Domain-Adjusted Regression or: ERM May Already Learn Features Sufficient for Out-of-Distribution Generalization

Elan Rosenfeld Pradeep Ravikumar Andrej Risteski

Paper:

Carnegie Mellon University
ML Machine Learning Department

Is the difficulty of distribution shift due to learning the wrong *features*?

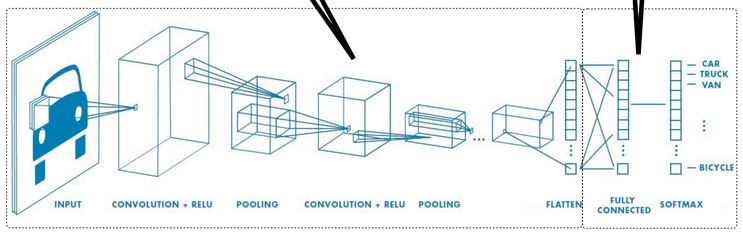
Or the wrong *classifier*?

TLDR: With a better linear classifier, features learned with standard ERM training can achieve substantially better accuracy even under massive distribution shift. We should focus more on learning simple, robust classifiers.

Part I: ERM Learns Better Features Than You Think

- Common belief: **DNNs fail under distribution shift because they learn the "wrong" features.**
- Proposed fixes brittle, **difficult to understand and optimize.**

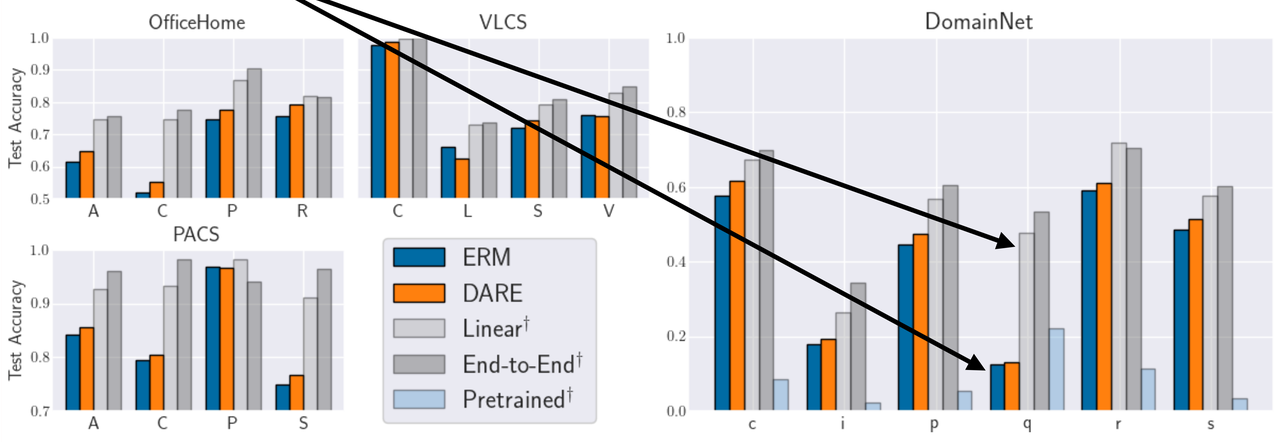
How good are the features we already have?



We train three networks:

1. **Standard ERM**; approximate state of the art
2. **End-to-end with test domain access**; "ideal" performance
3. **Freeze network** from (1) and **retrain just the last layer** with test domain access; lower bound on performance with ERM features

These two networks use *exactly the same features*

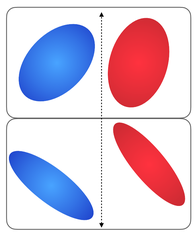


- Given small amount of test data, **just retraining the last linear layer** substantially beats SOTA.
- **Existing features are very good already**—no need to develop finicky end-to-end objectives.
- Prior robustness interventions ***do* perform better than ERM** when using frozen features. In other words, they succeed in **learning a more robust linear classifier**.

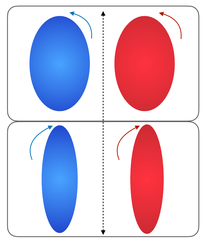
Part II: Domain-Adjusted Regression (DARE)

- Prior methods look for "invariant" representation: **throws away useful information!**
- Instead, model observations as **domain-specific transformations from shared representation space.**
- **We introduce and study a new latent variable model of distribution shift** which subsumes:
  - Covariate shift
  - Invariant/varying latent features [1]
  - Structured interventions on (some) causal DAGs [2]
- We propose a new objective (DARE) to handle this distribution shift.

Our Model: $\epsilon \sim p_e(\epsilon), \quad y = \mathbf{1}\{\beta^{*T}\epsilon + \eta \geq 0\}, \quad x = A_e\epsilon$



Prior work: Learn an invariant classifier. Low margin, ignores meaningful information



This work: Adjust each domain in a *unique, data-dependent* way, *then* learn a shared classifier.

- **Theorem 1:** Closed-form expression for the population minimizer of the DARE objective.
- **Theorem 2:** Derive exact DARE risk under **bounded distribution shift**. Also prove **risk is minimax.**
- **Theorem 3:** DARE risk approaches the minimax risk at a rate of $O\left(E^{-1/2}\right)$.
- **Theorem 4:** Given **unlabeled data from the test domain**, we derive finite-sample convergence bounds to Bayes-optimal risk for a new distribution.

See paper for more detailed theoretical analysis!

[1] The Risks of Invariant Risk Minimization. Rosenfeld et al. 2020
[2] Anchor Regression: Heterogeneous Data Meets Causality. Rothenhäusler et al. 2018