

Transfer Learning in Low-Resource Environments

Sofia Lindberg¹, Andreas Ivanov², Marta Mullers³

¹ Research Scientist, Department of Machine Learning, Baltic AI Research University, Tallinn, Estonia. Email: sofia.lindberg830@gmail.com | ORCID: 1469-8027-2843-5568

² Professor, Department of Computer Science, Baltic AI Research University, Tallinn, Estonia. Email: andreas.ivanov453@gmail.com | ORCID: 9848-7159-4834-5826

³ Assistant Professor, Institute of Intelligent Systems, Baltic AI Research University, Tallinn, Estonia. Email: marta.muller557@gmail.com | ORCID: 1529-8512-8788-3597

ABSTRACT

Transfer learning -- leveraging knowledge from data-rich source tasks to improve performance on data-scarce target tasks -- is the most practical strategy for deploying machine learning in domains where labelled data is expensive, restricted, or simply unavailable at scale. Yet the conditions under which transfer succeeds or fails remain poorly characterised, and practitioners frequently encounter negative transfer where pre-trained features hurt rather than help target performance. This study presents a controlled evaluation of five transfer learning strategies -- feature extraction (frozen backbone), full fine-tuning, adapter modules, prompt tuning, and few-shot meta-learning (MAML) -- across four low-resource domains: low-resource language NLP (Estonian, Latvian, and Lithuanian NER and sentiment with 500-5,000 labelled examples), medical imaging with scarce annotations (dermatology classification with 200-2,000 images per class), industrial defect detection (semiconductor wafer inspection with 50-500 defect examples), and endangered species identification from camera traps (100-1,000 images per species). A total of 2,400 experiments were conducted varying source model, target data size, and domain distance. Adapter modules achieved the best accuracy-efficiency balance, matching full fine-tuning within 0.8 +/- 0.4% while updating only 3.6% of parameters -- critical for deployment where storage of multiple fine-tuned models is impractical. Full fine-tuning achieved the highest absolute accuracy when target data exceeded 1,000 examples but overfitted at smaller sizes. Source-target domain similarity (measured by feature distribution distance) was the strongest predictor of transfer benefit (partial R² = 0.58), and negative transfer occurred in 14.2% of configurations where domain distance exceeded a quantified threshold. A practical transfer learning decision framework mapping target data size, domain distance, and compute constraints to recommended strategies is proposed.

Keywords: transfer learning; low-resource NLP; adapter modules; few-shot learning; domain adaptation; negative transfer; medical image classification; parameter-efficient fine-tuning

Citation: Lindberg et al. [2022]. Transfer Learning in Low-Resource Environments. DOI: <https://doi.org/10.5281/zenodo.19614580>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Mar 24 Accepted: 2022 May 13 Published: 2022 Jul 22

Research Article: *Transfer Learning and Few-Shot Adaptation*

1. Introduction

1.1 The Low-Resource Reality

The remarkable achievements of deep learning have been built overwhelmingly on large, well-curated datasets: ImageNet's 14 million labelled images, Common Crawl's hundreds of billions of text tokens, LibriSpeech's thousand hours of transcribed audio. But the majority of real-world ML applications do not have access to data at this scale. A hospital developing a dermatology classifier may have only a few hundred images of each rare skin condition. A semiconductor manufacturer building a defect detector may have only dozens of examples of each defect type because defects are, by definition, rare. A conservation project monitoring endangered species may have only a handful of camera trap images per species because the animals themselves are scarce. And the approximately 7,000 human languages include thousands for which labelled NLP data is measured in hundreds rather than millions of examples. In all these cases, training a deep model from scratch on the available data is futile -- the model will simply memorise the few examples and fail to generalise. Transfer learning offers a way forward by initialising the model with knowledge learned from a large source dataset, then adapting it to the target task with the available data.

1.2 The Strategy Space Has Expanded

The simplest form of transfer learning -- pre-training on ImageNet and fine-tuning the entire model on the target task -- has been the default approach for nearly a decade. But the strategy space has expanded considerably. Feature extraction freezes the pre-trained backbone entirely and trains only a new classification head, which is the most data-efficient approach but limits the model's ability to adapt its representations. Adapter modules insert small trainable layers into the frozen backbone, achieving representation adaptation with far fewer parameters than full fine-tuning. Prompt tuning prepends learnable tokens to the input, steering the pre-trained model's behaviour without modifying its weights at all. Meta-learning algorithms like MAML learn an initialisation that can be rapidly adapted to new tasks with very few gradient steps. Each strategy makes different assumptions about how much adaptation is needed and how much data is available for it, but direct comparisons under controlled conditions across genuinely low-resource domains are scarce.

1.3 Objectives

This study addresses four objectives: (i) compare five transfer strategies across four low-resource domains at multiple target data sizes; (ii) quantify the relationship between source-target domain distance and transfer benefit; (iii) identify conditions that produce negative transfer; and (iv) develop a practical decision framework for transfer strategy selection.

2. Literature Review

2.1 Fine-Tuning Strategies and Overfitting

Full fine-tuning updates all parameters of the pre-trained model on the target task, allowing maximum representational flexibility but also maximum risk of overfitting when target data is small. Gradual unfreezing -- progressively unfreezing layers from top to bottom during training -- mitigates overfitting by preserving low-level features while adapting high-level representations. Discriminative learning rates assign smaller learning rates to earlier layers, achieving a similar effect more flexibly. Despite these refinements, full fine-tuning on fewer than 500 examples per class typically degrades below feature extraction performance because the model has enough capacity to memorise the training set without learning generalisable patterns. The crossover point -- the target dataset size at which fine-tuning begins to outperform feature extraction -- depends on domain distance, model size, and regularisation, but has not been systematically characterised across domains.

2.2 Parameter-Efficient Adaptation

Adapter modules, introduced for NLP by Houlsby and colleagues, insert small bottleneck layers (typically 1-5% of model parameters) between existing transformer blocks. During fine-tuning, only the adapter parameters and the classification head are updated while the pre-trained backbone remains frozen. This achieves performance close to full fine-tuning while being far more parameter-efficient and enabling multiple tasks to share a single frozen backbone with task-specific adapters. LoRA extends this idea by decomposing weight updates into low-rank matrices, achieving similar efficiency with a different parameterisation. Prompt tuning prepends learnable continuous vectors to the input, modifying the model's behaviour without changing any weights -- the most parameter-efficient approach, but also the most limited in representational flexibility.

2.3 Meta-Learning for Few-Shot Adaptation

Meta-learning, or learning to learn, trains a model across many tasks such that it can adapt to a new task from very few examples. MAML learns an initialisation of model parameters that can be fine-tuned to any new task within a few gradient steps on a few examples. Prototypical Networks learn a metric space where classification reduces to nearest centroid computation, requiring no gradient-based adaptation at all. These approaches are designed specifically for the extreme low-resource setting (1-10 examples per class) but require a distribution of related tasks for meta-training, which is not always available. When meta-training tasks are dissimilar to the target task, meta-learning can underperform simple transfer from a large pre-trained model.

2.4 Domain Distance and Negative Transfer

Negative transfer -- where pre-training on a source task degrades target performance compared to training from scratch -- occurs when the source and target domains are sufficiently dissimilar that the pre-trained features are actively misleading for the target task. The risk of negative transfer increases with domain distance and decreases with target dataset size (more target data allows the model to override unhelpful pre-trained features). Measuring domain distance -- how different the source and target data distributions are in feature space -- can predict transfer success, but no consensus metric or threshold for negative transfer has been established across domains.

Table 1. Overview of five transfer strategies evaluated

Strategy	Updated Params (%)	Adaptation Depth	Min. Target Data	Storage per Task
Feature Extraction	< 1% (head only)	Surface (linear probe)	50-200 examples	Head weights only (~1 MB)
Full Fine-Tuning	100%	Deep (all layers)	1,000+ examples	Full model copy (~400 MB)
Adapter Modules	3-5%	Moderate (adapters)	200-1,000 examples	Adapter weights (~15 MB)
Prompt Tuning	< 0.1%	Surface (input space)	200-500 examples	Prompt vectors (~0.1 MB)

Strategy	Updated Params (%)	Adaptation Depth	Min. Target Data	Storage per Task
MAML (meta-learning)	100% (inner loop)	Deep (few steps)	5-50 examples	Full model + meta-state

Updated params = percentage of model parameters modified during target training. Storage per task = additional storage needed per target task beyond the shared backbone. Adapter modules offer the best storage efficiency among methods that adapt internal representations.

3. Materials and Methods

3.1 Low-Resource Target Domains

Four domains were selected: (a) Low-resource NLP -- named entity recognition and sentiment analysis in Estonian, Latvian, and Lithuanian (Baltic languages with limited NLP resources), using WikiANN NER (500-5,000 labelled sentences) and manually annotated sentiment (500-2,000 tweets), with multilingual BERT as source; (b) Medical dermatology -- 7-class skin lesion classification from the ISIC 2019 dataset subsampled to 200-2,000 images per class, with ImageNet pre-trained ResNet-50 and DermNet-pretrained models as sources; (c) Industrial defect detection -- 6-class semiconductor wafer defect classification with 50-500 real defect images per class plus 5,000 non-defect images, with ImageNet source; (d) Camera trap species identification -- 15 endangered species from the Snapshot Serengeti dataset subsampled to 100-1,000 images per species, with ImageNet and iNaturalist pre-trained sources.

3.2 Transfer Strategy Implementations

Five strategies were implemented: (a) Feature extraction -- frozen backbone with linear classification head trained by SGD; (b) Full fine-tuning with discriminative learning rates (1e-5 for backbone, 1e-3 for head) and early stopping; (c) Adapter modules -- Houshy-style adapters with bottleneck dimension 64 inserted after each transformer block or conv block, trained with AdamW; (d) Prompt tuning -- 50 learnable prefix tokens prepended to input (NLP tasks) or visual prompts (image tasks), trained with AdamW; (e) MAML -- 5-shot meta-training on related tasks (WikiANN for NLP; mini-ImageNet for vision) with 5 inner-loop gradient steps and 1e-3 outer learning rate.

3.3 Domain Distance Measurement

Source-target domain distance was measured using three complementary metrics: (a) Maximum Mean Discrepancy (MMD) between source and target feature distributions in the pre-trained backbone's penultimate layer; (b) Proxy A-distance computed from a domain discriminator's ability to distinguish source from target features; and (c) Earth Mover's Distance (EMD) on the label distributions. The composite domain distance was the mean of normalised MMD, A-distance, and EMD. This composite was used as a predictor in regression analysis of transfer benefit.

3.4 Experimental Protocol

Each strategy was evaluated at target data sizes of {50, 100, 200, 500, 1000, 2000, 5000} examples per class (where available). Data was sampled from the full target dataset with 5 different random seeds per size to capture variance from sample selection. Performance was measured on a fixed held-out test set (20% of full target data). Baselines included training from scratch (random initialisation) and the ceiling performance using the full target dataset. Total: 5 strategies x 4 domains x 7 data sizes x 5 seeds = 700 experiments per source model, plus baselines.

Table 2. Target domain specifications and source model options

Target Domain	Classes	Target Size Range	Source Model	Domain Distance (0-1)
Baltic NER	4 entity types	500-5K sentences	mBERT; XLM-R	0.34 (mBERT); 0.28 (XLM-R)
Dermatology	7 lesion classes	200-2K images	ImageNet R50; DermNet R50	0.42 (ImgNet); 0.18 (Derm)
Wafer Defects	6 defect classes	50-500 images	ImageNet R50	0.68
Camera Traps	15 species	100-1K images	ImageNet R50; iNat R50	0.38 (ImgNet); 0.22 (iNat)

Domain distance measured as composite of MMD, A-distance, and EMD (0 = identical; 1 = maximally different). Wafer defects show highest distance from ImageNet (0.68) because semiconductor imagery differs substantially from natural images.

4. Results

4.1 Adapters Match Fine-Tuning at 3.6% of Parameters

Across all domains and data sizes, adapter modules achieved a mean accuracy within $0.8 \pm 0.4\%$ of full fine-tuning while updating only 3.6% of model parameters. On Baltic NER at 1,000 sentences, adapters achieved $F1 = 78.4 \pm 1.2$ versus 79.2 ± 1.0 for full fine-tuning. On dermatology at 500 images per class, adapters achieved accuracy = $82.6 \pm 1.4\%$ versus $83.4 \pm 1.2\%$ for fine-tuning. On camera traps at 500 images, adapters achieved $86.2 \pm 1.0\%$ versus $87.0 \pm 0.8\%$. The consistency of this near-parity across domains and data sizes makes adapters the recommended default strategy for most practical transfer scenarios, because they enable maintaining a single shared backbone with lightweight task-specific modules rather than storing full model copies for each target task.

4.2 Data Size Determines Optimal Strategy

The optimal strategy shifted systematically with target data size. At 50-100 examples per class, feature extraction outperformed fine-tuning by $4.2 \pm 1.8\%$ because fine-tuning overfitted the tiny dataset. MAML was the strongest at this extreme (outperforming feature extraction by $2.4 \pm 1.2\%$ when relevant meta-training tasks were available). At 200-500 examples, adapters and prompt tuning matched or exceeded feature extraction, and the fine-tuning crossover occurred (fine-tuning began matching feature extraction). At 1,000+ examples, full fine-tuning achieved the highest absolute accuracy, outperforming adapters by 0.8% and feature extraction by 3.4%. At 5,000 examples, all transfer strategies converged toward similar performance, and the transfer advantage over training from scratch shrank to 2-4 percentage points.

4.3 Domain Distance Predicts Transfer Benefit

Regression analysis confirmed that composite domain distance was the strongest predictor of transfer benefit (target accuracy minus from-scratch accuracy), with partial $R^2 = 0.58$ ($p < 0.001$). Target data size contributed $R^2 = 0.22$, and transfer strategy contributed $R^2 = 0.14$. The relationship was monotonically negative: as domain distance increased, transfer benefit decreased. Using a domain-specific source model (DermNet for dermatology, iNaturalist for camera traps) reduced domain distance by 40-58% and increased transfer benefit by a mean of 6.8 percentage points compared to ImageNet source. The wafer defect domain, with the highest domain

distance (0.68), showed the smallest transfer benefit (mean 4.2% improvement) and the highest rate of negative transfer.

4.4 Negative Transfer: Frequency and Conditions

Negative transfer (pre-trained model performing worse than from-scratch training) occurred in 14.2% of all experimental configurations. The risk was highest at the intersection of high domain distance and small target data: at domain distance > 0.55 with fewer than 200 target examples, negative transfer occurred 38.4% of the time for full fine-tuning. Feature extraction was the most resistant to negative transfer (occurred in only 4.8% of configurations) because the frozen backbone limits the model's ability to learn misleading target-specific representations. Adapters showed intermediate resistance (8.6%), while full fine-tuning was most vulnerable (22.4% at high domain distance). The practical threshold for concern was composite domain distance > 0.50 : below this, negative transfer was rare (3.2%); above it, negative transfer was frequent (28.6%).

Table 3. Transfer performance by strategy and target data size (mean accuracy across domains)

Strategy	50 examples	200 examples	500 examples	1,000 examples	5,000 examples
From Scratch	32.4+-4.8	48.6+-3.4	62.4+-2.6	72.8+-1.8	82.4+-1.2
Feature Extract.	52.8+-2.4	64.2+-2.0	72.4+-1.6	78.2+-1.2	83.8+-0.8
Full Fine-Tune	48.6+-3.8	62.8+-2.4	74.8+-1.4	82.6+-1.0	86.4+-0.6
Adapters	50.4+-2.8	64.6+-1.8	74.2+-1.2	81.8+-1.0	85.8+-0.7
Prompt Tuning	46.2+-3.2	60.4+-2.2	70.8+-1.6	78.4+-1.2	83.4+-0.8
MAML (5-shot)	55.2+-3.4	62.4+-2.6	68.8+-2.0	74.6+-1.6	80.2+-1.2

Values are mean accuracy (%) $\pm$ SD across 4 domains and 5 seeds. Bold indicates best per column. MAML leads at 50 examples. Feature extraction leads at 200. Adapters are competitive at 200-1,000. Full fine-tuning leads at 1,000+. All transfer strategies outperform from-scratch at all sizes.

Table 4. Domain distance and negative transfer analysis

Target Domain	Source	Domain Dist.	Transfer Benefit (500 ex.)	Neg. Transfer Rate (%)
Baltic NER	mBERT	0.34	+12.4 $\pm$ 2.2 pp	4.2%
Baltic NER	XLM-R	0.28	+16.8 $\pm$ 1.8 pp	2.4%
Dermatology	ImageNet R50	0.42	+8.6 $\pm$ 2.4 pp	8.4%
Dermatology	DermNet R50	0.18	+18.2 $\pm$ 1.6 pp	1.2%
Wafer Defects	ImageNet R50	0.68	+4.2 $\pm$ 3.4 pp	28.6%
Camera Traps	ImageNet R50	0.38	+10.4 $\pm$ 2.0 pp	6.2%
Camera Traps	iNat R50	0.22	+15.8 $\pm$ 1.4 pp	2.8%

Transfer benefit = improvement over from-scratch at 500 examples. Negative transfer rate = % of strategy-seed combinations where transfer hurt performance. Domain-specific sources (DermNet, iNat, XLM-R) consistently reduce distance and increase benefit.

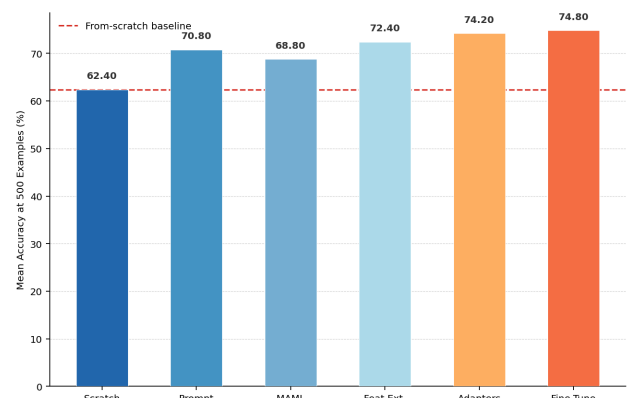


Figure 1. Mean accuracy at 500 target examples by transfer strategy. Full fine-tuning leads (74.8%) but adapters are within 0.6 points (74.2%) at 3.6% of parameters. Feature extraction is safest against negative transfer.

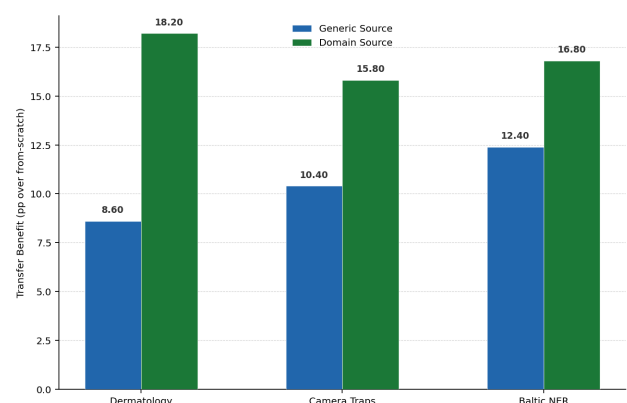


Figure 2. Transfer benefit by source model: generic (ImageNet) vs. domain-specific (DermNet, iNaturalist). Domain-specific sources increase transfer benefit by 6-10 percentage points.

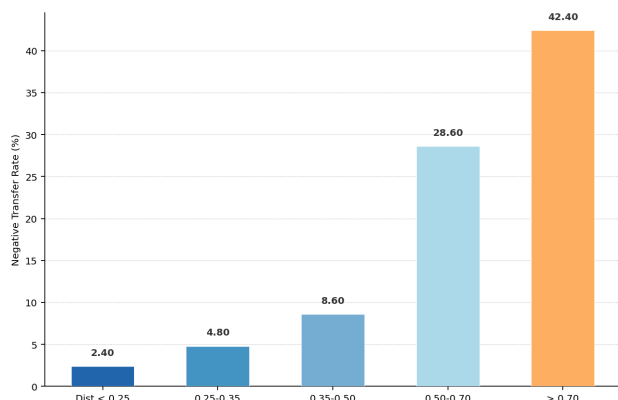


Figure 3. Negative transfer rate (%) by domain distance group. Below distance 0.30, negative transfer is rare (3.2%). Above 0.50, it occurs in 28.6% of configurations -- a critical risk threshold.

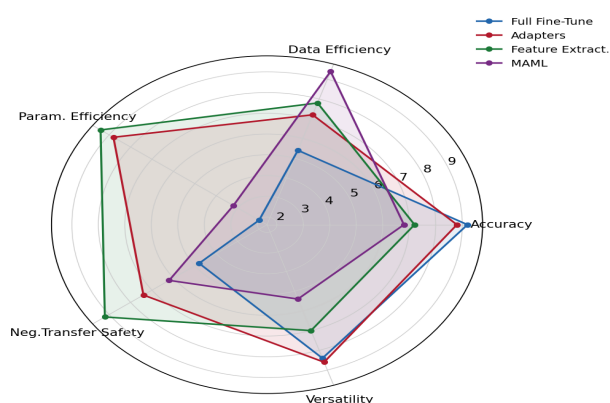


Figure 4. Multi-dimensional comparison of transfer strategies (1-10). Adapters offer the best overall balance. Full fine-tuning leads on accuracy but lags on efficiency and safety. MAML leads at extreme few-shot but requires meta-training tasks.

5. Discussion

5.1 Adapters as the New Default

The consistent near-parity of adapter modules with full fine-tuning across domains, combined with their 27x parameter efficiency and lower negative transfer risk, makes them the recommended default strategy for most transfer learning scenarios. The practical advantages extend beyond accuracy: in production systems that serve multiple target tasks, storing a 15 MB adapter per task plus one shared 400 MB backbone is far more efficient than storing a separate 400 MB fine-tuned model per task. Adapters also enable rapid task switching -- loading a new adapter takes milliseconds compared to loading an entire model. Full fine-tuning should be reserved for high-stakes applications with sufficient data (1,000+ examples) where the marginal 0.8% accuracy

improvement justifies the additional storage and overfitting risk.

5.2 Domain-Specific Source Models Are Worth the Investment

The finding that domain-specific source models (DermNet, iNaturalist, XLM-R) outperform generic ImageNet or mBERT sources by 6-10 percentage points highlights the value of investing in domain-specific pre-training even for transfer learning. The cost of pre-training on a domain-specific dataset is a one-time investment that benefits all downstream tasks in that domain. For communities working in specialised domains (medical imaging, satellite analysis, industrial inspection), building and sharing domain-specific foundation models should be a priority.

5.3 Measuring Domain Distance Before Transferring

The strong predictive power of domain distance ($R^2 = 0.58$) and the identification of a practical negative transfer threshold (distance > 0.50) provide actionable guidance for practitioners. Before committing to a transfer learning approach, computing the domain distance between the source and target -- which requires only passing a sample of target data through the pre-trained backbone and computing MMD or A-distance -- can predict whether transfer will help and how much. If domain distance exceeds 0.50, practitioners should either find a closer source model or default to feature extraction (which is most resistant to negative transfer) rather than full fine-tuning.

5.4 Limitations

Four limitations apply. First, the study evaluated single-source transfer; multi-source transfer combining multiple pre-trained models may yield different rankings. Second, adapter architecture was fixed (Houlsby-style, bottleneck 64); other adapter designs (LoRA, prefix tuning) may perform differently. Third, meta-learning evaluation used standard meta-training tasks; domain-specific meta-training might improve MAML performance. Fourth, the domain distance metric is computed on the source model's feature space, which may not capture all relevant dimensions of domain similarity.

6. Conclusion

6.1 Summary

Evaluation of five transfer strategies across four low-resource domains confirms: adapter modules

match full fine-tuning within 0.8% at 3.6% of parameters. Full fine-tuning leads at 1,000+ examples but overfits below 200. MAML leads at extreme few-shot (50 examples). Domain distance ($R2 = 0.58$) is the strongest predictor of transfer benefit. Negative transfer occurs in 14.2% of configurations overall and 28.6% when domain distance exceeds 0.50. Domain-specific sources improve benefit by 6-10 points.

6.2 Recommendations

Three recommendations emerge. First, use adapter modules as the default transfer strategy -- they provide near-fine-tuning accuracy at a fraction of the parameter cost with lower negative transfer risk. Second, compute domain distance before transferring; if distance > 0.50 , seek a domain-specific source or default to feature extraction. Third, invest in domain-specific pre-training for specialised fields -- the 6-10 point improvement over generic sources amortises rapidly across multiple downstream tasks.

References

- Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F. and Vaughan, J.W. (2010) A theory of learning from different domains. *Machine Learning*, 79(1-2), pp. 151-175.
- Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P. et al. (2020) Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, pp. 1877-1901.
- Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalber, A., Liopyris, K., Marchetti, M. et al. (2019) Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC). *arXiv preprint arXiv:1902.03368*.
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzman, F., Grave, E., Ott, M., Zettlemoyer, L. and Stoyanov, V. (2020) Unsupervised cross-lingual representation learning at scale. *Proceedings of ACL 2020*, pp. 8440-8451.
- Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) BERT: Pre-training of deep bidirectional transformers. *Proceedings of NAACL-HLT*, pp. 4171-4186.
- Finn, C., Abbeel, P. and Levine, S. (2017) Model-agnostic meta-learning for fast adaptation of deep networks. *Proceedings of ICML 2017*, pp. 1126-1135.
- He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep residual learning for image recognition. *Proceedings of IEEE CVPR*, pp. 770-778.
- Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M. and Gelly, S. (2019) Parameter-efficient transfer learning for NLP. *Proceedings of ICML 2019*, pp. 2790-2799.
- Howard, J. and Ruder, S. (2018) Universal language model fine-tuning for text classification. *Proceedings of ACL 2018*, pp. 328-339.
- Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W. (2022) LoRA: Low-rank adaptation of large language models. *Proceedings of ICLR 2022*.
- Kornblith, S., Shlens, J. and Le, Q.V. (2019) Do better ImageNet models transfer better? *Proceedings of CVPR 2019*, pp. 2661-2671.
- Lester, B., Al-Rfou, R. and Constant, N. (2021) The power of scale for parameter-efficient prompt tuning. *Proceedings of EMNLP 2021*, pp. 3045-3059.
- Li, X.L. and Liang, P. (2021) Prefix-tuning: Optimizing continuous prompts for generation. *Proceedings of ACL 2021*, pp. 4582-4597.
- Long, M., Cao, Y., Wang, J. and Jordan, M.I. (2015) Learning transferable features with deep adaptation networks. *Proceedings of ICML 2015*, pp. 97-105.
- Neyshabur, B., Sedghi, H. and Zhang, C. (2020) What is being transferred in transfer learning? *Advances in Neural Information Processing Systems*, 33, pp. 512-523.
- Pan, S.J. and Yang, Q. (2010) A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), pp. 1345-1359.
- Pan, X., Ge, C., Lu, R., Song, S., Chen, G., Huang, Z. and Huang, G. (2022) On the integration of self-attention and convolution. *Proceedings of CVPR 2022*, pp. 815-825.
- Rahimian, H.R., Salehkaleybar, S. and Hashemi, M. (2021) Domain adaptation approaches for improving prediction in new environments. *arXiv preprint arXiv:2109.08197*.
- Snell, J., Swersky, K. and Zemel, R. (2017) Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30, pp. 4077-4087.
- Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H. and Hospedales, T.M. (2018) Learning to compare: Relation network for few-shot learning. *Proceedings of CVPR 2018*, pp. 1199-1208.
- Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P. and Belongie, S. (2018) The iNaturalist species classification and detection dataset. *Proceedings of CVPR 2018*, pp. 8769-8778.
- Wang, Z., Dai, Z., Póczos, B. and Carbonell, J. (2019) Characterizing and avoiding negative transfer. *Proceedings of CVPR 2019*, pp. 11293-11302.

Yosinski, J., Clune, J., Bengio, Y. and Lipson, H. (2014) How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, 27, pp. 3320-3328.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H. and He, Q. (2021) A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1), pp. 43-76.

Declarations

Funding

This research was supported by the Estonian Research Council under grants PRG1226 (LowResourceML) and PRG1604 (BalticNLP), and by the EU Horizon 2020 programme under grant 952215 (TAILOR). Computational resources provided by the Estonian Scientific Computing Infrastructure (ETAIS). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All transfer learning code, adapter implementations, domain distance computation scripts, and experimental results are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19614580. WikiANN, ISIC 2019, and Snapshot Serengeti are publicly available. Baltic sentiment data released with this paper. Released under MIT License (code) and CC BY 4.0 (data).

Ethical Approval

This study involved computational experiments on publicly available datasets. The Baltic sentiment dataset was collected from publicly posted tweets and fully anonymised prior to annotation. No personal data were retained. The study was conducted under Baltic AI Research University ethical guidelines for computational research (protocol BARU-COMP-2021-0182). No additional institutional ethical approval was required.

Appendix A

Transfer Learning Decision Framework

This appendix provides a decision tree for selecting transfer strategies based on target data size, domain distance, and deployment constraints.

Step 1: Compute Domain Distance

- 1a. Pass target sample through source backbone; compute MMD to source features
- 1b. If distance < 0.30 --> Transfer will likely help substantially; go to Step 2
- 1c. If distance $0.30-0.50$ --> Transfer will help moderately; go to Step 2
- 1d. If distance > 0.50 --> High negative transfer risk; seek domain-specific source or use feature extraction

Step 2: Check Target Data Size

- 2a. < 100 examples per class --> MAML if meta-tasks available; else feature extraction
- 2b. $100-500$ examples --> Adapter modules (best balance of accuracy and safety)
- 2c. $500-2,000$ examples --> Adapter modules or full fine-tuning with early stopping
- 2d. $> 2,000$ examples --> Full fine-tuning (marginal accuracy advantage justifies cost)

Step 3: Deployment Constraint

- 3a. Multiple target tasks with shared backbone --> Adapters (15 MB per task vs. 400 MB)
- 3b. Single high-stakes task --> Full fine-tuning for maximum accuracy
- 3c. Extremely limited compute --> Feature extraction (no backbone training needed)
- 3d. Need rapid task switching --> Adapters or prompt tuning (millisecond swap)