Specialization after Generalization: Towards Understanding Test-Time Training in Foundation Models

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS
ETH zürich
Jonas Hübotter*, Patrik Wolf*, Alexander Shevchenko*, Dennis Jüni, Andreas Krause, Gil Kur
Learning & Adaptive Systems

Motivation

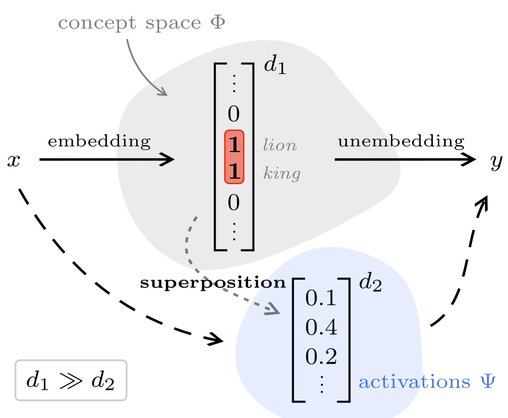
- In test-time training (TTT) each instance presents its own learning problem.
- TTT leads to strong empirical gains.
- But why? Prior explanations relied on **out-of-distribution** or **privileged data**.

We ask: Can TTT improve **in-distribution** while using only **already-seen data?**

Model

Linear representation hypothesis:

- Sparse high-dimensional concept space ϕ.
  - High-level (monosemantic) concepts.
- Target is **locally** linear in ϕ.
- Model learns a dense, lower-dimensional approximation Ψ of concept vectors ϕ.



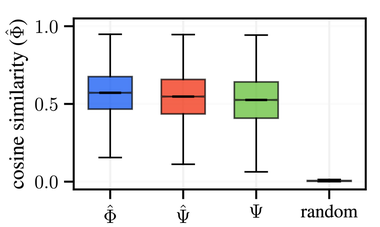
Key idea: Foundation models are *globally underparameterized*, but learning targets are *locally linear*.

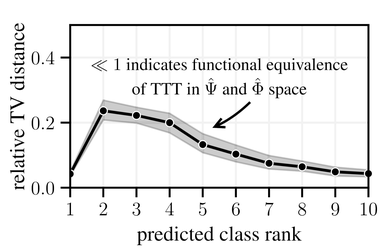
Test-time training ***specializes*** a model to the ***concepts*** relevant for a test instance.

How, when, and why does specialization work?

How does specialization behave?

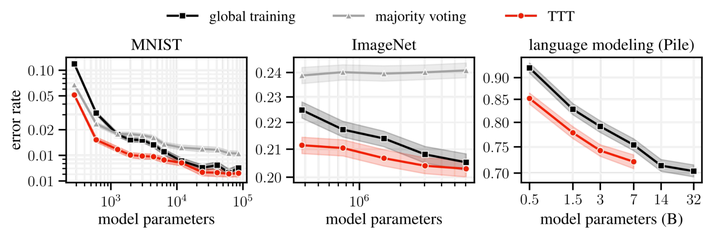
- **O1:** The learned features Ψ yield similar neighborhoods to those in the concept space ϕ.
- **O2:** Around the neighborhood of a test point, the ground truth function can be approximated by a sparse linear function in the concept space ϕ.
- **O3:** TTT implicitly implicitly adjust coefficients based on only a few concepts relevant to the test task.





When does specialization help?

1. Map test sample to feature space.
2. Identify nearest neighbors.
3. Fine-tune classification head.



TTT consistently outperforms global training and majority voting.

Takeaway. TTT locally improves predictions for underparameterized models, but its improvement diminishes as models become overparameterized.

Why may specialization help?

Let any x^* be any test point and Ψ locally sufficiently expressive to represent f. Then under Obs. 1–3, σ^2-subgaussian noise and regularity conditions, with high probability over the data sampling:

$$\left(f(x^*) - \langle \Psi(x^*), \hat{v}_{x^*}^{\mathrm{TTT}} \rangle\right)^2 \leq O\left(\frac{\sigma^2 s \log(d_1/s)}{k}\right)$$

Takeaway. TTT efficiently learns the meaning of exponentially many concepts from data, whereas global learning cannot disentangle them.