OT3L: Optimal Transport-based Targeted Transfer Learning for Efficient Domain Adaptation

Sayyed Farid Ahamed*§, Amit Chakraborty†§, Sandip Roy*‡, Bikash Chandra Singh‡, Sachin Shetty*
*Center for Secure & Intelligent Critical Systems, Old Dominion University, Virginia, USA
‡School of Cybersecurity, Old Dominion University, Virginia, USA
{saham001, sroy, bsingh, sshetty}@odu.edu
†Department of CSBS, Asansol Engineering College, Asansol 713305, India
idmeamit@gmail.com

***Abstract*—Machine learning models typically assume that training and testing datasets are drawn from the same distribution, a premise that often fails in real-world applications, thereby restricting model performance. To address this challenge, we introduce OT3L, an efficient targeted transfer learning approach that leverages the optimal transport distance (OTD) metric to improve knowledge transfer between source and target domains. Unlike conventional transfer learning methods that rely on random or heuristic-based fine-tuning of pre-trained models, our approach utilizes label-to-label OTD metrics to strategically select the most relevant samples from the target domain for model retraining. We implement a set difference, random, and mixed selection strategy that effectively balances diversity and similarity among target samples, enhancing the transfer of knowledge. The proposed OT3L approach is validated using the ImageNet source dataset across seven image classification benchmarks: COCO, CIFAR-10, CIFAR-100, FashionMNIST etc. Experimental results demonstrate that the targeted selection set, which includes both the highest and lowest OT distance samples, covers a broader range of class labels and achieves significant performance improvements, with gains ranging from 5.54% to 8.38% across various datasets.**

***Index Terms*—Transfer Learning, Optimal Transport Distance, Targeted Transfer Learning, Domain Adaptation, Fine-tuning.**

I. Introduction

In recent years, transfer learning has emerged as a powerful technique for addressing the challenges associated with training machine learning models, mostly in scenarios where labeled data is scarce or domain-specific [1]. The fundamental idea behind transfer learning is to influence knowledge gained from one domain (the source domain) to improve learning performance in a different yet related domain (the target domain). This approach is highly effective in various applications, including image recognition, natural language processing, medical diagnostics [2], [3], [4]. Despite its success, traditional transfer learning methods often meet difficulties when the source and target domains are significantly dissimilar [5]. In such cases, the feature distributions between the domains can vary considerably, leading to suboptimal performance when the model is applied directly [6].

The selection and adaptation of a pre-trained model for target domain adaptation is a critical step, as improper adjustments can result in issues such as bias and overfitting, further diminishing model performance [7]. To address these challenges, it is essential to carefully choose the source dataset and refine the pre-trained model to minimize the differences between the source and target domains. Several effective transfer learning strategies have been proposed to bridge these gaps, including domain adaptation [8], data augmentation [9], data selection [10], and addressing data shift [8].

Target-side adaptation, also known as target-oriented transfer learning, involves modifying source domain data to better align with the target domain [11]. Unlike source-side adaptation, where target data is adjusted to match the source, target-side adaptation leverages information from the target domain to enhance model performance on the target task. This approach is particularly effective when a pre-trained model underperforms on the target task due to significant domain gaps or insufficient target-specific data. However, it is crucial that the additional target domain data accurately reflects the target distribution, as non-representative data can degrade model performance [12]. Recently, a few research studies, such as those by Romero et al. [13] and Ahamed et al. [6], have demonstrated the effectiveness of target-side adaptation in improving performance across medical and satellite image datasets. In our earlier research works, we introduced automated targeted transfer learning (ATTL), a strategy designed to transfer knowledge between source and target domains while addressing the challenges of negative transfer [6], [14], [15]. ATTL resolves the transferability between source and target with minimal data requirements.

In this paper, **we aim to investigate two fundamental research inquiries on targeted transfer learning (TTL) based efficient domain adaptation strategies**. Firstly, *"How can the optimal transport distance metric be integrated with TTL to improve knowledge transfer between dissimilar source and target domains?"*. Secondly, *"How do sample selection strategies and fine-tuning affect domain adaptation and performance on across various datasets?"*

In this study, we leverage the optimal transport distance (OTD) metric to address distributional shifts between domains, enhancing knowledge transfer in a targeted transfer learning framework. By computing OTD between the source and target

§The authors have equal contributions.

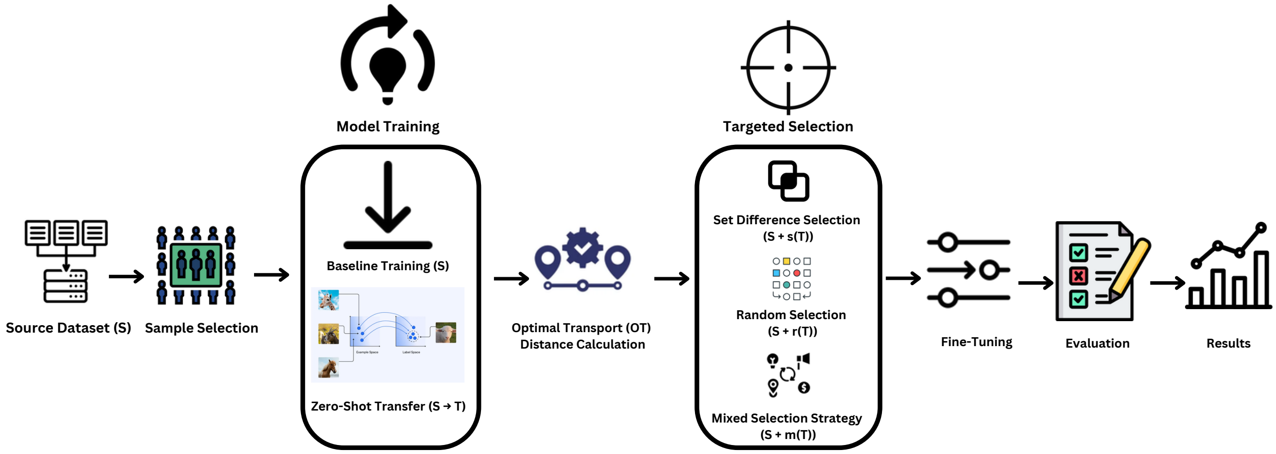

Fig. 1: Framework of the proposed optimal transport distance-based targeted transfer learning approach.

datasets, we identify a set difference set comprising samples with the highest OTD values. This set is then used to fine-tune the pre-trained model, bypassing the need for conventional retraining and yielding significant accuracy improvements. To further assess performance, we also generate a random selection set with the lowest OTD samples and develop a mixed selection strategy that balances both sets, covering a broader range of class labels. This mixed approach consistently outperforms individual selection methods. We rigorously evaluate the proposed approach on eight large-scale benchmark datasets, demonstrating its effectiveness.

In summary, the main contributions of this paper are as follows:

- **Introduction of OTD Metric in TTL:** We introduce OT3L, a targeted transfer learning method that employs label-to-label optimal transport distance (OTD) to enhance knowledge transfer across different domains.
- **New Sample Selection Strategy:** We develop a mixed strategy that effectively balances diversity and similarity among target samples, optimizing model retraining.
- **Extended Fine-Tuning:** We implement advanced fine-tuning for the target model, resulting in significant performance gains across various datasets.
- **Evaluation on Benchmark Datasets:** We rigorously evaluate the performance of OT3L using eight large-scale benchmark datasets to confirm its effectiveness.

The remainder of the paper is organized as follows: Section II describes the proposed OT3L approach. Section III outlines the research methodology. In Section IV, we present experimental results, including analysis and discussion, validating the approach across multiple datasets. Finally, Section V provides conclusions and discusses future research directions.

II. Proposed Approach

In this section, we introduce the proposed Optimal Transport Distance-based Targeted Transfer Learning (OT3L) approach. Figure 1 illustrates the framework, outlining the key steps of the approach.

Zero-shot learning: We train a machine learning model using the ImageNet dataset [16], which serves as the pre-trained model. This pre-trained model is then used as the source model in a transfer learning scenario. We considered multiple target datasets such as COCO [17], CIFAR-10 [18], CIFAR-100 [18], and others. According to the transfer learning approach, we utilized the pre-trained model to perform certain image classification target task. Initially, we apply zero-shot learning, i.e., deploy the source model directly to the target domain without any training in the target environment. Since the source model has never interacted with the target dataset, its perform poorly in the target domain. To avoid re-training in the target domain, transfer learning is the most obvious solution.

OTD-based targeted TL: Next, we propose a OTD-based targeted transfer learning approach to enhance model performance. First, we compute the *optimal transport distance* between datasets. For instance, consider CIFAR-10 as the target dataset. We measure the optimal transport distance between ImageNet and CIFAR-10. Using the OT distance metric, we create set difference sets, which consist of samples with the highest OT distance between the two datasets. In this experiment, we selected a minimal number of samples (1000) as the set difference set. The pre-trained model is then *fine-tuned* using this set difference set. Notably, we only fine-tuned the output and the last dense layer to limit the number of parameters involved in fine-tuning, thereby avoiding conventional re-training. This approach resulted in a significant improvement in accuracy.

Set difference, random and mixed-sample selection: To evaluate the effectiveness, we created a random selection set using the samples with the lowest OT distance between datasets, ensuring the number of samples is the same in both sets. Although the pre-trained model is re-trained in the target domain as in previous experiments, the accuracy does not improve as much as with the set difference set. In both the set difference and random selection sets, the pre-trained model is re-trained in the target environment with limited domain coverage, but the set difference set consistently outperforms the random selection set across all datasets. We also developed a mixed selection strategy using both selection sets to create a balanced selection process. Since the mixed selection sets cover more class labels (both highest and lowest OT distance samples), they outperform both individual selection sets.

To summarize, we propose a OTD-based targeted transfer learning (TTL) approach that enhances knowledge transfer between datasets. Our methodology involves strategically selecting target samples using OT distance strategy in order to maximize the model performance across diverse domains. Figure 1 depicts the proposed OT3L approach. The selected samples are then used for model training for efficient domain adaptation between datasets. The model is then undergoes fine-tuning, where the pre-trained model is re-training in target domain. The final step involves evaluation of the model's performance across different approaches to assess the effectiveness of the proposed approach.

III. Methodology

A. Datasets

We employed a diverse set of eight image datasets to calculate the effectiveness of the proposed OT3L approach, including ImageNet [16], COCO [17], CIFAR-10 [18], CIFAR-100 [18], FashionMNIST [19], EMNIST [20], MNIST [21], and KMNIST [22]. The table provides detailed information about those datasets. The datasets are chosen based on three different criteria: sample size, dimensionality, and labels. For example, EMNIST and MNIST have digit images, while CIFAR-10 and CIFAR-100 have a variety of object labels. Similarly, the sample sizes and classes vary between datasets. As a result, these datasets have very different feature distributions, making them an excellent testbed for the proposed method.

TABLE I: Image Datasets

Dataset	Dimension	#Samples	Class	Label
COCO	Variable	118k	80	Objects
CIFAR-10	32*32	60k	10	Various objects
CIFAR-100	32*32	60k	100	Various objects
FashionMNIST	28*28	70k	10	Fashion products
EMNIST	28*28	280K	10	Digits
MNIST	28*28	70k	10	Digits
KMNIST	28*28	70k	10	Japanese characters

B. Optimal Transport Distance (OTD)

The distance between two different distributions can be measured by using the principle of optimal transport. The fundamental idea behind optimal transport is to map every point in the source distribution to a corresponding point in the target distribution [23]. In order to achieve the most efficient transportation, the total distance between all paired points must be minimized. It is essential to define a distance metric between the sampled points from each distribution when applying OT to compare two probability distributions. Consider comparing, for instance, two datasets where each point is a combination of a feature vector and a label. Let's say that we have to find the distance between the pair (a, "kite"), where a is an ImageNet image of a "kite", and the pair (a', "five"), where a' is an EMNIST image of a "5". Although it is relatively simple to calculate the distances between similar samples using conventional methods, it is more complicated to determine the distance metrics for their labels.

The OT distance is a powerful metric for quantifying the dissimilarity between probability distributions, making it ideal for measuring the difference between corresponding classes in the source and target datasets. For two probability distributions $p(x)$ and $q(y)$, the OT distance $W(p, q)$ is defined as:

$$W(p,q) = \inf_{\gamma \in \Gamma(p,q)} \int_{X \times Y} c(x,y)\, d\gamma(x,y), \tag{1}$$

where $\Gamma(p, q)$ is the set of all possible joint distributions (couplings) between $p(x)$ and $q(y)$. $c(x, y)$ represents the cost function, typically the Euclidean distance.

We computed the OT distance between all class pairs across the source and target datasets. Classes with the highest mean OT distance were identified as the most dissimilar, while those with the lowest mean OT distance were considered most similar.

The fundamental goal of this distance metric is to be relevant even when the user has multiple sets of labels that do not correlate (for example, image to digit). In this situation, each label is represented as a geometric feature, similar to a vector, which gives a computational benefit [24].

C. Sample Selection for Targeted Transfer Learning

Based on the OT distance calculations, we employed two strategies for selecting samples from the target dataset.

Set Difference Set $(S + s(T))$: A random selection of samples from the target classes with the highest OT distance from the corresponding source classes. This selection aims to introduce the greatest diversity into the source dataset.

Random Selection Set $(S + r(T))$: A random selection of samples from the target classes with the lowest OT distance from the corresponding source classes. This selection aims to enhance the existing similarities between the source and target datasets.

To further enhance transfer learning, we implemented a mixed selection strategy $(S + m(T))$, where $m(T)$ is defined as:

$$m(T) = \alpha \cdot s(T) + (1 - \alpha) \cdot r(T), \tag{2}$$

where α is a mixing coefficient that determines the proportion of samples selected from the set difference and random selection sets. This approach allows for a flexible balance between diversity and similarity in the augmented source dataset.

D. Model Training and Transfer Learning

We followed a multi-phase training procedure to evaluate the impact of the selected samples on transfer learning.

Baseline Training (S): A baseline model was trained on the original source dataset S. The model's architecture consisted of a convolutional neural network (CNN) with multiple layers, including convolutional, pooling, and dense layers, followed by a softmax output layer.

Zero-Shot Transfer ($S \rightarrow T$): The baseline model was applied directly to the target dataset T without any additional training. The performance on the target dataset was recorded as the zero-shot baseline.

Transfer Learning with Adaptation ($S + s(T) \rightarrow T$) or ($S + r(T) \rightarrow T$): The model's output layer and the penultimate dense layer is fine-tuned using the selected samples from the target dataset to adapt the model to the target domain.

Let $f_\theta(x)$ represent the CNN model with parameters θ. The final output after softmax is:

$$\hat{y} = \text{softmax}(f_\theta(x)) \tag{3}$$

During fine-tuning, we minimize the cross-entropy loss:

$$L_{CE} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i), \tag{4}$$

where y_i is the true label and $\hat{y}_i$ is the predicted probability for the i-th sample.

IV. Experimental Evaluation

In this section, we conduct experiments to evaluate the proposed OT3L performance across a diverse set of benchmark image classification datasets. We consider ImageNet to be the source dataset for training the pre-trained model. As the target dataset we evaluate the performance of the proposed approach on seven target datasets, including COCO, CIFAR-10, CIFAR-100, FashionMNIST, EMNIST, MNIST, and KMNIST. The details of these datasets are presented in Section III. The findings are structured to highlight the impact of targeted selection strategies such as set difference selection, random selection, and mixed sample selection.

A. Results Analysis

1) Dataset Selection: Figure 2 depicts the OT distance between seven target datasets. This distance metric assists us in selecting appropriate source datasets for transfer learning. The figure is colour-coded based on their distance. For example, the CIFAR-10 dataset is very close to the CIFAR-100 dataset (3.75), as expected, but far away from the FashionMNIST dataset (5.20). As a result, if we use CIFAR-10 as the source dataset, CIFAR-100 will outperform the FashionMNIST dataset. Another interesting finding is that among MNSIT datasets, CIFAR-10 is closest to the EMNIST dataset, followed by MNIST and KMNIST. As a result, OT distance between datasets could provide important insights into source selection in a transfer learning scenario.

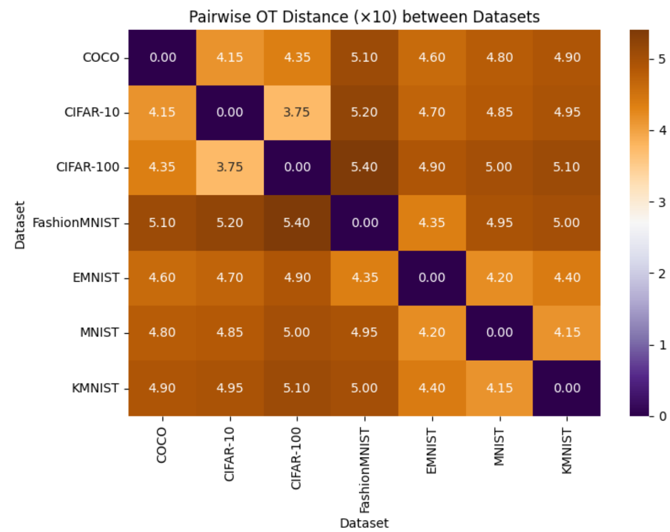

Fig. 2: OT distance between datasets.

2) Baseline Performance: The baseline performance of a model is determined by training the model solely on a source dataset (S), which is considered as pre-trained model, and then applying it directly to target datasets in a zero-shot transfer setting. In this experiment, we use ImageNet as source domain for all target domain datasets. Table II displays the zero-shot transfer learning performance across various target datasets. Deploying pre-trained model directly to the target dataset without re-training yields poor result. Therefore, domain shift techniques are crucial for improving transfer learning performance.

TABLE II: Zero-shot transfer learning

Baseline	Zero-shot Transfer on ImageNet $\rightarrow T$						
ImageNet	**DS1**	**DS2**	**DS3**	**DS4**	**DS5**	**DS6**	**DS7**
80.12	14.10	25.56	18.32	19.67	17.89	24.12	13.89

Notations- **DS1**: COCO; **DS2**: CIFAR-10; **DS3**: CIFAR-100; **DS4**: FashionMNIST; **DS5**: EMNIST; **DS6**: MNIST; **DS7**: KMNIST.

3) Targeted Sample Selection: We compute the targeted selection set by measuring the OT distances between source and target sets. Three primary targeted sets are considered to evaluate the OT3L process. In set difference (SD) selection, target samples are chosen based on their distinct distance from the source samples. In random selection (RS), samples are collected from the sets with lower OT distances. Furthermore, in mixed selection (MS) strategy, both higher and lower OT distance samples are combined to balance the target samples' diversity and similarity. This strategy provides a more comprehensive transfer learning approach. To validate our methodology, we apply an identical approach to all selections, including sample size, pre-trained model, and training procedure.

Table III compares the results of the zero-shot, set difference, random selection, and mixed selection strategies. According to the table, SD achieves an accuracy of 80.12% on the CIFAR-10 dataset, while RS slightly underperforms

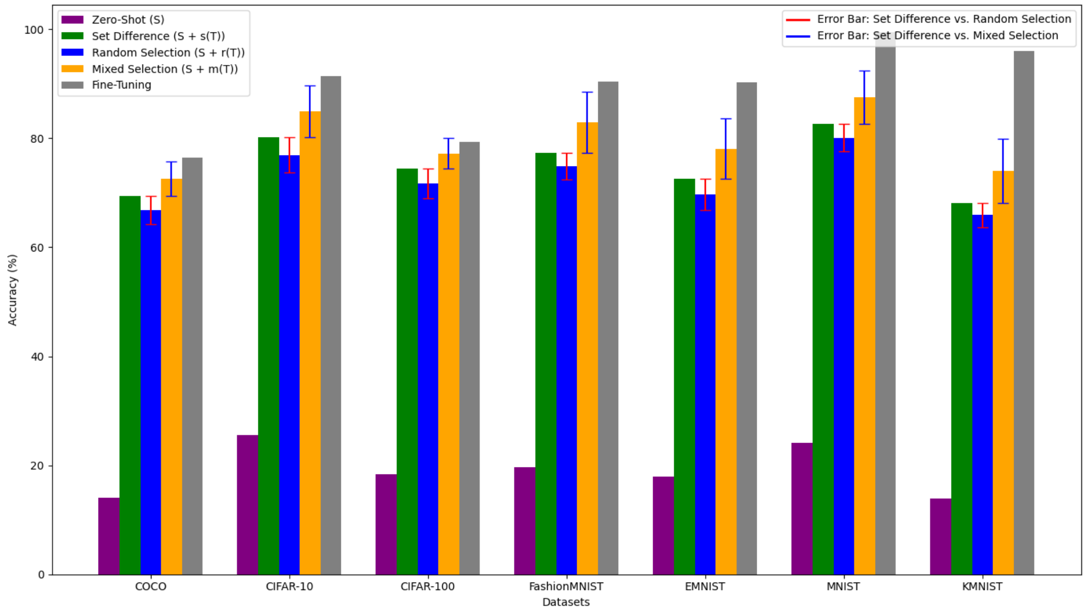

Fig. 3: Performance comparison across datasets.

TABLE III: Performance improvement with OTD-based targeted TL

Dataset	Zero-Shot (S)	SD $(S+s(T))$	RS $(S+r(T))$	MS $(S+m(T))$
COCO	14.10	69.45	66.81 (**-2.64%**)	72.6 (**+3.15%**)
CIFAR-10	25.56	80.12	76.92 (**-3.2%**)	84.92 (**+4.8%**)
CIFAR-100	18.32	74.45	71.68 (**-2.77%**)	77.22 (**+2.77%**)
FashionMNIST	19.67	77.32	74.86 (**-2.46%**)	82.9 (**+5.58%**)
EMNIST	17.89	72.56	69.72 (**-2.84%**)	78.10 (**+5.54%**)
MNIST	24.12	82.56	80.11 (**-2.45%**)	87.45 (**+4.89%**)
KMNIST	13.89	68.12	65.91 (**-2.21%**)	74.02 (**+5.9%**)

at 76.92%. In comparison, MS outperforms both with an accuracy of 84.92%. This indicates a significant improvement of 4.80% over SD and 8% over RS. Similarly, for the CIFAR-100 dataset, SD achieves an accuracy of 74.45%, while RS falls behind at 71.68%. MS again outperforms SD by 2.77% and RS by 5.54%, with an accuracy of 77.22%. Using OT-based set difference selection results in significant accuracy improvements. On the other hand, while the random selection improves performance through introducing target samples into the source, the improvements are less pronounced. The mixed selection strategy consistently outperforms both set difference and random selection sets.

4) Extended Fine-tuning: To enhance model adaptation to the given environment, we fine-tuned the pre-trained model using target dataset. This will allow the model to fully adapt to the target environment. Table IV shows that further fine-tuning results in significant accuracy improvements across datasets.

TABLE IV: Extended fine-tuning

Baseline	**Fine-tuning** (T)						
ImageNet	**DS1**	**DS2**	**DS3**	**DS4**	**DS5**	**DS6**	**DS7**
80.12	76.40	91.34	79.26	90.35	90.27	99.51	96

Notations- **DS1**: COCO; **DS2**: CIFAR-10; **DS3**: CIFAR-100; **DS4**: FashionMNIST; **DS5**: EMNIST; **DS6**: MNIST; **DS7**: KMNIST.

B. Discussion

When compared to baseline and zero-shot transfer methods, the proposed OT3L approach improves model performance significantly across a wide range of image datasets. Due to limited domain coverage, zero-shot transfer performs poorly in all datasets, as expected. However, the set difference results show that intentionally selecting a number of samples from the target can significantly improve the pre-trained model's performance in a specific environment. To validate this performance, we examined three selection methodologies with the same feature space. In this experiment, we found significant improvements in accuracy when employing the mixed selection strategy over the random selection strategy across multiple datasets. Compared to the random selection, the mixed selection strategy improved accuracy by 5.79% for COCO, 8% for CIFAR-10, 5.54% for CIFAR-100, 8.04% for FashionMNIST, 8.38% for EMNIST, 7.34% MNIST, and 8.11% for KMNIST. These results demonstrate the effectiveness of the targeted selection in enhancing model performance in target domain. Finally, the extended fine-tuning produces significant performance improvements.

V. Conclusion and future Scope

This paper introduces OT3L, an efficient targeted transfer learning approach that leverages optimal transport distance metrics to enhance knowledge transfer between source and target domains. Unlike traditional transfer learning methods, OT3L strategically selects target samples based on OTD, utilizing a combination of set difference, random, and mixed selection strategies. This allows for an effective balance between sample diversity and similarity, leading to significant performance improvements across various image classification benchmarks. The effectiveness of OT3L is validated across eight large-scale benchmark image classification datasets. The experimental results highlight OT3L's ability to achieve superior model performance, demonstrating its potential as a robust solution for domain adaptation challenges.

Future work could focus on automating the sample selection process, potentially incorporating techniques like active learning to dynamically select the most informative samples during training. Another avenue for research could involve exploring alternative fine-tuning strategies that adjust deeper layers of the model or even incorporating domain-specific knowledge into the fine-tuning process. This research could extend these findings to other domains and further refine the methodology to enhance its effectiveness.

Acknowledgement

This work is supported in part by DoD Center of Excellence in AI and Machine Learning (CoE-AIML) under Contract Number W911NF-20-2-0277 with the U.S. Army Research Laboratory, National Science Foundation under Grant No. 2219742 and Grant No. 2131001, and also supported in part by the Coastal Virginia Center for Cyber Innovation (CoVA CCI).

References

[1] 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.

[2] G. Capuano and J. J. Rimoli, "Smart finite elements: A novel machine learning application," *Computer Methods in Applied Mechanics and Engineering*, vol. 345, pp. 363–381, 2019.

[3] H. Malik, M. S. Farooq, A. Khelifi, A. Abid, J. N. Qureshi, and M. Hussain, "A comparison of transfer learning performance versus health experts in disease diagnosis from medical imaging," *IEEE Access*, vol. 8, pp. 139 367–139 386, 2020.

[4] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, "Transfer learning in natural language processing," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials*, 2019, pp. 15–18.

[5] P. Marcelino, "Transfer learning from pre-trained models," *Towards data science*, vol. 10, no. 330, p. 23, 2018.

[6] S. F. Ahamed, P. Aggarwal, S. Shetty, E. Lanus, and L. J. Freeman, "Attl: an automated targeted transfer learning with deep neural networks," in *2021 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2021, pp. 1–7.

[7] 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.

[8] 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.

[9] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[10] S. Ruder and B. Plank, "Learning to select data for transfer learning with bayesian optimization," *arXiv preprint arXiv:1707.05246*, 2017.

[11] J. Wang, Y. Chen, W. Feng, H. Yu, M. Huang, and Q. Yang, "Transfer learning with dynamic distribution adaptation," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 1, pp. 1–25, 2020.

[12] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, "Optimal transport for domain adaptation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 9, pp. 1853–1865, 2016.

[13] M. Romero, Y. Interian, T. Solberg, and G. Valdes, "Targeted transfer learning to improve performance in small medical physics datasets," *Medical physics*, vol. 47, no. 12, pp. 6246–6256, 2020.

[14] S. F. Ahamed, K. A. Islam, and S. Shetty, "Targeted transfer learning: Leveraging optimal transport for enhanced knowledge transfer," in *2024 International Conference on Computing, Networking and Communications (ICNC)*. IEEE, 2024, pp. 506–510.

[15] P. Aggarwal, S. F. Ahamed, S. Shetty, and L. J. Freeman, "Selective targeted transfer learning for malware classification," in *2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*. IEEE, 2021, pp. 114–120.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[18] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[19] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[20] G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik, "Emnist: Extending mnist to handwritten letters," in *2017 international joint conference on neural networks (IJCNN)*. IEEE, 2017, pp. 2921–2926.

[21] Y. LeCun, C. Cortes, C. Burges *et al.*, "Mnist handwritten digit database," 2010.

[22] T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha, "Deep learning for classical japanese literature," *arXiv preprint arXiv:1812.01718*, 2018.

[23] C. Villani *et al.*, *Optimal transport: old and new*. Springer, 2009, vol. 338.

[24] D. Alvarez-Melis and N. Fusi, "Geometric dataset distances via optimal transport," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 428–21 439, 2020.