Transfer Learning for Individual Treatment Effect Estimation

Ahmed Aloui *[1]   **Juncheng Dong** *[1]   **Cat P. Le**[1]   **Vahid Tarokh**[1]

[1] Department of Electrical and Computer Engineering, Duke University

Abstract

This work considers the problem of transferring causal knowledge between tasks for Individual Treatment Effect (ITE) estimation. To this end, we theoretically assess the feasibility of transferring ITE knowledge and present a practical framework for efficient transfer. A lower bound is introduced on the ITE error of the target task to demonstrate that ITE knowledge transfer is challenging due to the absence of counterfactual information. Nevertheless, we establish generalization upper bounds on the counterfactual loss and ITE error of the target task, demonstrating the feasibility of ITE knowledge transfer. Subsequently, we introduce a framework with a new Causal Inference Task Affinity (CITA) measure for ITE knowledge transfer. Specifically, we use CITA to find the closest source task to the target task and utilize it for ITE knowledge transfer. Empirical studies are provided, demonstrating the efficacy of the proposed method. We observe that ITE knowledge transfer can significantly (up to 95%) reduce the amount of data required for ITE estimation.

1 INTRODUCTION

Assessing the effects of treatments on people (i.e., the *Individual Treatment Effect* (ITE) estimation) is of significant interest to various research communities, such as those studying medicine and social policy making. In order to study the causal relationship between the outcome and the treatment, however, researchers must gather sufficient data samples from randomized control trials. This process can be both costly and time-consuming [Kaur and Gupta, 2020]. To this end, it is desirable to utilize knowledge from different but closely related problems with *transfer learning*. For instance, new vaccines must be developed for treatment when the viruses undergo mutation. Suppose the mutated viruses can be related to the known ones by a similarity measure. In that case, the effects of vaccine candidates can be quickly estimated based on this similarity with a small amount of data collected from the new scenario. Hence, this approach can notably accelerate the study.

While the recent progress in transfer learning is very promising [Wang and Deng, 2018, Alyafeai et al., 2020, Pan and Yang, 2010, Zhuang et al., 2021], a major challenge for transferring causal knowledge arises from non-causal (spurious) correlations to which the statistical learning models are vulnerable. For example, a classifier may learn to use the background colors to differentiate images of camels and horses, as these objects are frequently depicted against different colored backgrounds [Arjovsky et al., 2019, Geirhos et al., 2018, Beery et al., 2018]. In practice, the performance of the ITE estimation models can *never* be evaluated because the counterfactual data is inaccessible, as shown in Figure 1. This problem is known in the literature as *the fundamental problem of causal inference* [Rubin, 1974, Holland, 1986]. For instance, to compute the effect of vaccination on a person at some given time, that individual must both be administered the vaccine, and also remain unvaccinated, which is obviously absurd. This scenario is very different from the conventional supervised learning problems, where researchers often use a separate validation set in order to estimate the accuracy of the trained model.

The aforementioned challenge implies that much attention must be paid to selecting the appropriate source model in causal knowledge transfer. Additionally, similar scenarios to the given target task must be determined using a distance accounting for the *immeasurable* counterfactual losses in scenarios under consideration. In this work, we first present a lower bound and a set of generalization bounds for transfer learning between causal inference tasks in order to demonstrate both the difficulty and viability of causal knowledge transfer. While these theoretical bounds are informative, a

*Equal Contribution.

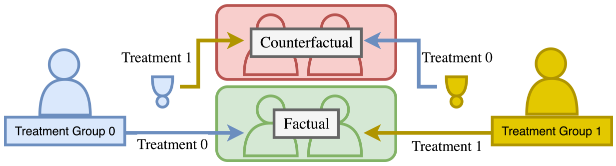

Figure 1: Inaccessibility to counterfactual data (e.g., a parallel universe where the treatments are reversed) makes transferring causal knowledge more challenging.

method is needed for selecting the optimal source model from multiple source tasks. This is discussed in Section 5, where we introduce a framework endowed with a new task affinity, namely the Causal Inference Task Affinity (CITA), tailored explicitly for causal knowledge transfer. This task affinity is used for selecting the "closest" source task. Subsequently its knowledge (e.g., trained models, source dataset) is utilized in the learning of the target task, as depicted in Figure 2. Our contributions are summarized below:

1. We establish a new lower bound to demonstrate the challenges of transferring ITE knowledge. Additionally, we prove new regret bounds for learning the counterfactual outcomes and ITEs of the target tasks in causal transfer learning scenarios. These bounds demonstrate the feasibility of transferring ITE knowledge by stating that the error of any source model on the target task is upper bounded by quantifiable measures related to *(i)* the performance of the source model on the source task and *(ii)* the *differences* between the source and the target causal inference tasks.
2. We introduce CITA, a task affinity for causal inference, which captures the symmetry of ITEs (i.e., invariance to the relabeling of treatment assignments under the action of the symmetric group). Additionally, we provide theoretical (e.g., Theorem F.3) and empirical evidence to show that CITA is highly correlated with the counterfactuals loss, which is *not measurable* in practice.
3. We propose an ITE estimation framework and a set of causal inference datasets *suitable for learning causal knowledge transfer*. The empirical evidence on the above datasets demonstrates that our methods can estimate the ITEs for the target task with significantly fewer (up to 95% reduction) data samples compared to the case where transfer learning is not performed.

2 RELATED WORK

Many approaches in transfer learning [Thrun and Pratt, 2012, Blum and Mitchell, 1998, Silver and Bennett, 2008, Sharif Razavian et al., 2014, Finn et al., 2016, Fernando et al., 2017, Rusu et al., 2016, Le et al., 2020] have been proposed, analyzed and applied in various machine learning applications. Transfer learning techniques inherently assume that prior knowledge in the selected source model helps with learning a target task [Pan and Yang, 2010, Zhuang et al., 2021]. In other words, these methods often do not consider the selection of the base task to perform knowledge transfer. Consequently, in some rare cases, transfer learning may even degrade the performance of the model Standley et al. [2020]. In order to avoid potential performance loss during knowledge transfer to a target task, *task affinity* (or task similarity) is considered as a selection method that identifies a group of closest base candidates from the set of the prior learned tasks. Task affinity has been investigated and applied to various domains (e.g., transfer learning [Zamir et al., 2018, Dwivedi and Roig, 2019, Wang et al., 2019], neural architecture search [Le et al., 2021, 2022a, Le et al., 2021], few-shot learning [Pal and Balasubramanian, 2019, Le et al., 2022b], multi-task learning [Standley et al., 2020], continual learning [Kirkpatrick et al., 2017, Chen et al., 2018]).

While transfer learning and task affinity have been investigated in numerous application areas, their applications to causal inference have yet to be thoroughly investigated. Neyman-Rubin Causal Model [Neyman, 1923, Donald, 2005] and Pearl's Do-calculus [Pearl, 2009] are popular frameworks for causal studies based on different perspectives. A central question in the Neyman-Rubin Causal Model framework is determining conditions for identifiability of causal quantities such as *Average* and *Individual Treatment Effects*. Previous work considered estimators for Average Treatment Effect based on various methods such as Covariate Adjustment [Rubin, 1978], weighting methods such as those utilizing propensity scores [Rosenbaum and Rubin, 1983], and Doubly Robust estimators [Funk et al., 2011]. With the emergence of Machine Learning techniques, more recent approaches to causal inference include the applications of decision trees[Wager and Athey, 2018, Athey and Imbens, 2016], Gaussian Processes [Alaa and Van Der Schaar, 2017], and Generative Modeling [Yoon et al., 2018] to ITE estimation. In particular, deep neural networks have successfully learned ITEs and estimated counterfactual outcomes by data balancing in the latent domain [Johansson et al., 2016, Shalit et al., 2017]. Please note that the transportation of causal graphs is another well-studied closely related field in the causality literature [Bareinboim and Pearl, 2012]. It studies transferring knowledge of causal relationships in Pearl's do-calculus framework. In contrast, in this paper, we are interested in transferring knowledge of ITE from a source task to a target task in the Neyman-Rubin framework using representation learning. A closely related problem to ours is the domain adaptation problem for ITE estimation, as explored in [Bica and van der Schaar, 2022, Vo et al., 2022, Aglietti et al., 2020]. These works primarily focus on situations where only the distribution of populations changes, leaving the causal functions unaltered. In our research, we provide theoretical analysis and empirical studies for the case where both the population distributions and the causal mechanisms can change.

3 MATHEMATICAL BACKGROUND

3.1 CAUSAL INFERENCE

Let $X \in \mathcal{X} \subset \mathbb{R}^d$ be the covariates (i.e., input features), $A \in \{0, \ldots, M\}$ be the treatment, and $Y \in \mathcal{Y} \subset \mathbb{R}$ be the factual (observed) outcome. For every $j \in \{0, \ldots, M\}$ we define Y_j to be the *potential outcome* [Rubin, 1974] that would have been observed if only the treatment $A = j$, $j \in \{0, 1, \cdots, M\}$ was assigned. In the medical context, for instance, X is the individual information (e.g., weight, heart rate), A is the treatment assignment (e.g., $A = 0$ if the individual did not receive a vaccine, and $A = 1$ if the individual is vaccinated), Y is the outcome (e.g., mortality data). A *causal inference dataset* is a collection of factual observations $D_F = \{(x_i, a_i), y_i\}_{i=1}^N$, where N is the number of samples. We assume these samples are independently drawn from the same factual distribution p_F. In a parallel universe, if the roles of the treatment and control groups were reversed, we would have observed a different set of samples D_{CF} sampled from the counterfactual distribution p_{CF}. In this work, we present our results for the binary case, i.e., $M = 1$. However, our approach can be easily extended to any positive integer $M < \infty$. In the binary case, the individuals who received treatments $A = 0$ and $A = 1$ are respectively denoted by the control and treatment groups.

Definition 3.1 (ITE)**.** The Individual Treatment Effect (ITE), referred to as the Conditional Average Treatment Effect (CATE) [Imbens and Rubin, 2015], is defined as:

$$\forall x \in \mathcal{X},\ \tau(x) = \mathbb{E}[Y_1 - Y_0 | X = x] \tag{1}$$

We assume that the data generation process respects the *overlap*, i.e. $\forall x \in \mathcal{X}, 0 < p(a = 1|x) < 1$, and *conditional unconfoundedness*, i.e. $(Y^1, Y^0) \perp\!\!\!\perp A|X$ [Robins, 1986]. These assumptions are sufficient conditions for the ITE to be identifiable [Imbens, 2004]. We also assume that the true causal relationship is described by a function $f(x, a)$, which can be expressed as an expected value in the non-deterministic case. By definition $\tau(x) = f(x, 1) - f(x, 0)$. Let $\hat{f}(x, a)$ denote a hypothesis that estimates the true function $f(x, a)$. Thus, the ITE function can then be estimated as $\hat{\tau}(x) = \hat{f}(x, 1) - \hat{f}(x, 0)$. We use $\ell_{\hat{f}}(x, a, y)$ to denote a loss function that quantifies the performance of $\hat{f}(\cdot, \cdot)$. A possible example is the L^2 loss defined as $\ell_{\hat{f}}(x, a, y) = (y - \hat{f}(x, a))^2$.

Definition 3.2 (Factual Loss)**.** For a hypothesis $\hat{f}$ and a loss function $l_{\hat{f}}$, the factual loss is defined as:

$$\epsilon_F(\hat{f}) = \int_{\mathcal{X} \times \{0,1\} \times \mathcal{Y}} l_{\hat{f}}(x, a, y)\ p_F(x, a, y) dx da dy \tag{2}$$

We also define the factual loss for the treatment ($a = 1$) and control ($a = 0$) groups respectively as:

$$\epsilon_F^{a=1}(\hat{f}) = \int_{\mathcal{X} \times \mathcal{Y}} l_{\hat{f}}(x, 1, y)\ p_F(x, y|a = 1) dx dy \tag{3}$$

and

$$\epsilon_F^{a=0}(\hat{f}) = \int_{\mathcal{X} \times \mathcal{Y}} l_{\hat{f}}(x, 0, y)\ p_F(x, y|a = 0) dx dy \tag{4}$$

Definition 3.3 (Counterfactual Loss)**.** The counterfactual loss is defined as:

$$\epsilon_{CF}(\hat{f}) = \int_{\mathcal{X} \times \{0,1\} \times \mathcal{Y}} l_{\hat{f}}(x, a, y)\ p_{CF}(x, a, y) dx da dy \tag{5}$$

We also define the counterfactual loss for the treatment ($a = 1$) and control ($a = 0$) groups respectively as:

$$\epsilon_{CF}^{a=1}(\hat{f}) = \int_{\mathcal{X} \times \mathcal{Y}} l_{\hat{f}}(x, 1, y)\ p_{CF}(x, y|a = 1) dx dy \tag{6}$$

and

$$\epsilon_{CF}^{a=0}(\hat{f}) = \int_{\mathcal{X} \times \mathcal{Y}} l_{\hat{f}}(x, 0, y)\ p_{CF}(x, y|a = 0) dx dy \tag{7}$$

The counterfactual loss corresponds to the expected loss value in a parallel universe where the roles of the control and treatment groups are exchanged.

Definition 3.4. The *Expected Precision in Estimating Heterogeneous Treatment Effect* (PEHE) is defined as:

$$\varepsilon_{PEHE}(\hat{f}) = \int_{\mathcal{X}} (\hat{\tau}(x) - \tau(x))^2\, p_F(x) dx. \tag{8}$$

Here, ε_{PEHE} [Hill, 2011] is often used as the performance metric for estimation of ITEs [Shalit et al., 2017, Johansson et al., 2016]. A critical connection between the factual loss (ϵ_F), the counterfactual loss (ϵ_{CF}), and ε_{PEHE} is that for small values of ϵ_F and ϵ_{CF} causal models have good performance (i.e., low ε_{PEHE}). However, the ε_{PEHE} is not directly accessible in causal inference scenarios because the calculation of $\tau(x)$ (i.e., the ground truth ITE values) requires access to the counterfactual values. In this light, we choose a hypothesis that instead optimizes an upper bound of ε_{PEHE} given in Equation 10.

3.2 REPRESENTATION LEARNING FOR ITE ESTIMATION

In this work, we consider The TARNet model Shalit et al. [2017] for causal learning. TARNet was developed as a framework to estimate ITEs using counterfactual balancing. It consists of a pair of functions (Φ, h) where $\Phi : \mathbb{R}^d \to \mathbb{R}^l$ is a representation learning function, and $h : \mathbb{R}^l \times \{0, 1\} \to$

$\mathbb{R}$ is a function learning the two potential outcomes functions in the representation space. The hypothesis learning for the true causal function is $\hat{f}(x,a) = h(\Phi(x),a)$ and the loss function $\ell_{\hat{f}}$ is denoted by $\ell_{(\Phi,h)}$. To ensure the similarity between the features of the treatment group and that of the control group in the representation space, TARNet uses the *Integral Probability Metric* in order to measure the distance between distributions, defined as:

$$\text{IPM}_G(p,q) := \sup_{g\in G}\left|\int_S g(s)(p(s)-q(s))ds\right| \quad (9)$$

where the supremum is taken over a given class of functions G. It follows from the Kantorovich-Rubinstein duality Villani [2009] that IPM reduces to the 1-Wassertein distance when G is the set of 1-Lipschtiz functions as is the case in our numerical experiments. Here, the TARNet model learns to estimate the potential outcomes by minimizing the following objective:

$$\begin{aligned}\mathcal{L}(\Phi,h) = &\frac{1}{N}\sum_{i=1}^{N} w_i \cdot \ell_{(\Phi,h)}(x_i,a_i,y_i)\\ &+ \alpha\cdot \text{IPM}_G\left(\{\Phi(x_i)\}_{i:a_i=0}, \{\Phi(x_i)\}_{i:a_i=1}\right)\end{aligned} \quad (10)$$

where $w_i = \frac{a_i}{2v} + \frac{1-a_i}{2(1-v)}$, $v = \frac{1}{N}\sum_{i=1}^{N} a_i$, and α is the *balancing weight* which controls the trade-off between the similarity of the representations in the latent domain and the model's performance on the factual data.

4 THEORETICAL FRAMEWORK

In this section, we provide learning bounds on the counterfactual loss of the target task, and ε_{PEHE} (i.e., the error in estimating ITE). These bounds are inspired by the work of Ben-David et al. [2010] in the non-causal setting. We use superscripts T and S to respectively denote quantities related to the target and source tasks. Let τ^T denote the individual treatment effect function of the target task. We consider the performance of a well-trained source model $\hat{f}^S : \mathcal{X}\times\{0,1\}\to\mathcal{Y}$ when applied to a target task:

$$\begin{aligned}&\varepsilon^T_{PEHE}(\hat{f}^S) =\\ &\mathbb{E}_{x\sim p_F^T}\left[\left(\tau^T(x) - [\hat{f}^S(x,1)-\hat{f}^S(x,0)]\right)^2\right]\end{aligned} \quad (11)$$

4.1 THE CHALLENGE OF ITE KNOWLEDGE TRANSFER

We first provide a lower bound on ε_{PEHE} that consists of both the factual and the counterfactual losses. This bound implies that good performance on the counterfactual data is a *necessary* condition for accurate estimation of ITE.

Theorem 4.1. *Let $\hat{f}^S$ be a model trained on a source task, and $u = p_F^T(A=1)$ then*

$$\epsilon_F^T(\hat{f}^S) + u\epsilon_{CF}^{T,a=0}(\hat{f}^S) \le \varepsilon^T_{PEHE}(\hat{f}^S) \quad (12)$$

According to the bound in Theorem 4.1, simply minimizing the factual loss of the target may not guarantee a good performance. Hence, choosing a source model with low (or zero) factual loss on the target task cannot perform well if the (immeasurable) counterfactual loss of the target becomes excessively high. In other words, the performance of the chosen source model can be arbitrarily inadequate, while its performance appears perfect on factual data.

While Theorem 4.1 has implied that causal knowledge cannot be transferred without any assumption, the learning bounds presented in the following section prove the viability of transferring causal knowledge under reasonable assumptions.

4.2 GENERAL LEARNING BOUNDS

The problem of ITE knowledge transfer can be expressed as two triples (p_F^S, p_{CF}^S, f^S) and (p_F^T, p_{CF}^T, f^T) where:

- p_F^S and p_F^T respectively denote the factual probability distribution of the source and target tasks.
- p_{CF}^S and p_{CF}^T respectively denote the counterfactual distribution of the source and target tasks.
- f^S and f^T respectively denote the underlying causal function of the source task and the target task.

We use the L_1 distance to measure the similarity between probability distributions, defined as:

$$V(p,q) = \int_S |p(s)-q(s)|ds. \quad (13)$$

Theorem 4.2. *For any hypothesis $\hat{f}$, we have:*

$$\begin{aligned}\epsilon_{CF}^T(\hat{f}) \le &\epsilon_F^S(\hat{f}) + V(p_F^T,p_F^S) + V(p_F^T,p_{CF}^T)\\ &+ \mathbb{E}_{(x,a)\sim p_F^S}[|f^S(x,a)-f^T(x,a)|]\end{aligned} \quad (14)$$

and

$$\begin{aligned}\varepsilon^T_{PEHE}(\hat{f}) \le &4\epsilon_F^S(\hat{f}) + 4V(p_F^T,p_F^S) + 2V(p_F^T,p_{CF}^T)\\ &+ 4\mathbb{E}_{(x,a)\sim p_F^S}[|f^S(x,a)-f^T(x,a)|]\end{aligned} \quad (15)$$

We note that the learning bounds consist of (1) the source factual loss, (2) the difference between the causal functions, and (3) a measure of similarities between probability distributions. However, the L_1 distance in Theorem 4.2 is intractable in practice. A more reasonable candidate distance is IPM distance as defined in Equation 9. The L_1 distance can be replaced with the IPM distance as demonstrated by the following Theorem 4.3.

Theorem 4.3. *Suppose that the function class G is stable under addition and multiplication and $\hat{f}, f^T \in G$, then*

$$\begin{aligned}\epsilon_{CF}^T(\hat{f}) \leq & \epsilon_F^S(\hat{f}) + \underset{G}{IPM}(p_F^T, p_F^S) + \underset{G}{IPM}(p_F^T, p_{CF}^T) \\ & + \underset{(x,a)\sim p_F^S}{\mathbb{E}}[|f^S(x,a) - f^T(x,a)|]\end{aligned}$$

and

$$\begin{aligned}\varepsilon_{PEHE}^T(\hat{f}) \leq & 4\epsilon_F^S(\hat{f}) + 4\underset{G}{IPM}(p_F^T, p_F^S) + 2\underset{G}{IPM}(p_F^T, p_{CF}^T) \\ & + 4 \underset{(x,a)\sim p_F^S}{\mathbb{E}}[|f^S(x,a) - f^T(x,a)|]\end{aligned}$$

4.3 BOUNDS FOR COUNTERFACTUAL BALANCING FRAMEWORKS

Suppose that we have a representation learning model (e.g., TARNet) $\hat{f}^S = (\Phi, h)$ trained on a source causal inference task. We apply the source model to a different target task. For notational simplicity, we denote $P(\Phi(X)|A = a)$ by $P(\Phi(X_a))$ for $a \in \{0, 1\}$. We make the following assumptions **A1, A2, A3**:

- **A1**: Φ is injective (thus $\Psi = \Phi^{-1}$ exists on Im(Φ)).
- **A2**: There exists a real function space G on Im(Φ) such that the function $r \mapsto \ell_{\Phi,h}^T(\Psi(r), a, y) \in G$.
- **A3**: There exists a function class G' on $\mathcal{Y}$ such that $y \mapsto \ell_{\Phi,h}(x, a, y) \in G'$.

The above theorem guarantees that causal knowledge can be transferred under reasonable assumptions. The following Lemma provides an upper bound on the counterfactual loss for transferring causal knowledge.

Lemma 4.4. *Suppose that Assumptions A1, A2, A3 hold. Then the counterfactual loss of any model (Φ, h) on the target task satisfies:*

$$\begin{aligned}\epsilon_{CF}^T(\Phi, h) \leq & \epsilon_F^{S,a=1}(\Phi, h) + \epsilon_F^{S,a=0}(\Phi, h) \\ & + \underset{G}{IPM}(P(\Phi(X_1^T)), P(\Phi(X_1^S))) \\ & + \underset{G}{IPM}(P(\Phi(X_0^T)), P(\Phi(X_0^S))) \\ & + \underset{G}{IPM}(P(\Phi(X_0^T)), P(\Phi(X_1^T))) + 2\gamma^*\end{aligned}$$

where

$$\gamma^* = \underset{x\sim p_F^S}{\mathbb{E}}\left[\underset{G'}{IPM}(P(Y_a^S|x), P(Y_a^T|x))\right] \qquad (16)$$

measures the fundamental difference between two causal inference tasks.

Theorem 4.5. *(Transferability of Causal Knowledge) Suppose that Assumptions A1, A2, A3 hold. The performance of source model on target task, i.e. $\varepsilon_{PEHE}^T(\Phi, h)$, is upper bounded by:*

$$\begin{aligned}\varepsilon_{PEHE}^T(\Phi, h) \leq & 2(\epsilon_F^{S,a=1}(\Phi, h) + \epsilon_F^{S,a=0}(\Phi, h) \\ & + \underset{G}{IPM}(P(\Phi(X_1^T)), P(\Phi(X_1^S))) \\ & + \underset{G}{IPM}(P(\Phi(X_0^T)), P(\Phi(X_0^S))) \\ & + \underset{G}{IPM}(P(\Phi(X_0^T)), P(\Phi(X_1^T)) + 2\gamma^*)\end{aligned}$$

Theorem 4.5 implies that good performance on the target task is guaranteed if (1) the source model has a slight factual loss (e.g., the first and second term in the upper bound) and (2) the distributions of the control and the treatment group features are similar in the latent domain (e.g., the last three terms in the upper bound). This upper bound provides a sufficient condition for transfer learning in causal inference scenarios, indicating the transferability of causal knowledge.

5 TASK-AWARE ITE KNOWLEDGE TRANSFER

In Section 4, the regret bounds indicate the transferability of causal knowledge between pair of causal inference tasks. In this section, we propose a causal inference learning framework (illustrated in Figure 2) capable of identifying the most relevant causal knowledge, when multiple sources exist, to train the target task. Note that although the generalization bounds are informative for understanding viability of transferring causal knowledge, they may not be the most constructive approach to select the best source task because the order of the upper bounds of errors is not necessarily the same as the order of the errors. To this end, we first propose a task affinity (CITA) that satisfies the symmetry property of causal inference tasks (see Sec 5.2) to find the closest source task to the target task. We observe that CITA strongly correlates with counterfactual loss. After obtaining the closest task using the computed task distances, its knowledge (e.g., trained model, bundled data) is utilized for training the target task.

5.1 TASK AFFINITY SCORE

Let (T, D) denote the pair of a causal inference task T and its dataset $D = (X, A, Y)$, where D consists of the covariates X, the corresponding treatment assignments A, and the factual outcomes Y. We formalize a *sufficiently well-trained* deep network representing a causal task-dataset pair (T, D) in Appendix (see Sec F). Here, all the previous tasks' models are assumed to be sufficiently well-trained to represent the corresponding tasks. Next, we recall the definitions of the Fisher Information matrix and the Task Affinity Score [Le et al., 2022b,a].

Definition 5.1 (Fisher Information Matrix)**.** For a neural network N_{θ_s} with weights θ_s trained on data D_s, a given test dataset D_t and the negative log-likelihood loss function $L(\theta, D)$, the Fisher Information matrix is defined as:

$$F_{s,t} = \mathbb{E}_{D \sim D_t}\Big[\nabla_\theta L(\theta_s, D) \nabla_\theta L(\theta_s, D)^T\Big] \tag{17}$$

Definition 5.2 (Task Affinity Score)**.** Let (T_s, D_s) and (T_t, D_t) denote the source and target task-dataset pairs, respectively. Let the source task be represented by the ε-approximation network N_{θ_s}. Let $F_{s,s}$ be the Fisher Information matrix of N_{θ_s} using the source data D_s. Let $F_{s,t}$ be constructed analogously using the target data D_t on N_{θ_s}. The distance from the source task T_s to the target task T_t is defined as:

$$d[s,t] = \frac{1}{\sqrt{2}} \|F_{s,s}^{1/2} - F_{s,t}^{1/2}\|_F, \tag{18}$$

where the norm is the Frobenius norm. It has been shown that $0 \leq d[s,t] \leq 1$, where $d[s,t] = 0$ denotes perfect similarity and $d[s,t] = 1$ indicates perfect dissimilarity. In Appendix (see Theorem F.3), we prove that under stringent assumptions, the order of task distances between candidate source tasks and the target task is preserved in a parallel universe where the roles of the control and treatment groups are exchanged.

5.2 CAUSAL INFERENCE TASK AFFINITY (CITA)

Symmetry of Causal Inference Tasks. We first observe that causal inference tasks present a unique symmetry. Specifically, causal inference tasks have multiple regression problems, one for each treatment group. Given a source task, if we alternate the treatment labels (i.e., 0 to 1 and 1 to 0), the treatment effect (i.e., $\mathbb{E}[Y_1 - Y_0|X]$) will be negated. Consequently, the non-symmetric task affinity measure [Le et al., 2022b] between the original task and the permuted task can still be considered. Moreover, the original model does not need to be retrained for transfer learning as we only need to permute the roles of output layers of the model to predict the individual treatment effects correctly for each group. In other words, the causal task affinity between these two permuted tasks must intuitively equal to zero. CITA lends itself to this property of causal inference tasks.

Properties of CITA. In this work, assume that all causal inference tasks under consideration have the same number of treatment labels, and each task is represented by a TARNet network. Let (T_s, D_s) and (T_t, D_t) be the source and target causal inference tasks, respectively. Here, $D_s = (X_s, A_s, Y_s)$, $D_t = (X_t, A_t, Y_t)$, and $A_s, A_t \in \{0, 1, \ldots, M\}$.

Consider the symmetric group $\mathbb{S}_{M+1}$ consisting of all permutations of labels $\{0, 1, \ldots, M\}$. For $\sigma \in \mathbb{S}_{M+1}$, let $A_{\sigma(t)}$ denote the permutation of the target treatment labels under the action of σ. Let $d_\sigma = \frac{1}{\sqrt{2}} \|F_{a,a}^{1/2} - F_{a,\sigma(t)}^{1/2}\|_F$ then

$$d_{sym}[s,t] = \min_{\sigma \in \mathbb{S}_{M+1}} (d_\sigma)$$

is the *label-invariant task affinity* between causal tasks T_s and T_t. Similar to the Task Affinity Score [Le et al., 2022b], the proposed distance is robust to the architectural choice of the representation networks (e.g., TARNet). In other words, the order of closeness of tasks found using this distance remains the same across different choices of network architecture or hyper-parameters. Notably, our approach involves a minimization problem with $(M+1)!$ candidates. However, in real-life applications, the number of treatments $M+1$ is often a a small number. It is nevertheless still important to note that it may be challenging to apply our method in the scenarios with large M or for a continuum of treatments.

Algorithm 1: Task-Aware ITE Knowledge Transfer

Data: Source tasks: $\mathcal{S} = \{\{(X_i, A_i, Y_i)\}, 1 \leq i \leq m\}$,
Target task: $T = (X_t, A_t, Y_t)$
Input: Causal Inference Models $N_{\theta_1}, N_{\theta_2}, \ldots, N_{\theta_m}$
Output: Causal Inference model for the target task T
Function `TAS(`$X_s, A_s, X_s, A_s, N_{\theta_s}$`)`**:**
  Compute $F_{s,s}$ using N_{θ_s} with X_s, A_s
  Compute $F_{s,t}$ using N_{θ_t} with X_t, A_t
  return $d[s,t] = \frac{1}{\sqrt{2}} \|F_{s,s}^{1/2} - F_{s,t}^{1/2}\|_F$
Function `Main`**:**
  ▷ Find the closest tasks in S
  for $i = 1, 2, \ldots, m$ **do**
    Train N_{θ_i} for source task i using (X_i, A_i, Y_i)
    Compute the distance from source task i to target task T:
    $d_i^+ = \texttt{TAS}(X_i, A_i, X_t, A_t, N_{\theta_i})$
    Compute the distance from source task i to target task T', where a''s treatments are inverted treatments of a:
    $d_i^- = \texttt{TAS}(X_i, A_i, X_t, 1 - A_t, N_{\theta_t})$
    CITA: $s_{sym_i} = \min(d_i^+, d_i^-)$
  return closest tasks: $i^* = \operatorname{argmin}_i d_{sym_i}$
  ▷ ITE Knowledge Transfer
  Fine-tune $N_{\theta_i^*}$ with the target task's data (X_t, A_t, Y_t)
return $N_{\theta_i^*}$

5.3 CAUSAL KNOWLEDGE TRANSFER FROM CITA

Based on CITA, we propose a task-aware causal inference learning framework, whose procedure is illustrated in Figure 2, that is capable of utilizing past experiences to learn

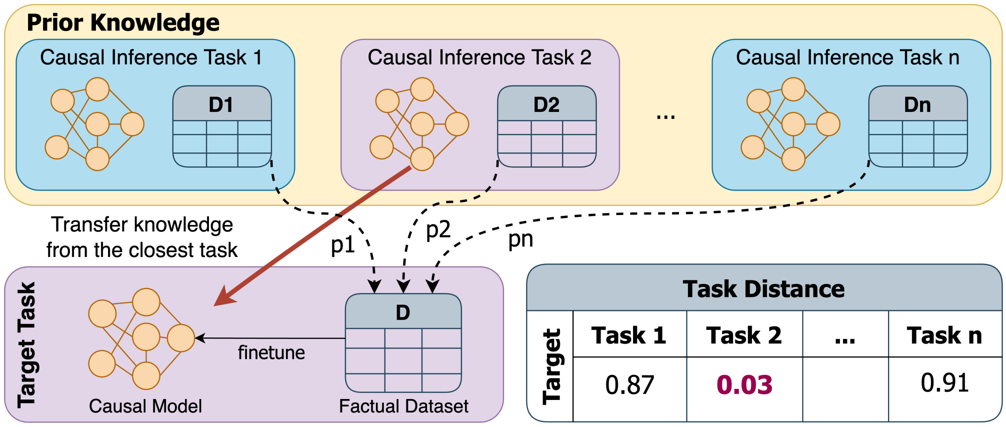

Figure 2: Overview of transfer learning in causal inference. Task affinity (CITA) is used to identify the closest task(s) from prior tasks. The models and datasets from the relevant prior tasks are transferred to the target task.

Table 1: Overview of the causal inference datasets constructed for Transfer Learning.

Name	Type	Task	CF Avail
IHDP	SEMI-SYNTHETIC	REG	YES
TWINS	REAL-WORLD	CLS	NO
JOBS	REAL-WORLD	CLS	NO
RKHS	SYNTHETIC	REG	YES
MOVEMENT	SYNTHETIC	REG	YES
HEAT	SYNTHETIC	CLS	YES
Name	**Subject**	**#Task**	**#Sample**
IHDP	HEALTH	100	747
TWINS	HEALTH	11	2000
JOBS	SOCIAL SCIENCES	10	619
RKHS	MATHEMATICS	100	2000
MOVEMENT	PHYSICS	12	4000
HEAT	PHYSICS	20	4000

REG/CLS: Regression/Classification, **CF Avail**: Counterfactual Data Availability

Table 2: The impact of causal knowledge transfer on the performance and the required size of the training dataset.

Dataset	IHDP	RKHS	MOVEMENT	HEAT
ORI Size	747	2000	4000	4000
TL Size	150	50	750	500
W/O TL(I)	0.61	0.68	0.021	6.7E-6
W/O TL (P)	0.97	0.96	0.025	1.4E-5
W TL (P)	0.65	0.46	0.011	4.2E-6
Data Gain	$> 80\%$	$> 95\%$	$> 80\%$	$> 85\%$
Perf Gain	$> 30\%$	$> 50\%$	$> 55\%$	$> 70\%$

ORI/TL Size: Number of data required without and with TL, **W/O TL (I & P)**: Performance achieved without TL (*ideal* & *practice*); (ideal) is the model with the lowest ε_{PEHE} (not attainable); (practice) is the model with the lowest training loss, **W TL (P)**: Performance with TL, **Data Gain**: Data Reduction with TL, **Perf Gain**: Error Reduction with TL.

the target task's ITE quickly. In particular, given multiple trained source tasks, the closest task is identified via causal inference task affinity (CITA). Subsequently, its knowledge (e.g., trained causal model, weights, parameters, initialization settings) is applied to learn the causal effect of the target task. Here, the source task's model is fine-tuned with the target task's data for estimating ITEs. In our experiments, we compare the performance of our method to training from scratch. We also compare our method to data-bundling, as illustrated in Figure 1 in the Appendix. The pseudo-code of the proposed framework is provided in Algorithm 1.

6 EXPERIMENTS

We first describe the datasets we have used for our empirical studies. Subsequently, we present empirical results that (1) show that CITA identifies the symmetries within causal inference tasks, (2) demonstrate the strong correlation between CITA and the counterfactual loss, and (3) quantify the gains of transfer learning for causal inference.

6.1 CAUSAL INFERENCE DATASETS

We present a representative family of causal inference datasets suitable for studying ITE knowledge transfer. Some of these are well-established datasets in the literature, while others are motivated by known causal structures in diverse areas such as social sciences, physics, health, and mathematics. Table 1 provides a brief description of the datasets used in our studies. A more detailed description is provided in Appendix (see Sec B). For each dataset, a number of corresponding causal inference tasks exist, which can be used to study transfer learning scenarios. Please note that we can

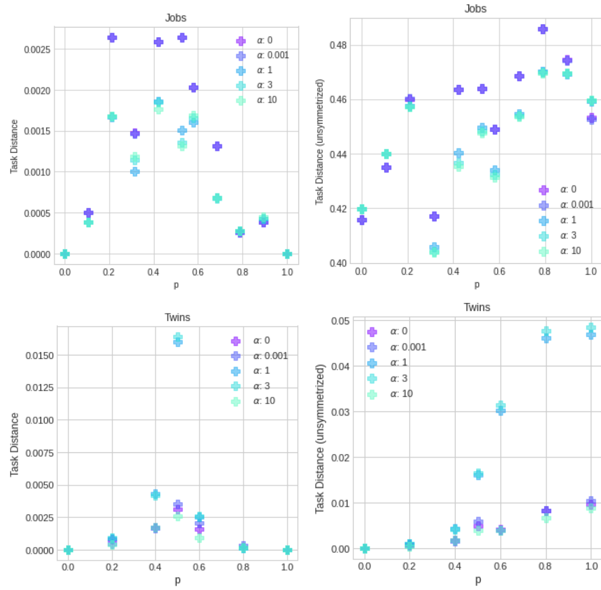

Figure 3: The symmetry of CITA. p (on the x-axis) denotes the probability of flipping treatment assignments of the original dataset. Left column: CITA; right column: non-symmetrized task affinity.

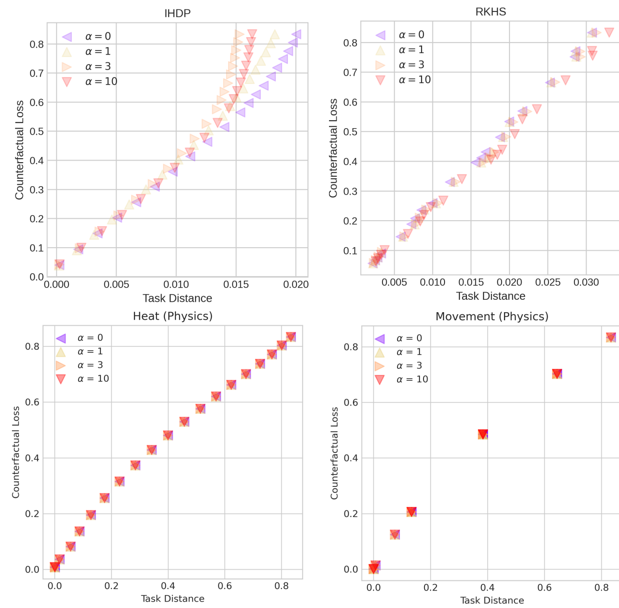

Figure 4: CITA vs. Counterfactual Error on causal inference datasets. CITA strongly correlates with the (immeasurable) counterfactual loss.

only access the counterfactual data of the synthetic/semi-synthetic datasets (i.e., IHDP, RKHS, Movement, Heat). We do not possess the counterfactual data of real-world datasets (i.e., Twins, Jobs).

6.2 THE SYMMETRY OF CITA

Our numerical results for the Jobs and Twins datasets verify that the CITA can capture the symmetries within causal inference problems. We flip treatment labels (0 and 1) with probability p (without any changes to the features and the outcomes) independently for each control and treatment data point. In Figure 3, we depict the trend of CITA between the original and the altered dataset by varying p, $p \in [0, 1]$. The symmetry of CITA is evident (with some deviation due to limited training data for calculating CITA). The altered dataset with $p = 1$ is the closest to the original dataset (as it should be) since we have completely flipped the treatment assignments. The altered dataset with $p = 0.5$ is the furthest (as it should be) since we have randomly shuffled the control and the treatment groups. We also compare CITA with the nonsymmetrized task affinity [Le et al., 2022b] on the Jobs and the Twins datasets. Figure 3 shows that CITA has successfully captured the symmetries within causal inference tasks. We observe that CITA demonstrates symmetry at $p = 0.5$, indicating the symmetry of the causal inference tasks. In contrast, the original nonsymmetrized task affinity fails to capture this symmetry property.

6.3 CITA AND THE COUNTERFACTUAL LOSS

In this experiment, we empirically show the strong correlation between CITA (which only uses available data) and counterfactual loss (which is impossible to measure directly except for synthetic datasets). In Figure 4, for different balancing weights α (see Equation 10), we give the correlation between CITA and counterfactual error on the IHDP, RKHS, Movement(Physics), and Heat(Physics) datasets (for which counterfactuals are known). It is both intuitively appealing and empirically observed that CITA and the counterfactual loss have a strong correlation: the model of a source task has a smaller counterfactual loss on the target data if the target task is closer (in terms of CITA). Note that the points in Figure 4 for different values of α (i.e., balancing weight) are extremely close. In other words, CITA highly relates to the counterfactual loss and is robust to hyper-parameters shift. This property is desirable, especially in causal inference scenarios where no validation data can be accessed to cross-validate the hyper-parameters. Additionally, it can also be observed that CITA trends are robust to variations in the balancing weight for all datasets.

6.4 COMPARISON OF PERFORMANCE WITH/WITHOUT TRANSFER LEARNING

This experiment aims to analyze the impact of transferring causal knowledge on the size of required training data. Here, we use Heat (Physics), Movement (Physics), IHDP, and RKHS datasets for which the counterfactual outcomes are available. We first fix a target causal inference task. For a wide range of balancing weights (α), we record the values of ε_{PEHE} for training the model from scratch while increasing the size of training datasets. In this process, the training datasets are slowly expanded such that smaller training sets are subsets of larger ones. We then report the minimum ε_{PEHE} achieved for each dataset size. We identify the closest source task to the target task and repeat the above process with a small amount of target task data. We then compare the performance with and without transfer learning to quantify the amount of data needed by transfer learning models to achieve the best possible performance without transferring causal knowledge. The results are summarized in Table 2, which demonstrates that transferring causal knowledge decreases the required training data in this setting by 75% to 95%.

7 CONCLUSION

In this paper, we provided theoretical analysis proving the transferability of causal knowledge and outlined the underlying challenges. We also proposed a method for ITE transfer learning. Specifically, we constructed CITA, a new task affinity tailored for causal inference tasks, to measure the similarity of causal inference tasks that captures the symmetry within them. Given a new causal inference task, we transferred the ITE knowledge from the closest task between all the available previously trained tasks. Simulations on a representative family of datasets provide empirical evidence for the gains of our method and the efficacy of CITA.

Acknowledgments

This work was supported in part by the National Science Foundation (NSF) under the National AI Institute for Edge Computing Leveraging Next Generation Wireless Networks Grant # 2112562.

References

Virginia Aglietti, Theodoros Damoulas, Mauricio Álvarez, and Javier González. Multi-task causal learning with gaussian processes. *Advances in Neural Information Processing Systems*, 33:6293–6304, 2020.

Ahmed M Alaa and Mihaela Van Der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in neural information processing systems*, 30, 2017.

Zaid Alyafeai, Maged Saeed AlShaibani, and Irfan Ahmad. A survey on transfer learning in natural language processing. *arXiv preprint arXiv:2007.04239*, 2020.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Elias Bareinboim and Judea Pearl. Transportability of causal effects: Completeness results. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 698–704, 2012.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

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.

Ioana Bica and Mihaela van der Schaar. Transfer learning on heterogeneous feature spaces for treatment effects estimation. *arXiv preprint arXiv:2210.06183*, 2022.

Avrim Blum and Tom. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT' 98*, 1998.

Shixing Chen, Caojin Zhang, and Ming Dong. Coupled end-to-end transfer learning with generalized fisher information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4329–4338, 2018.

Rubin Donald. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469): 322–31, 2005.

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

Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A Rusu, Alexander Pritzel, and Daan Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*, 2017.

Chelsea Finn, Xin Yu Tan, Yan Duan, Trevor Darrell, Sergey Levine, and Pieter Abbeel. Deep spatial autoencoders for visuomotor learning. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 512–519. IEEE, 2016.

Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7):761–767, 2011.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029. PMLR, 2016.

Simran Preet Kaur and Vandana Gupta. Covid-19 vaccine: A comprehensive status report. *Virus research*, 288:198114, 2020.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Cat P Le, Yi Zhou, Jie Ding, and Vahid Tarokh. Supervised encoding for discrete representation learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3447–3451. IEEE, 2020.

Cat P. Le, Mohammadreza Soltani, Robert Ravier, and Vahid Tarokh. Improved Automated Machine Learning from Transfer Learning. *arXiv e-prints*, art. arXiv:2103.00241, Feb. 2021.

Cat P Le, Mohammadreza Soltani, Robert Ravier, and Vahid Tarokh. Task-aware neural architecture search. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4090–4094. IEEE, 2021.

Cat P. Le, Mohammadreza Soltani, Juncheng Dong, and Vahid Tarokh. Fisher task distance and its application in neural architecture search. *IEEE Access*, 10:47235–47249, 2022a. doi: 10.1109/ACCESS.2022.3171741.

Cat Phuoc Le, Juncheng Dong, Mohammadreza Soltani, and Vahid Tarokh. Task affinity with maximum bipartite matching in few-shot learning. In *International Conference on Learning Representations*, 2022b.

Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10(1):1–51, 1923.

Arghya Pal and Vineeth N Balasubramanian. Zero-shot task transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2189–2198, 2019.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014.

Daniel L Silver and Kristin P Bennett. Guest editor's introduction: special issue on inductive transfer learning. *Machine Learning*, 73(3):215–220, 2008.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International Conference on Machine Learning*, pages 9120–9132. PMLR, 2020.

Sebastian Thrun and Lorien Pratt. *Learning to learn*. Springer Science & Business Media, 2012.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Thanh Vinh Vo, Pengfei Wei, Trong Nghia Hoang, and Tze Yun Leong. Adaptive multi-source causal inference from observational data. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1975–1985, 2022.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523): 1228–1242, 2018.

Aria Y Wang, Leila Wehbe, and Michael J Tarr. Neural taskonomy: Inferring the similarity of task-derived representations from brain activity. *BioRxiv*, page 708016, 2019.

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

Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

Amir R Zamir, Alexander Sax, William B Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2021. doi: 10.1109/JPROC.2020.3004555.