Freeze then Train: Towards Provable Representation Learning under Spurious Correlations and Feature Noise

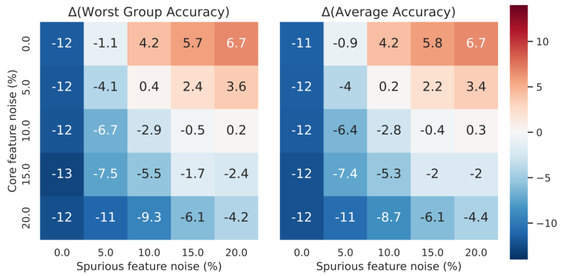
[paper] [code]

Haotian Ye[1], James Zou[2], Linjun Zhang[3] [1]Peking University, [2]Stanford University, [3]Rutgers University

1. Backgrounds

- **Goal**: learn a model in training environments **with spurious correlations**; performs well in test environments where the correlations are broken.
- **Observation**: ERM can learn the core features under spurious correlations, despite its bad test performance.
- **A new strategy "last layer retraining" (LLR)**: learn features in training environments, and retrain the last layer (linear probe) in test environments.
- **Advantage:**
  - Less demanding and more effective than OOD generalization;
  - More applicable and computationally efficient than general domain adaptation.
- However, this understanding is incomplete!

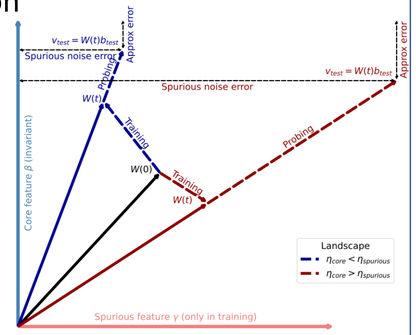
2. Findings & Motivations



- ERM only learns the core features when the noise of core features is smaller than that of spurious features.
- Why is noise so important for LLR performance?
- How to make LLR work under different noise conditions?

3. Intuitions

- LLR performance is determined by the quality of the learned features.
- ERM typically learns a mixture of different features.
- The proportion depends on the trade-off between information and noise: features with larger noise are used less.
- During LLR, when the proportion of the core feature is small, we suffer more to amplify it.



4. Frameworks & Theorems

- Our framework: a two-layers non-convex optimization.
- η_{core}, η_{spu} are core noise and spurious noise.
- **Main Theorem (informal)**:

$\eta_{core} < \eta_{spu}$ (Informal upper bound)

Under some assumptions, for any $\eta_{core} < \eta_{spu}$, any time t,

$$\ell_{te}(\mathbf{W}(t)) \leq \left(1 + \frac{\eta_{core}^2}{\eta_{spu}^2}\right) \mathrm{err}_{te}^* + \mathcal{O}(t^{-1}),$$

where err_{te}^* is the optimal testing error.

$\eta_{core} > \eta_{spu}$ (Informal lower bound)

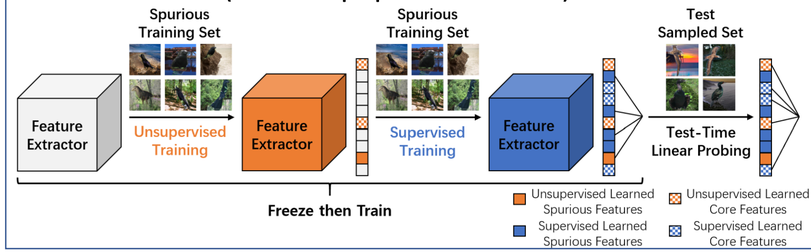
Under some assumptions, for any $\eta_{core} > \eta_{spu}$,

$$\lim_{t\to\infty} \frac{\ell_{te}(\boldsymbol{W}(t))}{\mathrm{err}_{te}^*} \geq 1 + \frac{\eta_{core}^2}{2\eta_{spu}^2}\left(1 \wedge \frac{1}{2\eta_{spu}^2 \left\|\Sigma^{-1}\right\|_2 \left\|\boldsymbol{W}_1^\dagger(\infty)\right\|_2^2}\right),$$

where $\boldsymbol{W}^\dagger$ is the Moore-Penrose inverse, and $a \wedge b$ takes the minimum over a, b.

5. Algorithms

- Takeaways: features learned in a supervised way are biased under certain noise conditions.
- Method: combine unsupervised method!
- Algorithm Freeze then Train: freezes certain salient features unsupervisedly and then trains the rest of the features supervisedly.
- We give theoretical guarantee on its performance in the framework. (see our paper for details)



6. Experiments

Dataset	η_{core} (%)	Worst Group Accuracy (%)					Average Accuracy (%)				
		ERM	IRM	CVaR-DRO	JTT	Ours	ERM	IRM	CVaR-DRO	JTT	Ours
Waterbirds	0	95.0	**95.3**	94.3	93.3	94.5	95.3	**95.5**	94.6	94.1	94.9
	2	93.6	**94.1**	93.8	89.7	93.6	94.2	**94.3**	94.0	90.7	94.2
	4	92.8	92.8	92.8	85.3	**92.9**	93.2	**93.5**	93.2	85.9	**93.5**
	6	90.8	91.5	77.8	86.8	**92.8**	91.3	91.8	77.8	87.1	**92.9**
	8	88.5	88.8	77.8	82.0	**92.7**	89.9	90.1	77.8	82.7	**93.0**
	10	87.6	87.9	77.8	78.6	**92.4**	89.4	89.4	77.8	78.9	**92.9**
	Mean	91.4	91.7	85.7	86.0	**93.1**	92.2	92.4	85.9	86.6	**93.6**
CelebA	0	95.0	95.2	92.9	94.4	**95.3**	**97.2**	**97.2**	96.0	96.7	**97.2**
	2	**95.2**	**95.2**	92.4	91.6	**95.2**	**97.2**	**97.2**	95.9	96.0	**97.2**
	4	94.5	94.2	91.9	92.7	**94.9**	97.1	97.0	95.5	96.4	**97.2**
	6	94.3	94.3	91.5	92.0	**94.4**	96.9	96.9	95.5	96.0	**97.0**
	8	93.7	93.8	91.4	91.4	**94.0**	**96.7**	**96.7**	95.4	95.7	**96.7**
	10	92.4	92.8	91.1	80.5	**93.1**	96.2	96.2	95.4	92.1	**96.3**
	Mean	94.2	94.2	91.9	90.4	**94.5**	**96.9**	**96.9**	95.6	95.5	**96.9**

Table 2: Test-time probing accuracy (%) for four methods on Waterbirds and CelebA, under different core noises η_{core}. **Bold** means the best accuracy across four methods. The "Mean" row stands for the average accuracy across η_{core}. We repeat all settings 10 times and average the numbers. For worst group accuracy, FTT (ours) can be competitive when η_{core} is small and outperform other algorithms by at most 4.5% when η_{core} increases. It can increase accuracy by 1.4% and 0.3% on Waterbirds and CelebA on average.

- FTT outperforms ERM, IRM, and DRO on (1) spurious correlations datasets; (2) distribution shift datasets.