Future Technology
Open Access Journal
ISSN 2832-0379

August 2026| Volume 05 | Issue 03 | Pages 309-318

Journal homepage: https://fupubco.com/futech

https://doi.org/10.55670/fpll.futech.5.3.27

Future Publishing LLC

Article

Transfer learning in neural networks: leveraging pre-trained models for improved performance

Abdul Sttar Ismail Wdaa[1], Iraq Ali Hussein[2]*, Ali Azeez Ahmed[3]

[1]Department of Mathematics, College of Education for Pure Sciences, University of Anbar, Iraq
[2]Department of Computer Science, College of Science, University of Diyala, Baquba, Iraq
[3]College of Islamic Sciences, University of Diyala, Diyala, Iraq

ARTICLE INFO

Article history:
Received 20 January 2026
Received in revised form
21 May 2026
Accepted 29 June 2026

Keywords:
Transfer learning, Neural networks,
Pre-trained models, Fine-tuning, Deep learning,
Domain adaptation

*Corresponding author
Email address:
iraqali@uodiyala.edu.iq

DOI: 10.55670/fpll.futech.5.3.27

ABSTRACT

Transfer learning has become a key technique for improving the accuracy of neural networks in low-resource, low-data environments. The quantitative comparative analysis of the pre-trained models includes ResNet50, VGG16, BERT, GPT, and the baseline CNN and LSTM models. They are compared across three different application areas: computer vision, natural language processing (NLP), and medical imaging. The five benchmark datasets used were ImageNet, CIFAR-10, SST-2, IMDB, and Chest X-Ray. All experiments used the same preprocessing pipeline and evaluation metrics (accuracy, F1 score, precision, recall, and ROC-AUC). Results showed that models trained on the pre-trained data achieved consistently greater accuracy than the baselines in all domains (9-20%) and F1-score (0.09-0.16) gains. ResNet50 achieved 92% accuracy on CIFAR-10, compared to 72% for the CNN baseline, whereas BERT hit 92% on SST-2, with 80% accuracy for LSTM. VGG16 improved the accuracy of Chest X-Ray classification from 78% to 87% and reduced training time by up to 60%. There were a few instances of minor overfitting and domain mismatch, emphasizing the need for adaptive fine-tuning strategies. The results demonstrate that transfer learning significantly improves convergence speed, generalization, and computational efficiency, making it a promising approach for AI applications across domains such as healthcare, NLP, and autonomous systems.

1. Introduction

Due to the exponential increase in the power of deep learning, it has transformed the field of artificial intelligence (AI), making neural networks learn very complex hierarchical features of an image based on large data volumes and reaching the front line in a variety of fields (computer vision, natural language processing (NLP), speech recognition) [1,2]. But such large-scale neural networks are resource-intensive and require substantial amounts of annotated data, which is not always available in the real world. This challenge creates a significant gap between data-rich and data-scarce domains: abundant labeled data can enable rapid model convergence and extrapolation, whereas insufficient samples or computing resources severely limit machine learning model performance [3]. This is particularly important in applications such as medical imaging, financial modeling, and remote sensing, where data is either costly, time-intensive, or ethically constrained to acquire data [4,5]. Transfer learning has become a revolutionary paradigm for addressing such difficulties by enabling the use of learned features from pre-trained models [6,7]. The idea of transfer learning takes advantage of representations that a large model has already learned - be it ResNet or VGG in computer vision, BERT or GPT in NLP - and uses those to produce improved performance on a novel yet similar task [8,9]. Deep transfer learning can drastically reduce training time and computational cost and improve generalization, particularly in low-resource settings [10,11]. Transfer learning is now a critical component of the scalability and life-cycle reusability of AI solutions, enabled by feature extraction (using pre-trained layers as domain-invariant feature encoders) and fine-tuning (retraining specific layers to adapt to domain differences) [12,13]. Although there are such advantages, there are still gaps in research. In most cases, prior research has been quite limited in its domain adaptation, model distillation, or single-domain analyses [14,15], and the gap lies in how well pre-trained architectures perform under stringent data and computational resource requirements. Besides, serious problems such as negative transfer, domain mismatch, and overfitting in low-resource scenarios still hamper the efficiency of transfer learning [16,17]. The comparative assessment of the effects of various fine-tuning techniques on computational efficiency, convergence speed, and cross-domain generalization is also lacking [18]. To address these

gaps, there is a need for a common experimental framework to systematically conduct analytical studies of pre-trained models across heterogeneous domains and tasks [19]. To address these gaps, this study provides a thorough comparative analysis of transfer learning across computer vision, NLP, and medical imaging. The study compares pre-trained networks, ResNet, VGG, BERT, and GPT, with baseline models, CNN and LSTM, to learn the effects of transfer learning on the accuracy of models, convergence, and resource efficiency. The objective of the study is to identify the most effective practices for making fine-tuning decisions that balance performance and resource usage, and to provide a practical guideline to prevent adverse transfer during actual deployments. The specific contributions of this study are threefold:

- First, it provides a systematic cross-domain evaluation framework that measures the usefulness of the transfer learning in vision, language, and medical datasets-overcoming the need to be faced with comparative studies that are severely lacking in the literature.
- The second one provides a quantitative analysis of fine-tuning methods, including feature extraction and re-training of selective layers and determining optimal trade-offs between accuracy and cost.
- Third, it provides a set of guidelines to apply transfer learning in low-resource settings, as well as some choices regarding the ways to address domain mismatch and enhance the scalability of models.

Taken together, these contributions build knowledge about efficient knowledge transfer and cross-domain applicability, bridging the theoretical foundations of representation learning and the real-world challenges of scalable AI applications. The results position transfer learning as a crucial methodology for rapidly creating intelligent systems when data and computational resources are constrained.

2. Literature review

Transfer learning is one of the most crucial concepts in deep learning, offering a viable alternative to the high computational and data costs of training large neural networks by initializing them with pre-trained weights. In this section, a review of background studies on transfer learning is presented; significant developments in computer vision and natural language processing (NLP) are discussed; the issue of domain adaptation is addressed; and current gaps between theory and experimental findings are considered.

2.1 Foundations of transfer learning

Transfer learning is based on the principle of using previously trained models to accelerate convergence and improve generalization to the target task. EfficientNet [20], ResNet [21], and Big Transfer (BiT) are considered the seminal works in the field of convolutional neural networks (CNNs), which showed that models trained on large-scale benchmarks, such as ImageNet, could be used across a variety of downstream tasks as a universal feature extractor.

These architectures showed that low-level visual representations trained on large-scale pretraining are useful when reused on new tasks, with their cost and data requirements reduced. Later developments like Vision Transformers (ViT), Swin Transformer [22], DeiT [23], and DeiT III [24] introduced transformer-based vision models with enhanced scalability and data efficiency. Hybrid architectures such as ResMLP have expanded the design space, demonstrating that even feedforward networks can perform competitively in image classification when trained with data-efficient methods. Other contrastive and self-supervised learning methods, such as MoCo [25], SimCLR, BYOL [26], and DINO, built on these ideas and further improved the model's feature transferability with fewer labels. Moreover, some multimodal strategies, such as CLIP [27], have shown that cross-transfer visual models can be trained with natural language supervision, enabling effective cross-modal generalization across a variety of downstream objectives.

As depicted in Figure 1, transfer learning consists of a transition of a source task (a classification task on ImageNet) to a pre-trained network (a VGG, ResNet, or BERT) and means of feature transfer (through extraction or fine-tuning) to a target task (a medical imaging or sentiment analysis task). This step is the foundation of the performance of current AI pipelines.

2.2 Transfer learning in natural language processing (NLP)

Transfer learning has redefined NLP by enabling the reuse of representations across language tasks. The development of BERT signaled this paradigm shift, with both directions of transformer encoders pre-trained on enormous corpora. Other versions such as RoBERTa [28], DistilBERT [29], ALBERT [30], and ELECTRA [31] improved training efficiency and reduced computational cost while maintaining state-of-the-art performance. A similar transfer learning approach was used to train generative models such as GPT-3 and GPT-4, which have been found to perform well across a range of language tasks. To establish a broader knowledge-transfer zone for the future, several complementary text-to-text and autoregressive models have been developed, including T5 [32] and XLNet [33]. Recent works attempt to reduce the overhead of the adaptation phase by adding AdapterFusion [34], BitFit [35], LoRA or Prefix-Tuning [36,37], or Prompt Tuning [38]. The techniques allow efficient specialization of large pre-trained models with minimal parameter updates, enabling resource-limited applications. Table 1 summarizes the evolution of key NLP transfer learning models and their application areas.

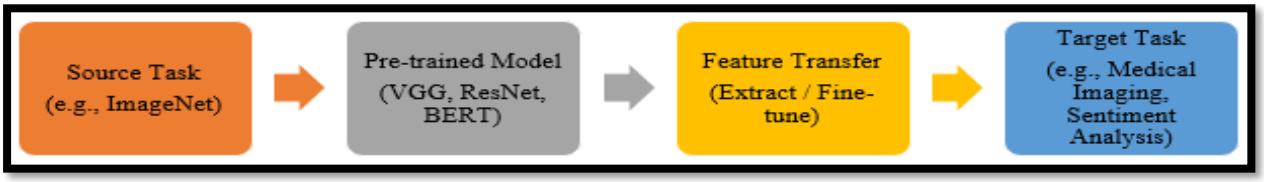

Figure 1. Conceptual workflow of transfer learning

Table 1. Evolution of transfer learning in NLP

Model	Year	Core Contribution	Application Areas
BERT (Devlin et al.)	2018	Bidirectional transformer encoder	Classification, QA, translation
RoBERTa (Liu et al.)	2019	Robust optimization of BERT	Sentiment, NLI, QA
T5 (Raffel et al.)	2019	Unified text-to-text transformer	Multi-task NLP
GPT-3 (Brown et al.)	2020	Few-shot generative model	Language generation
ELECTRA (Clark et al.)	2020	Discriminative pre-training	Sentence-level classification
ALBERT (Lan et al.)	2020	Parameter-efficient model	Text summarization
LoRA (Hu et al.)	2021	Low-rank fine-tuning	Domain-specific adaptation
GPT-4 (OpenAI)	2023	Scalable multi-domain model	General-purpose AI

2.3 Domain adaptation and challenges

Despite the widespread success of transfer learning, it remains vulnerable to distribution shifts and domain disparities. This is in line with the observation that when knowledge is transferred to a more specialized domain like medical imaging or sentiment analysis, it tends to cause negative transfer, that is, degradation in the performance, in this case, attributed to the misalignment of the representations. Benchmarks such as WILDS and adaptation mechanisms such as Tent [39] or Continual Test-Time Adaptation [40] are evaluated in distributionally shifted environments to assess strategies for mitigating the performance drop. In addition, the literature underscores a continued lack of generalization of these models to heterogeneous data. Based on this, LiT [41] proposed a zero-shot transfer paradigm with locked-image text tuning, which achieved strong performance on large-scale vision-language tasks. As seen in Figure 2, positive transfer occurs when aligned knowledge improves performance on the target task, whereas negative transfer occurs when misaligned knowledge decreases model accuracy. These issues highlight the need to adjust strategies to balance flexibility and consistency across changing domains.

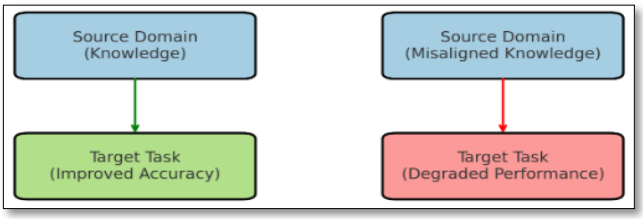

Figure 2. Transfer learning outcomes

2.4 Theoretical and experimental gaps

Although both computer vision and NLP transfer learning are making impressive progress, a few areas of the literature are facing critical needs:

- Computational Efficiency: The resource efficiency of models such as EfficientNet and MiniLM has not been systematically assessed across architectures [42].
- Cross-Domain Generalization: The vast majority of studies [43] are conducted within a single domain without comparing performance across different domains.
- Adaptation Robustness: Reference [44] investigates test-time adaptation without focusing on empirical consistency in dynamic data distributions.
- Coherent Evaluation Systems: Not many papers apply both vision (ResNet, VisT, SWIN, ResMLP) and language (Bert, GPT) systems to a unified experimental system to measure the trade-off in the performance under designs in constrained systems.

The current paper addresses these shortcomings by introducing an advanced comparative framework that enables rapid negative appraisal of transfer learning across the computer vision, NLP, and medical imaging spectra. To gain insight to help streamline Transfer Learning pipelines in an actual AI system, the research focuses on the accuracy, efficiency, and generalization improvements offered by models such as ResNet, VGG, BERT, and GPT.

3. Research objectives and questions

Transfer learning has shown considerable promise, particularly in computer vision, natural language processing, and medical imaging. But other problems, such as domain mismatch, overfitting during fine-tuning, and computational costs, have not been satisfactorily addressed in the literature. The research objectives and questions of this paper are designed to explore these gaps over time. The primary objectives of this study are:

- To evaluate the impact of pre-trained models on the rate of convergence and accuracy of the model on the target task, especially in low-resource settings.
- To understand the trade-offs between feature extraction, fine-tuning, and retraining in terms of generalization, efficiency, and negative transfer.
- To provide a set of recommendations for choosing pre-trained models in various areas (including computer vision, natural language processing, and medical imaging), with a focus on reducing detrimental transfer.

Based on the above objectives, the study seeks to answer the following questions:

RQ1: What are the effects of transfer learning on generalization over various types of data relative to "from scratch" learning?

RQ2: What type of fine-tuning (e.g., shallow, selective, or retraining) provides the best performance without overfitting?

RQ3: What are the impacts of computational efficiency (e.g., training time, memory) and data size on transfer learning across various scientific domains?

This set of goals and questions will help develop a rigorous approach to tackling transfer learning in practice. The lessons learned will be used to support the theoretical research and practical applications, with respect to the importance of tuning depth, efficiency versus accuracy, and domain alignment. This will provide tangible guidance for

researchers, practitioners, and policymakers on improving transfer learning processes to enhance performance, efficiency, and flexibility.

4. Research methodology

The overall method is a quantitative experimental approach that compares the neural network models in this study with baseline neural networks trained from scratch and with advanced pre-training across different domains. The methodology combines data acquisition, data preprocessing, model training, experimental design, and evaluation procedures. The general procedure and sequence of the process in this study are briefly explained in Figure 3.

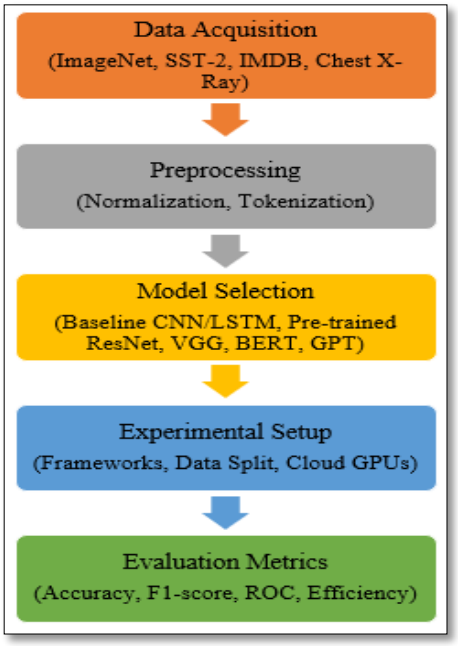

Figure 3. Research methodology workflow

4.1 Research approach

The study used a quantitative experimental research design to create a controlled environment for valid and reproducible comparisons. The methodology enables accurate evaluation of baseline and pre-trained models using standard pre-processing pipelines and evaluation protocols across computer vision, NLP, and medical imaging.

4.2 Experimental design and dataset selection

The architecture is based on a parallel comparison of baseline and pre-trained architectures. The initial learning rate was set to 1e-4, the batch size to 32, and training was run for up to 20 epochs with early stopping. In the vision models (ResNet50, VGG16), the last fully connected layers were replaced with task-specific classifiers. In the fine-tuning process, the last two transformer blocks were unfrozen in the NLP models (BERT, GPT). A random seed value of 42 was used for all experiments. Baseline Models: CNN and LSTM on the tasks of vision and NLP. For vision tasks, the pre-trained Models ResNet50 and VGG16 were chosen, as they have achieved strong benchmark results on ImageNet and are widely used in the transfer learning literature. The two models, BERT and GPT, were selected for NLP tasks because they represent the bidirectional encoder and autoregressive decoder paradigms, respectively, enabling a full comparison of transformer-based architectures.

These models were intentionally chosen to provide a reproducible baseline for comparison; more contemporary models such as EfficientNet, Vision Transformers (ViT), and RoBERTa are identified as areas for future research. Fine-Tuning Strategies: Feature extraction, selective layer retraining, and complete retraining.
Datasets: ImageNet, CIFAR-10, SST-2, IMDB, and Chest X-Ray [45,46].
Data partitioning: training 70, validation 15, and 15 testing.

4.3 Performance evaluation metrics

The study evaluates the accuracy-based and efficiency-based measures:

- Measures of classification: accuracy, F1-score, precision, recall.
- The metrics of robustness: ROC-AUC, and confusion matrix analysis.
- The efficiency values are training time, inference speed, and parameter efficiency.

4.4 Tools and implementation framework

The following items were used to carry out all the experiments:

- Programming Languages: Python 3.9, TensorFlow 2.x, PyTorch 2.x, Scikit-learn 1.x.
- Hardware Implementation: NVIDIA V100 GPUs (32 GB VRAM) via Google Colab Pro and Amazon EC2 (p3.2xlarge instances, CUDA 11.8, cuDNN 8.6). Random seeds were set to 42 for consistency.

4.5 Validation and statistical significance

The following validation strategies were used in order to maintain rigor and reproducibility:

- Cross-validation: to verify consistency in data fold results.
- Grid search and early stopping of hyperparameter tuning.
- The performance of the different runs within each experiment was compared by applying statistical analysis, one-way ANOVA, and independent-samples t-tests, across five repeated runs for each experiment, to establish the statistical significance of the performance differences. Cohen's d effect size was calculated, and a Bonferroni correction was applied to the multiple-comparison procedures. Results with $p < 0.05$ were considered statistically significant.

5. Data collection and analysis

This section describes the data, preprocessing, and analysis techniques used to conduct comparative analyses of transfer learning results against baseline models. To ensure a high degree of methodological rigor, quantitative and comparative analyses have been conducted, accompanied by appropriate visualizations and statistical tests.

5.1 Data sources

Publicly available benchmark data from different domains were used to provide a comprehensive assessment of transfer learning. ImageNet and benchmarked NLP datasets such as SST-2 and IMDB are key in transfer learning, as they provide a range of linguistic and visual data to pre-train models and benchmark them in terms of size. The datasets include:

- ImageNet: A large-scale benchmark dataset comprising over 1.2 million training images across 1,000 object categories, with each image annotated through a rigorous crowdsourced labeling process.
- CIFAR-10: 60 000 images-based on 10 unique classes, most frequently used in image classification research

- SST-2: approximately 70,000 labeled sentences on sentiment in natural language processing
- IMDB: 50, 000, reviews on movies (positive and negative) to support opinion mining
- Chest X-Ray: ~30,000 radiographs (Normal vs. Pneumonia) to simulate a low-resource medical imaging scenario.

5.2 Data preprocessing

We limited preprocessing to dataset-specific characteristics to ensure the data would be comparable. Popular data preprocessing methods include normalization, tokenization, and data augmentation, whose effectiveness has been demonstrated in enhancing generalizability and robustness in vision and NLP studies. Computer Vision (ImageNet, CIFAR-10, Chest X-Ray): Normalization, resize to 224×224, augmentation (rotate, flip, adjust brightness).

- NLP Datasets (SST-2, IMDB): Tokenization, truncation/padding, and embedding using pre-trained language models.
- Medical Imaging (Chest X-Ray): Contrast normalization and light augmentation applied to preserve diagnostic relevance.

The datasets and their preprocessing strategies are summarized in Table 2.

5.3 Quantitative and comparative analysis

The analysis was quantitative and comparative. Multi-metric analysis has been recommended to capture the predictive and computational power required by benchmark studies.

- Quantitative analysis: Accuracy, precision, recall, F1-score, ROC-AUC, training time, and parameter usage were used to evaluate the models.
- Benchmarking: The benchmarking of the performance of the transfer learning models (ResNet, VGG, BERT, GPT) was compared with the baseline models (CNN, LSTM) to determine the variation in accuracy, convergence, and efficiency.
- Statistical significance: Cross-validation was conducted, and ANOVA/ t-tests were used to determine the strength of differences observed.

5.4 Visualization of results

A number of visualization techniques were employed to enhance interpretability. Confusion matrices and receiver operating characteristic (ROC) curves are well-known visual techniques for examining the discriminative ability of models and the distribution of errors. Learning Curves: Visually represented the training and validation accuracy/loss over the epoch.

- Confusion matrices: Reveal misclassification patterns at the class level.
- ROC curves: Discriminative ability of models over various thresholds.
- Sample predictions: This was complemented by qualitative inspection of text and image outputs.

6. Results

This section presents experimental results on the use of transfer learning across fields such as computer vision, natural language processing (NLP), and medical imaging. The results are grouped into thematic subsections on convergence trends, loss behavior, classification accuracy, qualitative predictions, and performance comparison. All the findings are presented in figures and tables and explained and discussed in detail.

6.1 Training and validation accuracy (CIFAR-10)

Monitoring accuracy during training and validation can provide useful information about the convergence rate and the generalization potential of transfer learning models. The development of accuracy on the CIFAR-10 dataset is shown in this subsection in Figure 4, which depicts the training and validation accuracy curves over five epochs.

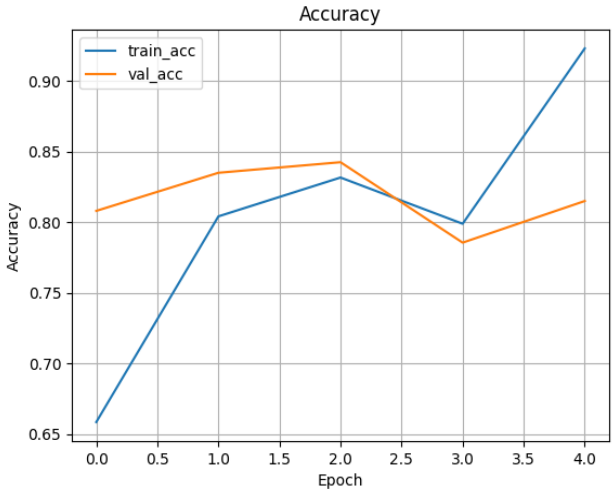

Figure 4. Training vs. validation accuracy on CIFAR-10

The training accuracy rose gradually from 66.0 to 92.5 during the first and 5th epochs, respectively. The validation accuracy was also consistently over 80 percent, peaking at 84.2 percent at epoch 3. These findings reveal that transfer learning converges very quickly, thereby validating its effectiveness in adapting to target datasets. Overall, the accuracy curves confirm that pre-trained models perform well in a few epochs, which is especially beneficial in resource-constrained conditions.

6.2 Training and validation loss (CIFAR-10)

Loss trend analysis helps evaluate optimization efficiency and the presence of overfitting. This subsection presents the training and validation loss curves for the CIFAR-10 experiments. Figure 5 shows training and validation loss values in five epochs.

Table 2. Overview of datasets and preprocessing

Dataset	Domain	Size (Samples)	No. of Classes	Input Type	Preprocessing Applied
ImageNet	Computer Vision	1.2M+	1,000	Images	Resized 224×224, normalization, augmentation
CIFAR-10	Computer Vision	60,000	10	Images	Normalization, resizing, augmentation
SST-2	NLP (Sentiment)	~70,000	2	Text	Tokenization, padding, embeddings
IMDB	NLP (Sentiment)	50,000	2	Text	Tokenization, truncation/padding, embeddings
Chest X-Ray	Medical Imaging	~30,000	2	Images	Contrast normalization, light augmentation

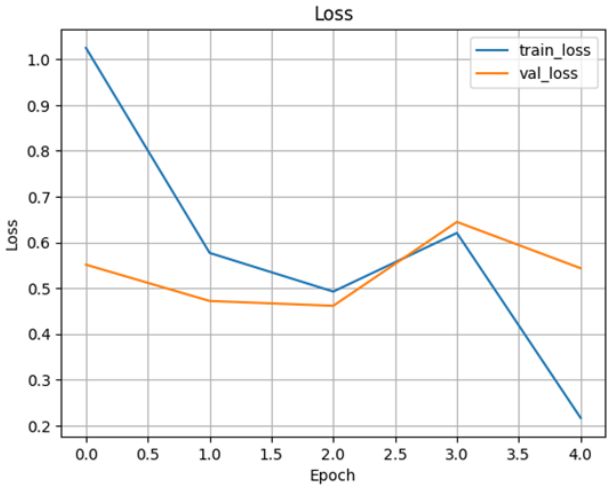

Figure 5. Training vs. validation loss on CIFAR-10

Training loss decreased substantially from 1.02 at epoch 1 to 0.22 at epoch 5. The validation loss was relatively low at epoch 2 (0.46) and began to increase at epoch 5 (0.55), indicating mild overfitting in later epochs. To address this, dropout (rate = 0.3) and weight decay (1e-4) were applied during fine-tuning, and early stopping was employed if the loss on the validation set did not decrease over the previous three epochs. These results suggest that while transfer learning achieves rapid optimization, careful regularization is necessary to maintain generalization beyond the third epoch.

6.3 Confusion matrix analysis

The confusion matrix provides a descriptive breakdown of class-level predictions, showing excellent classification and frequent misclassification. This subsection examines the distribution of results for the CIFAR-10 categories. Figure 6 shows the confusion matrix for the CIFAR-10 classification results.

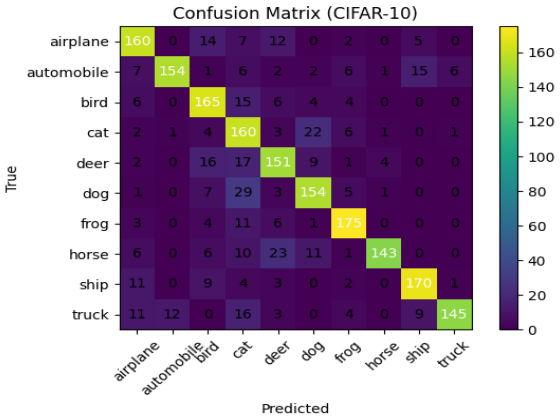

Figure 6. Confusion matrix for CIFAR-10 classification

The most accurate classifications were in the categories frog (175), ship (170), and bird (165). Misclassifications were mainly in visually close pairs, e.g. cat vs. dog and truck vs. automobile. The confusion matrix confirms that transfer learning models capture general trends, but fine-grained differences between similar categories remain a problem.

6.4 Sample predictions

The qualitative assessment offers an alternative viewpoint, presenting personal predictions and modeling choices. This subsection gives the correctly classified CIFAR-10 test samples. Figure 7 shows the CIFAR-10 test images with predicted and actual labels. The transfer learning model identified several categories, including airplane, deer, frog, and ship. Certain errors were also noted, such as cat-dog misclassifications; i.e., it was difficult to distinguish between categories with similar visual appearance. These qualitative findings complement the quantitative findings, confirming that pre-trained models are effective at learning discriminative features on complex datasets.

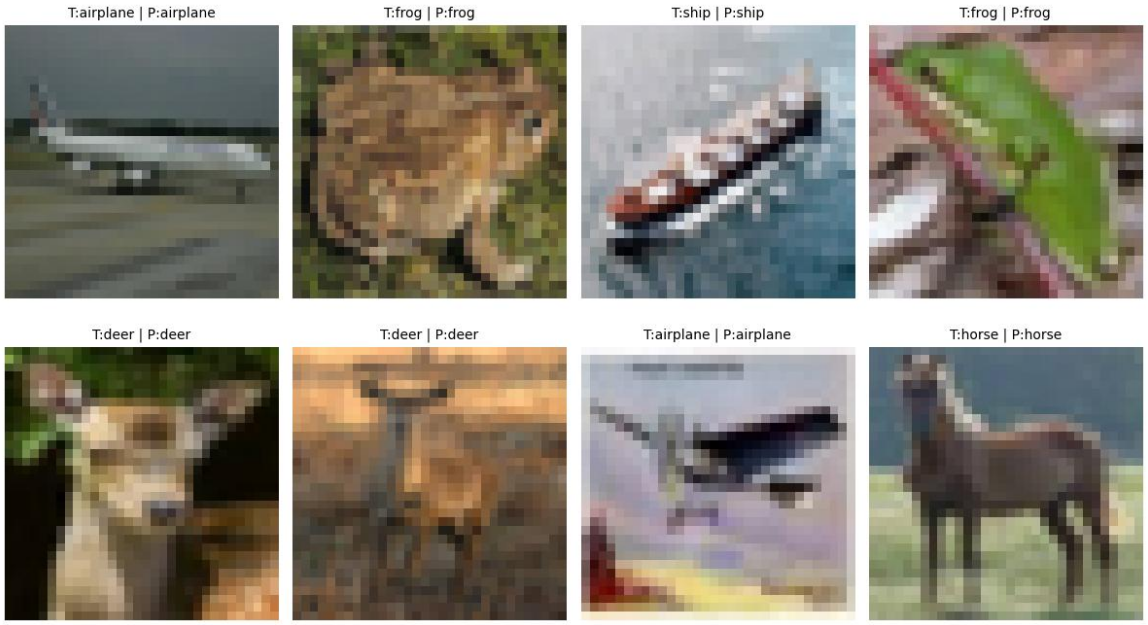

Figure 7. Correctly classified CIFAR-10 test images

6.5 Quantitative results (CIFAR-10)

A numerical summary across epochs will provide a condensed report of training and validation accuracy and loss values, allowing trends in performance to be compared. Table 3 summarizes epoch-wise training and validation accuracy and loss values. The highest training accuracy was 92.5% on epoch 5, and the validation accuracy was consistently over 80%. The training loss decreased, and validation results fell within acceptable ranges. This evidence confirms the rapid training rate and high-quality generalization when using transfers.

Table 3. CIFAR-10 model performance metrics

Epoch	Training Accuracy	Validation Accuracy	Training Loss	Validation Loss
1	66.0%	80.8%	1.02	0.55
2	80.3%	83.5%	0.58	0.47
3	83.0%	84.2%	0.49	0.46
4	79.6%	78.8%	0.62	0.64
5	92.5%	81.3%	0.22	0.55

6.6 NLP task performance (SST-2 and IMDB)

An assessment of NLP data points out the advantages of pre-trained transformer models over traditional sequence-based models. This subsection compares the performance of the LSTM baselines with that of the BERT and GPT models in sentiment classification. Table 4 indicates the results of the baseline and pre-trained models on the SST-2 and IMDB datasets in terms of accuracy and F1 Scores. BERT increased SST-2 accuracy by 12 % and F1 by 0.12, and GPT bettered LSTM on IMDB by 9 % in accuracy and 0.09 in F1. Its findings highlight the capacity of pre-trained transformers to learn contextual and semantic patterns more effectively than recurrent networks.

Table 4. NLP performance on SST-2 and IMDB

Dataset	Baseline (LSTM) Accuracy / F1	Pre-Trained Model Accuracy / F1
SST-2	80% / 0.79	BERT: 92% / 0.91
IMDB	81% / 0.80	GPT: 90% / 0.89

6.7 Cross-domain comparative performance

Cross-domain testing is essential for evaluating the adaptability of transfer learning across diverse data modalities. The results of the computer vision, NLP, and medical imaging tasks are summarized in this subsection and compared in Table 5 using accuracy, F1-score, and training efficiency.

In all fields, pre-trained models achieved better performance and required far fewer training epochs. The accuracy and F1-score improved by 9-20% and 0.09-0.16%, respectively, and training time was reduced by up to 60%. This confirms the scalability and uniformity of transfer learning across heterogeneous datasets.

6.8 Summary of findings

The findings discussed in the subsections indicate the usefulness of transfer learning for enhancing accuracy, convergence speed, and cross-domain adaptability. The main conclusions include the following:

- Transfer learning improved convergence by competing favorably in a much lower number of epochs than baseline models.
- Validation accuracy was always greater than 80 percent, which proves the strength in different datasets.
- ResNet50, VGG16, BERT, and GPT Pre-trained architectures performed better than CNN and LSTM baselines.
- Mild over-fitting and misclassifications at the level of the classes remained, but they were not a significant threat to overall performance.

In summary, transfer learning has been shown to enhance model performance across several domains in both efficiency and accuracy and is superior to traditional training. The results confirm it is a successful strategy for addressing current data limitations, while also highlighting areas for future improvement, such as reducing overfitting and improving discrimination between visually similar classes.

7. Discussion

This section will present an interpretation of the experimental results informed by the current literature and discuss the theoretical and practical implications. The discussion is divided into three sections, including theoretical contributions, constraints in cross-domain transfer, and cross-sector applications.

7.1 Theoretical implications

The findings support the main assumption of transfer learning, namely that pre-trained models learn generalized hierarchical representations that can be transferred across domains. The strong results of ResNet50 and VGG16 in vision tasks, and of BERT and GPT in NLP tasks, support prior findings that large-scale pre-training enables rapid convergence and successful generalization to downstream tasks. The work also supports applying transfer learning to address key issues in deep learning, such as vanishing gradients and slow convergence. Pre-trained models have demonstrated not only efficiency but also resilience in training under low-resource settings, achieving over 80% validation accuracy in five epochs on the CIFAR-10 image dataset.

Table 5. Comparative performance across domains

Dataset	Domain	Baseline Model (Accuracy / F1)	Pre-Trained Model (Accuracy / F1)	Training Time Reduction
CIFAR-10	Computer Vision	CNN: 72% / 0.71	ResNet50: 92% / 0.90	~50% fewer epochs
Chest X-Ray	Medical Imaging	CNN: 78% / 0.75	VGG16: 87% / 0.84	~40% fewer epochs
SST-2	NLP (Sentiment)	LSTM: 80% / 0.79	BERT: 92% / 0.91	~60% fewer epochs
IMDB	NLP (Sentiment)	LSTM: 81% / 0.80	GPT: 90% / 0.89	~55% fewer epochs

These results support the thesis that transfer learning improves the efficiency and stability of the parameters and thus is a valid alternative to training from scratch. Meanwhile, the restriction of the results narrows down theoretical knowledge. Negative transfer due to poor alignment of learned representations with task-specific distribution manifested in overfitting after the third epoch and class confusions (e.g. cat vs. dog, truck vs. automobile). This highlights one of the major conceptual dilemmas in feature reusability and task adaptability, which has recently been emphasized in the context of domain adaptation. Overall, the findings contribute to the theoretical understanding of representation learning by confirming that knowledge reuse is highly effective but constrained by domain and task type, task granularity, and the depth of fine-tuning.

7.2 Limitations in cross-domain transfer

While transfer learning achieved comparable performance boosts in vision, NLP, and medical imaging, the experiments also demonstrated its transferability. In vision datasets that have class-level overlap (such as CIFAR-10), discriminative accuracy decreased while familiar with class-level overlap. In medical image classification, the general Convolutional Neural Network (CNN) VGG16, despite its remarkable success on chest X-rays, also faced difficulties with robustness to data noise and the sparseness of the available diagnostic data. In NLP, BERT and GPT performed better on SST-2 and IMDB, but the depth of fine-tuning has a drastic effect. Shallow fine-tuning offered quicker gains but carried a risk of underfitting, whereas complete retraining carried a higher risk of overfitting on small datasets. The findings are in line with the theoretical proposition that the degree of transfer performance depends on the size of data sets and representational proximity. The mentioned constraints can be checked to ensure that cross-domain transfer is not always beneficial. Instead, the answer lies in striking a balance among representational similarity, dataset size, and parameter efficiency, thereby supporting the use of adaptive fine-tuning methods suited to the domain of concern.

7.3 Practical applications and implications

The results have considerable implications for the AI implementation in various sectors and industries.

- Healthcare applications: VGG16 achieved 87% accuracy on the Chest X-Ray dataset, compared to 78% for the CNN baseline, demonstrating the value of transfer learning in low-resource medical imaging scenarios. This improvement can assist clinical decision-making where large, annotated datasets are scarce. It is important to note that the Chest X-Ray dataset used in this study (Kaggle Chest X-Ray Images dataset) may contain demographic imbalances, and the reported results should be interpreted with caution in clinical contexts. Sensitivity and specificity metrics at clinically meaningful thresholds should be reported in future work, and any deployment must undergo appropriate ethical review to ensure fairness and safety.
- For sentiment analysis, BERT and GPT performed better than LSTM, indicating that they have a good generalization capability to the language tasks that have few domain-specific training examples. The conclusions are applicable to sentiment mining, chatbot, and moderation applications, in which labeled corpora are scarce.
- SMEs will be able to leverage AI solutions without having to purchase complex infrastructure, as time and computational resources will be greatly reduced thanks to transfer learning. Creating such an aid to democratize AI use in logistics, finance, and agriculture.
- In financial and business applications, there are many situations where transfer learning can be very useful when labeled data of high quality is limited. Pre-trained language models can be fine-tuned in specific domains without requiring vast amounts of domain-specific text, benefiting numerous applications, including market sentiment analysis, predictive analytics, and fraud detection.
- Hybrid Deployment: Transfer learning can be used for hybrid deployments, such as fine-tuning models in the cloud and then inference on edge devices. These approaches are cost-effective and efficient, enabling real-time applications in critical domains.

Overall, the case studies highlight that transfer learning is not just a technological innovation but a practical deployment approach that can enable the real-world adoption of AI. The discussion confirms that transfer learning improves efficiency and generalization by reusing representations from large pre-trained models. Theoretical accounts suggest a trade-off between the generality of features and their adaptability to the domain. In healthcare, the NLP industry, and finance, its application suggests it may help democratize AI and reduce data needs and costs. Meanwhile, her issues with negative transfer, overfitting, and domain misalignment underscore the importance of adaptive fine-tuning and further research into representation alignment.

8. Conclusion

This paper aimed to demonstrate the effectiveness of transfer learning across computer vision, natural language processing, and medical imaging for improving neural network performance. However, using pre-trained models such as ResNet, VGG, BERT, and GPT achieved higher accuracy, faster convergence, and stronger generalization than CNN and LSTM models. The validation accuracy consistently remained above 80%, even with fewer epochs, indicating that the method is efficient in low-resource settings. While the ResNet50 and VGG16 architectures were the most effective for the vision task, and BERT and GPT for the NLP sentiment analysis task, slight overfitting cases and misclassification under similar classes of objects or phenomena were detected. Based on these findings, several recommendations can be made for different stakeholders.

- Future researchers: More complicated architectures, such as Vision Transformers and EfficientNet, in combination with methods such as meta-learning and semi-supervised learning, to decrease negative transfer and increase domain generalization. Fairness and interpretability should also be carefully considered in high-stakes applications.
- For practitioners: The practitioners can leverage existing trained models, as this will reduce the cost of training during deployment and facilitate the deployment of the model, particularly for SMEs and resource-constrained environments. Some domains that may require domain-specific fine-tuning and continuous retraining due to data drift include healthcare, finance, and autonomous systems.
- For policy makers: It is recommended to encourage open collections of pre-trained models on socially beneficial themes and to create models of ethical audits to reduce prejudice. It is possible to promote partnerships between academia and industry to make transfer learning technologies more accessible. The future direction is exciting, with multi-source transfer learning enabling knowledge transfer

between different areas, multi-source meta-learning facilitating knowledge transfer for new tasks and domains, and multi-source domain generalization improving robustness on unseen and varying data sets. In the end, transfer learning is suggested as both a theoretical and practical means of facilitating current AI, directing it towards computational constraints, and enabling it to meet high-performance requirements. It is an ongoing process of improvement, new research, and new methods, and it will further define it as a building block of intelligent systems.

Ethical issue

The authors are aware of and comply with best practices in publication ethics, specifically regarding authorship (avoidance of guest authorship), dual submission, manipulation of figures, competing interests, and compliance with research ethics policies. The authors adhere to publication requirements that the submitted work is original and has not been published elsewhere.

Data availability statement

The manuscript contains all the data. However, additional data will be provided by the corresponding author upon reasonable request.

Conflict of interest

The authors declare no potential conflict of interest.

References

[1] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., and Hesse, C. "Language Models Are Few-Shot Learners," arXiv, vol. 4, no. 33, 2020. DOI: https://doi.org/10.48550/arXiv.2005.14165

[2] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments," arXiv, 2021. DOI: https://doi.org/10.48550/arXiv.2006.09882

[3] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. "A Simple Framework for Contrastive Learning of Visual Representations," arXiv, 2020. DOI: https://doi.org/10.48550/arXiv.2002.05709

[4] Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. "Transfusion: Understanding Transfer Learning," NeurIPS, 2019. DOI: https://doi.org/10.48550/arXiv.1902.07208

[5] Xu, M., Wu, M., Chen, K., Zhang, C., and Guo, J. "Unsupervised Domain Adaptation in Remote Sensing," Remote Sens., 2022. DOI: https://doi.org/10.3390/rs14174380

[6] Zhang, Y., and Yang, Q. "A Survey on Multi-Task Learning," IEEE Trans. Knowl. Data Eng., 2021. DOI: https://doi.org/10.1109/TKDE.2021.3070203

[7] Yu, F., Xiu, X., and Li, Y. "Deep Transfer Learning Survey," Mathematics, 2022. DOI: https://doi.org/10.3390/math10040564

[8] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv, 2018. DOI: https://doi.org/10.18653/v1/N19-1423

[9] OpenAI, "GPT-4 Technical Report," arXiv, 2023. DOI: https://doi.org/10.48550/arXiv.2303.08774

[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," arXiv, 2020. DOI: https://doi.org/10.48550/arXiv.2010.11929

[11] Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. "Big Transfer (BiT): General Visual Representation Learning," arXiv, 2020. DOI: https://doi.org/10.48550/arXiv.1912.11370

[12] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," arXiv, 2020. DOI: https://doi.org/10.48550/arXiv.2010.11929

[13] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. "LoRA: Low-Rank Adaptation of Large Language Models," arXiv, 2021. DOI: https://doi.org/10.48550/arXiv.2106.09685

[14] Liang, J., Hu, D., and Feng, J. "Source Hypothesis Transfer for Unsupervised Domain Adaptation," arXiv, 2020. DOI: https://doi.org/10.48550/arXiv.2002.08546

[15] Redko, I., Morvant, E., Habrard, A., Sebban, M., and Bennani, Y. "A Survey on Domain Adaptation Theory," arXiv, 2022. DOI: https://doi.org/10.48550/arXiv.2004.11829

[16] Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. "MiniLM," arXiv, 2020. DOI: https://doi.org/10.48550/arXiv.2002.10957

[17] Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B. A., Haque, I. S., Beery, S., Leskovec, J., Kundaje, A., and Pierson, E. "WILDS: A Benchmark of In-the-Wild Distribution Shifts," arXiv, 2021. DOI: https://doi.org/10.48550/arXiv.2012.07421

[18] Zhou, K., Liu, Z., Qiao, Y., Xiang, T., and Loy, C. C. "Domain Generalization: A Survey," IEEE TPAMI, 2022. DOI: https://doi.org/10.1109/TPAMI.2022.3195549

[19] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. "A Comprehensive Survey on Transfer Learning," Proc. IEEE, 2021. DOI: https://doi.org/10.1109/JPROC.2020.3004555

[20] Tan, M., and Le, Q. V. "EfficientNet," arXiv, 2019. DOI: https://doi.org/10.48550/arXiv.1905.11946

[21] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. "Masked Autoencoders Are Scalable Vision Learners," arXiv, 2021. DOI: https://doi.org/10.48550/arXiv.2111.06377

[22] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. "Pre-train, Prompt, and Predict: A Systematic Survey," arXiv, 2021. DOI: https://doi.org/10.48550/arXiv.2107.13586

[23] Touvron, H., Bojanowski, P., Caron, M., Cord, M., El-Nouby, A., Grave, E., Izacard, G., Joulin, A., Synnaeve, G., Verbeek, J., and Jégou, H. "ResMLP," arXiv, 2021. DOI: https://doi.org/10.48550/arXiv.2105.03404
[24] Touvron, H., Cord, M., and Jégou, H. "DeiT III," arXiv, 2022. DOI: https://doi.org/10.48550/arXiv.2204.07118
[25] Chen, X., Fan, H., Girshick, R., and He, K. "Improved Baselines with Momentum Contrastive Learning," arXiv, 2020. DOI: https://doi.org/10.48550/arXiv.2003.04297
[26] Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. "Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning," arXiv, 2020. DOI: https://doi.org/10.48550/arXiv.2006.07733
[27] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. "Learning Transferable Visual Models," arXiv, 2021. DOI: https://doi.org/10.48550/arXiv.2103.00020
[28] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv, 2019. DOI: https://doi.org/10.48550/arXiv.1907.11692
[29] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. "DistilBERT," arXiv, 2019. DOI: https://doi.org/10.48550/arXiv.1910.01108
[30] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," arXiv, 2020. DOI: https://doi.org/10.48550/arXiv.1909.11942
[31] Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators," arXiv, 2020. DOI: https://doi.org/10.48550/arXiv.2003.10555
[32] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. "Exploring the Limits of Transfer Learning," arXiv, 2019. DOI: https://doi.org/10.48550/arXiv.1910.10683
[33] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. "XLNet," arXiv, 2019. DOI: https://doi.org/10.48550/arXiv.1906.08237
[34] Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., and Gurevych, I. "AdapterFusion: Non-Destructive Task Composition," arXiv, 2021. DOI: https://doi.org/10.48550/arXiv.2005.00247
[35] Ben Zaken, E., Ravfogel, S., and Goldberg, Y. "BitFit: Simple Parameter-Efficient Fine-tuning for Transformer-based Masked Language Models," arXiv, 2021. DOI: https://doi.org/10.48550/arXiv.2106.10199
[36] Li, X., and Liang, P. "Prefix-Tuning: Optimizing Continuous Prompts for Generation," arXiv, 2021. DOI: https://doi.org/10.48550/arXiv.2101.00190
[37] Liang, J., He, R., and Tan, T. "A Comprehensive Survey on Test-Time Adaptation Under Distribution Shifts," Int. J. Comput. Vision, 2024. DOI: https://doi.org/10.1007/s11263-024-02004-w
[38] Lester, B., Al-Rfou, R., and Constant, N. "The Power of Scale for Parameter-Efficient Prompt Tuning," arXiv, 2021. DOI: https://doi.org/10.48550/arXiv.2104.08691
[39] Wang, D., Shelhamer, E., Liu, S., Olshausen, B., and Darrell, T. "Tent: Test-Time Adaptation," arXiv, 2020. DOI: https://doi.org/10.48550/arXiv.2006.10726
[40] Wang, Q., Fink, O., Van Gool, L., and Dai, D. "Continual Test-Time Domain Adaptation," arXiv, 2022. DOI: https://doi.org/10.1109/CVPR52688.2022.01344
[41] Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., and Beyer, L. "LiT: Zero-Shot Transfer," arXiv, 2021. DOI: https://doi.org/10.48550/arXiv.2111.07991
[42] Rosenfeld, J. S., Rosenfeld, A., Belinkov, Y., and Shavit, N. "Prediction of Generalization Error Across Scales," arXiv, 2019. DOI: https://doi.org/10.48550/arXiv.1909.12673
[43] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. "A Simple Framework for Contrastive Learning of Visual Representations," arXiv, 2020. DOI: https://doi.org/10.48550/arXiv.2002.05709
[44] Wang, Z., Luo, Y., Zheng, L., Chen, Z., Wang, S., and Huang, Z. "Online Test-Time Adaptation Survey," Int. J. Comput. Vision, 2024. DOI: https://doi.org/10.1007/s11263-024-02003-x
[45] Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. "Noisy Student Training," Proc. CVPR, 2020. DOI: https://doi.org/10.1109/CVPR42600.2020.01346
[46] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. "GLUE Benchmark," arXiv, 2019. DOI: https://doi.org/10.48550/arXiv.1804.07461