Project-Probe-Aggregate: Efficient Fine-Tuning for Group Robustness

Beier Zhu, Jiequan Cui, Hanwang Zhang, Chi Zhang.

CVPR Highlight

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
WESTLAKE UNIVERSITY

Background

- Adapting image-text foundation models remains challenging due to group robustness: low average test error but incur **high risk on certain groups**.
- Two trends: efficient fine-tuning, and no access to training group labels.
- For the first trend, we train linear probes. For the second trend, we follow the common **failure-based debiasing**, which first identifies minority groups by training a biased classifier. Then, a debiased model is trained using the inferred group labels.
- **Two contributions:** (1) Failure-based debiasing hinges on the biased model that overfits on spurious features. We enhance bias by projecting out the class proxies from the input features, using the remaining information for bias discovery. (2) We predict pseudo group labels and applies a group prior offset to correct for imbalance. We prove the loss minimizes the balanced group error without heavy hyperparameter tuning.

Method

Setup: A classification problem with instance $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ and labels $y \in \mathcal{Y} = [K]$. Data point $(\mathbf{x}, y)$ has an attribute $a(\mathbf{x}) \in \mathcal{A}$ but is inaccessible in training. $g \in \mathcal{G}$ is the combination of a and y: $g = (a, y)$ with total $|\mathcal{G}| = |\mathcal{A}|\times|\mathcal{Y}|$ groups. Given a VLM like CLIP, we compute $Z = [\mathbf{z}_1, \dots, \mathbf{z}_K]^\top \in \mathbb{R}^{K\times d}$ whose rows are the text embeddings of the K class names: $\mathbf{z}_j$ is derived from a prompt like ``a photo of a `[CLASS]`".

Goal: Learn a function f that minimizes the Balanced Group Error (BGE): $\frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} \mathbb{E}_{\mathbf{x}|g}[y \neq \operatorname{argmax}_{y'\in\mathcal{Y}} f(\mathbf{x})_{y'}]$

Algorithm: Project-Probe-Aggregate (PPA)

Step 1: Project out class proxies. Let $\Pi \in \mathbb{R}^{d\times d}$ be the projection operator onto the null space of Z:

$$\Pi = I - Z^\top(ZZ^\top)^{-1}Z.$$

The biased model is $f_b(\mathbf{x}) = W_b \Pi \mathbf{x}$. To account for class imbalance, we apply the loss adjustment loss to learn W_b. Denote $\boldsymbol{\pi} = [\pi_1, \dots, \pi_K]$ the class priors, we have

$$\ell_{\text{la}}(W_b, \mathbf{x}, y) = -\ln \frac{\exp(f_b(\mathbf{x}) + \ln \boldsymbol{\pi})_y}{\sum_{y'\in\mathcal{Y}} \exp(f_b(\mathbf{x}) + \ln \boldsymbol{\pi})_{y'}}.$$

Step 2: Probe with group target. We identify minority groups that f_b misclassifies:

$$\hat{a}(\mathbf{x}) = \mathbf{1}[y \neq \operatorname*{argmax}_{y'\in\mathcal{Y}} f_b(\mathbf{x})_{y'}],$$

where $\mathbf{1}[\cdot]$ is the indicator function. Each training sample is augmented as $(\mathbf{x}, y, \hat{a})$, with the group label $\hat{g} = (y, \hat{a})$. Our PPA uses $h_d(\mathbf{x}) = W_d\mathbf{x}$: $\mathcal{X} \to \mathbb{R}^{|\hat{\mathcal{G}}|}$ to predict pseudo group labels. Let $\hat{\boldsymbol{\beta}}$ denote the group priors and $\tau > 0$ (hyperparameter), we propose **group logit adjustment loss** to achieve BGE:

$$\ell_{\text{gla}}(W_d, \mathbf{x}, \hat{g}) = -\ln \frac{\exp(h_d(\mathbf{x}) + \tau \ln \hat{\boldsymbol{\beta}})_{\hat{g}}}{\sum_{g'\in\hat{\mathcal{G}}} \exp(h_d(\mathbf{x}) + \tau \ln \hat{\boldsymbol{\beta}})_{g'}}.$$

Step 3: Aggregate weights. The final debiased classifier aggregates the weights belonging to each class:

$$f_d(\mathbf{x})_y = \mathbf{w}_y^\top \mathbf{x}, \text{ where } \mathbf{w}_y^\top = \textstyle\sum_{g\in\hat{\mathcal{G}}(y)} W_{d,g}$$

Experimental Results

Table 2. **Evaluation of methods for improving group robustness of CLIP models across the Waterbirds, CelebA, and MetaShift benchmarks.** Best worst-group accuracy (WGA) of the methods without group labels are in **bold**.

		CLIP ResNet-50						CLIP ViT-L/14					
Group labels		Waterbirds		CelebA		MetaShift		Waterbirds		CelebA		MetaShift	
in train sets?	Method	WGA	Avg	WGA	Avg	WGA	Avg	WGA	Avg	WGA	Avg	WGA	Avg
✓	GroupDRO [32]	75.1	83.8	84.1	89.5	83.2	87.3	90.8	96.4	88.3	91.2	93.9	97.4
	S-CS [44]	77.5	83.2	75.2	80.4	81.2	89.8	89.1	95.7	86.1	89.3	92.3	97.1
	S-CL [44]	75.2	86.0	75.6	80.4	81.5	88.8	89.9	96.0	87.8	90.5	93.1	96.9
	DFR [17]	73.2	83.8	80.0	92.8	83.1	88.3	89.7	97.8	85.6	90.8	92.3	97.0
✗	Zero-Shot (ZS) [31]	54.2	92.4	55.0	88.0	86.2	95.4	26.5	88.2	27.0	85.9	93.2	96.2
	Group Prompt ZS [31]	46.4	91.7	53.4	73.5	84.6	95.2	25.4	85.8	66.9	83.1	93.9	96.7
	ERM [37]	7.9	93.5	11.9	94.7	75.4	94.4	65.9	97.6	28.3	94.7	84.6	96.7
	WiSE-FT [41]	49.8	91.0	85.6	88.6	86.2	95.4	65.9	97.6	80.0	87.4	93.9	97.2
	Orth-Cali [3]	74.0	78.7	82.2	84.4	86.2	94.8	68.8	84.5	76.1	86.2	92.7	96.2
	AFR [30]	48.4	89.3	53.4	94.3	76.9	86.8	73.4	88.2	70.0	85.2	90.3	97.1
	JTT [21]	61.7	90.6	60.2	79.9	78.5	89.4	83.6	97.3	75.6	93.3	91.2	94.2
	CnC [49]	61.2	87.1	63.9	90.3	78.3	87.1	84.5	97.5	79.2	89.3	92.2	94.7
	CA [48]	83.7	89.4	90.0	90.7	77.9	85.5	86.9	96.2	84.6	90.4	91.3	93.4
	CFR [47]	76.9	77.6	73.7	81.1	81.5	89.5	**88.2**	96.8	84.8	87.8	93.7	95.5
	PPA (ours)	**84.3**	88.3	**91.1**	92.1	**90.8**	94.7	87.2	94.6	**90.4**	91.0	**94.8**	96.8

Table 6. **Main component analysis.** We present the worst group accuracies using CLIP ResNet-50. "**Proj.**" and "**GLA**" stands for the projection operation in Eq. (6) and group logit adjustment loss in Eq. (9), respectively. "**GT**" means we use the ground-truth group labels for training debiased models.

	Proj.	**GLA**	**GT**	Waterbirds	CelebA	MetaShift
(a)				7.9	11.9	75.4
(b)	✓			54.4	29.4	86.2
(c)		✓		81.6	70.0	89.2
(d)	✓	✓		84.3	91.1	90.8
(e)		✓	✓	86.8	91.5	91.3

Theoretical Justification

Removal of Class Proxies Amplifies Model Bias

Let $\boldsymbol{c}$ denote the core features which are stable for predict target y and s be a spurious feature which is correlated with y in the training data, but the correlation fails during testing. n observed features are stacked as $C = [\mathbf{c}_1, \dots, \mathbf{c}_n]^\top \in \mathbb{R}^{N\times d}$ and $\mathbf{s} = [s_1, \dots, s_n]^\top \in \mathbb{R}^N$.

Full model: regression on the core and spurious feature.

$$\mathbf{y} = C\boldsymbol{\alpha} + \gamma\mathbf{s} + \boldsymbol{\varepsilon},$$

Projected model: Project C to obtain $\tilde{C} = C\Pi$, then regress on projected features $\tilde{C}$ and spurious features.

$$\mathbf{y} = \tilde{C}\tilde{\boldsymbol{\alpha}} + \gamma'\mathbf{s} + \boldsymbol{\varepsilon}',$$

where $\boldsymbol{\alpha}, \tilde{\boldsymbol{\alpha}}, \gamma$ and $\gamma'$are weights. $\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}'$ are noise terms. Denote $C_o = C(I - \Pi)$. Let $\mathbf{y}_o = C_o\boldsymbol{\alpha}$ denote the contribution of the projected-out core features. Define $M = I - \tilde{C}^\top(\tilde{C}\tilde{C}^\top)^{-1}\tilde{C}$, $\mathbf{r}_{\mathbf{y}_o} = M\mathbf{y}_o$ and $\mathbf{r}_\mathbf{s} = M\mathbf{s}$.

Proposition 1. *The weight of the spurious feature after projection is*

$$\gamma' = \gamma + \frac{\mathbf{r}_\mathbf{s}^\top \mathbf{r}_{\mathbf{y}_o}}{\mathbf{r}_\mathbf{s}^\top \mathbf{r}_\mathbf{s}}$$

Projecting out core features can make the model more susceptible to spurious feature: $\gamma' > \gamma$. Because the spurious feature s is naturally positively correlated with $\mathbf{y}_o$ in the space of M, *i.e.*, $\mathbf{r}_\mathbf{s}^\top \mathbf{r}_{\mathbf{y}_o} > 0$ (**Step 1**).

Group Classification and Aggregation Mitigate Spurious Correlation

Proposition 2. *Let $\mathcal{G}(y)$ denote the set of groups with class label y, i.e., $\mathcal{G}(y) = \{g = (y', a) \in \mathcal{G} | y' = y\}$. Let $\boldsymbol{\beta}$ denote the group priors, i.e., $\boldsymbol{\beta}_g = \mathbb{P}(g)$. The prediction:*

$$\operatorname*{argmax}_{y\in\mathcal{Y}} f^*(\mathbf{x})_y = \operatorname*{argmax}_{y\in\mathcal{Y}} \sum_{g\in\mathcal{G}(y)} (h(\mathbf{x}) - \ln \boldsymbol{\beta})_g$$

is Bayes optimal for minimizing the balanced group error.

Guided by the proposition, we enforce group prior offset while learning the group classifier $h_d(\mathbf{x})$ which is our group adjustment loss ℓ_{gla} (**Step 2**). Since h_d is linear, summing over the output-space is equivalent to aggregating in weight-space, eliminating the overhead of group inference (**Step 3**).