Theoretical Analysis

No Author Given

No Institute Given

In this section, we provide a solid theoretical foundation for the core mechanism of our framework: wise neighbor-domain selection. Our analysis is presented from two complementary perspectives. First, drawing from classical domain adaptation theory, we establish that our selective transfer strategy achieves a tighter generalization error bound. Second, leveraging information-theoretic principles, we fundamentally quantify the performance gap, thereby demonstrating the intrinsic superiority of our selection strategy in mitigating negative transfer.

A. Generalization Bound Analysis via Domain Discrepancy

We begin by adopting the standard framework from the theory of learning from different domains Ben-David et al. [2010]. Let $\mathcal{H}$ be a hypothesis space, with $\mathcal{D}_t$ and $\mathcal{D}_s$ representing the data distributions for the target and source domains, respectively. For any hypothesis $h \in \mathcal{H}$, its expected risk on the target domain, $\epsilon_t(h)$, is bounded by:

$$\epsilon_t(h) \leq \epsilon_s(h) + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t) + \lambda$$

Here, $d_{\mathcal{H}\Delta\mathcal{H}}(\cdot,\cdot)$ is the $\mathcal{H}\Delta\mathcal{H}$-divergence Acuna et al. [2021], quantifying the distributional discrepancy, and λ is the ideal joint risk. This bound underscores that minimizing the source-target discrepancy is paramount for effective knowledge transfer.

We analyze two distinct transfer strategies:

1. **Blind Full-Domain Transfer**: This approach aggregates knowledge from all K available source domains, $\{\mathcal{D}_{s_k}\}_{k=1}^{K}$. The effective source distribution becomes a mixture, $\mathcal{D}_{\text{all}} = \sum_{k=1}^{K} w_k \mathcal{D}_{s_k}$. Its generalization bound is:
$$\epsilon_t(h) \leq \epsilon_{\text{all}}(h) + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{\text{all}}, \mathcal{D}_t) + \lambda_{\text{all}}$$
2. **Our Selective Transfer**: Our framework intelligently selects an optimal subset of domains, $\mathcal{S}_{\text{sel}} \subset \{\mathcal{D}_{s_k}\}_{k=1}^{K}$. The effective source distribution is thus $\mathcal{D}_{\text{sel}} = \sum_{s_j \in \mathcal{S}_{\text{sel}}} w'_j \mathcal{D}_{s_j}$. The tighter bound is:
$$\epsilon_t(h) \leq \epsilon_{\text{sel}}(h) + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{\text{sel}}, \mathcal{D}_t) + \lambda_{\text{sel}}$$

Theorem 1. *Our selection mechanism, guided by a metric designed to approximate domain discrepancy, aims to minimize $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{sel}, \mathcal{D}_t)$. In the presence of irrelevant source domains, this strategy is guaranteed to yield a strictly smaller discrepancy term compared to blind transfer, i.e., $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{sel}, \mathcal{D}_t) < d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{all}, \mathcal{D}_t)$, thus achieving a tighter generalization bound and providing a theoretical safeguard against negative transfer.*

B. Information-Theoretic Quantification of the Performance Gap

We now delve deeper by quantifying the performance improvement of selective transfer over blind transfer.

Let us define the notation. Let $\mathcal{D} = \{1, \ldots, D\}$ be the set of all domains. For each domain $i \in \mathcal{D}$, let X^i and Y^i denote its feature data and labels, respectively. We assume an underlying joint distribution $p(X^1, \ldots, X^D, Y^1, \ldots, Y^D)$. For a target domain i, our model selects a subset of source domains $S^i \subseteq \mathcal{D}$ and learns a representation $Z^i = g^i(\{X^j\}_{j \in S^i})$.

First, we establish the fundamental connection between the cross-entropy loss and conditional entropy.

Lemma 1. (cf. Lemma 4 in Wen et al. [2025], Cover and Thomas [2006]) *For any data representation $X_{\mathcal{S}}$ from a set of domains $\mathcal{S} \subseteq \mathcal{D}$ and a target label Y^i, the infimum of the expected cross-entropy loss for any predictive function f is equal to the conditional entropy:*

$$\inf_f \mathbb{E}\left[\ell_{ce}(f(X_{\mathcal{S}}), Y^i)\right] = H(Y^i|X_{\mathcal{S}})$$

Proof. The expected cross-entropy loss can be decomposed into the Kullback-Leibler divergence and the conditional entropy:

$$\mathbb{E}[\ell_{ce}] = \mathbb{E}_{X_{\mathcal{S}}}[D_{KL}(p(Y^i|X_{\mathcal{S}}) \,||\, f(Y^i|X_{\mathcal{S}}))] + H(Y^i|X_{\mathcal{S}})$$

. Since $D_{KL} \geq 0$, the infimum is achieved when f perfectly matches the true conditional distribution, yielding the conditional entropy $H(Y^i|X_{\mathcal{S}})$.

We now present the main theorem, which provides a lower bound on the performance gain achieved by domain selection.

Theorem 2. *Let $Z^i = g^i(\{X^j\}_{j \in S^i})$ be the representation learned by our selective model. The performance gap between using all domains' raw data versus our learned selective representation is lower-bounded by:*

$$\inf_h \mathbb{E}[\ell_{ce}(h(Z^i), Y^i)] - \inf_{h'} \mathbb{E}[\ell_{ce}(h'(X^1, \ldots, X^D), Y^i)] \geq \Delta^i$$

where the information gap Δ^i is defined as:

$$\Delta^i = I(X^i; Y^i) - \min_{j \in S^i} I(X^j; Y^i)$$

assuming that the most informative domain for task i is domain i itself (i.e., $i = \arg\max_k I(X^k; Y^i)$).

Proof. Following the proof structure in Wen et al. [2025], we first apply Lemma 1 to convert the difference in optimal risks into a difference in conditional entropies:

$$\text{Gap} = H(Y^i|Z^i) - H(Y^i|X^1, \ldots, X^D)$$

Using the identity $H(A|B) = H(A) - I(A; B)$, this becomes:

$$\begin{aligned}\text{Gap} &= \big(H(Y^i) - I(Y^i; Z^i)\big) - \big(H(Y^i) - I(Y^i; X^1, \ldots, X^D)\big) \\ &= I(X^1, \ldots, X^D; Y^i) - I(Z^i; Y^i)\end{aligned}$$

By the data-processing inequality, since Z^i is a function of $\{X^j\}_{j\in S^i}$, we have $I(Z^i;Y^i) \leq I(\{X^j\}_{j\in S^i};Y^i)$. The proof in Wen et al. [2025] establishes the following chain of inequalities:

$$I(Z^i;Y^i) \leq \min_{j\in S^i}\{I(X^j;Y^i)\}$$

This provides a tractable, albeit conservative, upper bound on the information carried by the learned representation. Furthermore, the joint mutual information is always greater than or equal to the information from any single domain. Under the optimality assumption that domain i is the most relevant for predicting Y^i, we have:

$$I(X^1,\ldots,X^D;Y^i) \geq \max_{k\in\mathcal{D}}\{I(X^k;Y^i)\} = I(X^i;Y^i)$$

Combining these inequalities, we can lower-bound the performance gap:

$$\begin{aligned}\text{Gap} &= I(X^1,\ldots,X^D;Y^i) - I(Z^i;Y^i)\\ &\geq \max_{k\in\mathcal{D}}\{I(X^k;Y^i)\} - \min_{j\in S^i}\{I(X^j;Y^i)\}\\ &= I(X^i;Y^i) - \min_{j\in S^i}\{I(X^j;Y^i)\} = \Delta^i\end{aligned}$$

This completes the proof.

Remark. The insight from Theorem 2 is profound. The lower bound of the performance gap, Δ^i, depends directly on the quality of the selected domain set S^i. If the selection process includes a domain j with very low mutual information with the target Y^i (i.e., a poorly chosen domain), the term $\min_{j\in S^i} I(X^j;Y^i)$ will be small, making the gap Δ^i large. This signifies a substantial performance degradation compared to an oracle with full information. Our framework's goal is to select domains such that this minimum is maximized, thereby minimizing the information gap and approaching the optimal performance. This formally justifies the need for wise domain selection to prevent performance loss from negative transfer.

Bibliography

Daniel Acuna, Marc T. Law, and Sanja Fidler. f-domain-adversarial learning: A general framework for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, and Fernando Pereira. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

Yi Wen, Yue Liu, Derong Xu, Huishi Luo, Pengyue Jia, Yiqing Wu, Siwei Wang, Ke Liang, Maolin Wang, Yiqi Wang, Fuzhen Zhuang, and Xiangyu Zhao. Measure domain's gap: A similar domain selection principle for multi-domain recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25)*, 2025. Preprint arXiv:2505.20227.