Generalization Bounds for Deep Transfer Learning Using Majority Predictor Accuracy

Cuong N. Nguyen*, Lam Si Tung Ho†, Vu Dinh‡, Tal Hassner§, and Cuong V. Nguyen*
*Florida International University, USA, {cnguy049, vcnguyen}@cs.fiu.edu
†Dalhousie University, Canada, lam.ho@dal.ca
‡University of Delaware, USA, vucdinh@udel.edu
§Facebook AI Research, USA, talhassner@gmail.com

***Abstract*—We analyze new generalization bounds for deep learning models trained by transfer learning from a source to a target task. Our bounds utilize a quantity called the majority predictor accuracy, which can be computed efficiently from data. We show that our theory is useful in practice since it implies that the majority predictor accuracy can be used as a transferability measure, a fact that is also validated by our experiments.**

I. Introduction

Deep transfer learning, the problem of transferring representations (or features) learned by deep neural networks from one task to another, has become a crucial part for training deep learning models in practice [10], [13], [18]. Despite this fact, the current literature still lacks a theory for understanding the generalization of models obtained by deep transfer learning. In this paper, we close this gap between the theory and practice of deep transfer learning by proving novel generalization bounds for models learned through such transfer learning methods.

To prove the bounds, we develop the *Majority Predictor Accuracy* (MPA), a simple and easy-to-compute quantity defined as the accuracy of the classifier that returns the most probable target label conditioned on a given source label. Using the MPA, we can show that when the source and target data share the input set, the true risk of the transferred model is upper bounded by the sum of the empirical risk of the source model, $1 - \text{MPA}$, and an $\widetilde{\mathcal{O}}(1/\sqrt{n})$ sample complexity with high probability. We further extend this result to the more general setting where the source and target datasets contain different inputs. This extension is achieved by using dummy source labels, a technique previously developed for transferability estimation [12].

We also demonstrate the usefulness of our theoretical bounds in practice by showing empirically that the MPA can be used as a transferability measure, defined as a numeric score that can tell whether deep transfer learning would be effective when transferring between a given pair of source-target tasks. Specifically, our experiments on the large-scale CUB-200 dataset [17] show that the MPA scores are highly correlated with the actual accuracies of the transferred models with statistical significance, thus indicating that the MPA is a good measure of transferability.

To summarize, our paper makes the following contributions: (1) developing the new MPA score, (2) proving novel deep transfer learning bounds using the MPA, and (3) showing our bounds are practically useful through experiments.

Related Work. Transfer learning [10], [19] is a long-standing research area of machine learning. Several previous work has provided theoretical analysis and generalization bounds for transfer learning, especially under the domain adaptation setting, such as [1], [4]–[6], [11]; however, these results were not explicitly developed for deep learning and the settings commonly used in practice, where a learned representation is adapted to the new domain [13], [18]. Our paper, on the other hand, provides generalization bounds explicitly for these commonly used deep transfer learning settings. Furthermore, unlike these previous work, our bounds are useful in practice for understanding the transferability between different tasks, as demonstrated in our experiments.

Our work is also related to a recent attempt to develop transferability measures for deep transfer learning [2], [8], [12], [14]–[16], [19]. Transferability measures aim to estimate the effectiveness of deep transfer learning between tasks and have been used for model or task selection [2], [12], [16], [19], checkpoint ranking [8], and few-shot learning [15]. Although some theoretical properties were shown for these transferability measures [12], [15], [16], they only focused on the empirical risk instead of the true risk as in our paper.

II. Deep Transfer Learning: Formal Setting

Deep transfer learning [10] refers to the problem of transferring a learned deep neural network representation from a source task to a target task. In this section, we formalize the deep transfer learning setting considered in our paper. This setting is commonly used in practice for several large-scale deep learning models [13], [18].

Let $\mathcal{S} = \{(x_1, s_1), (x_2, s_2), \dots, (x_n, s_n)\}$ be a train dataset for a source classification task where each input-label pair $(x_i, s_i) \in \mathbb{R}^d \times [m_S]$ is drawn iid from a joint distribution $\mathbb{P}_{X,S}$, with $[n] = \{1, 2, \dots, n\}$ for any positive integer n. Consider a target classification task with a train set $\mathcal{T} = \{(x_1, t_1), (x_2, t_2), \dots, (x_n, t_n)\}$ where each target example $(x_i, t_i) \in \mathbb{R}^d \times [m_T]$ is drawn iid from $\mathbb{P}_{X,T}$. We will first consider this simple case where the source and target datasets share the same inputs $\{x_1, x_2, \dots, x_n\}$, with each x_i being a d-dimensional vector having the same marginal distribution $\mathbb{P}_X$ in both tasks. Here the source task has m_S

classes and the target task has m_T classes. The more general case with different input sets will be discussed in Section V.

In our deep transfer learning setting, we first train a source model $h \circ w$ using $\mathcal{S}$, where $w(x) \in \mathbb{R}^r$ is the r-dimensional representation (also called embedding or feature vector) of the input x extracted from the network w, and $h \circ w(x) = h(w(x)) \in [m_S]$ is the source label returned by the network h with the representation $w(x)$ as input. The functions w and h are usually called the *feature extractor* and the *head classifier* respectively. In deep learning, the whole model $h \circ w$ is a deep neural network with w being its parameters from the input up to some layer L, and h being the network parameters from layer L to the output. We obtain the optimal source model on $\mathcal{S}$ by minimizing the empirical risk:[1]

$$w^*, h^* = \underset{w,h \in \Omega_w \times \Omega_h}{\arg\min} \; \widehat{R}_{\mathcal{S}}(w, h), \tag{1}$$

where Ω_w and Ω_h are the spaces of all possible w's and h's respectively, and with $\mathbf{1}[\cdot]$ being the indicator function,

$$\widehat{R}_{\mathcal{S}}(w, h) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[s_i \neq h \circ w(x_i)]. \tag{2}$$

In practice, the optimal feature extractor w^* often learns generic feature representations (e.g., edges or shapes in images) that can be reused for several tasks, while the optimal head classifier h^* is often specialized for a particular source task. To transfer this trained model $h^* \circ w^*$ to a target task, the usual practice is to discard h^* and reuse w^* for the target task. Specifically, we will re-train a new head classifier k^* on the target dataset $\mathcal{T}$ using the features extracted from w^*.

In this paper, we allow the target head classifier to return real-valued scores for all target labels. That means for each example x, a head classifier k on the target task would take $w^*(x)$ as input and return $k \circ w^*(x) = k(w^*(x)) \in \mathbb{R}^{m_T}$, the scores (before softmax) for all target labels. We will consider the optimal target head classifier k^* obtained by minimizing the empirical risk with a given margin $\gamma \geq 0$:

$$k^* = \underset{k \in \Omega_k}{\arg\min} \, \widehat{R}_{\mathcal{T},\gamma}(w^*, k), \tag{3}$$

where Ω_k is the space of all k's, and for all $w \in \Omega_w$, $k \in \Omega_k$:

$$\widehat{R}_{\mathcal{T},\gamma}(w, k) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[k \circ w(x_i)_{t_i} < \gamma + \max_{t \neq t_i} k \circ w(x_i)_t],$$

with $k \circ w(x_i)_t$ being the t-th element of the vector $k \circ w(x_i)$. Here $\widehat{R}_{\bullet,\gamma}$ is a general version of the empirical risk in Eq. (2). The margin γ measures the gap between the prediction probability of the correct label and those of the other labels, and has often been used in generalization bounds for deep learning [3], [9].

Our paper shall prove generalization bounds for the optimal transferred model $k^* \circ w^*$. For this purpose, we introduce in the next section the majority predictor accuracy, a transferability measure that we will use for our bounds.

III. Majority Predictor Accuracy

The ***Majority Predictor Accuracy*** (MPA) is defined as the accuracy of the simple predictor (classifier) that maps each source label to the target label with maximal empirical conditional probability. Formally, given the source dataset $\mathcal{S}$ and the target dataset $\mathcal{T}$, the empirical joint distribution between all possible source-target label pairs $(s, t) \in [m_S] \times [m_T]$ is $\hat{P}(s, t) = \frac{1}{n} |\{i \in [n] : s_i = s \text{ and } t_i = t\}|$, the empirical marginal distribution over the source labels is $\hat{P}(s) = \sum_{t \in [m_T]} \hat{P}(s, t), \forall s \in [m_S]$, and the empirical conditional distribution of a target label t given a source label s is $\hat{P}(t|s) = \hat{P}(s, t)/\hat{P}(s)$, for all $(s, t) \in [m_S] \times [m_T]$.

To define the MPA, consider the following majority predictor f_{mp} that takes a source label $s \in [m_S]$ and simply returns a target label t that maximizes the empirical conditional probability $\hat{P}(t|s)$:

$$f_{\text{mp}}(s) = \underset{t \in [m_T]}{\arg\max} \, \hat{P}(t|s), \tag{4}$$

The MPA is then defined as the accuracy of f_{mp} on the target dataset, as stated in the following definition.

Definition 1. *The majority predictor accuracy MPA$(\mathcal{T}|\mathcal{S})$ of a target dataset $\mathcal{T}$ given a source dataset $\mathcal{S}$ is the accuracy of the majority predictor f_{mp} on the target dataset:*

$$\mathit{MPA}(\mathcal{T}|\mathcal{S}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[t_i = f_{mp}(s_i)]. \tag{5}$$

The MPA is very simple to compute, requiring only $O(n)$ computational time by looping through the datasets a few times to compute the empirical distributions, f_{mp}, and its accuracy. In Section VII, we will show that it can also be used as a transferability measure that estimates the effectiveness of transfer learning between two tasks. We now prove generalization guarantees for the transferred models in the next section.

IV. Bounds for Shared Training Inputs Setting

This section proves our generalization bounds for deep transfer learning based on the MPA where the source and target training sets are assumed to share the inputs. In particular, we bound the true risk of the transferred model $k^* \circ w^*$ on the target distribution $\mathbb{P}_{X,T}$, which is:

$$R_T(w^*, k^*) = \mathbb{P}(t \neq \arg\max_i k^* \circ w^*(x)_i)$$

for $(x, t) \sim \mathbb{P}_{X,T}$. We will prove the bounds for both fully connected neural networks and convolutional neural networks. For this purpose, we consider the head classifier $f_{\text{mp}} \circ h^*$ defined on any representation $w(x) \in \mathbb{R}^r$, where:

$$f_{\text{mp}} \circ h^*(w(x)) = f_{\text{mp}}(h^*(w(x))). \tag{6}$$

Throughout our paper, we will make an assumption that using a deep neural network as the target head classifier can achieve better empirical risk than using the naive classifier $f_{\text{mp}} \circ h^*$. This assumption is usually satisfied in practice because of the expressiveness of neural network models [20].

[1] Throughout our paper, we assume $\arg\min$ and $\arg\max$ follow any deterministic tie-breaking strategy.

Assumption 1. *For any datasets $\mathcal{S}$ any $\mathcal{T}$, there exists $\bar{\gamma} > 0$ and $\bar{k} \in \Omega_k$ such that $\widehat{R}_{\mathcal{T},\bar{\gamma}}(w^*, \bar{k}) \leq \widehat{R}_{\mathcal{T}}(w^*, f_{mp} \circ h^*)$.*

In this assumption, $\widehat{R}_{\mathcal{T}}(w^*, f_{\text{mp}} \circ h^*)$ is defined similarly as in Eq. (2). We also note that $\widehat{R}_{\bullet,\gamma}$ is non-decreasing in γ, so the assumption implies, for all $\gamma \in [0, \bar{\gamma}]$, $\widehat{R}_{\mathcal{T},\gamma}(w^*, \bar{k}) \leq \widehat{R}_{\mathcal{T},\bar{\gamma}}(w^*, \bar{k}) \leq \widehat{R}_{\mathcal{T}}(w^*, f_{\text{mp}} \circ h^*)$. Under this assumption, we first prove the following lemma relating the empirical risks of the optimal source and transferred models using the MPA.

Lemma 1. *Under Assumption 1, for any $\gamma \in [0, \bar{\gamma}]$, we have:*

$$\widehat{R}_{\mathcal{T},\gamma}(w^*, k^*) \leq \widehat{R}_{\mathcal{S}}(w^*, h^*) + 1 - \mathit{MPA}(\mathcal{T}|\mathcal{S}).$$

Proof. Consider the majority predictor f_{mp} defined in Eq. (4). We first split the data index set $[n]$ into two non-overlap sets:

$$I = \{i \in [n] : t_i = f_{\text{mp}}(s_i)\}, \text{ and}$$
$$\bar{I} = \{i \in [n] : t_i \neq f_{\text{mp}}(s_i)\}.$$

Here the set I (respectively, $\bar{I}$) contains indices of data points whose source-target label pairs are consistent (respectively, inconsistent) with f_{mp}. For any $\gamma \in [0, \bar{\gamma}]$, we have:

$$\begin{aligned}
&\widehat{R}_{\mathcal{T},\gamma}(w^*, k^*) \leq \widehat{R}_{\mathcal{T},\gamma}(w^*, \bar{k}) && \text{(def. of } k^*) \\
&\leq \widehat{R}_{\mathcal{T},\bar{\gamma}}(w^*, \bar{k}) && (\widehat{R}_{\mathcal{T},\gamma} \text{ is non-decreasing in } \gamma) \\
&\leq \widehat{R}_{\mathcal{T}}(w^*, f_{\text{mp}} \circ h^*) && \text{(assumption 1)} \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[t_i \neq f_{\text{mp}} \circ h^* \circ w^*(x_i)] && (\text{def. of } \widehat{R}_{\mathcal{T}}) \\
&= \frac{1}{n} \Big(\sum_{i \in I} \mathbf{1}[t_i \neq f_{\text{mp}} \circ h^* \circ w^*(x_i)] + \\
&\qquad \sum_{i \in \bar{I}} \mathbf{1}[t_i \neq f_{\text{mp}} \circ h^* \circ w^*(x_i)] \Big) && (\text{def. of } I \text{ and } \bar{I}) \\
&\leq \frac{1}{n} \Big(\sum_{i \in I} \mathbf{1}[t_i \neq f_{\text{mp}} \circ h^* \circ w^*(x_i)] + |\bar{I}| \Big) \\
&= \frac{1}{n} \Big(\sum_{i \in I} \mathbf{1}[f_{\text{mp}}(s_i) \neq f_{\text{mp}} \circ h^* \circ w^*(x_i)] + |\bar{I}| \Big) \\
&\leq \frac{1}{n} \sum_{i \in I} \mathbf{1}[s_i \neq h^* \circ w^*(x_i)] + \frac{|\bar{I}|}{n}. && (7)
\end{aligned}$$

By definition of $\widehat{R}_{\mathcal{S}}(w^*, h^*)$, we also have:

$$\begin{aligned}
&\widehat{R}_{\mathcal{S}}(w^*, h^*) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[s_i \neq h^* \circ w^*(x_i)] \\
&= \frac{1}{n} \Big(\sum_{i \in I} \mathbf{1}[s_i \neq h^* \circ w^*(x_i)] + \sum_{i \in \bar{I}} \mathbf{1}[s_i \neq h^* \circ w^*(x_i)] \Big) \\
&\geq \frac{1}{n} \sum_{i \in I} \mathbf{1}[s_i \neq h^* \circ w^*(x_i)].
\end{aligned}$$

Plug this into Eq. (7) and note that $\text{MPA}(\mathcal{T}|\mathcal{S}) = |I|/n$, we have: $\widehat{R}_{\mathcal{T},\gamma}(w^*, k^*) \leq \widehat{R}_{\mathcal{S}}(w^*, h^*) + |\bar{I}|/n = \widehat{R}_{\mathcal{S}}(w^*, h^*) + 1 - \text{MPA}(\mathcal{T}|\mathcal{S})$. □

As a remark, the bound in Lemma 1 gets tighter when $\text{MPA}(\mathcal{T}|\mathcal{S}) \to 1$, that is, when f_{mp} is more accurate. Using this lemma, we now prove the generalization bounds for the transferred model $k^* \circ w^*$. Section IV-A below proves the bound for fully connected neural networks, while Section IV-B proves the bound for convolutional neural networks.

A. Generalization Bound for Fully Connected Networks

In this section, we consider target models $k \circ w$ that are deep neural networks parameterized by $w = \{A^1, A^2, \ldots, A^L\}$ and $k = \{A^{L+1}, A^{L+2}, \ldots, A^{L_T}\}$ such that:

$$w(x) = \sigma_L(A^L \sigma_{L-1}(A^{L-1} \sigma_{L-2}(\ldots A^1(x)))), \text{ and}$$
$$k(w(x)) = A^{L_T} \sigma_{L_T-1}(A^{L_T-1} \sigma_{L_T-2}(\ldots A^{L+1}(w(x)))),$$

where L is the depth of the neural network w, L_T is the depth of the whole target network $k \circ w$, and $A^i \in \mathbb{R}^{W_i \times W_{i-1}}$ is the weight matrix at layer i with $W_0 = d$, $W_L = r$, and $W_{L_T} = m_T$. In the above formulas, $\sigma_i : \mathbb{R}^{W_i} \to \mathbb{R}^{W_i}$ is a non-linear activation function that is assumed to be 1-Lipschitz.

We do not make any assumption regarding the form or architecture of the source head classifier h, except for Assumption 1. Thus, our generalization bound in this section holds for all types of source head classifiers, including neural networks, logistic regression, support vector machines, etc. In practice, however, h is usually chosen as a logistic regression or neural network for ease of implementation and better accuracy.

Following the notations in [9], in our result, we write $\|A\|_{\text{Fr}}$, $\|A\|_{\sigma}$, and $\|A\|_{p,q}$ to denote respectively the Frobenius norm, the spectral norm, and the (p, q)-norm of a matrix A. We also write $A_{i,\bullet}$ to denote the i-th row of A. We let $\bar{W} = \max_{i=1}^{L_T} W_i$ be the maximum width of the target neural network. We now state and prove our generalization bound of the true risk $R_T(w^*, k^*) = \mathbb{P}(t \neq \arg\max_i k^* \circ w^*(x)_i)$ for this fully connected network setting in the theorem below.

Theorem 1. *Assume we are given some fixed reference matrices $M^1, M^2, \ldots, M^{L_T}$ representing the initialized weights of the target network. Under Assumption 1, with probability at least $1 - \delta$, for all margin $\gamma \in (0, \bar{\gamma}]$, we have:*

$$\begin{aligned}
&R_T(w^*, k^*) \leq \widehat{R}_{\mathcal{S}}(w^*, h^*) + (1 - \mathit{MPA}(\mathcal{T}|\mathcal{S})) + \\
&\qquad \widetilde{\mathcal{O}}\Big(\frac{\max_{i=1}^{n} \|x_i\|_{Fr} \mathcal{F}_{\mathcal{A}}}{\gamma \sqrt{n}} \log(\bar{W}) + \sqrt{\frac{\log(1/\delta)}{n}} \Big),
\end{aligned}$$

where $\mathcal{A} = (A^{*1}, A^{*2}, \ldots, A^{*L_T})$ is the weight matrices of the target fully connected neural network $k^* \circ w^*$ trained using the deep transfer learning procedure in Section II, and

$$\begin{aligned}
\mathcal{F}_{\mathcal{A}} := L_T \max\nolimits_i \|A^{*L_T}_{i,\bullet}\|_{Fr} \Big(\textstyle\prod_{i=1}^{L_T-1} \|A^{*i}\|_{\sigma} \Big) \\
\Big(\sum_{i=1}^{L_T-1} \frac{\|A^{*i} - M^i\|_{2,1}^{2/3}}{\|A^{*i}\|_{\sigma}^{2/3}} + \frac{\|A^{*L_T}\|_{Fr}^{2/3}}{\max_i \|A^{*L_T}_{i,\bullet}\|_{Fr}^{2/3}} \Big)^{3/2}.
\end{aligned}$$

Proof. Using Theorem 1 of [9], with probability at least $1 - \delta$, for all margin $\gamma > 0$, we have:

$$\begin{aligned}
&R_T(w^*, k^*) \leq \widehat{R}_{\mathcal{T},\gamma}(w^*, k^*) + \\
&\qquad \widetilde{\mathcal{O}}\Big(\frac{\max_{i=1}^{n} \|x_i\|_{\text{Fr}} \mathcal{F}_{\mathcal{A}}}{\gamma \sqrt{n}} \log(\bar{W}) + \sqrt{\frac{\log(1/\delta)}{n}} \Big).
\end{aligned}$$

Combining this with Lemma 1, we have, with probability at least $1-\delta$, for all margin $\gamma \in (0, \bar{\gamma}]$:

$$R_T(w^*, k^*) \leq \widehat{R}_{\mathcal{S}}(w^*, h^*) + (1 - \text{MPA}(\mathcal{T}|\mathcal{S})) + \widetilde{\mathcal{O}}\Big(\frac{\max_{i=1}^n \|x_i\|_{\text{Fr}}\, \mathcal{F}_{\mathcal{A}}}{\gamma\sqrt{n}} \log(\bar{W}) + \sqrt{\frac{\log(1/\delta)}{n}}\Big). \qquad \square$$

B. Generalization Bound for Convolutional Neural Networks

We now consider target models $k \circ w$ that are convolutional neural networks. For these models, the matrices $A^1, A^2, \ldots, A^{L_T}$ are the filter matrices of the convolutional layers. Following [9], for each filter matrix A^i, we can construct a corresponding larger convolutional matrix $\tilde{A}^i$ by repeating the weights of A^i as many times as the filter A^i is applied. The activation functions considered here are assumed to be either ReLU or max pooling.

In our result, $\bar{W}$ is the maximum number of neurons in a single layer before pooling, counting all the channels. For each layer i, w_i is the spacial width of the layer after pooling, and B_i is the maximum l_2 norm of any convolutional patch of the layer's activations over all inputs. For any layer $i \leq L_T - 1$, we also write $\|\tilde{A}^i\|_{\sigma'}$ to denote the maximum spectral norm of any matrix obtained by deleting, for each pooling window, all but one of the corresponding rows of $\tilde{A}^i$. For $i = L_T$, $\|\tilde{A}^{L_T}\|_{\sigma'} = \rho_{L_T} \max_j \|A^{L_T}_{j,\bullet}\|_{\text{Fr}}$, with ρ_{L_T} being the Lipschitz constant of the activation and pooling at layer L_T. More details of the notations can be found in [9].

Theorem 2 below shows our generalization bound for this convolutional neural network setting. Similar to Section IV-A, we do not restrict the form of the source head classifier h, so our result will also hold for all types of source head classifiers.

Theorem 2. *Assume we are given some fixed reference matrices $M^1, M^2, \ldots, M^{L_T}$ representing the initialized weights of the target network's filter matrices. Under Assumption 1, with probability at least $1-\delta$, for all margin $\gamma \in (0, \bar{\gamma}]$, we have:*

$$R_T(w^*, k^*) \leq \widehat{R}_{\mathcal{S}}(w^*, h^*) + (1 - \mathit{MPA}(\mathcal{T}|\mathcal{S})) + \widetilde{\mathcal{O}}\Big(\frac{\mathcal{G}_{\mathcal{A}}}{\sqrt{n}} \log(\bar{W}) + \sqrt{\frac{\log(1/\delta)}{n}}\Big),$$

where $\mathcal{A} = (A^{*1}, A^{*2}, \ldots, A^{*L_T})$ is the filter matrices of the target convolutional neural network $k^* \circ w^*$ trained using the deep transfer learning procedure in Section II, and $\mathcal{G}_{\mathcal{A}}^{2/3} := \sum_{i=1}^{L_T} T_i^{2/3}$, where for all $i \leq L_T - 1$,

$$T_i := B_{i-1} \|(A^{*i} - M^i)^\top\|_{2,1} \sqrt{w_i} \max_{U \leq L_T} \frac{\prod_{u=i+1}^{U} \|\tilde{A}^{*u}\|_{\sigma'}}{B_U},$$

and $T_{L_T} := B_{L_T - 1} \|A^{*L_T} - M^{L_T}\|_{Fr}/\gamma$.

Proof. This proof is similar to the proof of our Theorem 1 above, but replacing Theorem 1 of [9] by their Theorem 3, which states that with probability at least $1-\delta$, for all $\gamma > 0$:

$$R_T(w^*, k^*) \leq \widehat{R}_{\mathcal{T},\gamma}(w^*, k^*) + \widetilde{\mathcal{O}}\big(\frac{\mathcal{G}_{\mathcal{A}}}{\sqrt{n}} \log(\bar{W}) + \sqrt{\frac{\log(1/\delta)}{n}}\big).$$

The theorem holds by combining this with Lemma 1. $\square$

V. Bounds for Different Training Inputs Setting

Up until now, we have only considered the case where source and target datasets share the same input set. In this section, we extend our results to the case where the source and target datasets contain different inputs. Formally, let $\mathcal{S} = \{(x_1, s_1), (x_2, s_2), \ldots, (x_n, s_n)\}$ be the source dataset where $(x_i, s_i) \in \mathbb{R}^d \times [m_S]$ is drawn iid from a joint distribution $\mathbb{P}_{X,S}$, and $\mathcal{T} = \{(z_1, t_1), (z_2, t_2), \ldots, (z_p, t_p)\}$ be the target dataset where $(z_i, t_i) \in \mathbb{R}^d \times [m_T]$ is drawn iid from $\mathbb{P}_{Z,T}$. We also follow the deep transfer learning procedure in Section II and first train the optimal model $h^* \circ w^*$ on the source data $\mathcal{S}$ using Eq. (1). Then we freeze w^* and train the target head classifier k^* using Eq. (3) with the target dataset $\mathcal{T}$, where we will apply w^* to the target inputs $\{z_1, z_2, \ldots, z_p\}$ to get the representations $\{w^*(z_1), w^*(z_2), \ldots, w^*(z_p)\}$.

To prove the generalization bounds, we will consider a new source dataset $\tilde{\mathcal{S}} := \{(z_i, h^* \circ w^*(z_i))\}_{i=1}^p$ induced by $h^* \circ w^*$ and the target inputs $\{z_1, z_2, \ldots, z_p\}$. In essence, $\tilde{\mathcal{S}}$ contains the target inputs with "dummy" source labels generated by $h^* \circ w^*$. This technique of using these dummy labels was previously employed to develop the LEEP transferability measure [12], and is useful for proving our bounds as well. With the new source dataset $\tilde{\mathcal{S}}$, we consider the majority predictor f_{mp} constructed from $(\tilde{\mathcal{S}}, \mathcal{T})$, as well as the corresponding majority predictor accuracy $\text{MPA}(\mathcal{T}|\tilde{\mathcal{S}})$. We still keep Assumption 1 in Section IV, but adapt it to the new $h^* \circ w^*$ and f_{mp}. The following lemma is the analogue of Lemma 1 for the different inputs setting.

Lemma 2. *With the adapted Assumption 1, for any $\gamma \in [0, \bar{\gamma}]$, we have: $\widehat{R}_{\mathcal{T},\gamma}(w^*, k^*) \leq 1 - \mathit{MPA}(\mathcal{T}|\tilde{\mathcal{S}})$.*

Proof. From (5), (6), and the definition of $\tilde{\mathcal{S}}$, we have:

$$\text{MPA}(\mathcal{T}|\tilde{\mathcal{S}}) = \frac{1}{p} \sum_{i=1}^{p} \mathbf{1}[t_i = f_{\text{mp}} \circ h^* \circ w^*(z_i)].$$

By Assumption 1, for any $\gamma \in [0, \bar{\gamma}]$, we have

$$\widehat{R}_{\mathcal{T},\gamma}(w^*, k^*) \leq \widehat{R}_{\mathcal{T}}(w^*, f_{\text{mp}} \circ h^*) = \frac{1}{p} \sum_{i=1}^{p} \mathbf{1}[t_i \neq f_{\text{mp}} \circ h^* \circ w^*(z_i)] = 1 - \text{MPA}(\mathcal{T}|\tilde{\mathcal{S}}). \qquad \square$$

Similar to Section IV, we derive the following generalization bounds, which are analogues of Theorems 1 and 2. The proofs of these theorems are similar to those of Theorems 1 and 2, with Lemma 1 being replaced by Lemma 2.

Theorem 3. *Assume we are given some fixed reference matrices $M^1, M^2, \ldots, M^{L_T}$ representing the initialized weights of the target network. Under the adapted Assumption 1, with probability at least $1-\delta$, for all margin $\gamma \in (0, \bar{\gamma}]$, with $\mathcal{F}_{\mathcal{A}}$ defined as in Theorem 1, we have: $R_T(w^*, k^*) \leq 1 - \mathit{MPA}(\mathcal{T}|\tilde{\mathcal{S}}) + \widetilde{\mathcal{O}}\big(\frac{\max_{i=1}^p \|x_i\|_{Fr}\, \mathcal{F}_{\mathcal{A}}}{\gamma\sqrt{p}} \log(\bar{W}) + \sqrt{\frac{\log(1/\delta)}{p}}\big)$.*

Theorem 4. *Assume we are given some fixed reference matrices $M^1, M^2, \ldots, M^{L_T}$ representing the initialized weights of the target network's filter matrices. Under the adapted*

Assumption 1, with probability at least $1-\delta$*, for all margin* $\gamma \in (0, \bar{\gamma}]$*, with* $\mathcal{G}_{\mathcal{A}}$ *defined as in Theorem 2, we have:*

$$R_T(w^*, k^*) \le 1 - MPA(\mathcal{T}|\tilde{\mathcal{S}}) + \widetilde{\mathcal{O}}\big(\tfrac{\mathcal{G}_{\mathcal{A}}}{\sqrt{p}}\log(\bar{W}) + \sqrt{\tfrac{\log(1/\delta)}{p}}\big).$$

VI. Discussions

The technique used to prove our theorems is general and can be combined with other generalization bounds for deep neural networks. Although we proved our results using the norm-based bounds of [9], we emphasize that our proof technique can also be used with other generalization bounds for neural networks, such as those of [3].

The bounds in Theorems 1 and 2 depend on both the optimal source empirical risk $\widehat{R}_{\mathcal{S}}(w^*, h^*)$ and MPA$(\mathcal{T}|\mathcal{S})$. These bounds get better when $\widehat{R}_{\mathcal{S}}(w^*, h^*) \to 0$ and MPA$(\mathcal{T}|\mathcal{S}) \to 1$. The bounds in Theorems 3 and 4 do not contain the source empirical risk, since it has been indirectly measured in MPA$(\mathcal{T}|\tilde{\mathcal{S}})$ when we use $h^* \circ w^*$ to construct $\tilde{\mathcal{S}}$.

From our results, we can see that MPA$(\mathcal{T}|\mathcal{S})$ (or MPA$(\mathcal{T}|\tilde{\mathcal{S}})$ for the setting with different inputs) can be used as a transferability measure. Specifically, for well-trained and deep enough neural networks, it has been observed empirically [20] that $\widehat{R}_{\mathcal{S}}(w^*, h^*) \approx 0$. Furthermore, the $\widetilde{\mathcal{O}}(\cdot)$ terms in our theorems are near 0 for large enough n. In this case, our results imply that MPA$(\mathcal{T}|\mathcal{S}) \lessapprox 1 - R_T(w^*, k^*)$ or MPA$(\mathcal{T}|\tilde{\mathcal{S}}) \lessapprox 1 - R_T(w^*, k^*)$. This means that MPA$(\mathcal{T}|\mathcal{S})$ or MPA$(\mathcal{T}|\tilde{\mathcal{S}})$ lower bounds the expected accuracy of the transferred model $k^* \circ w^*$, and thus can be used as a transferability measure. We now validate this observation empirically.

VII. Experiments

We show the usefulness of our theoretical bounds in practice by empirically illustrating the ability of MPA as a transferability measure on the large-scale Caltech-UCSD Birds-200 dataset [17], which contains 11,788 images of 200 bird species labeled with 312 binary attributes. We keep the train-test split as provided in dataset, with 5,994 train images and 5,794 test images. We pick 4 attributes *Curved Bill*, *Iridescent Wings*, *Brown Upper Parts* and *Olive Under Parts* for training source models, and randomly choose 100 different attributes as target tasks. Regarding the model architecture, we use ResNet18 [7] without the last fully connected layer as the feature extractor w. In all tests, we train our source model $h^* \circ w^*$ and the transferred model $k^* \circ w^*$ using the cross-entropy loss with batch size 32 and run the stochastic gradient descent optimizer with momentum for 40 epochs. The initial learning rate is set at 0.01 and is divided by 10 every 10 epochs.

Following the settings in [12], [16], we estimate the correlations between the MPA scores and the actual test accuracies of the transferred models to evaluate the relationship between these two quantities. High correlations mean the MPA score is a good measure for comparing test accuracies of the transferred models, and thus is a good transferability measure. For the 4 source tasks above with 100 randomly chosen target tasks, our experimental results give the following Pearson correlation coefficients: 0.9534 (Curved Bill), 0.9452 (Iridescent Wings), 0.9484 (Brown Upper Parts), and 0.9611 (Olive Under Parts). These coefficients show that the MPA scores and the test accuracies are highly positive correlated with statistical significance ($p < 10^{-4}$), which clearly indicates that the MPA is a reliable transferability measure for estimating the performance of transferred models.

VIII. Conclusion

We proved novel generalization bounds for transfer learning of deep neural networks using a new quantity, the majority predictor accuracy, that can be computed easily and efficiently from data. We showed the usefulness of our bounds in practice by demonstrating that the majority predictor accuracy can be used for estimating the effectiveness of deep transfer learning.

References

[1] Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. In *ICLR*, 2018.

[2] Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas Guibas. An information-theoretic approach to transferability in task transfer learning. In *ICIP*, 2019.

[3] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *NeurIPS*, 2017.

[4] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

[5] Shai Ben-David and Reba Schuller. Exploiting task relatedness for multiple task learning. *Learning theory and kernel machines*, pages 567–580, 2003.

[6] John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. In *NeurIPS*, 2007.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[8] Jiaji Huang, Qiang Qiu, and Kenneth Church. Exploiting a zoo of checkpoints for unseen tasks. In *NeurIPS*, 2021.

[9] Antoine Ledent, Waleed Mustafa, Yunwen Lei, and Marius Kloft. Norm-based generalisation bounds for deep multi-class convolutional neural networks. In *AAAI*, 2021.

[10] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, 2017.

[11] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT*, 2009.

[12] Cuong V Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In *ICML*, 2020.

[13] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In *CVPR Workshops*, 2014.

[14] Yang Tan, Yang Li, and Shao-Lun Huang. OTCE: A transferability metric for cross-domain cross-task representations. In *CVPR*, 2021.

[15] Xinyi Tong, Xiangxiang Xu, Shao-Lun Huang, and Lizhong Zheng. A mathematical framework for quantifying transferability in multi-source transfer learning. In *NeurIPS*, 2021.

[16] Anh T Tran, Cuong V Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *ICCV*, 2019.

[17] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical report, Caltech, 2010.

[18] Paul N Whatmough, Chuteng Zhou, Patrick Hansen, Shreyas Kolala Venkataramanaiah, Jae-sun Seo, and Matthew Mattina. FixyNN: Efficient hardware for mobile computer vision via transfer learning. In *MLSys*, 2019.

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

[20] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.