (instead of maximum likelihood) given the feature set extracted by the pre-trained source models. Considering the fact that previous methods largely relied on classifier outputs and their sub-optimal performance in practical scenarios like cross-domain settings, OTCE [47] proposed an optimal transport framework to compute domain difference (based on feature space) and task difference (based on label space) to estimate transferability. This method leveraged the source model's latent representations in addition to classifier outputs with no explicit assumptions on the source and target datasets. All the above works are, however focused on estimating transferability from a single source model to a target dataset.

Transferability Estimation Metrics (Multi-Source Ensembles): Agostinelli et al[1] recently proposed the first work on extending transferability estimation to select source model ensembles in [1], specifically focused on semantic segmentation. This work extends LEEP [36] to ensembles, and shows promising results in the considered settings. Our work builds on this effort in multiple ways: (i) instead of solely relying on classifier outputs for estimating transferability [36, 1, 48], we also consider the domain mismatch in the latent feature representation space; (ii) beyond looking at the individual model's outputs in an ensemble, we also consider the interactions and correlation between the model outputs; (iii) we make no assumptions on the source and target data distributions; and (iv) while [1] focused on segmentation, we show our method's results on classification, segmentation and domain adaptation tasks. We also show results on multiple pre-training strategies while previous works [36, 58, 48, 47] mostly focus on fully-supervised pre-training strategies. Our proposed metric can also be viewed as a submodular function, which allows us to leverage ranking-based greedy optimization strategies to make it efficient in practice.

Ensemble Learning. Learning ensembles of models has been popular in machine learning to increase overall task performance, decrease prediction variance, prevent over-fitting, and increase out-of-distribution robustness [7, 18, 56, 38]. More recent efforts in training ensembles of neural network models have focused on speeding up their training [49, 52], leveraging a single model's capacity to train multiple subnetworks whose predictions are ensembled to improve robustness [19], or studying mixture-of-experts paradigms which bring together thousands of subnetworks for large language models [43]. We clarify that our work focuses rather on selecting model ensembles from a larger source model pool via estimating transferability without explicitly training ensembles themselves. One can view our work as a step before ensemble learning when there is a larger model pool and only few models can be ensembled. As stated in [1], this setting is commonly encountered by a practitioner in the real-world across application domains.

3. Background and Preliminaries

Notations: Given M source datasets, we denote the r^{th} source dataset as $D_{s^r}=\{(x^i_{s^r}, y^i_{s^r})\}_{i=1}^{n^r} \sim P_{s^r}(x,y)$ and target dataset as $D_t=\{(x^i_t, y^i_t)\}_{i=1}^{m} \sim P_t(x,y)$ where, $x^i_{s^r} \in \mathcal{X}_{s^r}, x^i_t \in \mathcal{X}_t$, $y^i_{s^r} \in \mathcal{Y}_{s^r}$, and $y^i_t \in \mathcal{Y}_t$. Note that we do not restrict the label spaces $P(\mathcal{Y}_{s^r})$ and $P(\mathcal{Y}_t)$ to span the same category set. We base our study on a domain-agnostic and task-agnostic setting.

Transferability Estimation for Ensembles: For every source dataset D_{s^r}, we assume there exists a pre-trained model on that dataset denoted by (θ_{s^r}, h_{s^r}) where θ is the feature extractor, and h is the classifier head. M represents the collection of such source models. As stated earlier, we focus on a multiple source model selection setting (i.e. ensembles) where our metric provides a transferability estimation (TE) score $\alpha^{M_e \to t}$ for a given subset of models M_e from the source pool M. When correlated to the accuracy $A^{M_e \to t}$ (i.e. fine-tuned accuracy of the ensemble on the target test set), this TE score provides the reliability of the transferability estimate. Following [1], we calculate the ensemble accuracy by fine-tuning individual models in subset M_e (both θ and h) on the target train set and averaging their predictions on the target test set.

Submodularity in TE for Ensembles. The main idea of **TE** involves choosing optimal source models for a given target dataset. Apart from performance & computation trade-offs, a crucial motivation to select a subset of models is to mitigate risk of negative transfer. Fig 2 herein shows that opting for all models in the ensemble could lead to a decrease in overall performance compared to selecting a smaller set of models. This can be due to the detrimental impact of weak or non-transferable models in the ensemble, highlighting the importance of carefully combining models to ensure optimal performance. Further, finding an optimal ensemble for a given target dataset requires checking all possible combinations of different source models for a particular ensemble size. This exhaustive process is an NP-hard problem. In this paper, we propose a submodular approach to rank the available models in the source pool according to the performance gain they would yield if added to the subset pool of the ensemble and select the top k models, where k is the required size of the ensemble. While submodular subset selection is popular in different machine learning settings [4, 24, 53], to the best of our knowledge, this is the first such use for transferability estimation. To this end, we first formally define submodularity below.

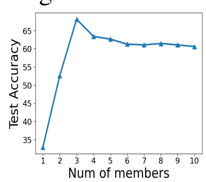

Figure 2. Test accuracies on Caltech101 with varying subsets of models (chosen randomly)

Definition 3.1. *Let Ω be a set and $\mathcal{P}(\Omega)$ be the power set of Ω, then a submodular function is a set function $f : \mathcal{P}(\Omega) \to \mathbb{R}$. The submodular function follows the property of diminishing returns, i.e. adding a new element to a smaller set produces a larger increase in f compared to a larger set. Mathematically, if for all $X, Y \subseteq \Omega$, where $X \subseteq Y$ and for all $v \in \Omega \setminus Y$, the property follows:*

$$f(X + v) - f(X) \geq f(Y + v) - f(Y) \tag{1}$$

A key benefit of posing a problem as one of submodular subset selection is that a greedy approach can be leveraged to efficiently identify a solution of required subset size that is reasonably close to the optimal solution. Nemhauser [34] showed that the quality of the subset chosen greedily cannot be worse than $1 - e^{-1}$ of the optimal value. This makes submodularity an attractive approach for usage in the field of TE for ensembles as we can rank the models in the source pool and select an ensemble of desired size. Further details on how to greedily select the models are discussed later in this paper.

Evaluation Criteria. As stated earlier, the reliability of a **TE** method is obtained by measuring the correlation between $\alpha^{M_e \to t}$ and $A^{M_e \to t}$. Previous works [58, 36, 1, 47, 48] measure this correlation using different techniques such as Pearson Correlation Coefficient (PCC), Kendall τ (KT) [25] and Weighted Kendall τ (WKT) [50]. We report results for all these correlation measures to be comprehensive in our analysis.

4. OSBORN: Transferability Estimation Metric for Model Ensemble Selection

In order to design a reliable transferability estimation approach for model ensembles, we propose the Optimal Transport-based Submodular Transferability metric (OSBORN), which considers three factors: domain difference, task difference, and inter-model cohesion. Inspired by earlier efforts on single-source transferability estimation [47], we consider both classifier output and distance in the latent representation space in our approach. Besides, since our focus is on model ensembles, we consider inter-model relationships in this metric. We now describe each of these quantities.

Minimize Domain Difference (W_D). In order to minimize the latent space mismatch between the source and target datasets, similar to [47], we choose Wasserstein distance and Optimal Transport (OT) to compute this mismatch owing to its advantages in capturing the geometries of underlying data. Mathematically, the p-Wasserstein distance is given as follows:

$$W_p(\beta, \gamma) = \left(\inf_{\pi \in \Pi(\beta, \gamma)} \int D(x, z)^p d\pi(x, z) \right)^{1/p} \tag{2}$$

where, $p \geq 1$, β, γ are continuous or discrete random variables in a complete and separable space S, $D(.,.) : S \times S \to \mathbb{R}^+$ is a distance or a cost function between two points x and z, $\pi(\beta, \gamma)$ is the coupling matrix which can also be understood as the joint probability distributions with marginals β and γ. In particular, in this work, we use the 1-Wasserstein distance, also called the Earth Mover Distance, to calculate the domain difference between source and target latents as:

$$W_D(\theta_s, x_t) = \sum_{i,j=1}^{m,n} ||\theta_s(x_s^i) - \theta_s(x_t^j)||_2^2 \pi_{ij}^*, \tag{3}$$

where $|| \cdot - \cdot ||_2^2$ is the distance or cost metric, π^* is the optimal coupling matrix of size $m \times n$ obtained by solving the optimal transport (OT) problem using the Sinkhorn algorithm [9, 47]. Note that $\theta_s(.)$ is the feature extractor belonging to the source model. Intuitively, if the latent space of the source dataset is closely aligned with that of the target dataset, it is easier for the model to transfer.

Minimize Task Difference (W_T). In order to measure the difference between a source task and the given target task, we use the mismatch between the model/classifier's outputs for source and target data forward-propagated through the source model. We use the conditional entropy (CE) of the predicted labels $\hat{y}_t \in \mathcal{Y}_s$ of the target dataset samples given their ground truth labels $y_t \in \mathcal{Y}_t$. The predicted labels are obtained by forward-propagating the target samples x_t through the corresponding source model θ_s. Let $\hat{Y}_t$ be a random variable that takes values in the range of $\mathcal{Y}_s$; and Y_t be a random variable that takes values in the range of $\mathcal{Y}_t$, then W_T can be calculated as:

$$W_T(\theta_s, x_t) = H(\hat{Y}_t | Y_t) = -\sum_{\hat{y}_t \in \mathcal{Y}_s} \sum_{y_t \in \mathcal{Y}_t} \hat{P}(\hat{y}_t, y_t) \log \frac{\hat{P}(\hat{y}_t, y_t)}{\hat{P}(y_t)} \quad (4)$$

where $\hat{P}(\hat{y}_t, y_t)$ is the joint distribution of predicted and ground truth target labels and $\hat{P}(y_t)$ is the marginal distribution of the ground truth labels. These quantities can be easily computed using the optimal coupling matrix (obtained in Eqn 3) as follows:

$$\hat{P}(\hat{y}_t, y_t) = \sum_{i,j:\hat{y_t^i}=\hat{y}_t, y_t^j = y_t} \pi^*_{ij}, \quad (5)$$

The marginal distribution can be obtained from the joint distribution as follows:

$$\hat{P}(y_t) = \sum_{\hat{y}_t \in \mathcal{Y}_s} \hat{P}(\hat{y}_t, y_t), \quad (6)$$

Intuitively, similar tasks will result in a low W_T value. Using W_T i.e CE alone represents empirical transferability according to [48]. However, in [47], it is experimentally shown that using only CE is insufficient in a domain-agnostic setting, which motivates us to combine this with W_D to account for feature representation space mismatch.

Minimize Model Disagreement (Cohesiveness W_C). For an ensemble, it is important that the individual models reinforce the predictions of each other and have less disagreement amongst themselves to have overall good performance. To understand the cohesiveness of an ensemble, we use Conditional Entropy to capture the amount of disagreement between models in the subset of models M_e. Mathematically, we represent W_C as:

$$W_C(M_e, x_t) = \sum_{m_i, m_j \in M_e} H(m_i(x_t) | m_j(x_t)) \quad (7)$$

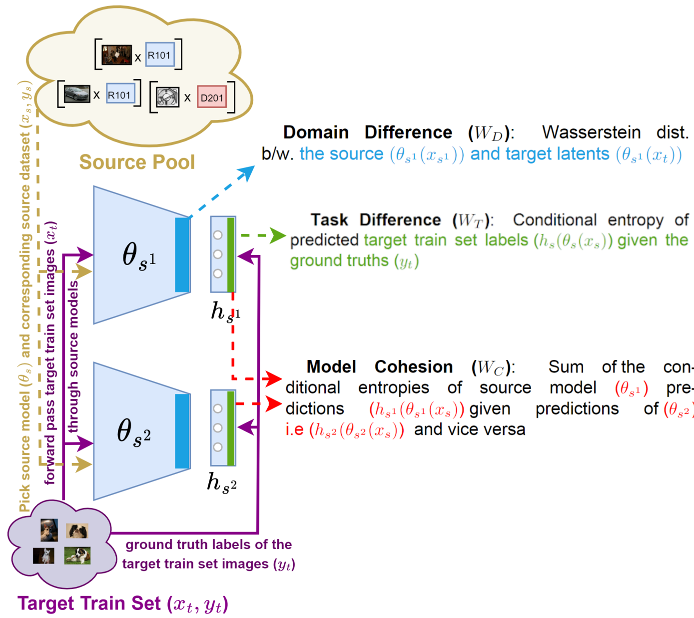

Figure 3. Overview of our method for estimating the transferability for ensembles.

Intuitively, we want a high cohesiveness and less disagreement among the models to reinforce the ensemble's predictive belief, i.e. a low W_C value, and to avoid scenarios where models vote out each other's predictions.

Bringing the quantities together, we define OSBORN for a subset of models M_e of our source pool M as follows. Our metric collectively accounts for domain difference, task difference and model cohesion. Ref. Fig. 3 for the overview.

$$\text{OSBORN} = \sum_{m_i \in M_e} [W_D(m_i, x_t) + W_T(m_i, x_t)] + W_C(M_e, x_t) \quad (8)$$

A model ensemble that obtains a low OSBORN score will have better transferability to a target dataset. Our experiments show that a simple combination of these three quantities (with no weighting co-efficients) outperforms existing methods in all our experiments. In our ablation studies and analysis, we study the contribution of each OSBORN component as well as the effect of weighting each component differently.

Submodular Subset Selection in OSBORN. As stated earlier, we show that the proposed OSBORN metric translates to a submodular optimization problem, which allows us to rank and pick models efficiently from the source pool. While the aforementioned quantities were written from a *minimization* perspective (for clarity and ease of understanding), to pose this as a submodular *maximization* problem, we consider the corresponding scoring function to be maximized as:

$$f(M_e) = -\sum_{m_i \in M_e} [W_D(m_i, x_t) + W_T(m_i, x_t)] - W_C(M_e, x_t) \tag{9}$$

The value of our set function is a transferability estimate designed such that it is highly correlated to the fine-tune accuracy (see Table 3 & 4), thus enabling us to select models without expensive fine-tuning.

Theorem 4.1. *The scoring function $f(X)$, as defined in Equation 9, is a submodular function.*

Proof. Let X_1 and X_2 be two sets such that $X_1 \subseteq X_2 \subseteq M$. If we consider an unselected model instance $v \in M \backslash X_2$. The gain in the score is obtained by appending v to the set X_1, and this is calculated as:

$$\begin{aligned} f(X_1 \cup v) - f(X_1) = &- [W_D(v, x_t) + W_T(v, x_t)] \\ &- \sum_{m_i \in X_1} H(m_i(x_t) \mid v(x_t)) \\ &- \sum_{m_j \in X_1} H(v(x_t) \mid m_j(x_t)) \end{aligned} \tag{10}$$

Similarly, the gain obtained by set X_2 is given by:

$$\begin{aligned} f(X_2 \cup v) - f(X_2) = &- [W_D(v, x_t) + W_T(v, x_t)] \\ &- \sum_{m_i \in X_2} H(m_i(x_t) \mid v(x_t)) \\ &- \sum_{m_j \in X_2} H(v(x_t) \mid m_j(x_t)) \end{aligned} \tag{11}$$

As we have $X_1 \subseteq X_2$, the number of terms in the summation of Equation 11 will be greater than or equal to that of Equation 10. Since entropy is always a non-negative value, we can say that

$$\begin{aligned} &- \sum_{m_i \in X_1} H(m_i(x_t) \mid v(x_t)) - \sum_{m_j \in X_1} H(v(x_t) \mid m_j(x_t)) \geq \\ &- \sum_{m_i \in X_2} H(m_i(x_t) \mid v(x_t)) - \sum_{m_j \in X_2} H(v(x_t) \mid m_j(x_t)) \end{aligned}$$

This implies that

$$f(X_1 \cup v) - f(X_1) \geq f(X_2 \cup v) - f(X_2) \tag{12}$$

We can see that Equation 12 satisfies the condition in Definition 3.1. This completes the proof. ☐

Submodular Optimization using Greedy Maximization. Since our set function $f(M_e)$ (mentioned in Eq. 9) is submodular, it exhibits monotonicity, i.e. the set with maximum gain is always the entire source pool M. However, since we want to select a subset of models i.e. ensemble set from the source pool M, we impose a cardinality constraint. Formally, we aim to select the set M_e of size at most k that maximizes the gain:

$$\max_{M_e : |M_e| = k} f(M_e) \tag{13}$$

This problem is however NP-hard, but we use the greedy maximization strategy to find a near-optimal set of models M_e for the target dataset. In practice, we pre-calculate pair-wise domain difference W_D and task difference W_T between each source and target datasets. Then, we calculate the model cohesion term W_C for adding each model m_i to the set of already selected models M_e. Using these three quantities pertaining to m_i, we calculate the gain achieved by adding it to the set M_e as $f(M_e \cup m_i) - f(M_e)$ and greedily pick the model with the highest gain and add it to the set M_e. We continue this iteration till we achieve the ensemble set size of k. Once the target samples are forward-propagated through the source models, the quantities in our metric can be computed independently for each source model, thus making our overall computations parallelizable.

Considering M_e^* as the optimal ensemble set, it is well-known from [34] that such a greedy approach has a performance guarantee of at least 63% of the optimal ensemble set, i.e.

$$f(M_e) \geq \left(1 - \frac{1}{e}\right) f(M_e^*) \tag{14}$$

In practice, we observe that we see that the avg. accuracy of the ensemble selected by greedy (76.315%) in a fully-supervised setting is, 95.56% of the avg. accuracy of the optimal ensemble(79.857%). Similarly for self-supervised setting, the avg. accuracy of the ensemble selected by greedy (79.857%) is, 93.50% of the avg. accuracy of the optimal ensemble(84.962%), as shown in Table 2. More details on the experiments are presented in Sec 5.

Target Dataset	**Ensemble Accuracy (Fully Supervised)**	
	Greedy	**Optimal**
Oxford102Flowers	90.720	91.697
Caltech101	68.533	75.333
StanfordCars	69.692	72.540
Average	76.315	79.857
	Ensemble Accuracy (Self Supervised)	
Oxford102Flowers	86.935	95.604
Caltech101	88.800	90.000
StanfordCars	62.604	69.282
Average	79.446	84.962

Table 2. Comparison of the target test set accuracies achieved by fine-tuned ensembles selected using the greedy optimization of OSBORN vs the optimal ensembles. We clearly observe that our approach empirically gives significantly stronger performances than the theoretical guarantee.

5. Experiments and Results

Experimental Setup. We follow the same experimental setup as the previous work on source model ensemble selection [1] to evaluate our transferability metric in the multiple source model setting. Given a total of M models in the source pool, our objective is to select an ensemble model by choosing k models from the source pool. We follow [1] in setting k to 3 for fairness of comparison. We also conducted a study to evaluate this on the Oxford-IIIT Pets dataset, and found that maximum accuracy is gained for an ensemble of size 3 (see Fig 4), which further reinforces our choice for conducting experiments.

Classification Datasets. For the classification tasks, we consider 11 widely-used datasets including CIFAR-10 [29], CIFAR-100 [29], Caltech-101 [13], Stanford Cars [28], Oxford 102 Flowers [37], Oxford-IIIT Pets [40], Imagenette [22], CUB200 [54], FashionMNIST [55], SVHN [35], Stanford Dogs [26]. These datasets are popularly used in many transfer learning tasks. Out of these 11 datasets, we set Caltech-101 [13], Stanford Cars [28], Oxford 102 Flowers [37], Oxford-IIIT Pets [40], Stanford Dogs [26] as our target datasets and estimate transferability using OSBORN.

Model Architectures (Fully-supervised). For this setting, we consider 2 source model architectures ResNet-101 [21] and DenseNet-201 [23], keeping in mind the model diversity and capacity. We take these models from the open-sourced PyTorch Library [41]. Initially, both the models are initialized with the fully-supervised ImageNet weights [31], and then we train them on the 11 classification datasets to prepare our source model pool.

Model Architectures (Self-supervised). For this setting, we consider ResNet-50 [21] as our source model architecture but initialize it with weights obtained from two self-supervised pre-training strategies, namely BYOL [17] and MoCov2 [8]. We have two variants of ResNet-50 models to produce enough diversity. And as done in the previous case, we train these two models on the 11 classification datasets to prepare our source model pool. We use multiple pre-trained SSL models to build our pool. However, finetuning is done in a fully-supervised fashion. Our motivation here was to study if OSBORN can estimate transferability reliably across multiple pre-training settings.

Training Setup for Source Models (Classification Tasks). For all classification tasks, we train the source models using a cross-entropy loss and optimize it using Stochastic Gradient Descent (SGD) with momentum. Given these details, the most important hyperparameters are learning rate, batch size and weight decay. We train the models with a grid search of learning rate in ($1e{-}1$, $1e{-}2$, $1e{-}3$, $1e{-}4$), batch size in $(32, 64, 128)$, and weight decay in ($1e{-}3$, $1e{-}4$, $1e{-}5$, $1e{-}6$, 0) to pick the best hyperparameters. All our experiments are written in PyTorch and are conducted on a single Tesla V-100 GPU. For the fully-supervised pre-trained setting, we initialize the models with ImageNet weights. In the case of a self-supervised pre-trained setting, we initialize the models using BYOL or MoCov2 (on ImageNet) weights. For our experiments on the multi-domain DomainNet dataset, we initialize our models with ImageNet weights.

Training Setup for Source Models (Semantic Segmentation Tasks). We train the source models using a pixel-wise cross-entropy loss and optimize it using Stochastic Gradient Descent (SGD) with momentum. The most important hyperparameters herein are learning rate, batch size and weight decay. We train the models with a grid search of learning rate in ($1e{-}1$, $1e{-}2$, $1e{-}3$, $1e{-}4$), batch size in $(32, 64, 128)$, and weight decay in ($1e{-}3$, $1e{-}4$, $1e{-}5$, $1e{-}6$, 0), and pick the best hyperparameters. All these experiments are also written in PyTorch and conducted on a single Tesla V-100 GPU. We initialize source models using the COCO pre-trained weights.

Implementation of Source Models and Baselines. We use open-source models available via the PyTorch Library for classification and semantic segmentation tasks. We use the PyTorch Lightning Library to obtain model weights for a self-supervised pre-training setting. We use the code released by the respective papers for calculating OTCE [47], MS-LEEP, E-LEEP, IoU-EEP and SoftIoU-EEP [1] scores.

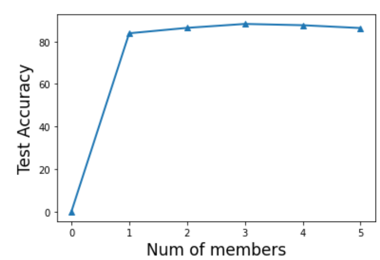

Figure 4. Test accuracy on the Oxford-IIIT Pets dataset compared to the ensemble size. We observed a similar trend across other datasets as well.

Evaluating Ensemble Performance. We follow the protocol in [1] for computing ground truth accuracies of ensembles. We finetune (both feature extractor and classifier of) all the source models present in the ensemble using the target training set. Then, we individually make predictions using the source models on the target test set and average them to get the final ensemble prediction. We note that no target-trained models are in the source pool. We compare this final prediction with the ground-truth label and calculate the classification accuracy. Note that we need to fine-tune all source models only once and can re-use their predictions on the test set across all ensemble combinations. As stated earlier, we report Pearson Correlation Coefficient (PCC), Kendall τ (KT) and Weighted Kendall τ (WKT) in our results.

Evaluation on Fully-Supervised Pre-Trained Models. We herein compare our OSBORN with the baseline metrics, i.e. MS-LEEP and E-LEEP, in terms of three correlation

Target Dataset	Weighted Kendall's τ			Kendall's τ			Pearson		
	MS	E	Ours	MS	E	Ours	MS	E	Ours
Oxford102Flowers	0.086	-0.019	**0.616**	0.138	0.074	**0.400**	0.230	0.164	**0.456**
OxfordIIITPets	0.414	0.393	**0.558**	0.346	0.326	**0.453**	0.504	0.500	**0.666**
StanfordDogs	0.326	0.323	**0.477**	0.244	0.242	**0.427**	0.398	0.407	**0.604**
Caltech101	0.435	0.409	**0.565**	0.240	0.231	**0.335**	0.353	0.341	**0.486**
StanfordCars	0.115	0.018	**0.486**	0.137	0.071	**0.368**	0.256	0.163	**0.549**
Average	0.275	0.225	**0.540**	0.221	0.190	**0.367**	0.348	0.315	**0.552**

Table 3. Comparison of different ensemble transferability estimation metrics for fully-supervised models (classification tasks). The best results are indicated in bold. Note: MS: MS-LEEP, E: E-LEEP, Ours: OSBORN.

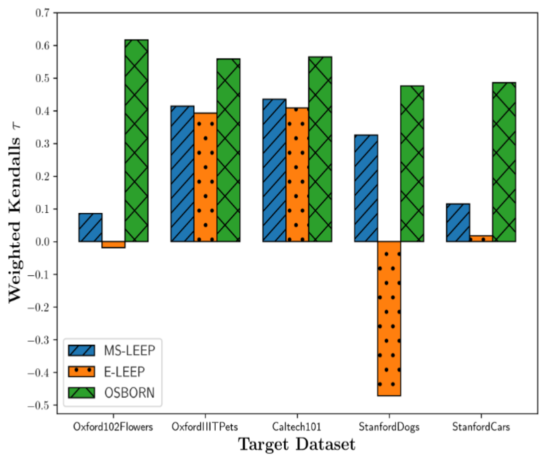

Figure 5. Comparison of OSBORN over 5 target datasets interms of Weighted Kendalls's τ. We can see that our metric constantly outperforms the baselines across every dataset by a large margin.

metrics, WKT, KT, and PCC[2]. The correlation values are reported in Table 3. Averaged across five target datasets, OSBORN improves 96.36% over MS-LEEP and 140% over E-LEEP in terms of WKT; improves 66.06% over MS-LEEP and 93.16% over E-LEEP in terms of KT; improves 58.62% over MS-LEEP and 75.23% over E-LEEP in terms of PCC. We can visually see the overall performance of our metric outperforming the existing baselines significantly in Fig 5.

Evaluation on Self-Supervised Pre-Trained Models. We compare the performance of our method with the baseline methods, i.e. MS-LEEP and E-LEEP. We present the experimental results regarding different correlation coefficients in Table 4. Note that we use the Frobenius norm regularizer while solving the OT problem because it gave us better results when compared to using other regularizers. In the appendix, we report results without any regularizers and compare them with the Frobenius norm variant. Table 4 shows that, averaged across five target datasets, OSBORN improves 268.69% over MS-LEEP and 231.82% over E-LEEP in terms of WKT; improves 442.10% over MS-LEEP and 379.07% over E-LEEP in terms of KT; improves 527.27% over MS-LEEP and 392.86% over E-LEEP in terms of PCC.

Performance of Selected Ensembles. Table 2 reports the ensemble accuracy of the models selected through OSBORN. For completeness of this discussion, we also report the same results for OSBORN without greedy maximization as well as for MS-LEEP and E-LEEP in Table 5. Following [1], we first calculate the OSBORN value for every ensemble candidate and pick the ensemble that bags the highest value. We follow a similar strategy with MS-LEEP and E-LEEP to pick the best model according to their values. To compute the ensemble accuracy, we used the individual models fine-tuned on the target train set and got their predictions on the target test set. We average these predictions and compare them with the ground truth labels to obtain overall accuracy. We observe that the ensemble selected by OSBORN achieves the highest test accuracy across all datasets. In the case of both fully supervised and self-supervised settings, the baseline methods, i.e. MS-LEEP and E-LEEP, select the same ensembles (despite having different correlation values) in every case, which is why they obtain the same ensemble accuracy.

Scaling Number of Models in Ensemble. As shown earlier in this section (Fig 4), we found the performance to saturate after an ensemble size of 3 in the datasets considered in this work as well as in [1]. On the other hand, we also observe unsurprisingly that the cost of ensemble selection can go up significantly as the ensemble size increases. We show the cost performance of models selected for the Caltech101 dataset in Fig 6. Despite the increasing trend, we note that the time taken is still in the order of seconds, which makes the proposed OSBORN metric practical and relevant.

Ablation Studies. We conducted additional experiments to understand the influence of each component in OSBORN (included in the Appendix). In general, while simple addition of the three quantities in OSBORN without any weights outperformed previous methods, we observed that these can be finetuned through grid search over a larger range of values to get even better transferability estimates. This however varies with the target dataset. On Caltech101 as the target dataset, we noticed that giving more weightage to W_D

[2] Our baselines MS-LEEP and E-LEEP use custom proprietary model architectures that are not publicly available. We hence followed the authors' code and guidelines in using their method on the models used in our work, and picked the best-performing hyperparameters for the results corresponding to their baselines shown in this work.

Target Dataset	Weighted Kendall's τ			Kendall 's τ			Pearson		
	MS	E	Ours	MS	E	Ours	MS	E	Ours
Oxford102Flowers	-0.080	-0.090	**0.549**	-0.035	-0.050	**0.336**	-0.077	-0.090	**0.306**
OxfordIIITPets	0.555	**0.574**	0.357	0.221	**0.229**	0.139	0.201	0.212	**0.232**
StanfordDogs	0.089	0.132	**0.170**	0.014	0.029	**0.110**	0.132	0.159	**0.236**
Caltech101	0.290	0.311	**0.488**	0.195	0.228	**0.308**	0.248	0.287	**0.374**
StanfordCars	-0.359	-0.377	**0.260**	-0.207	-0.221	**0.139**	-0.285	-0.289	**0.232**
Average	0.099	0.110	**0.365**	0.038	0.043	**0.206**	0.044	0.056	**0.276**

Table 4. Comparison of different ensemble transferability estimation metrics for self-supervised pre-trained models (classification tasks). The best results are indicated in bold. Note: MS: MS-LEEP, E: E-LEEP and Ours: OSBORN.

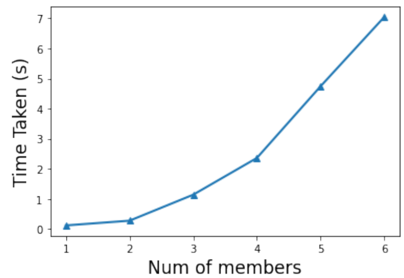

Figure 6. Cost performace of model selection for the Caltech101 dataset.

Target Dataset	Ensemble Accuracy (Fully Supervised)		
	MS-LEEP	E-LEEP	Ours
Oxford102Flowers	85.347	85.347	**89.865**
Caltech101	**68.533**	**68.533**	**68.533**
StanfordCars	48.623	48.623	**62.915**
Average	67.501	67.501	**73.771**
	Ensemble Accuracy (Self Supervised)		
Oxford102Flowers	88.278	88.278	**93.040**
Caltech101	**86.933**	**86.933**	**89.333**
StanfordCars	6.056	6.056	**61.820**
Average	60.422	60.422	**80.598**

Table 5. We compare the target test set accuracies achieved by fine-tuned model ensembles picked by MS-LEEP, E-LEEP and OSBORN.

compared to the other two terms (W_T and W_C) achieved higher correlation scores, as shown in Fig 7. This could be because of the wide variety of images in this dataset. W_D measures the latent space mismatch between such varied images with the source datasets (which may not have overlapping images/representation with the target set), which benefits in this case. More detailed analysis is provided in the Appendix.

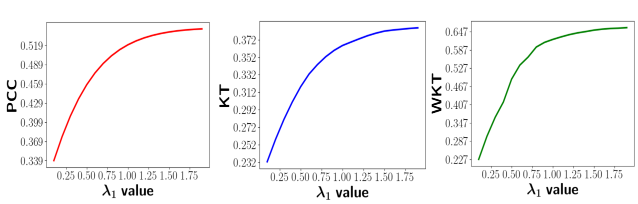

Figure 7. W_D weightage vs. correlation comparison for Caltech101. We set weights for W_T and W_C as 1.

6. Conclusions

In this paper, we propose a novel optimal transport-based transferability estimation metric, OSBORN, carefully designed for ensembles that consider multiple factors, such as the mismatch in the latent space, label space, and the cohesiveness amongst the individual models in the ensemble. We show that the proposed metric can be treated as a submodular optimization problem, allowing us to leverage a greedy strategy for source model ensemble selection. We show experimentally that our metric outperforms the existing metrics MS-LEEP and E-LEEP across tasks on multiple correlation metrics. Future directions include increasing the computational efficiency of this method, as well as studying its applicability to other tasks and problem settings.

Acknowledgements

This work was partly supported by KLA and the Department of Science and Technology, India through the DST ICPS Data Science Cluster program. We would like to thank the authors of [1] for insightful discussions. Further, we thank the anonymous reviewers for their valuable feedback that improved the presentation of this paper.

References

[1] Andrea Agostinelli, Jasper Uijlings, Thomas Mensink, and Vittorio Ferrari. Transferability metrics for selecting source model ensembles. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7926–7936, 2022. 1, 2, 3, 4, 7, 8, 9

[2] Khurshed Ali, Chih-Yu Wang, and Yi-Shin Chen. Leveraging transfer learning in reinforcement learning to tackle competitive influence maximization. *Knowledge and Information Systems*, 64:2059–2090, 2022. 2

[3] Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. Factors of transferability for a generic convnet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9):1790–1802, 2016. 1

[4] Francis Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning*, 6(2-3):145–373, 2013. 2, 4

[5] Adrià Puigdomènech Badia, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, Bilal Piot, Steven Kapturowski, Olivier Tieleman, Martín Arjovsky, Alexander Pritzel, Andew Bolt,

and Charles Blundell. Never give up: Learning directed exploration strategies. *8th International Conference on Learning Representations, 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020. 2

[6] Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas J. Guibas. An information-theoretic approach to transferability in task transfer learning. In *2019 IEEE International Conference on Image Processing, 2019, Taipei, Taiwan, September 22-25, 2019*, pages 2309–2313. IEEE, 2019. 3

[7] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 2004. 1, 3

[8] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020. 7

[9] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. 5

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. 2

[11] Kshitij Dwivedi and Gemma Roig. Representation similarity analysis for efficient task taxonomy & transfer learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12379–12388, 2019. 2

[12] Linus Ericsson, Henry Gouk, and Timothy M. Hospedales. How well do self-supervised models transfer? In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5410–5419, 2021. 2

[13] 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. 7

[14] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997. 1

[15] Mudasir A. Ganaie, Minghui Hu, Mohammad Tanveer, and Ponnuthurai N. Suganthan. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151, 2022. 2

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. 1

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. 7

[18] L.K. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990. 3

[19] Marton Havasi, Rodolphe Jenatton, Stanislav Fort, Jeremiah Zhe Liu, Jasper Snoek, Balaji Lakshminarayanan, Andrew M Dai, and Dustin Tran. Training independent subnetworks for robust prediction. *International Conference on Learning Representations*, 2020. 3

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9726–9735, 2020. 2

[21] 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*, pages 770–778, 2016. 1, 7

[22] Jeremy Howard and Sylvain Gugger. Fastai: A layered api for deep learning. *Information*, 11(2), 2020. 7

[23] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2261–2269, 2017. 7

[24] Rishabh Iyer and Jeff Bilmes. Submodular optimization with submodular cover and submodular knapsack constraints. volume 26 of *Advances in neural information processing systems*, page 2436–2444, Red Hook, NY, USA, 2013. Curran Associates Inc. 4

[25] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 06 1938. 2, 4

[26] 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, June 2011. 7

[27] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2019. 2

[28] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 7

[29] Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012. 7

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012. 1

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural net-

works. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 1, 7

[32] Thomas Mensink, Jasper Reinout Robertus Uijlings, Alina Kuznetsova, Michael Gygli, and Vittorio Ferrari. Factors of influence for transfer learning across diverse appearance domains and task types. *Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[33] Basil Mustafa, Aaron Loh, Jana von Freyberg, Patricia MacWilliams, Megan Wilson, Scott Mayer McKinney, Marcin Sieniek, Jim Winkens, Yuan Liu, Peggy Bui, Shruthi Prabhakara, Umesh Telang, Alan Karthikesalingam, Neil Houlsby, and Vivek Natarajan. Supervised transfer learning at scale for medical imaging. *ArXiv*, abs/2101.05913, 2021. 2

[34] George Nemhauser, Laurence Wolsey, and M. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14:265–294, 12 1978. 2, 4, 6

[35] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Ng. Reading digits in natural images with unsupervised feature learning. *Twenty-fifth Conference on Neural Information Processing Systems Workshop on Deep Learning and Unsupervised Feature Learning*, 01 2011. 7

[36] Cuong V Nguyen, Tal Hassner, C. Archambeau, and Matthias W. Seeger. Leep: A new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning*, 2020. 1, 3, 4

[37] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008. 7

[38] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. *Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift*. Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019. 3

[39] Michal Pándy, Andrea Agostinelli, Jasper R. R. Uijlings, Vittorio Ferrari, and Thomas Mensink. Transferability estimation using bhattacharyya class separability. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, abs/2111.12780:9162–9172, 2021. 1

[40] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505, 2012. 7

[41] 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 H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2, 7

[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. 1

[43] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *5th International Conference on Learning Representations, 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. 3

[44] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 1

[45] Jie Song, Yixin Chen, Xinchao Wang, Chengchao Shen, and Mingli Song. Deep model transferability from attribution maps. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 3

[46] Jie Song, Yixin Chen, Jingwen Ye, Xinchao Wang, Chengchao Shen, Feng Mao, and Mingli Song. Depara: Deep attribution graph for deep knowledge transferability. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3921–3929, 2020. 3

[47] Yang Tan, Yang Li, and Shao-Lun Huang. Otce: A transferability metric for cross-domain cross-task representations. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15774–15783, 2021. 1, 3, 4, 5, 7

[48] Anh Tran, Cuong Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *2019 IEEE/CVF International Conference on Computer Vision*, pages 1395–1405, 2019. 1, 3, 4, 5

[49] Matias Valdenegro-Toro. Deep sub-ensembles for fast uncertainty estimation in image classification. *Bayesian Deep Learning Workshop at Thirty-third Conference on Neural Information Processing Systems, 2019*, abs/1910.08168, 2019. 3

[50] Sebastiano Vigna. A weighted correlation index for rankings with ties. *Proceedings of the 24th International Conference on World Wide Web*, 2015. 2, 4

[51] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001*, volume 1, pages I–I. Ieee, 2001. 1

[52] Xiaofang Wang, Dan Kondratyuk, Eric Christiansen, Kris M. Kitani, Yair Movshovitz-Attias, and Elad Eban. Wisdom of committees: An overlooked approach to faster and more accurate models. In *International Conference on Learning Representations*, 2022. 2, 3

[53] Kai Wei, Rishabh K. Iyer, and Jeff A. Bilmes. Submodularity in data subset selection and active learning. In Francis R. Bach and David M. Blei, editors, *International Conference on Machine Learning*, volume 37, pages 1954–1963, 2015. 4

[54] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 7

[55] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, 08 2017. 7

[56] Yongquan Yang, Haijun Lv, and Ning Chen. A survey on ensemble learning under the era of deep learning. *Artificial Intelligence Review*, nov 2022. 3

[57] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Conference on Neural Information Processing Systems*, 2014. 2

[58] Kaichou You, Yong Liu, Mingsheng Long, and Jianmin Wang. Logme: Practical assessment of pre-trained models for transfer learning. In *International Conference on Machine Learning*, 02 2021. 1, 3, 4

[59] Amir R. Zamir, Alexander Sax, William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3712–3722, 2018. 2

[60] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 6230–6239, 2017. 1