ORIENT: Submodular Mutual Information Measures for Data Subset Selection under Distribution Shift

Anonymous Author(s)
Affiliation
Address
email

Abstract

Real-world machine-learning applications require robust models that generalize well to distribution shift settings, which is typical in real-world situations. Domain adaptation techniques aim to address this issue of distribution shift by minimizing the disparities between domains to ensure that the model trained on the source domain performs well on the target domain. Nevertheless, the existing domain adaptation methods are computationally very expensive. In this work, we aim to improve the efficiency of existing supervised domain adaptation (SDA) methods by using a subset of source data that is similar to target data for faster model training. Specifically, we propose ORIENT, a subset selection framework that uses the submodular mutual information (SMI) functions to select a source data subset similar to the target data for faster training. Additionally, we demonstrate how existing robust subset selection strategies, such as GLISTER, GRADMATCH, and CRAIG, when used with a held-out query set, fit within our proposed framework and demonstrate the connections with them. Finally, we empirically demonstrate that SDA approaches like d-SNE, CCSA, and standard Cross-entropy training, when employed together with ORIENT, achieve a) faster training and b) better performance on the target data.

1 Introduction

The recent success of deep learning frameworks in applications such as image classification [10], speech recognition [21], and object detection [14] stems primarily from the availability of large amounts of labeled data. However, due to enormous labeling costs and the need for specialists in specific domains such as medical imaging, it is not always possible to obtain vast amounts of labeled data in all situations. On the contrary, training deep models on limited amounts of data can lead to poor performance due to overfitting [2]. Consequently, where obtaining large amounts of labeled data is difficult for the target domain, closely related domain (source domain) is used to train the model. However, this may result in a deep model with suboptimal performance on target domain as a result of the distribution shift [1, 6, 5, 50], i.e., change of data distribution from source domains to target domains. A change in the data distribution often renders the features learned by the model on the source domain irrelevant for the target domain.

To address the problem of distribution shift, many domain adaptation [54, 42] and domain generalization [41] techniques have been proposed in recent years. Domain adaptation methods assume that some target domain information is available (usually target data), whereas domain generalization methods do not. Domain adaptation (DA) methods can be categorized into unsupervised [16, 13, 33, 52] (using unlabeled target data), semi-supervised [17, 59, 44, 48] (using labeled and unlabeled target data), and supervised [40, 39, 58, 38, 20] (using labeled target data) methods. UDA methods that do not require any labeled target data assume access to large volumes of unlabeled target data. When providing the same amount of target data, SDA methods are more effective than UDA methods [40]. As it is not difficult to obtain a few samples (as less as 2 examples per class) of labeled target data in

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

practice, SDA methods are an attractive approach in scenarios with limited target data. Hence, in this work we are explicitly focused on the SDA setting.

Recently, various effective SDA methods are proposed [40, 58, 38]. However, they are usually compute intensive. For example, using one of the state-of-the-art SDA method, d-SNE [58], to train a ResNet50 model [19] on the Office-Home dataset [53] with 3022 samples of the a source data and 338 samples a target data for 300 epochs takes > 18 hours using a `GTX 1080 Ti` GPU. To put that in perspective, training a ResNet50 model on the larger CIFAR100 dataset with 45000 training samples for 300 epochs on the same machine for standard cross-entropy loss takes only 14 hours. The increase in training time not only increases energy consumption and CO2 emissions but also restricts the usefulness of SDA methods in resource-constrained environments. We address this problem by seeking an answer to the following question: ***Can we improve the efficiency of SDA methods by training on subsets of source data to ensure faster adaptation and reduction in training time?***

To this end, we propose ORIENT, a subset selection framework that utilizes Submodular Mutual Information (SMI) measures [22, 18] to select subsets of source data that are similar to target data for training. Our key insight is that by training the model on data points that are similar to target data, we can speed up the training. The ORIENT framework, illustrated in Figure 1, complements existing SDA methods and can be effectively combined with any of them, enabling us to achieve a reduction in training times, energy costs, and CO2 emissions.

The key contributions of our paper include: (1) ORIENT, an SMI-based subset selection framework for efficient and effective supervised domain adaptation. (2) In Section 3, we illustrate that ORIENT unifies three existing subset selection methods: CRAIG [36], GLISTER [26], and GRAD-MATCH [25]. Specifically, these methods fit into ORIENT's framework when used with a held-out validation set from the target domain. (3) We empirically show the effectiveness of ORIENT when used with existing SDA methods like CCSA [40], d-SNE [58], and Standard Cross-entropy training on two real-world datasets: Office-31 [43] and Office-Home [53]. Our experiments also demonstrate that model learn better class discrimination features with ORIENT, resulting in better classification performance.

1.1 Related work

SDA When the labeled training data is scarce for a domain τ (or a task), a powerful way to boost performance is by pretraining a model on a related domain and fine-tuning it in the target domain [15, 60, 9]. I SDA techniques have been investigated for various distribution shift settings where the target domain ($\tau^t = \{X^t, Y^t\}$) has different marginal or conditional distributions than the source domain ($\tau^s = \{X^s, Y^t\}$). In this work, we consider *covariate shift* [1], wherein the marginal distibution is different $\big($i.e. $P(X^s) \neq P(X^t)\big)$, but the conditional distribution remains the same $\big($i.e. $P(Y^s|X^s) = P(Y^t|X^t) = P(Y|X)\big)$. A notable body of work for covariate shift SDA has focused on identifying a latent feature space that is domain-invariant [40, 39, 58, 38, 20]. This line of work train a network consisting of a feature extractor and a classifier end-to-end. The feature extractor learns a non-linear transformation of the samples from different domains to a shared latent space, and the classifier learns to assign class labels. Different researchers have proposed different optimization objectives and loss functions to encourage domain confusion and class separability.

Tzeng et al. [51] designed a domain-confusion loss to optimize for domain invariance and match the distribution over classes in the source domain to the soft label in the target domain. Motiian et al. [40] propose a classification and contrastive semantic alignment loss (CCSA). CCSA uses contrastive loss with Siamese Network to encourage the same class samples from different distributions to be nearby in the latent space (semantic alignment loss), and the different class samples from different distributions to be far apart (separation loss). Few-shot adversarial domain adaptation (FADA) [39] proposes to learn similar latent space with adversarial training. Domain-adaptation using stochastic neighborhood embedding (d-SNE) [58] proposes to maximizes the smallest distance between the samples of different classes and minimizes the largest distance between the samples of the same class while being domain invariant. Another approach, Second- or Higher-order Transfer of Knowledge (So-HoT) [30], aims to align the with-in class scatters and maintain separation of between-class scatters. Finally, Morsing et al. [38] propose domain adaptation using graph embedding (DAGE) and show that CCSA and d-SNE can also be expressed as graph embedding methods [20].

Subset Selection methods Submodular function [34, 12] is one of the effective approach of data subset selection which is used in variety of applications [7]. For example, it has been employed for efficient training in speech recognition [56, 55], machine translation [29], computer vision [23], supervised classification [36, 26, 25], semi-supervised classification [27], active learning methods [57, 47, 3, 26], and hyper-parameter optimization [28]. Another widely used method for selecting subsets is coreset construction. The coresets [11] are weighted subsets of the data that approximate the semantic characteristics of the entire dataset (e.g., loss, marginal probability, etc.). A number of recent coreset selection-based methods [37, 26, 25, 27] have demonstrated the promise to efficiently and robustly train deep models. Coresets are used for other deep learning applications like continual learning, active learning and data summarization [8, 49].

Subset selection plays a significant role in robust learning under realistic scenarios, such as imbalances or rare classes, out-of-distribution(OOD) data, or redundancy. Axelrod et al. [4] propose a simple cross-entropy based data selection methods for effective domain adaptation in statistical machine learning. Kothawade et al. [31] employed SMI measures for active learning in realistic scenarios tackling imbalance, OOD, and redundancy. PRISM [32] employed SMI measures for targeted data summarization. GLISTER [26] and GRADMATCH [25] shows that using a clean held-out validation set with SMI mitigates the label noise and class imbalance in the training set. Mirzasoleiman et al. [37] showed the data samples with clean labels cluster together, whereas the ones with noisy labels spread out in the gradient space, thus making k-medoid clustering an effective method for reducing noise in labeled data. Furthermore, Killamsetty et al. [27] tackles OOD and imbalance in the unlabeled data in semi-supervised learning setting through data subset selection. We show how these previous subset selection strategies including a modified version of CRAIG [36] were also using instantiations of SMI measures in Section 3.

2 Preliminaries

In this section, we introduce submodular functions, SMIs and SDA loss functions.

Submodular Functions: Let D be a set of n data points $D = \{1, \cdots, n\}$ and $f : 2^D \rightarrow \mathbb{R}$ be a set function returning real-value for any subset of set D. A function f is *submodular* [12] if $f(A \cup x) - f(A) \geq f(B \cup x) - f(B), \forall x \in D, \forall A \subseteq B \subseteq D$ and $x \notin B$. Monotone Submodular functions, when maximized for a cardinality constraint using a simple greedy algorithm, admit a constant factor of $1 - \frac{1}{e}$. This can be done in near-linear time using a stochastic-greedy algorithm.

Submodular Mutual Information (SMI): The SMI between two sets $A, B \subseteq D$, *instantiated* with a submodular function f, is defined as $I_f(A; B) = f(A) + f(B) - f(A \cup B)$ [22]. In this work, we aim to select data points $A \subset D$ that maximize the SMI between A and B. Intuitively, this allows selection of data points that are *similar* to B while being diverse.

SDA Loss Functions: For supervised domain adaptation, specialized loss functions are introduced as described in the previous section. Here we highlight two state-of-the-art domain adaptation loss functions, CCSA and d-SNE, which are used in our experiments.

In CCSA, the classifier is modeled as a composition of two functions–$h \circ g$. Here, $g : X \rightarrow Z$ is a feature extractor that transforms the input from the feature space X to an embedding space Z, and $h : Z \rightarrow Y$ is a predictor function. The CCSA loss for supervised domain adaptation is defined as,

$$\mathcal{L}_{CCSA}(h \circ g) = \mathcal{L}_{\text{CE}}(h \circ g) + \mathcal{L}_{SA}(g) + \mathcal{L}_{S}(g),$$

where $\mathcal{L}_{\text{CE}}(h \circ g)$ is the *cross-entropy loss* for multi-class classification, $\mathcal{L}_{SA}$ is a *semantic alignment loss* encouraging the samples from different domains but same label to map nearby in the embedding space, and $\mathcal{L}_S$ is a *separation loss* encouraging the samples from different domains and different labels to map far away in the embedding space.

The d-SNE loss function for SDA is defined as

$$\mathcal{L}_{d\text{-SNE}}(h \circ g) = \tilde{\mathcal{L}}(g) + \alpha \mathcal{L}^s_{\text{CE}}(h \circ g) + \beta \mathcal{L}^t_{\text{CE}}(h \circ g),$$

where $\mathcal{L}^s_{\text{CE}}$ and $\mathcal{L}^t_{\text{CE}}$ are the cross-entropy loss on the source and the target domain examples, respectively, and $\tilde{\mathcal{L}}$ minimizes the largest distance between the samples from different domains with same label and maximizes the smallest distance between the samples from different domain with different label in the embedding space. SDA loss functions are further elaborated in Section A.4.

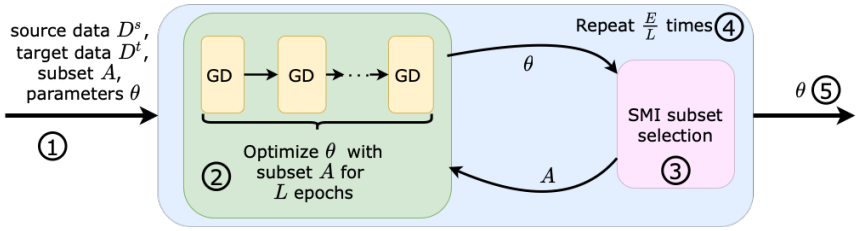


Figure 1: Illustration of ORIENT framework. **(1)** Given the target data D^t and source data D^s, a subset $A \subseteq D^s$ and model parameters θ are randomly initialized. **(2)** Model parameters θ are optimized for L epochs on the subset A with any gradient descent (GD) based method. **(3)** Subset A is updated using the SMI measure and current model parameters θ after every L^{th} epoch. **(4)** Model is trained for E epochs, with subset selection after every L interval. **(5)** The final model parameters are returned after E epochs.

Name	$f(A)$	$I_f(A; D^t)$
FLMI	$\sum_{i \in D^t} \max_{j \in A} S_{ij}$	$\sum_{i \in D^t} \max_{j \in A} S_{ij} + \eta \sum_{i \in A} \max_{j \in D^t} S_{ij}$
LOGDETMI	$\log\det(S_A)$	$\log\det(S_A) - \log\det(S_A - \eta^2 S_{A,D^t} S_{D^t}^{-1} S_{A,D^t}^T)$
GCMI	$\sum_{i \in A, j \in D^s} S_{ij} - \lambda \sum_{i,j \in A} S_{ij}$	$2\lambda \sum_{i \in A, j \in D^t} S_{ij}$
COM	Appendix Eq. 3	$\eta \sum_{i \in A} \psi(\sum_{j \in D^t} S_{ij}) + \sum_{j \in D^t} \psi(\sum_{i \in A} S_{ij})$

Table 1: Instantiations of submodular mutual information functions [32].

3 ORIENT

For our SDA setting, we assume a small amount of labeled data (as less as 2 examples per class) is available from the target domain τ^t and sufficient labeled data is available from the source domain τ^s. We denote the source dataset as $D^s = (x_i, y_i)_{i=1}^{|D^S|}$ and the target dataset as $D^t = (x_i, y_i)_{i=1}^{|D^t|}$. The source dataset is used for training the model $\mathcal{M}$ using an SDA loss function $\mathcal{L}$. We will introduce other necessary notations in the remainder of the paper, as necessary. A complete list of notations is presented in Appendix (A.1).

First, we extend the SMI measure to incorporate data subsets from different domains. For a source domain subset $A \subseteq D^s$ and the target domain data set D^t, we denote SMI measure as $I_f(A; D^t)$. Essentially, assuming that both the domains form a set $\mathcal{D}$, and $A \subseteq D^s \subset \mathcal{D}$ and $D^t \subset \mathcal{D}$. With S denoting the similarity matrix defined over a subset A and D^t, Table 1 presents the mathematical expression of the SMI functions for different instantiations of submodular functions (f) [32]. We defer the readers to Kothawade et al. [32] for more details. Note that the FLMI, GCMI and COM only need the pairwise similarities of the data points in A and D^t. Consequently, the similarity kernel is of size $|A| \times |D^t|$, making it very efficient to optimize.

For efficient SDA, we want to select a subset A of the source domain τ^s such that training a model on A results in a classifier that is proficient on the target domain τ^t. To this effect, we use the gradients $\boldsymbol{\chi}$ of the current model to represent the data points and use it to compute the similarity matrix. Specifically, we define the pairwise similarity between two data points as the cosine similarity between the gradients,

$$S_{ij} = \frac{\chi_i \, \chi_i}{|\chi_i| \, |\chi_j|} \tag{1}$$

$$\text{where, } \chi_i = \nabla_\theta \mathcal{L}(x_i, y_i) \text{ and } \chi_j = \nabla_\theta \mathcal{L}(x_j, y_j)$$

Given a set size b, we select a *diverse* subset A of the source data D^s that is *similar* to the target data D^t by maximizing the following SMI measure,

$$\underset{A \subseteq D^s, |A| \leq b}{\arg\max} \; I_f(A; D^t) \tag{2}$$

$I_f(A; D^t)$ is an instance of cardinality constrained monotone submodular maximization for all SMI functions except for LogDetMI. Therefore, we can achieve a $1 - \frac{1}{e}$ approximation using the lazy greedy algorithm [35]. Even though LogDetMI is not submodular, previous works [32, 31] have reported good empirical performance using the lazy greedy algorithm for maximization of the LogDetMI function. Following the footsteps of Kothawade et al. [32, 31], we also use the lazy greedy algorithm to maximize the LogDetMI function. However, as the number of parameters in the deep learning model can be extremely high, our gradients χ can be very high dimensional. Following the success of targeted subset selection approaches, GLISTER [26], SIMILAR [31], CRAIG [37] and BADGE [3], we circumvent this problem by using last-layer gradient approximations to represent the data point in the similarity matrix S. Further, we select a new subset every L epochs as the model is trained, and the subsets chosen are adapted accordingly.

Algorithm: We present the complete training procedure of ORIENT in Algorithm 1. We first randomly initialize the training subset A and the model parameters θ in **line 1** and **2**, respectively. For each epoch, we update the model parameters by optimizing a gradient-descent (GD) based loss $\mathcal{L}$ on the current training subset A in **line 4**. After a fixed interval of L epochs, we update the subset A (**line 5–11**). To do this, we compute the gradients $\boldsymbol{\chi}^s$ and $\boldsymbol{\chi}^t$ for all the datapoints in D^s and D^t, in **line 6** and **7**,respectively. Next, we use the gradients to compute the similarity matrix S in **line 8**, as described in Equation 1. In **line 9** we instantiate the SMI function I_f with S and update the subset A in **line 10** using Equation 2. Our implementation is made available online[1].

Algorithm 1 ORIENT

Input: source domain data D^s, query data D^t, batch size b, total epochs E, subset selection interval L, SMI function I_f.

Output: Final model parameters θ

1: $A \leftarrow \text{RandomSubset}(D^s)$ ▷ *Randomly initialize a subset*
2: θ ▷ *initial model parameters*
3: **for** epoch e in E **do** ▷ *for each epoch*
4: $\quad\theta \leftarrow \text{mini-batch-gd}(\theta, \mathcal{L}(A))$ ▷ *mini-batch gradient descent*
5: $\quad$**if** $e \mod L == 0$ **then** ▷ *perform subset selection every L epochs*
6: $\quad\quad\boldsymbol{\chi}^s \leftarrow \nabla_\theta \mathcal{L}(D^s)$ ▷ *compute gradients for source data*
7: $\quad\quad\boldsymbol{\chi}^t \leftarrow \nabla_\theta \mathcal{L}(D^t)$ ▷ *compute gradients for query data*
8: $\quad\quad S \leftarrow \textsc{Similarity}(\boldsymbol{\chi}^s, \boldsymbol{\chi}^t)$ ▷ *compute the similarity matrix*
9: $\quad\quad$Instantiate I_f with S
10: $\quad\quad A \leftarrow \underset{A \subseteq D^s, |A| \leq b}{\arg\max} \; I_f(A; D^t)$ ▷ *update subset A*

Figure 1 illustrates the ORIENT workflow. The general framework of ORIENT can incorporate any gradient decent (GD) based SDA technique. It can train any model with loss function $\mathcal{L}$ by using a subset of the source data, selected every L epochs. In our experiments, we demonstrate the effectiveness of our method in conjunction with standard cross-entropy loss as well as two state-of-the-art domain adaptation losses, CCSA [40] and d-SNE [58].

Connections to Previous Work

ORIENT generalizes three approaches, GLISTER [26], GRADMATCH [25], and a slightly adapted version of CRAIG [36], which were initially proposed for efficient and robust subset selection. Specifically, these subset selection approaches maximize a submodular function corresponding to a different instantiations of SMI functions for subset selection. Following theorems summarize the connection of SMI functions with these subset selection approaches.

Theorem 1 *When the outer level loss of the discrete bi-level optimization problem of* GLISTER *is hinge loss, logistic loss, and perceptron loss, then the optimization problem becomes an instance of Concave over Modular (COM) SMI function.*

Theorem 2 *When the optimization problem of* GRADMATCH *is used to match gradients of a held-out validation set, it becomes an instance of the summation of GCMI and a diversity function.*

[1]Anonymized link: https://anonymous.4open.science/r/6845_Orient

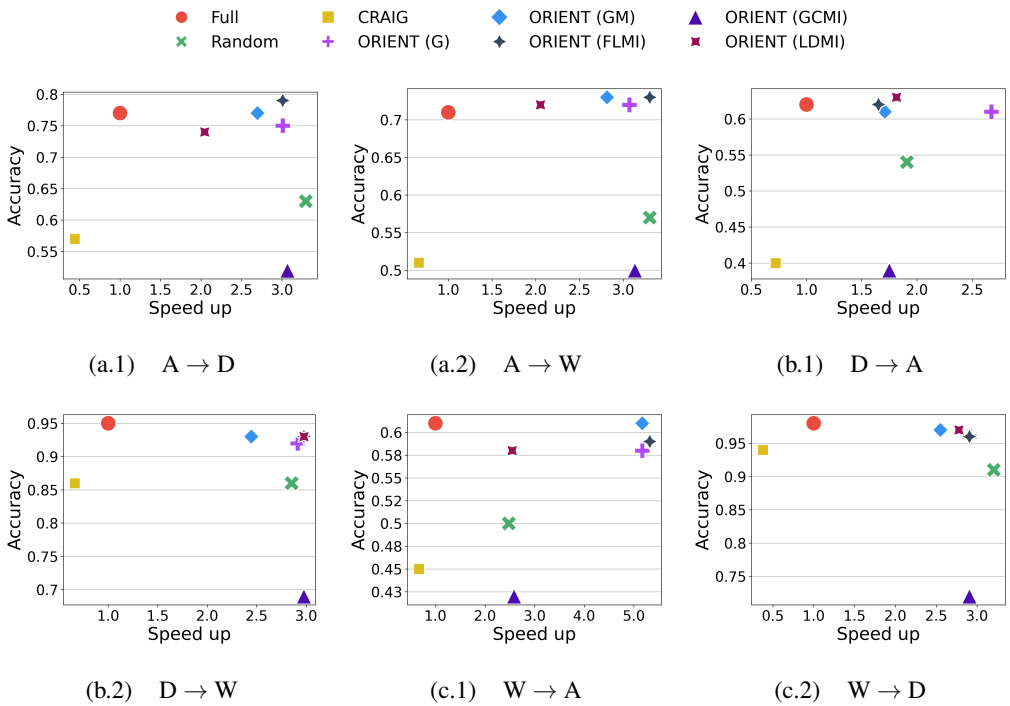


Figure 2: Speed up vs prediction accuracy on three domains of Office31 dataset: Amazon (A), DSLR (D), and Webcam (W). X-axis represents the speed up by the model, i.e. the ratio of the time taken to train on complete source dataset (Full) to the time taken by the model. Y-axis represents the prediction accuracy of the model on the target domain.

Theorem 3 *When the optimization problem of* CRAIG *is adapted to match gradients of a held-out validation set, it becomes an instance of the FLMI function with* $\eta = 0$.

We present the proofs of these theorems in Appendix (A.3). In conclusion, SMI measures have been used in previous subset selection strategies for robust learning to deal with the class imbalance and noisy labels in training data sets and achieved comparable performance to current state-of-the-art robust supervised learning approaches.

4 Experimental Evaluation

Our experiments explicitly aim at answering the following questions,

Q1: Does the proposed ORIENT approach substantially reduce the training time while maintaining comparable performance to training on complete source dataset?

Q2: Can the proposed ORIENT approach augment the existing domain adaptation approaches?

We evaluate our proposed approach on two domain adaptation datasets: Office-31 [43] and Office-Home [53]. Office-31 dataset consists of 4110 images of 31 object categories from three different domains: Amazon (A), DSLR (D), and Webcam (W). For our experiments we used all the images from the source domain in the training set (D^s) and two examples of each category in the query set (D^t). Office-Home dataset consists of 10812 images of 65 object categories from four different domains: Art (A), Clipart (C), Product (P), and Real-world (R). Here, the target data (D^t) consists of about 20% of the images from the target domain. We use $L = 20$, that is, we sample the subset after every 20 epochs and use $b = 0.3\%$, that is, we sample 30% of the source domain for training.

To answer the first question, we use ORIENT to train a ResNet50 architecture using the stochastic gradient descent (SGD) algorithm with the momentum of 0.9 and the weight decay ratio of $5e4$. We compare our approach against three baselines: **Full**—training on the source data and the target data,

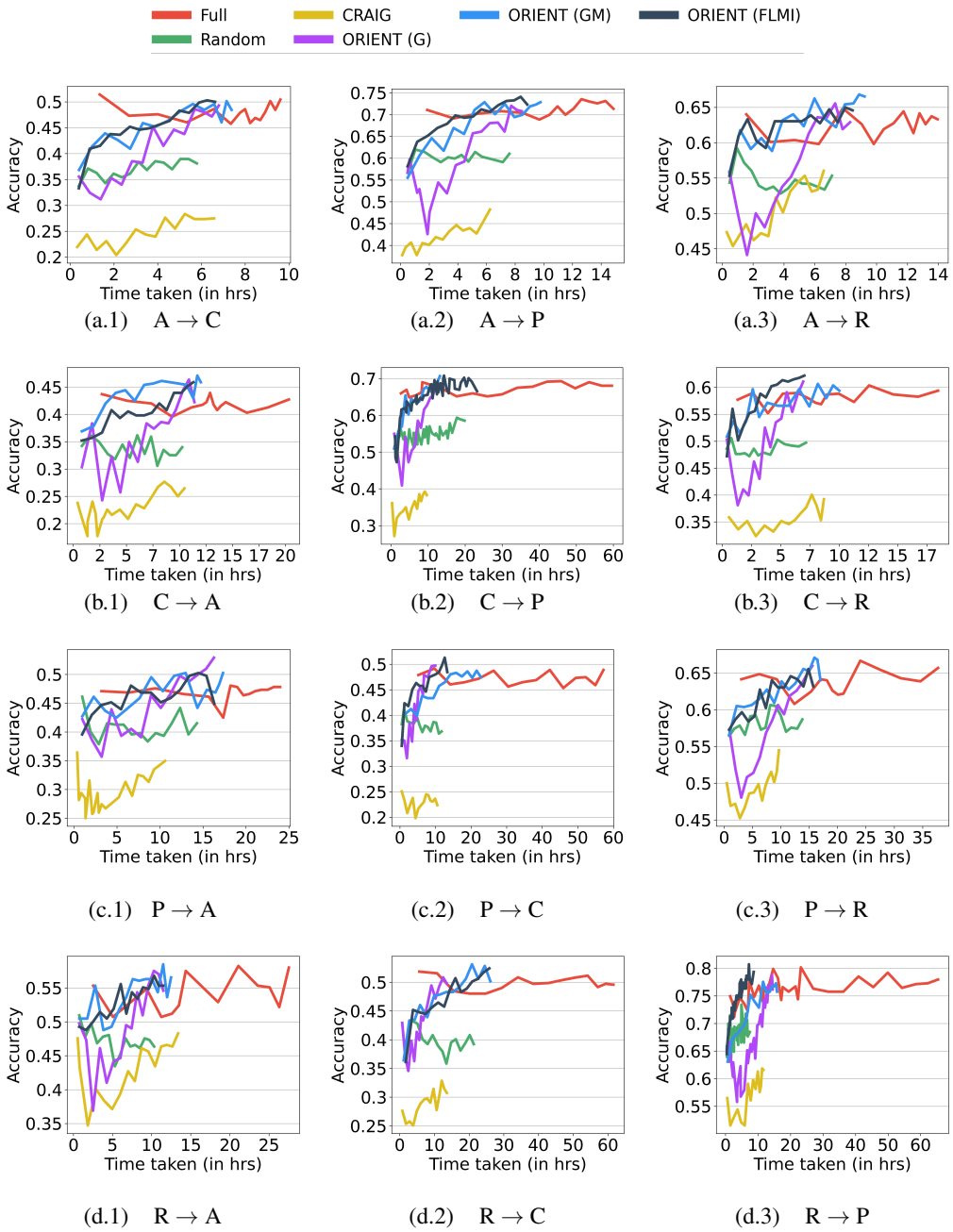


Figure 3: Convergence curves on four domains of Office-Home dataset: Art (A), Clipart (C), Product (P), and Real World (R). X-axis presents the training time in hours and Y-axis presents the prediction accuracy on the target domain.

i.e. $D^s \cup D^t$; **Random**—a random subset of the source dataset and the target set, i.e. $R \cup D^t, R = \text{RandomSubset}(D^s)$; and **CRAIG**—a coreset of the source dataset, i.e $\text{coreset}(D^t)$, without any information of the target dataset. We use standard categorical cross-entropy loss function (for multi-class classification) and five different instantiations of the submodular mutual information function: Facility Location Mutual Information (**ORIENT (FLMI)**), Graph Cut Mutual Information (**ORIENT (GCMI)**), Log Determinant Mutual Information (**ORIENT (LDMI)**), GLISTER (**ORIENT (G)**), and GradMatch (**ORIENT (GM)**).

		A → D	A → W	D → A	D → W	W → A	W → D
CCSA	Full	0.78	0.72	0.55	**0.93**	0.55	**0.97**
	Random	0.76	0.72	0.54	0.81	0.54	0.92
	ORIENT	0.77	**0.76**	0.55	0.89	0.55	**0.96**
d-SNE	Full	0.77	0.69	0.53	**0.93**	0.54	**0.98**
	Random	0.76	0.68	0.53	0.86	0.53	0.94
	ORIENT	0.78	0.71	**0.55**	0.90	0.56	**0.97**

Table 2: Test accuracy for office-31 with SDA methods

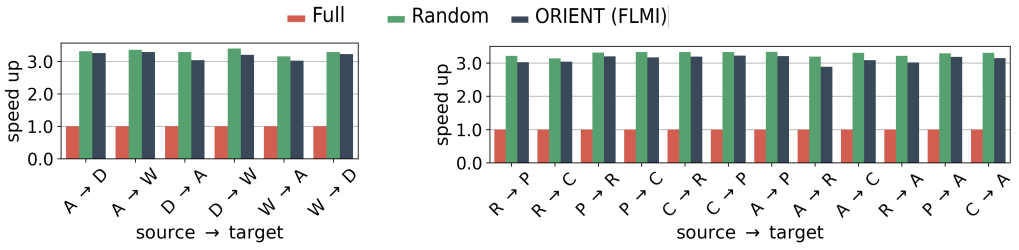
		R → P	R → C	P → R	P → C	C → R	C → P	A → P	A → R	A → C	R → A	P → A	C → A
CCSA	Full	0.74	**0.55**	0.62	0.46	0.56	0.67	0.70	**0.64**	0.48	0.57	**0.49**	0.41
	Random	0.71	0.47	0.62	0.45	0.54	0.66	0.69	0.57	0.48	0.5	0.47	0.45
	ORIENT	**0.78**	**0.54**	**0.65**	**0.5**	**0.61**	**0.71**	0.71	**0.65**	**0.51**	**0.59**	**0.54**	**0.47**
d-SNE	Full	0.77	**0.53**	**0.62**	**0.50**	**0.60**	0.71	**0.72**	**0.63**	0.49	**0.52**	0.44	0.40
	Random	0.75	0.50	0.60	0.45	0.57	0.69	0.68	0.59	0.49	0.46	0.43	0.40
	ORIENT	0.77	**0.52**	**0.63**	**0.50**	**0.60**	0.71	**0.71**	0.61	0.51	**0.52**	0.44	0.42

Table 3: Test accuracy for Office-Home with SDA methods

Figure 2 presents the scatter plot of the speed up against the prediction accuracy of Office-31 dataset using cross-entropy loss.. In all our experiments, we use 0.3 fraction of the source data as the subset and see $> 2.5\times$ speed up. Figure 3 presents the convergence curves of the Office-Home dataset using cross-entropy loss. It is evident from these charts that our proposed approach ORIENT can achieve better or comparable performance to the Full training in significantly less amount of time, indicating better efficiency. This analysis helps us answer our Q1. ORIENT reduces the training time substantially, the speed up achieved is reciprocal to the fraction of the subset used, without trading the performance. Rather, in some combinations, we observe that ORIENT outperforms Full.

Figure 4 presents the bar plots of the speed up when ORIENT(FLMI) is used in conjunction with d-SNE loss function on Office-31 and Office-Home datasets. Here, we see a consistent $3\times$ speed up as compared to full model training with d-SNE loss. These bar plots demonstrate that our proposed approach ORIENT can augment existing domain adaptation approaches and substantially reduce the training times. Additionally, Tables 2 and 3 present the test accuracy while using d-SNE and CCSA loss in conjunction with ORIENT(FLMI) on Office-31 and Office-Home datasets, respectively. It's evident that augmenting existing domain adaptation approaches with ORIENT achieves better or comparable performance to the Full training. These two observations help us answer our Q2 in the affirmative. Precise values of prediction accuracy and training time for all the experiments are provided in Appendix (A.5). Further, in Appendix Appendix (A.7) and Appendix (A.8), we analyze the effect of different subset sizes and subset sampling frequencies.

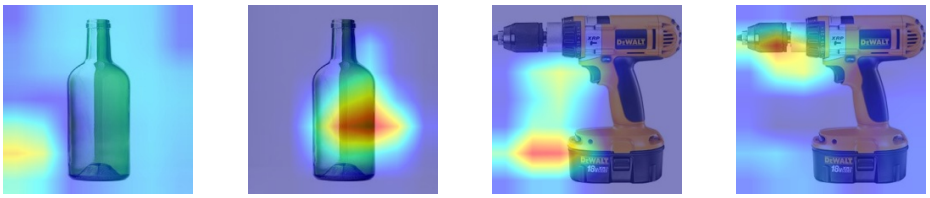
Figure 5 presents the GradCam [46] class-activation maps of trained models on the Office-Home dataset ($P \rightarrow R$ setting) using d-SNE loss for both Full and ORIENT (FLMI). These activation maps show that the model trained with the ORIENT framework learned better class discrimination features



(a) Office-31 dataset

(b) Office-Home dataset

Figure 4: Speed up achieved by combining d-SNE with ORIENT



(a) d-SNE + Full (b) d-SNE + ORIENT (FLMI) (c) d-SNE + Full (d) d-SNE + ORIENT (FLMI)

Figure 5: GradCam [46] activation maps of the models learned using d-SNE + Full and d-SNE + ORIENT (FLMI) on the Office-Home dataset with "Product" as the source domain and "Real World" as the target domain. As evidenced by class activation maps, the ORIENT framework enabled the model to learn more effective class discriminative features than Full data training.

than the model trained with Full. This might explain why the ORIENT framework performs better sometimes than Full.

Comparison of different SMI functions: Domain adaptation results on Office31 (Figure 2) and OfficeHome (Figure 3) datasets show that ORIENT(GCMI) performs suboptimally compared to ORIENT using other SMI functions in terms of target domain accuracy. Although ORIENT(LDMI) achieves reasonable target domain accuracy, it is computationally expensive and does not achieve the best performance-speedup trade-off. In comparison to the ORIENT using remaining SMI functions, i.e., ORIENT(GM), ORIENT(G), and ORIENT(FLMI), ORIENT(FLMI) consistently achieves the best performance versus speed-up trade-off. We further present synthetic experiments to provide intuitions on how different SMI functions selects data subsets in Appendix A.6

To summarize, our experiments on two adaptation datasets suggest that our proposed approach ORIENT can substantially reduce the training time while maintaining comparable performance to training on complete source data. Additionally, our experiments using d-SNE and CCSA loss functions suggest that ORIENT can be used in conjunction with existing SDA methods to achieve significant speed ups in training time, while maintaining or improving the prediction accuracy. In particular, two instantiations of our proposed approach, ORIENT(FLMI) and ORIENT(GM) consistently achieve comparable or better performance compared to Full training while being $\sim 3\times$ faster to train.

5 Conclusion, Limitations, and Broader Impact

We introduce ORIENT, a subset selection framework based on SMI functions for supervised domain adaptation. The submodularity of SMI functions allows us to use scalable greedy algorithms to select the data subsets efficiently. In addition, we demonstrate how ORIENT is a unified framework that integrates previous approaches based on subset selection for robust learning. Empirically, we show that ORIENT is very effective for SDA. Specifically, it achieves $\sim 3\times$ speed up over existing SDA approaches like d-SNE and CCSA while achieving comparable or better performance. Our findings confirm that ORIENT has a significant social impact by making existing SDA algorithms significantly faster and more energy-efficient, reducing CO2 emissions and energy consumption incurred during training, thus contributing to a Green AI [45]. The main limitation of our work is that although ORIENT significantly reduces the training time, it requires more memory to store the similarity kernel required for subset selection. This makes running ORIENT harder on devices with low memory.

References

[1] Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. ISSN 0378-3758.

[2] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. C. Courville, Y. Bengio, and S. Lacoste-Julien. A closer look at memorization in deep networks. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 233–242. PMLR, 2017.

[3] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *ICLR*, 2020.

[4] A. Axelrod, X. He, and J. Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL `https://aclanthology.org/D11-1033`.

[5] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *NIPS*, pages 137–144. MIT Press, 2006.

[6] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

[7] J. Bilmes. Submodularity in machine learning and artificial intelligence, 2022. URL `https://arxiv.org/abs/2202.00132`.

[8] Z. Borsos, M. Mutný, M. Tagliasacchi, and A. Krause. Data summarization via bilevel optimization, 2021. URL `https://arxiv.org/abs/2109.12534`.

[9] B. Chu, V. Madhavan, O. Beijbom, J. Hoffman, and T. Darrell. Best practices for fine-tuning visual classifiers to new domains. In *ECCV Workshops (3)*, volume 9915 of *Lecture Notes in Computer Science*, pages 435–442, 2016.

[10] D. C. Ciresan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In *CVPR*, pages 3642–3649. IEEE Computer Society, 2012.

[11] D. Feldman. Core-sets: Updated survey. In *Sampling Techniques for Supervised or Unsupervised Tasks*, pages 23–44. Springer, 2020.

[12] S. Fujishige. *Submodular functions and optimization*. Elsevier, 2005.

[13] Y. Ganin and V. S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1180–1189. JMLR.org, 2015.

[14] R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann. Generalisation in humans and deep neural networks. In *NeurIPS*, pages 7549–7561, 2018.

[15] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587. IEEE Computer Society, 2014.

[16] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, pages 2066–2073. IEEE Computer Society, 2012.

[17] Y. Guo and M. Xiao. Cross language text classification via subspace co-regularized multi-view learning. In *ICML*. icml.cc / Omnipress, 2012.

[18] A. Gupta and R. Levin. The online submodular cover problem. In *SODA*, pages 1525–1537. SIAM, 2020.

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016.

[20] L. Hedegaard, O. A. Sheikh-Omar, and A. Iosifidis. Supervised domain adaptation: A graph embedding perspective and a rectified experimental protocol. *IEEE Trans. Image Process.*, 30: 8619–8631, 2021.

[21] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson. Super-human multi-talker speech recognition: A graphical modeling approach. *Comput. Speech Lang.*, 24(1):45–66, 2010.

[22] R. K. Iyer, N. Khargonkar, J. A. Bilmes, and H. Asnani. Generalized submodular information measures: Theoretical properties, examples, optimization algorithms, and applications. *IEEE Trans. Inf. Theory*, 68(2):752–781, 2022.

[23] V. Kaushal, R. K. Iyer, S. Kothawade, R. Mahadev, K. Doctor, and G. Ramakrishnan. Learning from less data: A unified data subset selection and active learning framework for computer vision. In *WACV*, pages 1289–1299. IEEE, 2019.

[24] K. Killamsetty, D. Bhat, G. Ramakrishnan, and R. Iyer. CORDS: COResets and Data Subset selection for Efficient Learning, March 2021. URL `https://github.com/decile-team/cords`.

[25] K. Killamsetty, D. Sivasubramanian, G. Ramakrishnan, A. De, and R. K. Iyer. GRAD-MATCH: gradient matching based data subset selection for efficient deep model training. In *ICML*, volume 139, pages 5464–5474. PMLR, 2021.

[26] K. Killamsetty, D. Sivasubramanian, G. Ramakrishnan, and R. K. Iyer. GLISTER: generalization based data subset selection for efficient and robust learning. In *AAAI*, pages 8110–8118. AAAI Press, 2021.

[27] K. Killamsetty, X. Zhao, F. Chen, and R. K. Iyer. RETRIEVE: coreset selection for efficient and robust semi-supervised learning. In *NeurIPS*, pages 14488–14501, 2021.

[28] K. Killamsetty, G. S. Abhishek, Aakriti, A. V. Evfimievski, L. Popa, G. Ramakrishnan, and R. Iyer. AUTOMATA: Gradient based data subset selection for compute-efficient hyper-parameter tuning, 2022. URL `https://arxiv.org/abs/2203.08212`.

[29] K. Kirchhoff and J. A. Bilmes. Submodularity for data selection in machine translation. In *EMNLP*, pages 131–141. ACL, 2014.

[30] P. Koniusz, Y. Tas, and F. Porikli. Domain adaptation by mixture of alignments of second-or higher-order scatter tensors. In *CVPR*, pages 7139–7148. IEEE Computer Society, 2017.

[31] S. Kothawade, N. Beck, K. Killamsetty, and R. K. Iyer. SIMILAR: submodular information measures based active learning in realistic scenarios. In *NeurIPS*, pages 18685–18697, 2021.

[32] S. Kothawade, V. Kaushal, G. Ramakrishnan, J. A. Bilmes, and R. K. Iyer. PRISM: a rich class of parameterized submodular information measures for guided subset selection. *AAAI*, 36, 2022.

[33] M. Liu and O. Tuzel. Coupled generative adversarial networks. In *NeurIPS*, pages 469–477, 2016.

[34] L. Lovász. Submodular functions and convexity. In *ISMP*, pages 235–257. Springer, 1982.

[35] M. Minoux. Accelerated greedy algorithms for maximizing submodular set functions. In *Optimization techniques*, pages 234–243. Springer, 1978.

[36] B. Mirzasoleiman, J. A. Bilmes, and J. Leskovec. Coresets for data-efficient training of machine learning models. In *ICML*, volume 119, pages 6950–6960. PMLR, 2020.

[37] B. Mirzasoleiman, K. Cao, and J. Leskovec. Coresets for robust training of neural networks against noisy labels. *Advances in Neural Information Processing Systems*, 33, 2020.

[38] L. H. Morsing, O. A. Sheikh-Omar, and A. Iosifidis. Supervised domain adaptation using graph embedding. In *ICPR*, pages 7841–7847. IEEE, 2020.

[39] S. Motiian, Q. Jones, S. M. Iranmanesh, and G. Doretto. Few-shot adversarial domain adaptation. In *NIPS*, pages 6670–6680, 2017.

[40] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto. Unified deep supervised domain adaptation and generalization. In *ICCV*, pages 5716–5726. IEEE Computer Society, 2017.

[41] K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *ICML (1)*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 10–18. JMLR.org, 2013.

[42] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Process. Mag.*, 32(3):53–69, 2015.

[43] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV (4)*, volume 6314 of *Lecture Notes in Computer Science*, pages 213–226. Springer, 2010.

[44] K. Saito, D. Kim, S. Sclaroff, T. Darrell, and K. Saenko. Semi-supervised domain adaptation via minimax entropy. *ICCV*, 2019.

[45] R. Schwartz, J. Dodge, N. Smith, and O. Etzioni. Green ai. *Communications of the ACM*, 63:54 – 63, 2020.

[46] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74.

[47] O. Sener and S. Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*. OpenReview.net, 2018.

[48] A. Singh. CLDA: contrastive learning for semi-supervised domain adaptation. In *NeurIPS*, pages 5089–5101, 2021.

[49] R. Tiwari, K. Killamsetty, R. Iyer, and P. Shenoy. GCR: Gradient coreset based replay buffer selection for continual learning, 2021.

[50] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, 2011. doi: 10.1109/CVPR.2011.5995347.

[51] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *ICCV*, pages 4068–4076. IEEE Computer Society, 2015.

[52] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *CVPR*, pages 2962–2971. IEEE Computer Society, 2017.

[53] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5385–5394. IEEE Computer Society, 2017.

[54] M. Wang and W. Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312: 135–153, 2018.

[55] K. Wei, Y. Liu, K. Kirchhoff, C. D. Bartels, and J. A. Bilmes. Submodular subset selection for large-scale speech training data. In *ICASSP*, pages 3311–3315. IEEE, 2014.

[56] K. Wei, Y. Liu, K. Kirchhoff, and J. A. Bilmes. Unsupervised submodular subset selection for speech data. In *ICASSP*, pages 4107–4111. IEEE, 2014.

[57] K. Wei, R. K. Iyer, and J. A. Bilmes. Submodularity in data subset selection and active learning. In *ICML*, volume 37, pages 1954–1963. JMLR.org, 2015.

[58] X. Xu, X. Zhou, R. Venkatesan, G. Swaminathan, and O. Majumder. d-sne: Domain adaptation using stochastic neighborhood embedding. In *CVPR*, pages 2497–2506. Computer Vision Foundation / IEEE, 2019.

[59] T. Yao, Y. Pan, C. Ngo, H. Li, and T. Mei. Semi-supervised domain adaptation with subspace learning for visual recognition. In *CVPR*, pages 2142–2150. IEEE Computer Society, 2015.

[60] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *NeurIPS*, pages 3320–3328, 2014.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 3 for theoretical results and Section 4 for empirical evaluation of our claims.
   (b) Did you describe the limitations of your work? [Yes] We discuss limitations of our proposed approach in Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We discuss the potential societal impact of our approach in Section 5.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We've ensured that our paper conforms to the guidelines to the best of our knowledge.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Sections 3 for theoretical results and A.3 for details
   (b) Did you include complete proofs of all theoretical results? [Yes] See A.3.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See C.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Please refer to Section 4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Please refer to A.5 for precise results for all our experiments.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Please refer to Section A.5
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Please refer to 4 dataset and model architecture citations.
   (b) Did you mention the license of the assets? [Yes] Please refer to Section D.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Please refer to Section C for our code.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] Please refer to Section D for licenses for the datasets used.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] Our data does not use personally identifiable information or offensive content. Please refer to Section D and Section A.5 for details regarding the data used.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] Our work does not use research involving human subjects.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] Our work does not use research involving human subjects.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] Our work does not use research involving human subjects.