Active Few-Shot Fine-Tuning

Jonas Hübotter, Bhavya Sukhija, Lenart Treven, Yarden As, Andreas Krause

Paper Code

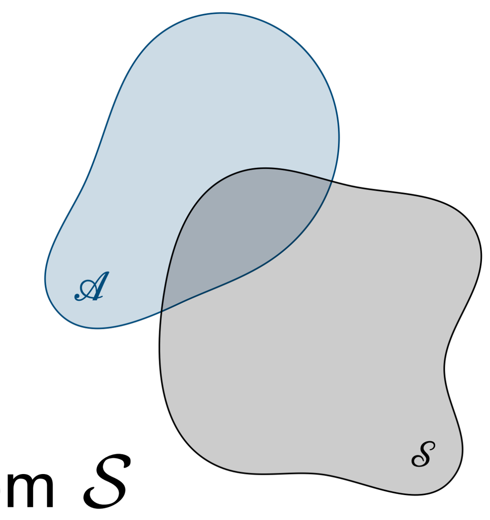
ETH zürich

Background

- Active learning is a powerful paradigm for data selection
- Many fine-tuning tasks have lots of structure that can be identified by large pre-trained models and used to effectively select data for fine-tuning
- So far: the growing literature on fine-tuning has primarily focused on architectures and optimization algorithms

How can we select the *right* data for fine-tuning a large NN to a specific task?

Transductive Active Learning



- Sample space $\mathcal{S} \subseteq \mathcal{X}$ (train set)
- Target space $\mathcal{A} \subseteq \mathcal{X}$ (test set)
- Unknown function f over $\mathcal{X}$

Goal: Learn f within $\mathcal{A}$ by sampling from $\mathcal{S}$

Can we exploit the (unknown) **latent structure** for fine-tuning?

- Embeddings $\phi(\cdot)$ generated by the NN capture the latent structure of f (e.g., neural tangent embds.)
- Approximate NN by a Gaussian process with kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x})^\top \phi(\boldsymbol{x}')$, which encodes "similarity"
- The marginal variance $\sigma_n^2(\boldsymbol{x})$ of the Gaussian process after n samples is a proxy for the approximation error at $\boldsymbol{x}$ after fine-tuning on these n samples

New goal: Reduce uncertainty $\sigma_n^2(\boldsymbol{x})$ at $\boldsymbol{x} \in \mathcal{A}$

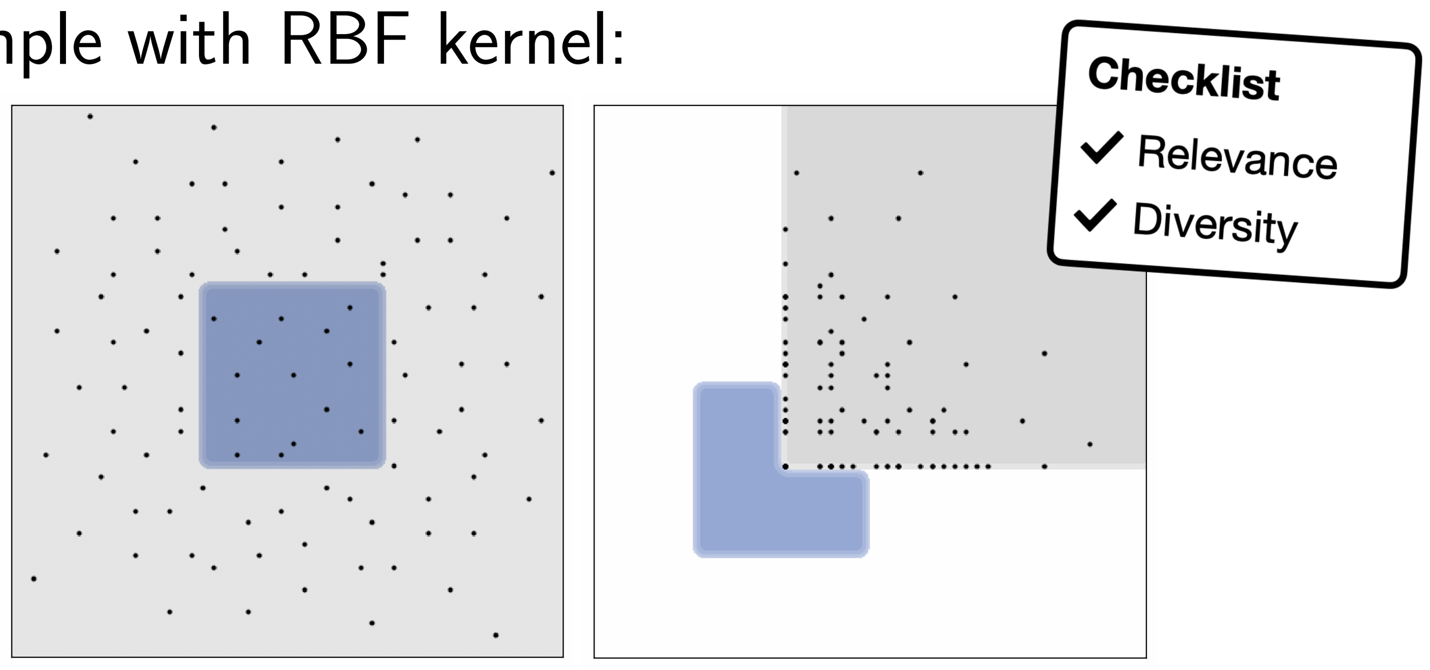
Information-based Transductive (Active) Learning

Proposal: Select samples to minimize posterior "uncertainty" within $\mathcal{A}$

Quantifying "uncertainty" by entropy yields **ITL**:

$$\begin{aligned} \boldsymbol{x}_n &= \operatorname*{arg\,min}_{\boldsymbol{x} \in \mathcal{S}} \mathrm{H}[\boldsymbol{f}(\mathcal{A}) \mid D_{n-1}, (\boldsymbol{x}, y)] \\ &= \operatorname*{arg\,max}_{\boldsymbol{x} \in \mathcal{S}} \mathrm{I}(\boldsymbol{f}(\mathcal{A}); (\boldsymbol{x}, y) \mid D_{n-1}) \end{aligned}$$

Example with RBF kernel:



Theory: Convergence Guarantees

How much can be learned about $\mathcal{A}$ from $\mathcal{S}$?

Generalization bound for ITL. For every $\boldsymbol{x} \in \mathcal{A}$:

$$\sigma_n^2(\boldsymbol{x}) \leq \underbrace{\mathrm{Var}[f(\boldsymbol{x}) \mid \boldsymbol{f}(\mathcal{S})]}_{\text{irreducible}} + \underbrace{C \log(n)/\sqrt{n}}_{\text{reducible}}$$

Approximation error bound for ITL. If $f \in \mathcal{H}_k(\mathcal{X})$ then for every $\boldsymbol{x} \in \mathcal{A}$ with probability $1 - \delta$:

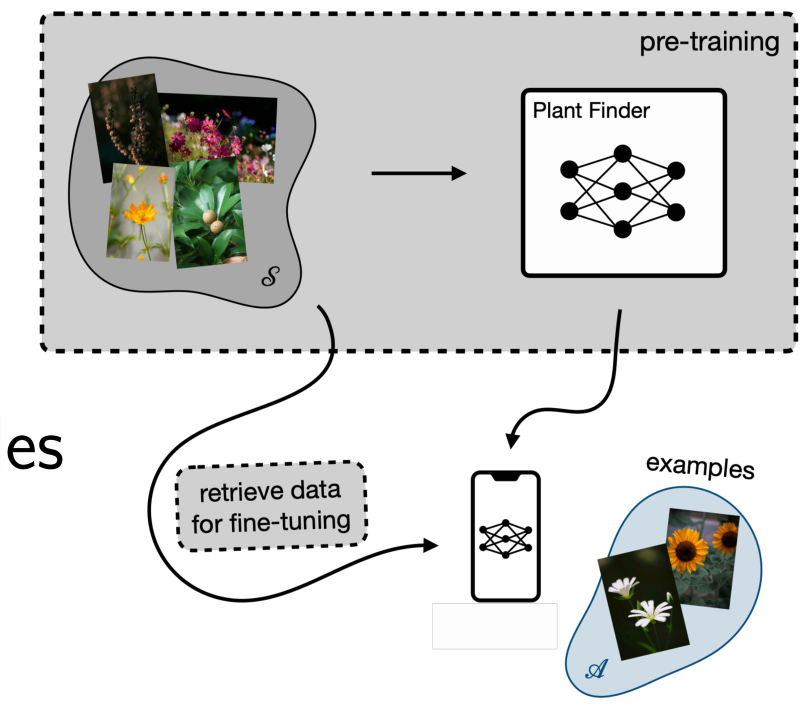
$$|f(\boldsymbol{x}) - \mu_n(\boldsymbol{x})| \leq \beta_n(\delta) \left[\text{irreducible} + C \log(n)/\sqrt{n}\right]$$

where $\mu_n(\boldsymbol{x})$ is the prediction and $\beta_n(\delta)$ the CI width

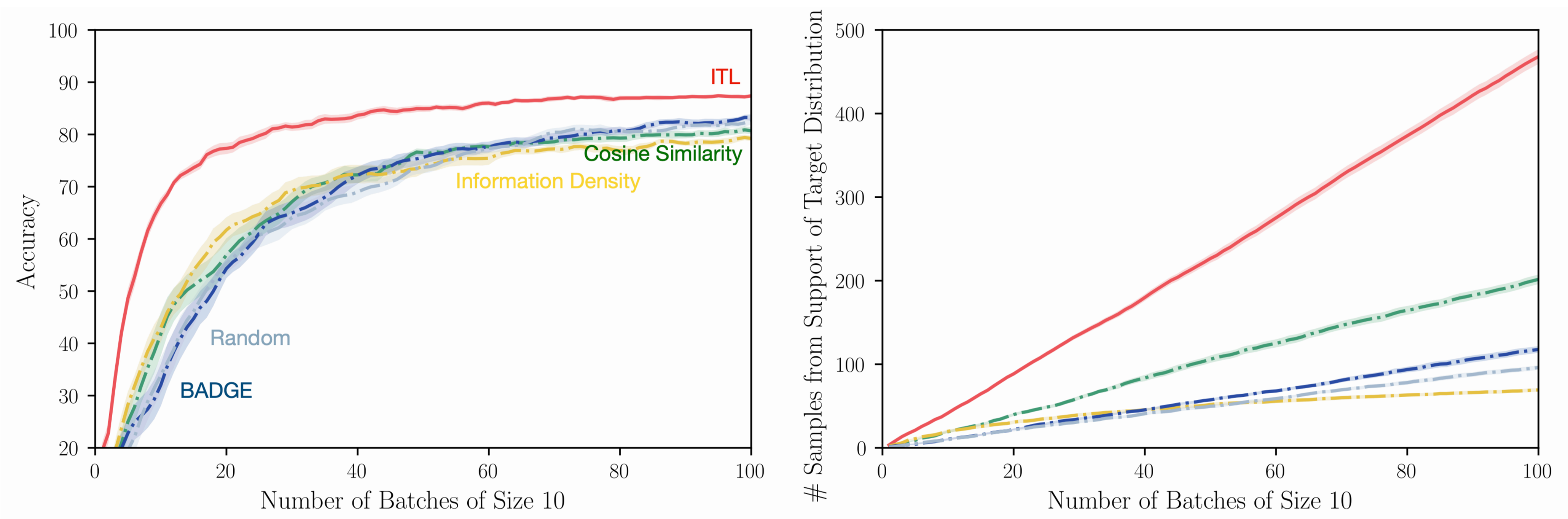
Practice

A motivating example: fine-tuning a plant classifier on a user's local biome

⇝ Find informative (that is, relevant & diverse) examples for the user's biome



Simplified example: fine-tuning on CIFAR-100



Cosine Similarity $\measuredangle$*:* only relevance $\operatorname*{arg\,max}_{\boldsymbol{x} \in \mathcal{S}} \sum_{\boldsymbol{x}' \in \mathcal{A}} \underbrace{\measuredangle(\phi(\boldsymbol{x}), \phi(\boldsymbol{x}'))}_{\mathrm{Cor}[f(\boldsymbol{x}), f(\boldsymbol{x}') \mid D_{n-1}]}$	*Information Density:* only relevance
BADGE: only diversity	***ITL:*** **relevance + diversity**

⇝ ITL generalizes $\measuredangle$ to query & batch sizes larger than 1

⇝ ITL synthesizes approaches to retrieval & coverage

Key Takeaways

- Retrieving the **right** examples for fine-tuning can lead to substantial performance gains
- Transductive active learning is a powerful paradigm for learning under resource constraints such as limited compute time & limited access
- ITL can be used as a simple drop-in replacement for random data selection