**International Journal of Academic Engineering Research (IJAER)**
**ISSN: 2643-9085**
**Vol. 10 Issue 6 June - 2026, Pages: 97-104**

# Transfer Learning and Domain Adaptation: Supervised Learning Across Distribution Shifts

**Virendra Tank**

Shri Mahaveer College, Jaipur, India

vtank87@gmail.com 0009-0003-9126-0982

***Abstract:*** *Transfer learning and domain adaptation are two important paradigms in contemporary machine learning, allowing models to utilize knowledge from source domains to work well in target domains with only little or no labeled data. In this review, we present the theoretical framework, methodological development and practical applications of transfer learning under distribution shifts. We investigate pre-trained basis models, domain adaptation methods such as adversarial and optimal transportation approaches, the few-shot and zero-shot paradigms of learning, as well as their consideration in low-resource conditions. The paper reviews the latest progress of this fast developing technology, challenges and prospect directions are illuminated.*

**Keyword:** BERT, GPT series, Vision Transformers, CLIP, Adversarial methods (DANN), optimal transport approaches, Meta-learning (MAML), prompt-based learning, Cross-lingual transfer, medical imaging, remote sensing

## 1. Introduction

The classical assumption used in supervised learning that the training and test data are sampled from the same distribution does not hold in practice in many real-world problems [1]. Domain distribution shifts are a major challenge with models needing to generalize well to new environmental conditions without loss of performance [2]. Transfer learning tackles this problem by transferring knowledge from some source domains to guide the learning in target domains, especially when no or little labeled data is avail- able [3].

The emergence of large-scale pre-trained models has brought disruptive innovation to the field of transfer learning and excelled in various tasks [4]. These base models learn knowledge-rich representations from a large scale dataset and could be used to tackle more task-specific downstream tasks with only light fine-tuning applied [5]. At the same time, domain adaptation methods have progressed to address more diverse forms of distribution change - not only covariate shift, but also label shift and beyond [6].

In this article, we survey the field of transfer learning and domain adaptation emphasizing supervised models dealing with distribution shifts. We survey theoretic underpinning, methodological innovations and practical applications, offering a comprehensive synthesis of this important topic for modern machine learning.

## 2. Theoretical Foundations

### 2.1 Distribution Shifts and Domain Divergence

Distribution shifts happen if feature and label distribution do not match between source and target domains [7]. In summary, for provided source domain data $D_S \sim P_S(X, Y) $ and target domain data $D_T \sim P_T(X, Y) $, the domain transfer learning problem is to use source data $D_S$ to enhance the performance on target domain data distribution $P_T$ when $P_S eq P_T$[8].

Among them, there exist a few different types of distribution shifts described in the literature. Covariate shift is the case when $P_S(X) eq P_T(X)$, but they have the same label distribution, that is $P_S(Y|X) = P_T(Y|X)$[9]. Label shift refers to the phenomenon where it has changes in label distribution and fixed conditional feature distributions [10]. More challenging cases have the feature change, as well as a change in conditional distribution, which needs one to learn more intelligent strategies [11] to adapt.

Domain discrepancy measures the gap between source and target distributions. H-divergence has theoretical guarantees on target domain error in terms of source domain error and the divergence to the domain [12]. Other distances include Maximum Mean Discrepancy (MMD) which measures the difference between distributions in reproducing kernel Hilbert spaces [13] and Wasserstein distance based on optimal transport theory [14].

### 2.2 PAC Learning Framework for Transfer Learning

The PAC learning framework has also been extended to transfer learning [15]. Ben-David et al. [16] derived basic bounds indicating that the error of the target domain is related to three quantities: source domain error, divergence across domains and optimal joint hypothesis. This theoretical framework informs when transfer learning can be helpful and shapes the creation of domain adaptation algorithms [17].

## 3. Pre-trained Foundation Models

### 3.1 Language Models

The transformer mechanism introduced a new era in the field of NLP, due to their excellent ability to capture the long-range dependencies between word sequences and contextual information [18]. In [19], BERT (Bidirectional Encoder Representations from Transformers) introduced masked language modeling as a pre-training task, whose target is to learn bidirectional representations from unlabeled text. This pre-training strategy allows BERT to learn rich language representations which can be gloved and transferred to other tasks through fine-tuning [20].

The GPT (Generative Pre-trained Transformer) series of models showcased the potential for large scale autoregressive language modeling [21]). The size-other GPT-3 with 175B parameters demonstrated impressive few-shot and zero-shot performance, tackling tasks from few task-specific examples [22]. These models can pick up a wide range of linguistic patterns and world knowledge during pre-training, which help improve performance in new domains and tasks [23].

More recent models such as InstructGPT and ChatGPT, which are also fine-tuned for instruction following at generation time by reinforcing through human feedback [24], extends the language model's ability to model preferences of humans. These models achieve better zero-shot transfer by using natural language explanations without task oriented fine-tuning [25].

### 3.2 Vision Transformers

Vision Transformers (ViT) introduced the transformer models to computer vision and considered them as sequences of patches [26]. These models attain state-of-the-art (SOTA) performance on a wide spectrum of vision tasks when pre-trained from scratch on large scale data such as ImageNet-21k or JFT-300M [27]. The ViTs leverage the self-attention mechanism to distill global image relations, for learning transferable visual presentations [28].

Later other models have been proposed, such as DeiT (Data-efficient image Transformers), which introduced distillation for ViT training with small amount of data [29], and Swin Transformer that uses hierarchical pervasions on shifted windows for better efficiency[30]. These models have become a cornerstone for transfer learning in computer vision with applications ranging from object detection to medical image analysis [31].

### 3.3 Multimodal Models

CLIP (Contrastive Language-Image Pre-training), which learned aligned representations of images and text, was the first work to scale vision-language pre-training[32]. CLIP, which is trained on 400 million image-text pairs, exhibits impressive zero-shot transfer to diverse vision tasks via the use of natural language supervision [33]. This allows novel visual concepts to be modeled with no need for task-specific training data, and enjoying a flexible adaptation capability to new concepts visual [34].

Other multimodal architectures are DALL-E to generate images from text[35], Flamingo designed for few-shot, vision and language tasks[36] or most recently unified models like GPT-4V handling both visual and textual inputs [37]. These models are a syntactical departure towards general purpose artificial intelligence that is capable of transferring knowledge across modalities [38].

## 4. Domain Adaptation Techniques

### 4.1 Adversarial Domain Adaptation

Adversarial domain adaptation uses adversarial training to learn representations invariant across domains [39]. The domain adversarial neural network (DANN) added a gradient reversal layer for enforcing the feature extractor to create representations that are indistinguishable across domains [40]. A domain discriminator discriminates features based on their source domain and the feature extractor is trained adversarially to fool the discriminator [41].

Conditional Adversarial Domain Adaptation: Conditional adversarial domain adaptation generalizes this framework to align class-conditional distributions instead of marginal distribution [42]. It handles cases where categories are skewed in the domains, mapping semantically similar instances to the same representation [43]. Follow-up work further incorporated multiple discriminators, attentions and curriculums to enhance adaptation performance [44].

### 4.2 Optimal Transport for Domain Adaptation

Optimal transport theory offers a principled way for source and target distributions alignment [45]. The Wasserstein distance, which represents the cost of transporting probability mass between distributions, is a geometrically interpretable measure of domain discrepancy [46]. Optimal transport domain adaptation minimizes this distance to learn the aligned representations [47].

Joint distribution optimal transport (JDOT) is a generalization of optimal transport to factorized joint distributions, preserving the label information in the process [48]. This technique jointly matches both feature and label distributions, making sure semantically similar samples are aligned between two domains [49]. Moreover, computational efficiency has also been enhanced with sliced Wasserstein distance and entropic regularization [50].

### 4.3 Self-Training and Pseudo-Labeling

Self-training iteratively creates pseudo-labels for the unknown target domain unlabeled data, gradually adding in high-confident predictions to the training set [51]. Such a semi-supervised method allows the model to be adapted to target domains without manual annotation [52]. Confidence thresholding and uncertainty estimation guide a reliable selection of pseudo-labels preventing the accumulation of errors [53].

Recent developments include regularization based on consistency between model predictions of perturbed versions of the input [54] and co-training to make use of more than one view or model to improve pseudo-labeling quality [55]. MixMatch and FixMatch have shown success in such approaches using pseudo-labeling along with data augmentation and consistency regularization [56].

## 5. Few-Shot and Zero-Shot Learning

### 5.1 Meta-Learning Approaches

Meta-learning (or learning to learn) is about fast adaptation on a new task in few-shot setting [57]. Model-Agnostic Meta-Learning (MAML) learns initialization parameters which allow the model to be fine-tuned quickly for new tasks [58]. Through training the model to be adaptable to other tasks, MAML learns models that generalize well with few examples [59].

The Prototypical network learns from metric spaces where classification is done by doing nearest class prototyping [60]. Such a method is especially suitable for few-shot classification as it could generalize to new classes by calculating the prototypes based on only few examples [61]. Relation Network and Matching Network are alternative meta-learning architectures with good few-shot learning performance [62].

### 5.2 Prompt-Based Learning

Prompt-based learning recasts tasks into language models as problems of language modeling, allowing large scale language models to achieve few-shot and zero shot inference. [63] Formats with prompts, such as natural language task descriptions or examples, allow models to use their pretraining without new parameters [64]. This strategy has performed well for GPT-3 and its peers, achieving strong performance with very little task-specific data [65].

Prompt engineering and prompt tuning have recently become active research topics on whether how to design effective prompts or learn continuous prompt representations [66]. Historical reasoning: Chain-of-thought prompting boosts reasoning by training models to generate intermediate historical reasons [67]. These methods have enhanced zero-shot and few-shot performance levels on a wide variety of tasks [68].

### 5.3 Zero-Shot Learning via Semantic Embedding

Zero-shot learning can be defined as the ability to classify unseen classes using semantic descriptions or attributes vectors [69]. By learning corresponding mappings between visual features and semantic embedding, models are able to recognize novel categories with no training examples [70]. In (attribute-based) approaches, classes are represented as either binary or continuous attributes that allow for class generalization through composition [71].

Cross-modal transfer methods, such as CLIP, learn joint spaces of image and text to be used for zero-shot visual recognition [72]. Upon training on a large image caption pairs, such models acquire the ability of cross-modal correlation between visual and language information in a flexible zero-shot transfer scenario with natural language descriptions [73].

## 6. Cross-Domain Supervised Learning Challenges

### 6.1 Negative Transfer

Transfer is negative if performance of the target task solution degrades due to the introduction of knowledge from the source domain [74]. Such a behavior occurs when the source and target domains express very little relation, or if the adaptation methods give too much importance to domain alignment as opposed to task learning [75]. Detecting and alleviating negative transfer still a challenge that needs to be addressed, (and has been well-studied in the literature) [76], careful source domains selection [77].

According to transfer learning theory, closeness (e.g., task similarity or distributional divergence) between the domains should have a positive impact on the effectiveness of transfer [77]. Selective transfer and multi-source domain adaptation methods aim to discover the most relevant source knowledge and meanwhile avoid those poisonous transfers [78].

### 6.2 Catastrophic Forgetting

When fine-tuning pre-trained models to novel tasks, catastrophic forgetting can lead to a performance reduction on original tasks [79]. This problem is especially common in the context of continuous learning and sequential domain adaption [80]. Strategies to prevent forgetting are elastic weight consolidation that identifies important parameters and keeps it unchanged [81], progressive neural network that separate devoted resources for each task[82] etc.

Replay methods use memory buffers to store data from old tasks and interleave replay with training on new tasks, thereby maintaining performance [83]. The knowledge distillation methods keep the learned representation by forcing the new models to mimic the predictions of original model [84].

### 6.3 Domain Shift Detection and Quantification

One also needs to have the means of detecting distribution shifts at deployment time for trustworthy model performance [85]. UQ methods assist in the task of flagging when models see out-of-distribution input [86]. Conformal prediction provides distribution free uncertainty guarantees, and so reliable inference under distribution shift [87].

Online adaption approaches continuously refine models when new data are observed, and the algorithmic behavior is dynamic in response to changing distributions [88]. Test-time adaptation methods adapt the weights or predictions of a model with unlabeled test data, enhancing its robustness to new environments without requiring source domain data [89].

## 7. Applications in Low-Resource Settings

### 7.1 Cross-Lingual Transfer

Cross-lingual transfer can make it possible to perform NLP in the low resources languages using models trained on high resources languages [90]. Multilingual pretrained models such as mBERT and XLM-R learn to share representations among languages, which enables zero-shot (or few-shot) transfer across languages [91]. These models have brought NLP down to masses by making sophisticated language understanding available for hundreds of languages [92].

Translation-based methods generate training data for low-resource languages by translating from high-resource languages [93].Code-switching and mixed-language pre-training additionally enhances cross-lingual transfer by introducing models to multilingual data [94].

### 7.2 Medical Image Analysis

Medical imaging is also a field where there are only a few annotated samples, as labeling can be expensive and because the process of annotation may invade privacy [95]. Contrasting to the domain shift challenge, transfer learning from natural images has showed promising results even with domain difference for the task of medical image classification [96]. Pre-trained models that learn general visual features (such as edges and textures) apply to medical settings as well [97].

Medical imaging modality and institution specific Distribution shifts are tackled by methods of domain adaptation [98]. Self-supervised learning based on unlabeled medical images is an alternative pre-training approach to discover domain-specific features [99]. Very few-shot learning allows developing diagnostic model for rare diseases of which the training set is intrinsically low [100].

### 7.3 Remote Sensing and Satellite Imagery

Satellite image analysis is challenged by restricted annotated data, sensor transferability, and geographic distribution drifts [101]. Large-scale vision models, such as land use classification and crop monitoring [102], have resulted in progress because of transfer learning. Domain adaptation is used to address the effect of sensor variations and seasonality, which induce distribution changes [103].

With the advent of massive unlabeled satellite image archives, self-supervised learning has become a powerful pre-training method [104]. These methods learn representations of Earth surface features that generalize well to downstream tasks with little supervision [105].

## 8. Future Directions and Open Challenges

There are many promising prospective directions for further research. The trend towards larger foundation models is not slowing down, with questions around emergent abilities and best practices for transfer [106]. Efficient adaptation methods which are computationally and data efficient would be critical for practical use [107]. Theoretical understanding of when and why transfer learning works (and fails) is not yet fully known, thereby encouraging further foundational research [108].

To be robust to distribution shift we need to get better at measuring uncertainty and adaptation [109]. Fairness and bias issues occur when models trained on source domains maintain or even exacerbate biases in target domains [110]. Establishing fair cross-domain transfer learning strategies is an ethical conflict to be addressed [111].

The problem of lifelong adaptation and continuous learning is still an open challenge, for which we need systems that can accumulate knowledge over multiple domains without catastrophically erasing it [112]. More generally, multi-task and even multi-domain learning approaches which share knowledge between similar tasks may be expected to enhance sample efficiency [113]. Lastly, interpretable and explainable transfer learning decision-making is important for trustworthiness of AI system [114].

## 9. Conclusion

Transfer learning and domain adaptation have revolutionized machine learning, allowing us to learn effectively under distribution shifts with only weak supervision. Pre-trained base models are powerful beginnings for a wide range of applications and sophisticated adaptation approaches deal with different distribution shifts. Few-shot learning and zero-shot learning knowledge-enabled paradigms extend the strength of data efficiency, and applications in low-resource settings show impact in real world.

While much has been achieved, there is still a long way to go. Negative transfer, catastrophic forgetting and other type of shift detection still require further research. Theoretical development needs to keep pace with empirical advances in order to furnish principled input into the design of methods. As the space of foundational models rapidly expand and diverge, transfer learning will increasingly become a core part of constructing flexible and general purpose AI systems.

The future of transfer learning is to have more robust, efficient, and interpretable algorithms that are capable of dealing with various distribution shifts without sacrificing fairness and safety. By solving these challenges, the community has the opportunity to turn this vision into reality and build AI systems that continually improve with experience and generalize knowledge across domains and tasks.

## References

[1] Quionero-Candela, J., et al. (2009). Dataset shift in machine learning. MIT Press.
[2] Moreno-Torres, J. G., and so on. (2012). Towards a unified analysis of dataset shift in classification. Pattern Recognition, 45(1), 521-530.
[3] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 1345-1359.
[4] Bommasani, R., et al. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.
[5] Brown, T., et al. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems 33, pages 1877–1901.
[6] Kouw, W. M., & Loog, M. (2019). (DA without target labels) A survey on domain adaptation without target domain data to train the model. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(3), 7-785.
[7] Sugiyama, M., & Kawanabe, M. (2012). Machine learning in non-stationary environments. MIT Press.
[8] Weiss, K., et al. (2016). A survey of transfer learning. Journal of Big Data, 3(1), 1–40.
[9] Shimodaira, H. (2000). Weighting by the log-likelihood to improve predictive inference under covariate shift. Journal of Statistical Planning and Inference, 90(2), 227-244.
[10] Lipton, Z., et al. (2018). Mitigating label shift by supervised loss representation learning. In International Conference on Machine Learning, 3122-3130.
[11] Gretton, A., et al. (2009). Kernel mean matching adversarial covariate shif t. Dataset shift in machine learning, 3(4), 5.
[12] Ben-David, S., et al. (2007). Exploring representations for domain adaptation. in Advances in Neural Information Processing Systems, 19.
[13] Gretton, A., et al. (2012). A kernel two-sample test. Journal of Machine Learning Research, 13(1), 723–773.
[14] Arjovsky, M., et al. (2017). Wasserstein generative adversarial networks. In International Conference on Machine Learning, 214–223.
[15] Valiant, L. G. (1984). A theory of the learnable. Communications of the ACM, 27(11), 1134–1142.
[16] Ben-David, S., et al. (2010). A theory of learning from multiple domains. Machine Learning, 79(1), 151-175.
[17] Mansour, Y., et al. (2009). Domain adaptation: Learning bounds and algorithms. Conference on Learning Theory.
[18] Vaswani, A., et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, 30.
[19] Devlin, J., et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. North American Chapter of the Association for Computational Linguistics, 4171-4186.
[20] Sun, C., et al. (2019). Fine-tuning BERT for text classification. Chinese Computational Linguistics, 194-206.
[21] Radford, A., et al. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 9.
[22] Brown, T., et al. (2020). Language models are few-shot learners. In Proceedings of the 33nd Conference on Neural Information Processing Systems (pp. 1877-1901).
[23] Bommasani, R., et al. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.
[24] Ouyang, L., et al. (2022). Training Language Models To Follow Instructions With Human Feedback. In Advances in Neural Information Processing Systems (pp. 27730-27744).
[25] Wei, J., et al. (2022). Fine-tuned language model is zero-shot learner. International Conference on Learning Representations.
[26] Dosovitskiy, A., et al. (2021). A picture is worth 16x16 words: transformers for image recognition at scale. International Conference on Learning Representations.
[27] Steiner, A., et al. (2021). How to train your ViT? Data, Augmentation, and Regularization for Transformers in vision. arXiv preprint arXiv:2106.10270.
[28] Raghu, M., et al. (2021). Do vision transformers perceive in the way that convolutional neural networks would? Mutation-based loss for learning on point clouds.

[29] Touvron, H., et al. (2021). Distilling task-specific attention from pre-trained image transformers. ICLR: International Conference on Learning Representations 10347–10357.
[30] Liu, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. International Conf. on Computer Vision, 10012-10022.
[31] Matsoukas, C., et al. (2021). Is it the dawn of transformers replacing CNNs for medical images? arXiv preprint arXiv:2108.09038.
[32] Radford, A., et al. (2021). Learning to learn from natural language via visual imitation. International Conference on Machine Learning, 8748-8763.
[33] Ramesh, A., et al. (2022). On the fly image generation with Text-CLIP latents. arXiv preprint arXiv:2204.06125.
[34] Girdhar, R., et al. (2023). ImageBind: Tightening Imagined Worlds with Sweet Games. Computer Vision and Pattern Recognition, 15180-15190.
[35] Ramesh, A., et al. (2021). Zero-shot text-to-image generation. International Conference on Machine Learning, 8821-8831.
[36] Alayrac, J. B., et al. (2022). Flamingo: Model-base one/help-shot learning through few-shot. In Advances in Neural Information Processing Systems (Vol. 35, pp. 23716-23736.).
[37] OpenAI (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.
[38] Driess, D., et al. (2023). PaLM-E: A multimodal Embodied Language Model. arXiv preprint arXiv:2303.03378.
[39] Y. Ganin and V. Lempitsky (2015)unsupervised domain adaptation by backpropagation. Unsupervised domain adaptation by backpropagation. Proceedings of the International Conference on Machine Learning, 1180-1189.
[40] Ganin, Y., et al. (2016). Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(1), 2096-2030.
[41] Tzeng, E., et al. (2017). Adversarial discriminative domain adaptation. Computer Vision and Pattern Recognition, 7167–7176.
[42] Long, M., et al. (2018). Conditional adversarial domain adaptation. In Neural Information Processing Systems.
[43] Saito, K., et al. (2018). Unsupervised Domain Adaptation with Category-Shifted Output Distribution Universal Maximum mean Discrepancy for unsupervised domain-adaptation. In: CVPR) (Computer Vision and Pattern Recognition, pp. 3723-3732
[44] Xu, R., et al. (2019). Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In 14th International Conference on Computer Vision Vol.
[45] Villani, C. (2009). Optimal transport: Old and new. Springer.
[46] Cuturi, M. (2013). Sinkhorn distances: lightspeed computation of optimal transport. In Advances in Neural Information Processing Systems, 26.
[47] Courty, N., et al. (2017). Optimal transport for domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39 (9), 1853-1865.
[48] Courty, N., et al. (2017). Poincaré-and-Wasserstein-Based Distribution Alignment for Pose Normalization in Domain Adaptation. In Advances in Neural Information Processing Systems 30.
[49] Damodaran, B. B., et al. (2018). DeepJDOT: Deep joint distribution optimal transport for unsupervised domain adaptation. European Conference on Computer Vision, 447-463.
[50] Kolouri, S., et al. (2019). Generalized sliced Wasserstein distances. Advances in Neural Information Processing Systems 32.
[51] Lee, D. H. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. Workshop on Challenges in Representation Learning, ICML, 3(2).
[52] Zou, Y., et al. (2019). Confidence regularized self-training. In International Conference on Computer Vision (pp. 5982-5991).
[53] Arazo, E., et al. (2020). Pseudolabeling and confirmation bias in deep semi-supervised learning. International Joint Conference on Neural Networks, 1-8.
[54] Laine, S., and Aila, T., 2017. Temporal ensembling for semi-supervised learning. International Conference on Learning Representations.
[55] Blum, A., and Mitchell, T. (1998). Co-training using labeled and unlabeled data. Conference on Learning Theory, 92-100.
[56] Sohn, K., et al. (2020). FixMatch: Simplified semi-supervised learning with consistency and confidence. In Advances in Neural Information Processing Systems 33 (pp. 596-608).
[57] Hospedales, T., et al. (2021). A survey of meta-learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(9), 5149-5169.
[58] Finn, C., et al. (2017). Model-agnostic meta-learning: Fast adaptation of deep networks. International conference on machine learning, 1126-1135.
[59] Nichol, A., et al. (2018). On first-order meta-learning algorithms. arXiv preprint arXiv:1803.02999.
[60] Snell, J., et al. (2017). Prototypical networks for few-shot learning. In: Proceedings 31st International Conference on Machine Learning.
[61] Sung, F., et al. (2018). Learning to compare: Relation network for few-shot learning. Computer Vision and Pattern Recognition, 1199-1208.

[62] Vinyals, O., et al. (2016). Matching networks for oneshot learning. 1 Dinesh Raghu, Jongyeon Lee 2 and Samy Bengio (Joint equal contribution) Google Brain dineshraghu, rainyspring spring1986 Abstract In the case of graph learning problems, structured regularizing norms such as the graph Laplacian can be highly beneficial.
[63] Liu, P., et al. (2023). Pre-training, prompt and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9), 1-35.
[64] Schick, T., & Schütze, H. (2021). Cloze driven learning for few-shot novel writing and natural language understanding. European Association for Computational Linguistics, 255-269.
[65] Zhao, Z., et al. (2021). Stand-alone Meta-language Model: Few-shot learning from scaled down language model. International Conference on Machine Learning, 12697-12706.
[66] Lester, B., et al. (2021). The scale of parameter-efficient prompt tuning. In: Proceedings of the Empirical Methods in Natural Language Processing, 3045-3059.
[67] Wei, J., et al. (2022). Prompting chain-of-thought induces reasoning in large language models. In Advances in Neural Information Processing Systems (Vol. 35, pp. 24824-24837).
[68] Kojima, T., et al. (2022). Zero-Shot Reasoning Zero-shot reasoning is a particular case of the more general problem of zero-shot learning [18, 5] and large language models are zero-shot reasoners. In NIPS Advances in Neural Information Processing Systems 35, 22199-22213.
[69] Xian, Y., et al. (2018). Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly. IEEE transactions on pattern analysis and machine intelligence, 41(9), 2251-2265.
[70] Lampert, C. H., et al. (2009). Detecting unseen classes by attributions transfer. Pattern Recognition and Computer Vision, 951-958.
[71] Farhadi, A., et al. (2009). Describing objects by their attributes. Computer Vision and Pattern Recognition, 1778-1785.
[72] Radford, A., et al. (2021). Learning visual models using language as a supervision. Proceedings of the International Conference on Machine Learning, 8748-8763.
[73] Jia, C., et al. (2021). Large-scale vision and vision-language representation learning with noisy text supervision. In International Conference on Machine Learning, 4904–4916.
[74] Rosenstein, M. T., et al. (2005). To transfer or not to transfer. NIPS Transfer Learning Workshop, 2.
[75] Wang, Z., et al. (2019). Characterizing and avoiding negative transfer. Computer Vision and Pattern Recognition, 11293-11302.
[76] Ge, Y., et al. (2014). Leveraging task relatedness for multi-task learning. arXiv preprint arXiv:1404.6033.
[77] Zamir, A. R., et al. (2018). Taskonomy: Disentangling task transfer learning. Computer Vision and Pattern Recognition, 3712-3722.
[78] Zhao, H., et al. (2020). Invariant Representations for Adversarial Domain Adaptation. International conference on machine learning, 7523-7532.
[79] Kirkpatrick, J., et al. (2017). Catastrophic forgetting in connectionist networks. Proceedings of the National Academy of Sciences, 114(13), 3521-3526.
[80] McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: Causes, consequences and solutions. In Mussen, P.H. (ed.), Handbook of child psychology (pp.
[81] Schwarz, J., et al. (2018). When there is very little parameter overlap, it is unlikely that this small amount of function adaptation will be sufficient to solve a new problem. (Machine Learning) International Conference on, 4528-4537.
[82] Rusu, A. A., et al. (2016). Progressive neural networks. arXiv preprint arXiv:1606.04671.
[83] Rebuffi, S. A., et al. (2017). iCaRL: Incremental classifier and representation learning. Computer Vision and Pattern Recognition, 2001{\textendash}2010.
[84] Hinton, G., et al. (2015). Knowledge distillation in a neural network. arXiv preprint arXiv:1503.02531.
[85] Ovadia, Y., et al. (2019). Can you trust that your model is actually unsure? Predictive uncertainty estimation under dataset shift. Proceedings of the 32 nd NIPS.
[86] Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. International Conference on Machine Learning, 1050-1059.
[87] Angelopoulos, A. N., & Bates, S. (2021). An easygoing introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511.
[88] Wang, D., et al. (2021). Tent: Fully test-time adaptation by entropy minimization. International Conference on Learning Representations.
[89] Sun, Y., et al. (2020). Test-time training with self-supervision for distribution shifting generalization. International Conference on Machine Learning, 9229-9248.
[90] Ruder, S., et al. (2019). A survey on cross-lingual word embedding models. Journal of Artificial Intelligence Research, 65, 569--631.
[91] Conneau, A., et al. (2020). Massively multilingual pretrained language models for zero-shot cross-lingual transfer and beyond. Association for Computational Linguistics, 8440-8451.

[92] Hu, J., et al. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. International Conference on Machine Learning, 4411-4421.
[93] Artetxe, M., and Schwenk, H. (2019). SPINE: A Semi-supervised Method for Identifying Negative Pairs in Embeddings for Solving Analogies. Transactions of the Association for Computational Linguistics 7, 597–610.
[94] Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. In Advances in Neural Information Processing Systems 32.
[95] Litjens, G., et al. (2017). A survey of deep learning for scientific discovery. Medical Image Analysis, 42, 60-88.
[96] Tajbakhsh, N., et al. (2016). Are we training enough? How much data and what kind of data are needed by deep learning models to become proficient at medical image analysis? IEEE Transactions on Medical Imaging, 35(5), 1299-1312.
[97] Raghu, M., et al. (2019). Transfusion: Learning to transfer knowledge for recognition of action content and context. Advances in Neural Information Processing Systems, vol. 32.
[98] Guan, H., & Liu, M. (2021). A survey on domain adaptation for medical image analysis. IEEE Transactions on Biomedical Engineering, 69(3), 1173-1185.
[99] Chen, T., et al. (2020). Simple framework for contrastive learning of visual representations. Machine Learning-International Conference on Machine Learning, 1597–1607.
[100] Zhou, H. Y., et al. (2021). Few-shot learning in medical image analysis: A review. arXiv preprint arXiv:2112.09958.
[101] Cheng, G., et al. (2020). Deep learning with satellite images for disaster mapping. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13 (9), 3735-3756.
[102] Tuia, D., et al. (2016). A survey of recent advances in domain adaptation for the classification of remote sensing data. IEEE Geoscience and Remote Sensing Magazine, 4(2), 41-57.
[103] Persello, C., and Bruzzone, L. (2014). Domain adaptation in remote sensing image scene classification. IEEE Trans Geosci Remote Sens 52(12):7180–7190110741.
[104] Caron, M., et al. (2021). What can help vision transformers with limited data: A closer look into self-supervised learning. International Conference on Computer Vision, 9650-9660.
[105] Ayush, K., et al. (2021). Geography-aware self-supervised learning. International Conference on Computer Vision, 10181-10190.CiteGet rights and content)paren.Open Archive DOIContributor Information0Copyright Published by Elsevier Ltd. All rights reserved./getPDF.pdfKeywordsComputer visionSegmentationAbstractThis paper proposes a hierarchical framework to segment the medical image with partially labelled data.
[106] Kaplan, J., et al. (2020). If ranked lists of words are not available, then the cross-entropy is used as a proxy to other such list-based as it can provide word-rankings that are necessary for learning neural language models. arXiv preprint arXiv:2001.08361.
[107] Hu, E. J., et al. (2022). LoRA: Large-lm makes small-lm better. International Conference on Learning Representations.
[108] Neyshabur, B., et al. (2020). What is learned and transferred in transfer learning? In Advances in Neural Information Processing Systems 33 (pp. 512--523).
[109] Hendrycks, D., & Dietterich, T. (2019). Big self-supervised models to corruptions: do they learn from errors? International Conference on Learning Representations.
[110] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification, 2018. Conference on Fairness, Accountability and Transparency, 77-91.
[111] Xu, D., et al. (2021). Fairness in domain adaptation. arXiv preprint arXiv:2103.11648.
[112] Parisi, G. I., et al. (2019). Lifelong learning with dynamically expandable networks: A review. Neural Networks, 113, 54-71.
[113] Crawshaw, M. (2020). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. arXiv preprint arXiv:2009.09796.
[114] Ribeiro, M. T., et al. (2016). "Why should I trust you?" Interpreting the decisions of any classifier. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144.