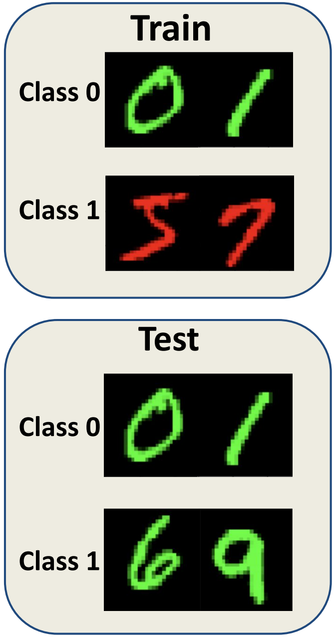
Rich Feature Construction for the Optimization-Generalization Dilemma

Jianyu Zhang[1] David Lopez-Paz[2] Léon Bottou[3,1]

[1]New York University, New York, NY, USA.
[2]Facebook AI Research, Paris, France.
[3]Facebook AI Research, New York, NY, USA.

Background

- Out of Distribution Generalization (OoD)
  - **Test & Train** have different distributions
  - In many settings, there exists **multiple different training environments**
  - The **invariant feature** (e.g. digits) works consistently on all training environments.
  - The **spurious feature** (e.g. color) doesn't.
- OoD Generalization Algorithms
  - Finding the **invariant feature** by adding penalties.



ColoredMNIST [1]

Optimization-generalization Dilemma

- **Dilemma:** A **strong generalization goal** in OoD, e.g. seeking an invariant representation (IRM), leads to an **optimization difficulty**.

Illustration of the Dilemma on ColoredMNIST

- Test the influence of **network initialization** on 9 OOD methods.
- All 9 methods depend on choosing the right initialization.
- **The OoD penalties are too strong to optimize reliably!**

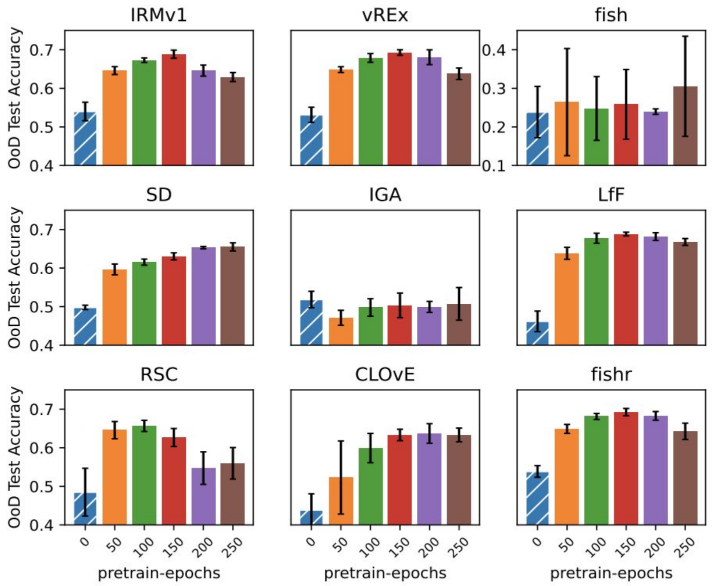


Figure 1. Test performance of nine penalized OoD methods as a function of the number of epochs used to pre-train the neural network with ERM. The final OoD testing performance is very dependent on choosing the right number of pretraining epochs, illustrating the challenges of these optimization problems.

Illustration of the Dilemma on ColoredMNIST

- How about learning from a “perfect” initialization? (where only the invariant feature is well learned)
- **No methods** can maintain the OoD performance.
- **OoD penalties are too weak to enforce invariance constraints!**

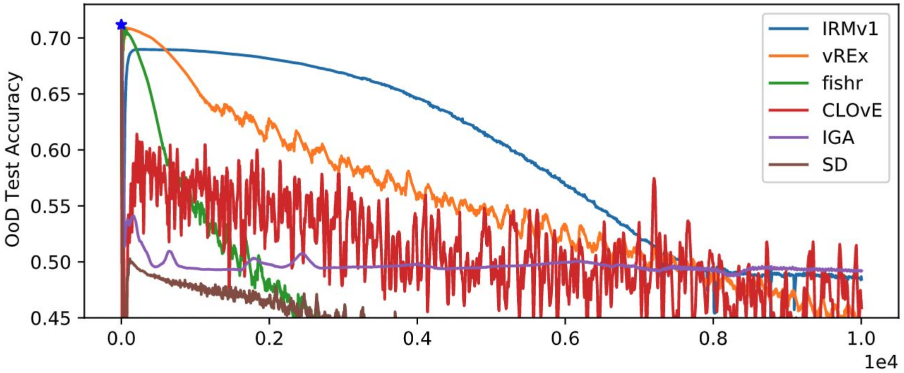


Figure 2: Test performance of OoD methods as a function of training epochs. Six OoD methods are trained from a ‘perfect’ initialization where only the robust feature is well learned. The blue star indicates the initial test accuracy.

Optimization-generalization Dilemma

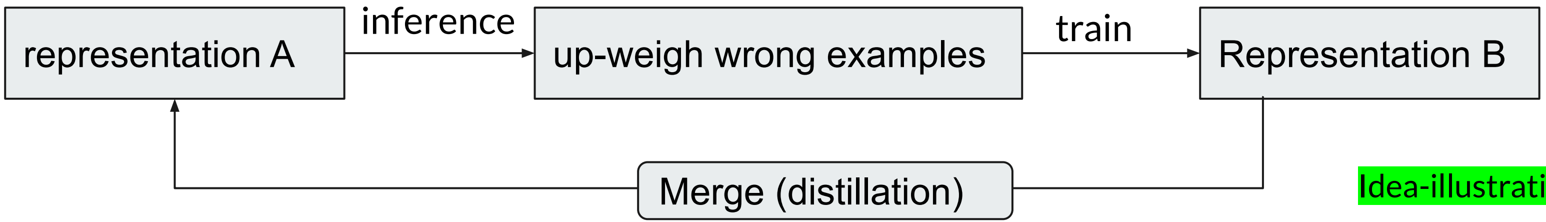
- The OoD problems associated with current OoD algorithms are **highly non-convex than usual**.
- Optimization becomes **super hard**.

Rich Feature Construction (called Bonsai)

- **Core Idea:** starting from a **representation with rich features** reduces the optimization difficulty.

Rich Feature Construction (called Bonsai)

- **Bonsai** creates such a rich representation by **impeding the learning process**.



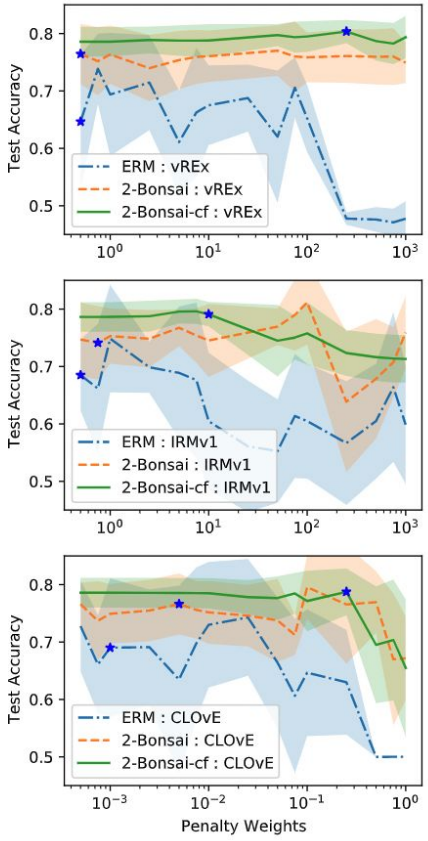
Idea-illustration version

- A distributionally robust optimization version avoids the **heuristic reweighting** and saves the **distillation time**.

Practical version

Camelyon17 tumor classification

- **Training set** contains tumor/non-tumor images from **three hospitals. Test set** comes from **another hospital.**
- Three OoD methods, vREx[2], IRMv1[1], CLOvE[3], are trained either on a **ERM pretrained** representation or the proposed **Bonsai rich** representation.
- "-cf": only the top layer classifier is trainable.



Reference

1. Martin Arjovsky, Leon Bottou, Ishaan Gulrajani, D. L.-P. (2020). Invariant Risk Minimization. 1–31.
2. Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Priol, R. Le, & Courville, A. (2020). Out-of-Distribution Generalization via Risk Extrapolation (REx).
3. Wald, Yoav, Amir Feder, Daniel Greenfeld, and Uri Shalit. "On calibration and out-of-domain generalization." Advances in neural information processing systems 34 (2021): 2215-2227.