On the Generalization Gain of Transfer Learning

1st Chenming Cao
Department of Data Science and Artificial Intelligence
The Hong Kong Polytechnic University
Hong Kong SAR
chenm.cao@connect.polyu.hk

2nd Xiaoming Xue
School of Petroleum Engineering
China University of Petroleum (East China)
Qingdao, China
xuexm@upc.edu.cn

3rd Liang Feng
College of Computer Science
Chongqing University
Chongqing, China
liangf@cqu.edu.cn

4th Yao Hu
Department of Data Science and Artificial Intelligence
The Hong Kong Polytechnic University
Hong Kong SAR
echo-yao.hu@polyu.edu.hk

5th Kay Chen Tan
Department of Data Science and Artificial Intelligence
The Hong Kong Polytechnic University
Hong Kong SAR
kctan@polyu.edu.hk

***Abstract*—Transfer learning (TL) has proven highly effective at improving generalization, yet theoretical insights lag behind its rapid methodological advances. This study attempts to reduce this research gap by establishing a unified theoretical framework for analogy-based TL, grounded in a commonly used concept known as *similarity*. We systematically map the three core subprocesses, i.e., retrieval, adaptation and evaluation, to the fundamental challenges in TL: what, how and when to transfer. Building upon this framework, we derive two theorems governing the gain in generalization performance of analogy-based TL: (1) the theorem of unconditionally non-negative gain, which ensures safety against negative transfer through accurate evaluation; and (2) the theorem of conditionally positive gain, contingent upon either large-scale source retrieval or effective adaptation. Furthermore, we discuss the compatibility of this framework with the No Free Lunch theorem, clarifying that the efficacy of TL arises from the alignment of analogy-based inductive biases with the non-uniform distribution of problems, rather than a violation of the theorem. These insights theoretically rationalize the paradigm shift from seeking universal transferability metrics to exploiting specialized inductive biases.**

***Index Terms*—transfer learning, similarity, analogical reasoning, no free lunch theorem.**

I. Introduction

Transfer Learning (TL) [1, 2] has emerged as a pivotal paradigm to extend the success of machine learning [3, 4] by transferring knowledge from related source problems to a target problem [5, 6]. Despite rapid algorithmic advancements [7, 8], a unified theoretical framework explaining their efficacy remains unestablished. Current theoretical explorations are largely fragmented. Classical statistical learning bounds rely on domain divergence measures [9], algorithmic stability ensures convergence in restricted settings [10, 11], and information-theoretic metrics quantify transferability [12, 13]. While these frameworks offer valuable explanations for constrained settings, such as representation learning [14] or deep linear dynamics [15], they have yet to be synthesized into a unified theory that systematically examines the generalization gain of knowledge transfer.

Heuristic TL algorithms inherently mimic human cognitive reasoning [16–18], offering a valuable lens to bridge the gap between empirical success and theoretical interpretability. Contemporary TL methodologies implicitly employ diverse reasoning paradigms, including analogy (e.g., instance-based reweighting [19]), induction (e.g., meta-learning [20] and pre-training [21]), deduction (e.g., zero-shot learning [22]), abduction (e.g., inverse reinforcement learning [23]), and causality (e.g., invariant learning [24]). In this study, we show particular interest in analogy-based knowledge transfer due to its extensive use in the literature. Although preliminary explorations exist in specific domains like evolutionary optimization [25], a generalized theoretical framework within machine learning remains unestablished. We posit that analogical reasoning, the cognitive process of recognizing structural parallels between contexts, serves as the cornerstone underpinning successful knowledge transfer. By formalizing this, we systematically link the three core subprocesses of analogy to TL's fundamental issues: *retrieval* determines what to transfer by quantifying source relevance [21, 26]; *adaptation* resolves how to transfer by transforming knowledge to bridge context discrepancies [27, 28]; and *evaluation* decides when to transfer to verify validity and prevent negative transfer [29].

Building upon this framework, we derive two theorems governing the generalization gain of analogy-based TL: (1) unconditionally non-negative gain, providing a safety guarantee against negative transfer through accurate evaluation; and (2) conditionally positive gain, contingent upon large-scale source retrieval or effective adaptation. Furthermore, we interpret TL through the lens of the No Free Lunch (NFL) theorem [30], demonstrating that TL's superiority is not unconditional but relies on aligning analogy-based inductive biases with the target problem's latent structure. This insight theoretically explains why transfer strategies succeed in specific contexts but fail in others.

The remainder of this paper is organized as follows. Section II introduces the definition of TL and analogical reasoning.

This work was supported in part by the Research Grants Council of the Hong Kong SAR (Grant No. C5052-23G, PolyU15229824, SRFS2526-5S04), and The Hong Kong Polytechnic University (Project IDs: P0051130, P0060651, P0058445). (Corresponding author: Yao Hu)

Section III constructs the theoretical foundations. Section IV presents the generalization gain theorems and NFL implications. Section V concludes the paper.

II. Preliminaries

This section establishes the mathematical foundations for our analysis. We first define TL and subsequently outline the framework of analogical reasoning, constructing the logical basis for the theories presented in Section III.

A. Transfer Learning

TL is a learning paradigm that aims to improve the generalization performance of a learner on a target problem by leveraging knowledge extracted from relevant source problems. To provide a rigorous formulation, a domain $\mathcal{A} = \{\mathcal{X}, P(X)\}$ consists of a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$. Given a specific domain, a task $\mathcal{T} = \{\mathcal{Y}, P(Y|X)\}$ consists of a label space $\mathcal{Y}$ and a conditional probability distribution $P(Y|X)$. We define a problem $\mathcal{P}$ as the pair $\mathcal{P} = \{\mathcal{A}, \mathcal{T}\}$.

Formally, given a source problem $\mathcal{P}_{src} = \{\mathcal{A}_{src}, \mathcal{T}_{src}\}$ and a target problem $\mathcal{P}_{tgt} = \{\mathcal{A}_{tgt}, \mathcal{T}_{tgt}\}$ where $\mathcal{P}_{src} \neq \mathcal{P}_{tgt}$, the goal of TL is to improve the estimation of $P_{tgt}(Y|X)$ using the knowledge derived from $\mathcal{P}_{src}$. The condition $\mathcal{P}_{src} \neq \mathcal{P}_{tgt}$ implies a discrepancy arising from either domain divergence ($\mathcal{A}_{src} \neq \mathcal{A}_{tgt}$) or task divergence ($\mathcal{T}_{src} \neq \mathcal{T}_{tgt}$). The implementation of effective TL algorithms typically necessitates addressing three issues [16]:

- What to Transfer: Quantifying the relevance of knowledge components derived from $\mathcal{P}_{src}$ to estimate their potential utility for $\mathcal{P}_{tgt}$.
- How to Transfer: Devising algorithmic mechanisms to operationalize knowledge transfer, aimed at improving its utility for $\mathcal{P}_{tgt}$.
- When to Transfer: Assessing the transferability (i.e., the utility of knowledge) to determine the timing of transfer, thereby safeguarding $\mathcal{P}_{tgt}$ against negative transfer.

B. Analogical Reasoning

Analogical reasoning is a cognitive process that infers the latent utility of a source candidate for a target problem based on their similarity. In the computational realm, analogy-based TL capitalizes on such similarity to facilitate knowledge transfer. The specific conceptualization of similarity manifests differently across contexts: in visual domain adaptation, it is formulated as distributional proximity [31]; similarly, in natural language processing, it takes the form of semantic correspondence [32]. Formally, let w denote the observable property, and let $s(w)$ be a valid similarity metric. For notational brevity, we refer to this joint similarity simply as s where context permits. The fundamental logic of analogy-based TL can be articulated as follows:

Rationale: *If two problems $\mathcal{P}_{src}$ and $\mathcal{P}_{tgt}$ share high similarity in terms of an observable property w, and $\mathcal{P}_{src}$ possesses knowledge v associated with w, then the knowledge v is likely to yield high utility for $\mathcal{P}_{tgt}$.*

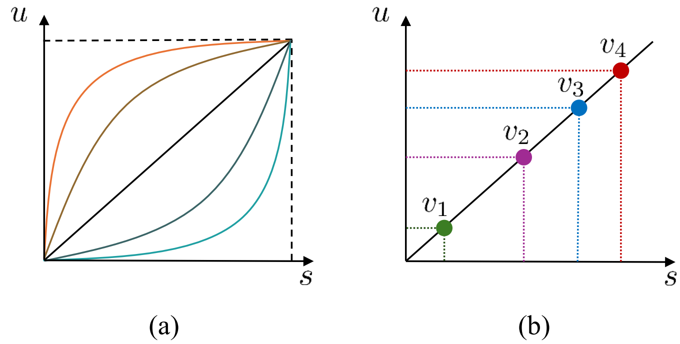


Fig. 1: Illustration of the retrieval subprocess: (a) different curvatures of f; (b) the retrieval process.

This rationale hinges on a positive correlation between the observable property w (e.g., feature distributions $P(X)$) and the latent utility $u(v)$, defined as the generalization performance on the target problem. While observable similarity is computationally accessible, the true utility $u(v)$ is typically unknown prior to transfer. To operationalize this rationale, the reasoning process is decomposed into three sequential subprocesses:

- Retrieval: Quantifying the similarity between the observable properties of the source and target to identify candidates likely to afford high utility.
- Adaptation: Adapting the source knowledge to improve its observable similarity to the target, thereby enhancing its compatibility and expected utility.
- Evaluation: Assessing the predicted utility against a baseline to verify the validity of the transfer, serving as a gatekeeper to safeguard against negative transfer.

III. Analogy-Based Knowledge Transfer

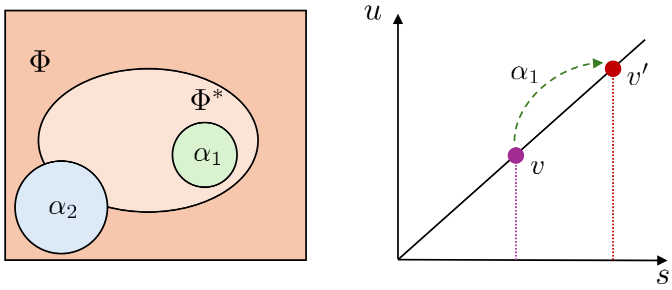
In this section, we construct a unified theoretical framework for analogy-based TL. We postulate that the heuristic success of TL can be formalized through mathematical principles governing the three cognitive subprocesses, providing a solid logical basis for understanding effective transfer mechanisms.

A. What to Transfer: Retrieval

The retrieval subprocess addresses what to transfer by bridging the informational gap between observable characteristics and latent transfer efficacy. Since the true utility $u(v)$ is typically unknown, retrieval operates on the premise that the similarity $s(w)$ serves as a reliable proxy.

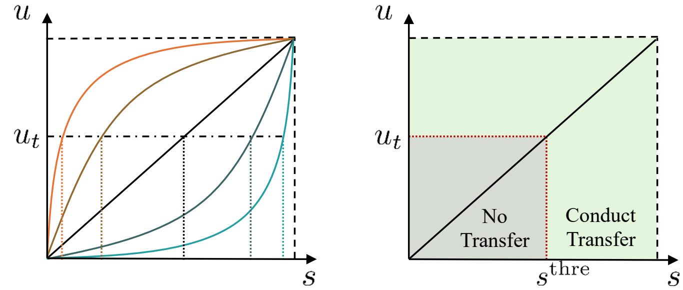
Principle 1. *Subject to the existence of an intrinsic dependency between the observable property w and the transferable knowledge v, the latent utility $u(v)$ of knowledge derived from $\mathcal{P}_s$ is monotonically increasing with the similarity $s(w)$ between the observable property w of $\mathcal{P}_{src}$ and $\mathcal{P}_{tgt}$.*

To formalize this, let S and U denote the continuous spaces of similarity and utility, respectively. We postulate a mapping function $f : S \to U$ such that $u = f(s)$. The validity of Principle 1 rests upon two premises regarding f:



(a) (b)

Fig. 2: Illustration of the adaptation subprocess: (a) algorithmic efficacy in hypothesis space; (b) utility enhancement.



(a) (b)

Fig. 3: Illustration of the evaluation subprocess: (a) threshold determination; (b) the gating mechanism.

- Bijectivity: $f : S \to U$ is bijective, ensuring a strict one-to-one correspondence between observable similarity and latent utility.
- Continuity: f is continuous over S ($\lim_{s \to s_0} f(s) = f(s_0)$), guaranteeing that infinitesimal similarity perturbations do not cause abrupt utility discontinuities.

Theorem 1. *Given the premises of bijectivity and continuity, the function $f : S \to U$ is strictly monotonic. Consequently, the rank of the latent utility u is strictly preserved by the observable similarity $s(w)$.*

Proof. Assume the bijective, continuous function $f : S \to U$ is not strictly monotonic. There exist $s_1 < s_2 < s_3$ where the order is not preserved. Without loss of generality, assume $f(s_2) > \max(f(s_1), f(s_3))$. Choose γ such that $\max(f(s_1), f(s_3)) < \gamma < f(s_2)$. By the Intermediate Value Theorem, there exist $s_a \in (s_1, s_2)$ and $s_b \in (s_2, s_3)$ satisfying $f(s_a) = f(s_b) = \gamma$. Since $s_a \neq s_b$, this contradicts the bijectivity of f. Thus, f must be strictly monotonic. Given the rationale that higher similarity correlates with higher utility, f is strictly increasing. □

Remark 1. *The bijective mapping f is an idealized abstraction. In practice, latent confounders render the similarity-utility relationship stochastic, causing local violations of rank preservation. However, given a strong underlying correlation between w and u, the global monotonic trend persists. This insight theoretically justifies the empirical efficacy of top-k ensembles and soft-attention mechanisms in absorbing such local ranking noise.*

As visualized in Fig. 1, while the explicit analytical form of f may vary, monotonicity guarantees that the ordinal ranking of knowledge candidates is strictly preserved. This rank preservation theoretically justifies using similarity as a proxy to prioritize source candidates. It underpins strategies like direct source selection [33, 34], ensemble-based continuous reweighting [19, 35], and adaptive knowledge representation selection [21, 36].

B. How to Transfer: Adaptation

The adaptation subprocess addresses how to transfer by mitigating context discrepancies, aiming to transform source knowledge to enhance compatibility with $\mathcal{P}_{tgt}$. Guided by the monotonicity established in Principle 1, we synthesize the core logic of this process into the following principle:

Principle 2. *The utility of knowledge v derived from a source problem $\mathcal{P}_{src}$ for a target problem $\mathcal{P}_{tgt}$ can be enhanced by transforming the knowledge v such that its associated observable property w exhibits higher similarity to the target.*

The validity of Principle 2 relies on two preconditions regarding the hypothesis space of transformation functions Φ:

- Realizability: The set of valid adaptation functions $\Phi^* = \{\phi \in \Phi \mid s(w') > s(w)\}$ is non-empty, where w' denotes the observable property of the adapted knowledge $v' = \phi(v)$, theoretically allowing for similarity enhancement.
- Reachability: An algorithm exists that consistently returns a candidate $\phi^* \in \Phi^*$, making enhancement achievable.

Theorem 2. *Given the realizability and reachability of an adaptation set Φ, the utility of source knowledge v can be enhanced by adapting the knowledge using $\phi \in \Phi$ that increases the similarity of its observable property to $\mathcal{P}_{tgt}$.*

Proof. Realizability guarantees an adaptation function $\phi \in \Phi$ yielding $s(w') > s(w)$. Since $f : S \to U$ is strictly increasing (Theorem 1), $s(w') > s(w)$ strictly implies $u(v') > u(v)$. Reachability ensures ϕ is algorithmically discoverable, guaranteeing utility improvement. □

Remark 2. *A fundamental trade-off exists between realizability and reachability via the capacity of Φ. Expanding Φ guarantees realizability but compromises reachability due to intractable optimization and degenerate transformations. Conversely, restricting Φ improves reachability but risks misspecification. Thus, effective adaptation requires a meticulous balance between expressivity and structural inductive biases to reliably discover utility-enhancing transformations.*

Fig. 2 illustrates this dynamic: algorithm α_1 satisfies reachability by guaranteeing $\phi^* \in \Phi^*$, mapping the knowledge state along the increasing curve $u = f(s)$. This asserts that adaptation fundamentally improves observable similarity to secure utility gain. It rationalizes explicit feature alignment (e.g., TCA [27], CORAL [37]) and instance reweighting (e.g., KMM [38]), where assigned weights align distributions to mitigate covariate shift.

C. When to Transfer: Evaluation

The evaluation subprocess addresses the critical decision-making challenge of when to transfer, which verifies whether problem proximity is sufficient to guarantee a positive outcome over the isolated learning baseline. Guided by the similarity-utility correlation, we explicitly state the principle governing this decision as follows:

Principle 3. *The potential impact (positive or negative) of transferring knowledge v from $\mathcal{P}_{src}$ to $\mathcal{P}_{tgt}$ can be assessed based on the observable similarity between the two problems.*

The validity of Principle 3 is mathematically guaranteed by the existence of a decision boundary in the similarity space, as formalized below:

Theorem 3. *There exists a critical similarity threshold s^{thre} that strictly demarcates positive transfer from negative transfer. Source knowledge contributes to a generalization gain if and only if its observable similarity to the target problem exceeds this threshold.*

Proof. Let $u_{tgt} \in U$ denote the baseline utility achieved on the target problem without transfer. Knowledge v is useful if $u(v) > u_{tgt}$. Since f is strictly increasing, its inverse $f^{-1} : U \to S$ is also strictly increasing. Defining the threshold $s^{\text{thre}} = f^{-1}(u_{tgt})$, the usefulness condition equates to:

$$f(s(w)) > u_{tgt} \iff s(w) > f^{-1}(u_{tgt}) \equiv s^{\text{thre}}. \quad (1)$$

Thus transfer yields a positive gain if the measured similarity $s(w)$ exceeds s^{thre}. □

Remark 3. *A precise threshold s^{thre} requires an accurate baseline utility u_{tgt}, which is empirically challenging. Estimation variance blurs this boundary, softening the strict safety guarantee and risking either negative transfer or unwarranted rejection. This inherent fuzziness suggests that rigid binary decisions may be structurally vulnerable, indicating the potential value of probabilistic gating or continuous confidence margins in practice.*

As shown in Fig. 3, s^{thre} acts as a hard decision boundary. Enforcing transfer only when $s(w) > s^{\text{thre}}$ theoretically guarantees a positive gain. This justifies safe transfer strategies, underpinning explicit analytical metrics (e.g., $\mathcal{H}\Delta\mathcal{H}$-divergence [40], LEEP [41]) and implicit gating mechanisms in multi-task architectures [42, 43], which suppress irrelevant features below the utility boundary.

IV. Analyses of Generalization Gain

Building upon the theoretical framework established in Section III, this section provides an analysis of the generalization gain Δ, defined as the net improvement in generalization performance relative to the isolated learning baseline:

$$\Delta = u(v) - u_{tgt}. \quad (2)$$

By leveraging the logical consistency of the three subprocesses, we derive two fundamental theorems governing the bounds of Δ. First, we prove that the evaluation mechanism imposes a theoretical lower bound, ensuring an *unconditionally non-negative gain*. Second, we demonstrate that large-scale source retrieval or effective adaptation provide sufficient conditions for a *conditionally positive gain*. Finally, we analyze the framework's compliance with the NFL theorem.

A. Unconditionally Nonnegative Gain

A pervasive concern in TL is negative transfer, where external knowledge degrades performance relative to the isolated learning baseline. To theoretically preclude this risk, explicitly incorporating an evaluation subprocess is a prerequisite. Within the proposed framework, this subprocess functions as a critical safeguard. By enforcing the similarity-based decision boundary derived in Theorem 3, we prove that this evaluation mechanism is the sufficient condition to ensure the resulting generalization gain is bounded from below by zero.

Theorem 4. *For any TL strategy incorporating an evaluation mechanism governed by s^{thre}, the generalization gain Δ is unconditionally non-negative, i.e., $\Delta \geq 0$.*

Proof. The evaluation mechanism enforces conditional transfer execution based on the observable similarity $s(w)$. If $s(w) > s^{\text{thre}}$, the transfer is executed; according to Theorem 3, this implies $u(v) > u_{tgt}$, yielding a strictly positive generalization gain $\Delta = u(v) - u_{tgt} > 0$. Conversely, if $s(w) \leq s^{\text{thre}}$, the transfer is rejected and the system defaults to the isolated learning baseline, resulting in a zero gain $\Delta = u_{tgt} - u_{tgt} = 0$. Encompassing both mutually exclusive scenarios, the generalization gain unconditionally satisfies $\Delta \geq 0$. □

Remark 4. *Building upon the evaluation fuzziness (Remark 3), the unconditional mathematical guarantee of $\Delta \geq 0$ is empirically vulnerable to estimation noise. To robustly uphold this safety bound in practice, algorithms could rely on an asymmetrically conservative decision boundary, effectively inflating the threshold. This highlights a fundamental theoretical trade-off: prioritizing empirical safety typically entails sacrificing marginal positive transfer opportunities, explaining why safe-transfer mechanisms generally yield conservative overall gains.*

Theorem 4 establishes the theoretical imperative for integrating evaluation mechanisms into TL systems. By strictly adhering to this validation, the operational paradigm of TL shifts from a risky endeavor to a structurally safe enhancement process. The worst-case scenario reduces to isolated learning, while executed transfers guarantee a net positive improvement. This principle finds broad application in contemporary safe transfer methodologies [29, 42], which employ gating strategies to ensure robustness and allow aggressive source exploration without performance regression.

B. Conditionally Positive Gain

While evaluation provides a safety net against negative transfer, achieving a strictly positive gain is conditional. It necessitates strategic investment: the system must either possess a sufficiently large source pool to statistically guarantee high-utility knowledge, or employ a powerful algorithm to appropriately adapt the knowledge. We formalize these sufficient conditions for positive gain as follows:

Theorem 5. *The analogy-based transfer framework yields a strictly positive gain if the system satisfies at least one of the following conditions:*

1) *Large-Scale Retrieval: The set of available source candidates $\mathcal{V}$ is sufficiently large. As $|\mathcal{V}| \to \infty$, the probability of identifying a source naturally satisfying $s(w) > s^{thre}$ approaches 1.*
2) *Effective Adaptation: The hypothesis space Φ contains a valid transformation satisfying $\exists v \in \mathcal{V}, \exists \phi^\dagger \in \Phi$ such that $s(w') > s^{thre}$, and the employed algorithm is capable of identifying $\phi^\dagger$ to guarantee $\Delta > 0$.*

Proof. We analyze the condition $u(v) > u_{tgt}$ through two complementary lenses. First, regarding source scale, assume candidates in $\mathcal{V}$ are i.i.d. samples drawn from a broad task distribution $\mathcal{D}_{\text{task}}$. Let $\epsilon = P(s(w) > s^{\text{thre}}) > 0$ denote the probability that a random source naturally satisfies the threshold condition. The probability of realizing a positive gain within a retrieval pool of size $N = |\mathcal{V}|$ is $P(\Delta > 0) = 1 - (1-\epsilon)^N$. As $N \to \infty$, this probability converges to 1, statistically guaranteeing a positive gain.

Alternatively, the guarantee relies on the function space Φ. The expressive capacity of Φ theoretically ensures a transformation $\phi^\dagger$ capable of satisfying the threshold condition, while the algorithm's effectiveness ensures this solution is practically identifiable. By applying $\phi^\dagger$, the system adapts knowledge to satisfy $s(w') > s^{\text{thre}}$, strictly ensuring $\Delta > 0$. □

Remark 5. *The statistical guarantee for large-scale retrieval presumes source problems are i.i.d. samples from a broad distribution. In reality, empirical task spaces are highly skewed or clustered. For targets in sparse out-of-distribution regions, the probability ϵ of retrieving a naturally valid source approaches zero. This structural sparsity demonstrates that retrieval scale alone cannot overcome severe shifts, theoretically reinforcing the complementary role of robust adaptation.*

Theorem 5 elucidates the strategic trade-off between source scale and algorithmic complexity. The first condition rationalizes foundation models [44], where extensive source pools maximize the probability of locating sources with high inherent similarity, diminishing reliance on complex adaptation. Conversely, the second condition grounds deep domain adaptation [31], leveraging high-capacity networks and advanced objectives [45] to actively minimize distribution shifts in data-constrained scenarios.

C. Compliance with the No Free Lunch Theorem

Seemingly, precluding negative transfer while securing conditional improvement implies universal superiority over isolated learning, which seemingly contradicts the premise of the NFL theorem. However, this superiority reflects a conditional advantage derived from strategically aligning the algorithm's inductive bias with the distribution of problems of interest [30, 46].

The NFL theorem posits zero expected performance difference between algorithms when averaged over a uniform distribution $\mathcal{D}$ of all possible problems Ω. Let $\Delta(\mathcal{P}, \beta)$ denote the generalization gain on instance $\mathcal{P} \in \Omega$ given inductive bias β. The NFL implies:

$$\mathbb{E}_{\mathcal{P}\sim\mathcal{D}}[\Delta(\mathcal{P}, \beta)] = 0. \tag{3}$$

In practice, we target a subset of problems governed by a non-uniform distribution $\tilde{\mathcal{D}}$. The empirical success of analogy-based TL arises because the inductive bias β aligns with the structure of problems in $\tilde{\mathcal{D}}$. Consequently, the expected gain is strictly positive over the distribution of interest:

$$\mathbb{E}_{\mathcal{P}\sim\tilde{\mathcal{D}}}[\Delta(\mathcal{P}, \beta)] > 0. \tag{4}$$

This positive expectation reflects the effective incorporation of domain-specific insights rather than an NFL violation.

We exemplify this dependency through evaluation. While Theorem 4 guarantees safety, this hinges on the validity of observable similarity $s(w)$ as a proxy for latent utility $u(v)$. The metric $s(\cdot)$ constitutes a strong inductive bias. For instance, a geometric distance metric may align effectively with physical motion tasks but remain orthogonal to semantic reasoning. If $s(\cdot)$ fails to align with the problem structure, the evaluation mechanism may yield incorrect judgments, nullifying the safety guarantee.

These insights suggest that pursuing a universal transferability metric is impractical. Effective TL algorithm design entails carefully engineering retrieval metrics, adaptation functions, and evaluation mechanisms to maximize alignment with the specific distribution of target problems.

V. Conclusion

In this study, we have established a unified theoretical framework for analogy-based TL to narrow the gap between the rapid empirical success of heuristic algorithms and the lack of a unified theoretical understanding. We formalized analogical reasoning within the context of machine learning, systematically mapping its three core subprocesses to the fundamental challenges of what, how, and when to transfer. Beyond these definitions, our analysis offers constructive design insights for analogy-based TL systems, advocating for the integration of metrics capturing deep structural topology in retrieval, alignment-based objectives in adaptation, and gating mechanisms for safe transfer.

Building upon this framework, we derived two theorems governing the generalization gain of analogy-based TL. First, we demonstrated that an unconditionally non-negative gain is strictly guaranteed by integrating an evaluation mechanism to prevent negative transfer. Second, we proved that a conditionally positive gain can be guaranteed, contingent upon either large-scale source retrieval or effective adaptation.

Crucially, we reconciled the relationship between TL and the NFL theorem. Our analysis demonstrates that the efficacy of analogy-based TL over isolated learning does not violate the theorem; rather, it is a conditional advantage stemming from the strategic alignment of analogy-based inductive biases with the non-uniform distribution of problems. This insight theoretically rationalizes the paradigm shift from seeking universal transferability metrics to designing specialized inductive biases, underscoring that effective transfer naturally hinges on the precise alignment of these priors with the specific distribution of target problems.

In future work, we plan to extend these foundations by incorporating causal inference to address the critical challenge of transferring invariant mechanisms across non-stationary environments.

References

[1] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, p. 9, 2016.

[2] C. Yang, Y.-m. Cheung, J. Ding, and K. C. Tan, "Concept drift-tolerant transfer learning in dynamic environments," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 8, pp. 3857–3871, 2021.

[3] R. Ren, T. Hung, and K. C. Tan, "A generic deep-learning-based approach for automated surface inspection," *IEEE Transactions on Cybernetics*, vol. 48, no. 3, pp. 929–940, 2017.

[4] R. Liu, Z.-A. Huang, Y. Hu, Z. Zhu, K.-C. Wong, and K. C. Tan, "Spatial–temporal co-attention learning for diagnosis of mental disorders from resting-state fmri data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 8, pp. 10 591–10 605, 2023.

[5] B. Neyshabur, H. Sedghi, and C. Zhang, "What is being transferred in transfer learning?" *Advances in Neural Information Processing Systems*, vol. 33, pp. 512–523, 2020.

[6] Y. Hu, Z.-A. Huang, R. Liu, X. Xue, X. Sun, L. Song, and K. C. Tan, "Source free semi-supervised transfer learning for diagnosis of mental disorders on fmri scans," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 13 778–13 795, 2023.

[7] C. Guokun, M. Jiang, G. Xing, H. Weizhen, G. Shihui, and K. C. Tan, "Online bagging for anytime transfer learning," in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2019, pp. 941–947.

[8] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[9] Z. Wang, "Theoretical guarantees of transfer learning," *arXiv preprint arXiv:1810.05986*, 2018.

[10] I. Kuzborskij and F. Orabona, "Stability and hypothesis transfer learning," in *International Conference on Machine Learning*. PMLR, 2013, pp. 942–950.

[11] M. Perrot and A. Habrard, "A theoretical analysis of metric hypothesis transfer learning," in *International Conference on Machine Learning*. PMLR, 2015, pp. 1708–1717.

[12] X. Wu, J. H. Manton, U. Aickelin, and J. Zhu, "On the generalization for transfer learning: An information-theoretic analysis," *IEEE Transactions on Information Theory*, 2024.

[13] Y. Bao, Y. Li, S.-L. Huang, L. Zhang, L. Zheng, A. Zamir, and L. Guibas, "An information-theoretic approach to transferability in task transfer learning," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 2309–2313.

[14] N. Tripuraneni, M. Jordan, and C. Jin, "On the theory of transfer learning: The importance of task diversity," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7852–7862, 2020.

[15] A. K. Lampinen and S. Ganguli, "An analytic theory of generalization dynamics and transfer learning in deep linear networks," *arXiv preprint arXiv:1809.10374*, 2018.

[16] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[17] D. Gentner, "Structure-mapping: A theoretical framework for analogy," *Cognitive Science*, vol. 7, no. 2, pp. 155–170, 1983.

[18] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behavioral and Brain Sciences*, vol. 40, p. e253, 2017.

[19] Y. Yao and G. Doretto, "Boosting for transfer learning with multiple sources," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1855–1862.

[20] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1126–1135.

[21] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[22] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 453–465, 2013.

[23] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proceedings of the Twenty-First International Conference on Machine Learning*, 2004, p. 1.

[24] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[25] X. Xue, L. Feng, Y. Feng, R. Liu, K. Zhang, and K. C. Tan, "A theoretical analysis of analogy-based evolutionary transfer optimization," in *2025 IEEE Congress on Evolutionary Computation (CEC)*, 2025, pp. 1–8.

[26] J. Li, T. Sun, Q. Lin, M. Jiang, and K. C. Tan, "Reducing negative transfer learning via clustering for dynamic multiobjective optimization," *IEEE Transactions on Evolutionary Computation*, vol. 26, no. 5, pp. 1102–1116, 2022.

[27] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2010.

[28] L. Chen, H.-L. Liu, K. C. Tan, and K. Li, "Transfer learning-based parallel evolutionary algorithm framework for bilevel optimization," *IEEE Transactions on Evolutionary Computation*, vol. 26, no. 1, pp. 115–129, 2021.

[29] Z. Wang, Z. Dai, B. Póczos, and J. Carbonell, "Characterizing and avoiding negative transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 293–11 302.

[30] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," *Neural Computation*, vol. 8, no. 7, pp. 1341–1390, 1996.

[31] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International Conference on Machine Learning*. PMLR, 2015, pp. 97–105.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[33] M. J. Afridi, A. Ross, and E. M. Shapiro, "On automated source selection for transfer learning in convolutional neural networks," *Pattern Recognition*, vol. 73, pp. 65–75, 2018.

[34] H. Larsson, J. Taghia, F. Moradi, and A. Johnsson, "Source selection in transfer learning for improved service performance predictions," in *2021 IFIP Networking Conference (IFIP Networking)*. IEEE, 2021, pp. 1–9.

[35] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation with multiple sources," *Advances in Neural Information Processing Systems*, vol. 21, 2008.

[36] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *International Conference on Machine Learning*. PMLR, 2014, pp. 647–655.

[37] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[38] J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, and A. Smola, "Correcting sample selection bias by unlabeled data," *Advances in Neural Information Processing Systems*, vol. 19, 2006.

[39] M. Sugiyama, S. Nakajima, H. Kashima, P. Buenau, and M. Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," *Advances in Neural Information Processing Systems*, vol. 20, 2007.

[40] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine Learning*, vol. 79, no. 1, pp. 151–175, 2010.

[41] C. Nguyen, T. Hassner, M. Seeger, and C. Archambeau, "Leep: A new measure to evaluate transferability of learned representations," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7294–7305.

[42] J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi, "Modeling task relationships in multi-task learning with multi-gate mixture-of-experts," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1930–1939.

[43] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert, "Cross-stitch networks for multi-task learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3994–4003.

[44] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[45] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[46] S. Hanneke and S. Kpotufe, "A no-free-lunch theorem for multitask learning," *The Annals of Statistics*, vol. 50, no. 6, pp. 3119–3143, 2022.