Advances in Computational Sciences and Technology
ISSN 0973-6107 Volume 18, Number 2 (2025) pp. 83-94

https://dx.doi.org/10.37622/ACST/1.8.25.83-94

Foundations and Frontiers of Transfer Learning in NLP : A Comprehensive Review

Mrs. Vaishali Suryawanshi[1] **Dr. Abhijeet Kaiwade**[2]
Research Scholar, *Research Guide,*
vaishali.suryawanshi2005@gmail.com[1] *kaiwade@gmail.com*[2]
Yashaswi Education Society's International Institute of Management Science , Savitribai Phule Pune University , Pune, India.

Abstract

The field of Natural Language Processing (NLP) has been transformed by Transfer Learning (TL) since it permits re-purposing of models trained on extensive data for narrower tasks with little extra supervision. Within the last decade, the evolution of TL has been the more advanced embedding reuse with transformer based models pre-trained on multilingual, multimodal, or task agnostic corpora. This paper aims to provide a review of TL in NLP combining recent literature. We cover the theory, pre-trained models, applications, benchmark datasets, and TL in NLP paradigms. Moreover, we provide a critical focus on adaptations for the low-resource situation, cross-domain adaptations, and the challenge of explainability. The final part of the paper discusses trends such as instruction tuning, quantum transfer, and learning with fewer parameters that are more efficient.

Keywords— Transfer Learning, NLP, Pretrained Language Models, Fine-tuning, Few-shot Learning, Multilingual NLP.

1. Introduction

The most prominent breakthrough in this evolution was the introduction of **BERT** (Bidirectional Encoder Representations from Transformers) [22], which showcased the power of contextualized word embeddings and masked language modeling. This was soon followed by **GPT**, **RoBERTa**, **T5**, and numerous other architectures [23], [12], each pushing the boundaries of transferability and generalization. The strength of these models lies in their ability to capture syntactic, semantic, and contextual nuances across varied tasks without task-specific design.

Transfer learning (TL) has become a phenomenal innovation in Natural Language Processing (NLP) with impact in model planning, training, and deployment. In the past, each NLP system was trained individually for a given task and it drew from vast streams of annotated data. Such an approach was expensive and time consuming,

especially in the case of scarce resources. However, the use of TL with deep neural networks and pretrained language models has allowed ease of retrieval of knowledge from one task or domain for use in another. [5], [13], [17].

The most prominent breakthrough in this evolution was the introduction of BERT (Bidirectional Encoder Representations from Transformers) [22]. It demonstrated the usefulness of contextualized word embeddings alongside masked language modeling. Subsequently, BERT was followed by GPT, RoBERTa, T5, and many others [23, 12]. These models strengthened the application of TL and generalization. These models are powerful as they understand the syntax, semantics, and context of language and the models are able to use this information across varied tasks seamlessly More recent studies have delved into TL for cross-lingual learning [9], for translating in turbulent environments [19], for Alzheimer's disease detection in specifically tailored domains [10], and programming tasks like code searching [1], [2]. TL is now a principal strategy in modern NLP systems, as highlighted in [15], [14], and [17].

As the field evolves, these advancements suggest a wider implementation in AI; for example, transfer learning becomes increasingly multimodal, as with CLIP [3] or even quantum-aware [20].

2. Basics of Transfer Learning

Transfer learning refers to the process of improving the learning process of a specific target task by utilizing knowledge gained from a related source task. This approach is very effective in the field of NLP because of the interconnections and unified structures in languages compared to tasks and fields. [5], [16].

2.1. Formal Definition

Let a source domain and task be denoted by $\mathcal{D}_S$ and $\mathcal{T}_S$. TL seeks to improve target task $\mathcal{T}_T$ in target domain $\mathcal{D}_T$ by leveraging knowledge of $o\mathcal{D}_S$ and $\mathcal{T}_S$. $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$ [13].

2.2. Categories of Transfer Learning

- **Inductive Transfer Learning:** Labeled data is available for the target task. Most current NLP use cases ((e.g., using BERT for NER) falls under this category. [22], [16].
- **Transductive Transfer Learning**: English and Spanish sentiment classification serve as an example. Cross-lingual NLP is a great example of this. [9].
- **Unsupervised Transfer Learning:** Focus is on representation learning. This approach applies when there is no labeled data in the source or target domain. [18].

2.3. The Benefits of Transfer Learning for NLP

A major factor in TL's success in NLP is its capacity to reuse distributed representations of language. Models developed for one task (like language modeling) can frequently be optimized to perform well for another (like sentiment analysis) because linguistic structures like syntax and semantics generalize across tasks [15], [22].

Additionally, richer representations than static word vectors are made possible by contextual embeddings. This is accomplished by unsupervised pretraining on sizable corpora in models such as ELMo, BERT, and T5. [5].

2.4. Restrictions and Difficulties

Despite its strength, TL is not always successful:

• **Negative Transfer:** Performance may suffer when source and target domains differ [14].

• **In low-data regimes, overfitting**: Poor generalization can result from fine-tuning with very little data [15].

• Cost of computation: Pretraining big models like GPT-3 demands extensive resources [23].

Despite these issues, recent advancements like adapter layers, prompt tuning, and federated fine-tuning aim to mitigate these limitations [19], [20].

3. NLP Transfer Learning Paradigms

Depending on how the pretrained knowledge is applied again, different paradigms are used to achieve transfer learning in NLP. These consist of:

3.1. Transfer Learning Based on Features

Before transformers became popular, this was the most common approach. In this case, contextual embeddings like ELMo or word embeddings like GloVe or word2vec were handled as static input features for downstream models [13]. They were helpful, but not very flexible.

3.2. Optimizing Pretrained Models

Fine-tuning has become the norm with models like BERT, RoBERTa, T5, and others. The target task is used to update the entire model or specific layers in this paradigm [22], [12]. On datasets like GLUE or SQuAD, fine-tuning frequently yields state-of-the-art results and allows for deeper alignment with the characteristics of the target task.

3.3. Learning through Progressive Transfer

By permitting lateral connections between pretrained and newly created task-specific subnetworks, progressive neural networks (PNNs), as assessed in [4], reduce catastrophic forgetting. The reusability of previously acquired knowledge is enhanced by this architecture.

3.4. Meta-learning and multitasking

In this context, models are trained on several tasks at once or use meta-learning techniques to quickly adjust to new tasks. Multi-task models require less target task data for fine-tuning and have better generalization, as demonstrated by works such as [14].

3.5. Adapter-based and prompt-based learning

Both prompt tuning and instruction tuning are examples of parameter-efficient paradigms in which only the input prompt or a tiny adapter module is updated, while model parameters stay fixed [19], [20]. These methods facilitate effective domain adaptation and lessen the memory footprint.

3.6. Cross-lingual and unsupervised paradigms

While cross-lingual transfer (e.g., using XLM-R or mT5) enables leveraging knowledge across languages, as highlighted in [8], [9], and [21], unsupervised learning has been essential to the success of models such as GPT-3 [23].

Language models that have been trained (PLMs)

In NLP, trained language models, or PLMs, are now the mainstay of contemporary transfer learning. These models are initially trained using tasks like masked language modeling (MLM), causal language modeling (CLM), or denoising autoencoding on large unsupervised corpora (like Wikipedia and Common Crawl). They are refined on particular downstream tasks following pretraining, frequently resulting in state-of-the-art performance.

a) **Table 1.** Key Pretrained Language Models Overview

Model	Architecture	Highlights	Year	Ref
BERT	Encoder	Masked LM + Next Sentence Prediction	2019	[22]
GPT	Decoder	Left-to-right, generative pretraining	2018	[23]
RoBERTa	Encoder	No NSP, dynamic masking, longer training	2019	[13]
XLNet	Permutation-based	Combines BERT with autoregression	2019	[11]
T5	Encoder-Decoder	Unified text-to-text format	2020	[4]
mT5	Multilingual T5	Trained on 100+ languages (mC4 corpus)	2021	[12]
DeBERTa	Disentangled attention	Separates position and content encoding	2021	[14]
ELECTRA	Discriminator-based	Replaces MLM with replaced token detection	2020	[17]

Model	Architecture	Highlights	Year	Ref
DistilBERT	Compressed BERT	40% smaller, faster inference	2019	[15]

Each of these models introduced efficiency gains or architectural innovations that made them appropriate for particular use cases, such as generative capabilities, low-resource deployment, or multilingual tasks.

4.1. BERT and Its Variants

BERT (Bidirectional Encoder Representations from Transformers) [22] used MLM and next sentence prediction (NSP) to model context in both directions. This model is based on a transformer encoder and was trained on BookCorpus and English Wikipedia. BERT changed the game in NLP by doing better than older models at things like answering questions (SQuAD), classifying sentences (GLUE), and NLI (MNLI).

RoBERTa is a better version of BERT that removes the NSP objective and uses dynamic masking and longer training to get big performance gains [13].

DistilBERT and TinyBERT are smaller, lighter versions that aim to lower inference time and memory use while keeping the same level of accuracy. They are good for mobile or edge devices [15].

4.2. GPT and models that use autoregression

GPT and its successors, GPT-2 and GPT-3 [23], use a unidirectional decoder-only architecture and are trained to model language in a way that makes sense. GPT is better than BERT at tasks that involve creating things, like dialogue, summaries, and completing text. Its success in few-shot and zero-shot learning made people want to learn more about prompt-based transfer.

DialoGPT [8] fine-tunes GPT-2 on conversational data, which makes open-domain dialogue systems work better. GPT-3 has 175 billion parameters, which lets it generalize with just a few examples without needing to be fine-tuned. This trend continues with GPT-4 and InstructGPT.

4.3. Models for encoding and decoding

T5 (Text-to-Text Transfer Transformer) [4] sees every task as a text-to-text problem and does very well at summarizing, answering questions, and translating. It was trained on the C4 corpus ahead of time with a denoising goal.

BART is another encoder-decoder model that uses both the MLM and autoregressive loss, which makes it good for tasks that involve generating sequences [7].

mT5 is an extension of T5 that works on multilingual tasks by pretraining on mC4 (101 languages). It has been shown to work well in cross-lingual transfer learning [12].

4.4. Vision-Language and Multimodal Models

Using joint training on image-text pairs, models such as FLAVA and CLIP [3] show off the effectiveness of cross-modal transfer learning. These models demonstrate how

language comprehension can inform visual learning and vice versa, even though they are not strictly NLP.

5. Use of Transfer Learning in Natural Language Processing

Numerous NLP tasks from various industries are supported by transfer learning. Key applications enhanced by PLMs are listed below:

5.1. Text Classification

Transfer learning is very helpful for tasks like sentiment analysis, spam detection, and intent classification. For instance, BERT performs noticeably better than CNN or LSTM-based models when fine-tuned on IMDB or Yelp datasets [5, 14]. For lightweight deployment in real-time classification systems, DistilBERT and RoBERTa are frequently utilized.

5.2. NER, or Named Entity Recognition

NER models have historically used CRFs or BiLSTMs, but transfer learning has made BERT-style models the standard [1], [13]. In specific domains, such as legal and medical documents, refined BERT models identify entities like names, locations, and dates.

5.3. Reading comprehension and question answering

High F1 and EM scores are obtained when BERT and XLNet are optimized on SQuAD, HotpotQA, or NewsQA. In extractive and abstractive QA, T5 and BART also exhibit strong performance. [4], [7].

5.4. Interpretation by Machines

Low-resource language translation is made possible by multilingual pretrained models such as mBART, mT5, and XLM-R [9], [21]. For neural machine translation (NMT) with low adaptation overhead, studies like [19] present trivial transfer learning techniques.

5.5. Chatbots and Dialogue Systems

For open-domain conversational agents, models such as BlenderBot, DialoGPT, and GPT-2 are utilized. These models can be tailored to particular fields, such as customer service, education, and healthcare, thanks to transfer learning [8].

5.6. Low-resource adaptation and cross-lingual natural language processing

In multilingual and cross-lingual NLP, transfer learning has revolutionized the field [9], [18]. Through multilingual pretraining and shared embeddings, models can generalize to languages with little labeled data by utilizing high-resource languages like English.

5.7. Healthcare and Biomedical NLP

TL is used to detect cognitive diseases, classify medical records, and predict diagnoses. For example, [10] offers a TL method for identifying Alzheimer's disease from speech transcripts, demonstrating encouraging accuracy despite the paucity of data.

5.8. Code Search and Software Engineering

Transfer learning has expanded to include understanding and retrieving codes in addition to natural language. [1], [2] show that pre-trained models such as CodeBERT

or GraphCodeBERT greatly enhance code search, code summarization, and code clone detection performance.

6. Datasets and Benchmarks

Transfer learning success in NLP is closely tied to the availability of high-quality pretraining corpora and standardized benchmarks for evaluation.

b) 6.1. Pretraining Corpora

- **C4** (Colossal Clean Crawled Corpus) for T5 [4]
- **BookCorpus + Wikipedia** for BERT [22]
- **OpenWebText + WebGPT** for GPT [23]
- **mC4 + CC100** for mT5 and multilingual models [12]

c) 6.2. Downstream Benchmarks

- **GLUE**: A suite of tasks including SST-2, MNLI, and QQP. Models like RoBERTa and T5 achieve near-human performance [6], [13].
- **SuperGLUE**: Harder successor to GLUE, used to evaluate instruction tuning and few-shot models [19].
- **SQuAD (1.1 & 2.0)**: Popular QA benchmark. BERT and XLNet dominate leaderboards [1].
- **XQuAD / MLQA**: Cross-lingual QA datasets used for evaluating multilingual transfer [23].
- **CodeSearchNet**: Benchmark for code retrieval, used in [1], [2].

6.3. Measures of Evaluation

• F1-Score and Accuracy: Classification tasks
• ROUGE and Exact Match (EM): QA and summary
• ChrF, TER, and BLEU: Machine translation
• MAP and Mean Reciprocal Rank (MRR): IR and code search tasks

These metrics and datasets offer a strong basis for standardized assessment, promoting model comparability and reproducibility.

7. Challenges and Limitations

Even though transfer learning has become widely used, there are still a number of practical, ethical, and technical issues that arise in various NLP tasks and domains..

d) 7.1 Key Challenges

Challenge	Description	References
Negative Transfer	Source and target tasks/domains are too dissimilar, leading to degraded performance.	[14], [17]
Catastrophic Forgetting	Fine-tuning on new tasks can overwrite learned knowledge from pretraining.	[4], [12]
Bias and Fairness	Pretrained models often encode social biases from their training data.	[20], [23]

Challenge	Description	References
Data Privacy	Large models may memorize and unintentionally leak sensitive data.	[10], [13]
Interpretability	Models are black-box and difficult to explain, especially in high-stakes applications.	[11], [14]
Computational Cost	Pretraining requires massive GPU/TPU resources and energy, limiting accessibility.	[3], [23]
Low-Resource Domains	Transfer does not always yield improvements in very low-data or niche domains.	[8], [9], [19]

These challenges motivate research into more efficient, fair, and adaptable models that preserve the benefits of transfer learning without introducing new risks.

8. Recent Advancements and Trends

The past few years have seen rapid innovation in how transfer learning is applied, especially in low-resource, multilingual, and prompt-based setups. Below is a summary of major advances:

e) 8.1 Key Advancements in Transfer Learning

Trend / Method	Impact / Contribution	Example Models / Papers
Prompt Engineering	Enables zero-shot/few-shot generalization using natural language prompts	GPT-3, InstructGPT [19], [23]
Adapter Tuning	Efficiently adds task-specific modules without changing base PLM parameters	Prefix-Tuning, AdapterFusion [20]
Multilingual Transfer	Cross-lingual performance improves using mT5, XLM-R, or mBART	mT5, XLM-R [8], [12]
Knowledge Distillation	Compresses large PLMs into lightweight	DistilBERT, TinyBERT [15]

Trend / Method	Impact / Contribution	Example Models / Papers
	deployable models	
Multimodal Learning	Uses joint image-text training to generalize across modalities	CLIP, Flamingo [3]
Quantum Transfer Learning	Applies quantum circuits and states for language modeling and classification	QNLP, QTL [20]
Progressive Networks	Enables continual learning without forgetting prior tasks	PNNs [4], [17]

These developments are shaping how models are trained and deployed, making transfer learning more adaptable, scalable, and cost-efficient.

9. Future Directions

Research in language translation for NLP is quickly advancing, and new possibilities are opening up.

9.1 Research Outlook

Research Focus	Objective
Low-resource and minority languages	Build robust multilingual systems that adapt to underrepresented linguistic settings.
Federated and private fine-tuning	Ensure user data remains local to protect privacy.
Green NLP and efficiency	Reduce model size, training cost, and carbon footprint.
Explainability and interpretability	Develop transparent models for healthcare, law, and education.
Task-agnostic pretraining objectives	Unify NLP tasks under generalized text-to-text or contrastive frameworks.

Research Focus	Objective
Continual and lifelong learning	Train models that accumulate knowledge without forgetting.
Hybrid Neuro-symbolic systems	Integrate logical reasoning with deep learning models for robust inference.

A critical future challenge will be balancing **scalability, fairness, and explainability** in transfer learning systems.

10. Conclusion

In NLP, transfer learning has emerged as a fundamental technique that opens up new avenues for classification, generation, translation, and even cross-modal reasoning. In practically every NLP benchmark, models like BERT, GPT, and T5 have outperformed conventional systems thanks to extensive pretraining and fine-tuning techniques.

The theoretical foundations, practical paradigms, and real-world applications of TL in NLP were presented in this review along with a discussion of its difficulties, current trends, and potential future applications. Transfer learning is expected to continue to be a key component of intelligent language systems as the field moves toward more understandable, effective, and accessible models.

References

[1] Y. Guo, X. Peng, L. Nie, et al., "On the Effectiveness of Transfer Learning for Code Search," *IEEE Transactions on Software Engineering*, 2022. [Online]. https://ieeexplore.ieee.org/document/9835142

[2] Z. Liu, Y. Chen, et al., "Using Transfer Learning for Code-Related Tasks," *IEEE Access*, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9797060

[3] A. Radford, J. W. Kim, C. Hallacy, et al., "Learning Transferable Visual Models From Natural Language Supervision," 2021. [Online]. Available: https://www.semanticscholar.org/reader/6f870f7f02a8c59c3e23f407f3ef00dd1dcf8fc4

[4] M. Moeed and D. Agerer, "An Evaluation of Progressive Neural Networks for Transfer Learning in Natural Language Processing," 2021. [Online]. Available: https://www.semanticscholar.org/paper/1e330152fdbb9199d221e93d1bca03d787637662

[5] S. J. Pan and Q. Yang, "A Survey on Transfer Learning for Speech and Language Processing," *IEEE Signal Processing Magazine*, vol. 33, no. 6, pp. 14–23, 2016. [Online]. Available: https://ieeexplore.ieee.org/document/7415532

[6] Y. Wang, Y. Sun, et al., "Supervised Contextual Embeddings for Transfer Learning in Natural Language Processing Tasks," 2022. [Online]. Available: https://www.semanticscholar.org/reader/648311a631a04f24f0351122dee77bb9c625f92f

[7] H. Guo, L. Shi, et al., "Research on Transfer Learning Technology in Natural Language Processing," in *Advances in Computer Science and Engineering*, Springer,

2021. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-15-8599-9_55
[8] L. Hedderich, J. Lange, and D. Adelani, "A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios," in *Proc. NAACL-HLT*, 2021. [Online]. Available: https://aclanthology.org/2021.naacl-main.201/
[9] M. Pikuliak, M. Šimko, and M. Bieliková, "Cross-lingual learning for text processing: A survey," *Expert Systems with Applications*, vol. 164, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0957417420305893
[10] Z. Tan, X. Zhou, et al., "A Transfer Learning Method for Detecting Alzheimer's Disease Based on Speech and Natural Language Processing," *Frontiers in Public Health*, vol. 10, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpubh.2022.772592
[11] Y. Zhang, S. Li, et al., "Analysis on Transfer Learning Models and Applications in Natural Language Processing," *High Scientific Education and Technology*, vol. 16, 2023. DOI: https://doi.org/10.54097/hset.v16i.2609
[12] A. Umar, S. Kumar, et al., "Unlocking Transfer Learning's Potential in Natural Language Processing: An Extensive Investigation and Evaluation," *IEEE Access*, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10743260
[13] S. Zhuang, X. Qi, et al., "A Decade Survey of Transfer Learning (2010–2020)," *IEEE Transactions on Artificial Intelligence*, vol. 1, no. 1, pp. 1–16, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9336290
[14] D. Bhosale and P. Borkar, "Transfer Learning in Natural Language Processing," *IEEE*, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10912895
[15] A. Alyafeai and M. Alshaibani, "A Survey on Transfer Learning in Natural Language Processing," 2021. [Online]. Available: https://www.semanticscholar.org/paper/5c09b7be6a7d4b0cd34fd95c4f783dc5c266edf3
[16] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, "Transfer Learning in Natural Language Processing," 2019. [Online]. Available: https://ruder.io/nlp-transfer-learning
[17] C. Tan, F. Sun, et al., "A Comprehensive Survey on Transfer Learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 9, pp. 4952–4973, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9134370
[18] S. J. Pan and Q. Yang, "A Survey of Transfer Learning," *Journal of Big Data*, vol. 3, 2016. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-016-0043-6
[19] A. Siddhant and Z. Kale, "Trivial Transfer Learning for Low-Resource Neural Machine Translation," in *Proc. EMNLP Workshop on Machine Translation*, 2018. [Online]. Available: https://aclanthology.org/W18-6325/
[20] A. Savytskyi, N. Zein, et al., "Quantum Transfer Learning for Acceptability Judgements," *arXiv preprint arXiv:2401.07777*, 2024. [Online]. Available: https://arxiv.org/abs/2401.07777
[21] W. Wang, X. Wang, et al., "Transfer Learning Based on Lexical Constraint Mechanism in Low-Resource Machine Translation," *Computers and Electrical*

Engineering, vol. 100, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0045790622001483
[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL*, 2019. DOI: 10.18653/v1/N19-1423.
[23] A. Radford, J. Wu, R. Child, et al., "Language Models are Unsupervised Multitask Learners," *OpenAI Blog*, 2019. [Online]. Available: https://www.bibsonomy.org/bibtex/1b926ece39c03cdf5499f6540cf63babd