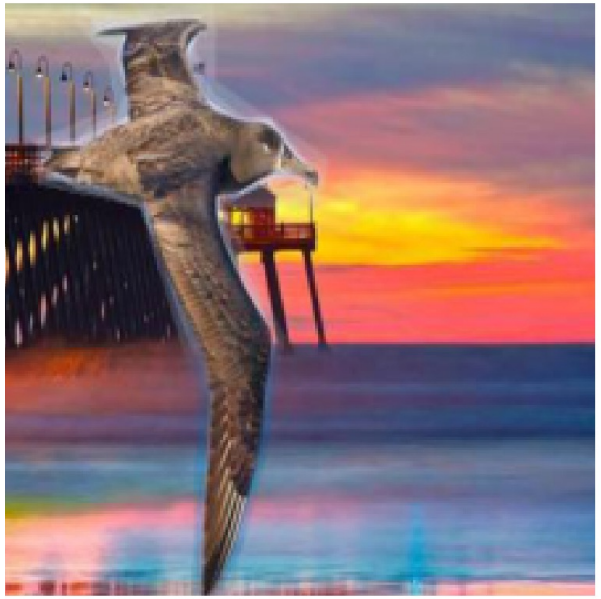
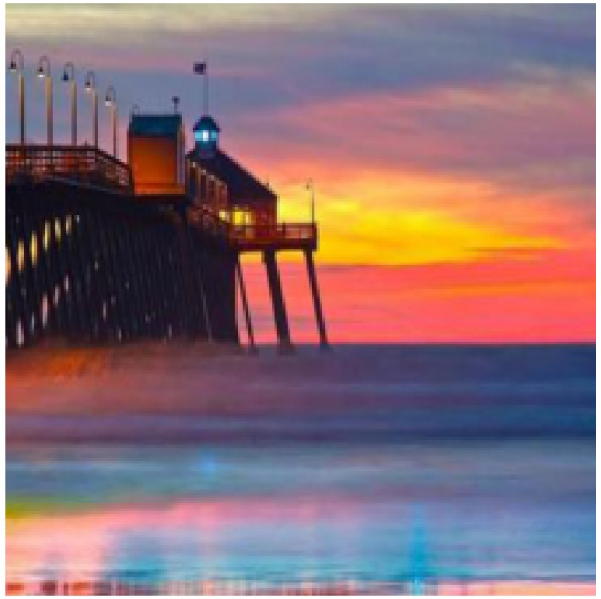
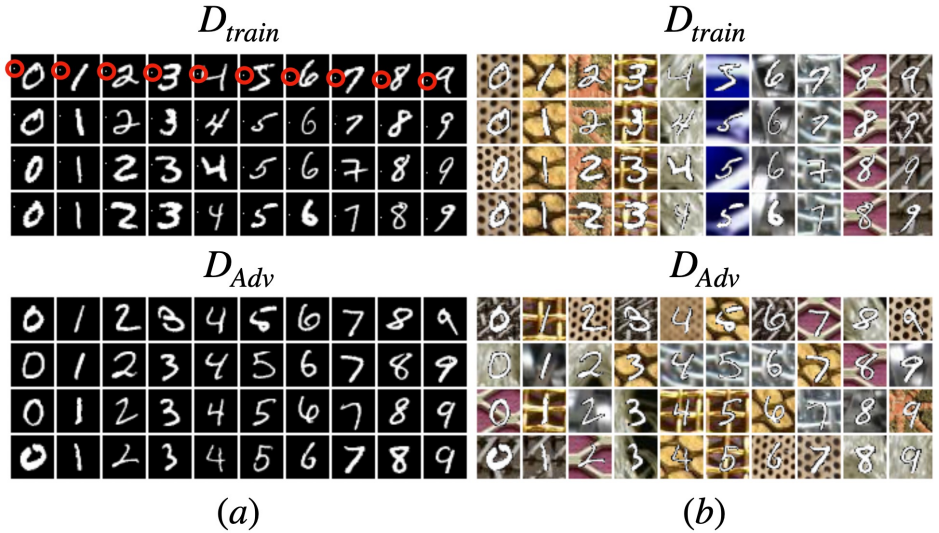
Обзор аналогов DFR





Vulnerability

Jörn-Henrik Jacobsen, Jens Behrmann, Richard Zemel, and Matthias Bethge. Excessive invariance causes adversarial vulnerability. ICLR 2019



% Error	$\mathbf{D_{Train}}$	$\mathbf{D_{Adv}}$
(a) CE ResNet	00.00	73.80
(a) CE fi-RevNet	00.00	57.09
(a) iCE fi-RevNet	00.02	34.73
(a) Difference	**00.02**	**38.33**
(b) CE ResNet	00.00	87.83
(b) CE fi-RevNet	00.18	73.71
(b) iCE fi-RevNet	00.53	59.99
(b) Difference	**00.53**	**27.84**

Definition 5 (Independence cross-entropy loss). *Let $F_\theta : \mathbb{R}^d \to \mathbb{R}^d$ a bijective network with parameters $\theta \in \mathbb{R}^{p_1}$ and $\tilde{F}_\theta(x) = softmax(F_\theta(x)_{1,\dots,C})$. Furthermore, let $D_{\theta_{nc}} : \mathbb{R}^{d-C} \to [0,1]^C$ be the nuisance classifier with $\theta_{nc} \in \mathbb{R}^{p_2}$. Then, the independence cross-entropy loss is defined as:*

$$\min_{\theta} \max_{\theta_{nc}} \mathcal{L}_{iCE}(\theta, \theta_{nc}) = \underbrace{\sum_{i=1}^{C} -y_i \log \tilde{F}_\theta^{z_s}(x)_i}_{=:\mathcal{L}_{sCE}(\theta)} + \underbrace{\sum_{i=1}^{C} y_i \log D_{\theta_{nc}}(F_\theta^{z_n}(x))_i}_{=:\mathcal{L}_{nCE}(\theta,\theta_{nc})} .$$

Group DRO

Стандартный DRO позволяет оптимизировать задачу, где Q - это какое-либо распределение, на котором мы хотим получать хорошее качество

$$\min_{\theta \in \Theta} \left\{ \mathcal{R}(\theta) := \sup_{Q \in \mathcal{Q}} \mathbb{E}_{(x,y)\sim Q} [\ell(\theta; (x, y))] \right\}$$

В Group DRO множество Q обобщается смесью из набора групп $\mathcal{G} = \{1, 2, \ldots, m\}$

$$\hat{\theta}_{\text{DRO}} := \arg\min_{\theta \in \Theta} \left\{ \hat{\mathcal{R}}(\theta) := \max_{g \in \mathcal{G}} \mathbb{E}_{(x,y)\sim \hat{P}_g} [\ell(\theta; (x, y))] \right\}$$

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization.

JTT

Наблюдение: простые модели учат легкие корреляции с фиктивными признаками

Алгоритм:

1. Обучаем стандартный ERM T шагов

$$\hat{f}_{\text{id}}$$

2. Находим объекты, классифицированные неверно

$$E = \{(x_i, y_i) \text{ s.t. } \hat{f}_{\text{id}}(x_i) \neq y_i\}$$

3. Учим модель со следующей функцией ошибки

$$J_{\text{up-ERM}}(\theta, E) = \left(\lambda_{\text{up}} \sum_{(x,y)\in E} \ell(x, y; \theta) + \sum_{(x,y)\notin E} \ell(x, y; \theta) \right)$$

4. Гиперпараметры подбираются по лоссу худшей группы

JTT

Для картинок ResNet50, для текстов BERT

Dataset	Worst-group Recall	Worst-group Precision	Worst-group Empirical Rate
Waterbirds	87.5%	19.1%	1.2%
CelebA	94.7%	9.4%	0.9%
MultiNLI	67.1%	2.2%	1.0%
CivilComments	96.9%	7.8%	0.9%

JTT

Method	Group labels in train set?	Waterbirds		CelebA		MultiNLI		CivilComments-WILDS	
		Avg Acc.	Worst-group Acc.	Avg Acc.	Worst-group Acc.	Avg Acc.	Worst-group Acc.	Avg Acc.	Worst-group Acc.
ERM	No	97.3%	72.6%	95.6%	47.2%	82.4%	67.9%	92.6%	57.4%
JTT (Ours)	No	93.3%	**86.7%**	88.0%	**81.1%**	78.6%	**72.6%**	91.1%	**69.3%**
Group DRO (Sagawa et al., 2020a)	Yes	93.5%	91.4%	92.9%	88.9%	81.4%	77.7%	88.9%	69.9%

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information.

CnC

Алгоритм:

1. Обучаем стандартный ERM

2. Генерируем контрастные батчи

Algorithm 2 Sampling two-sided contrastive batches

Require: Number of positives M and number of negatives N to sample for each batch.

Initialize set of contrastive batches $B = \{\}$

for $x_i \in \{x_i \in X : \hat{y}_i = y_i\}$ **do**

  Sample $M - 1$ additional "anchors" to obtain $\{x_i\}_{i=1}^{M}$ from $\{x_i \in X : \hat{y}_i = y_i\}$

  Sample M positives $\{x_m^+\}_{m=1}^{M}$ from $\{x_m^- \in X : \hat{y}_m^- = \hat{y}_i,\ y_m^- \neq y_i\}$

  Sample N negatives $\{x_n^-\}_{n=1}^{N}$ from $\{x_n^- \in X : \hat{y}_n^- = \hat{y}_i,\ y_n^- \neq y_i\}$

  Sample N negatives $\{x_n'^-\}_{n=1}^{N}$ from $\{x_n'^- \in X : \hat{y}_n'^- = \hat{y}_1^+, y_n'^- \neq y_1^+\}$

  Update contrastive batch set: $B \leftarrow B \cup \left(\{x_i\}_{i=1}^{M}, \{x_m^+\}_{m=1}^{M}, \{x_n^-\}_{n=1}^{N}, \{x_n'^-\}_{n=1}^{N}\right)$

end for

Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. Correct-ncontrast: A contrastive approach for improving robustness to spurious correlations.

CnC

Supervised contrastive loss стремится сблизить z к z+, а не z-

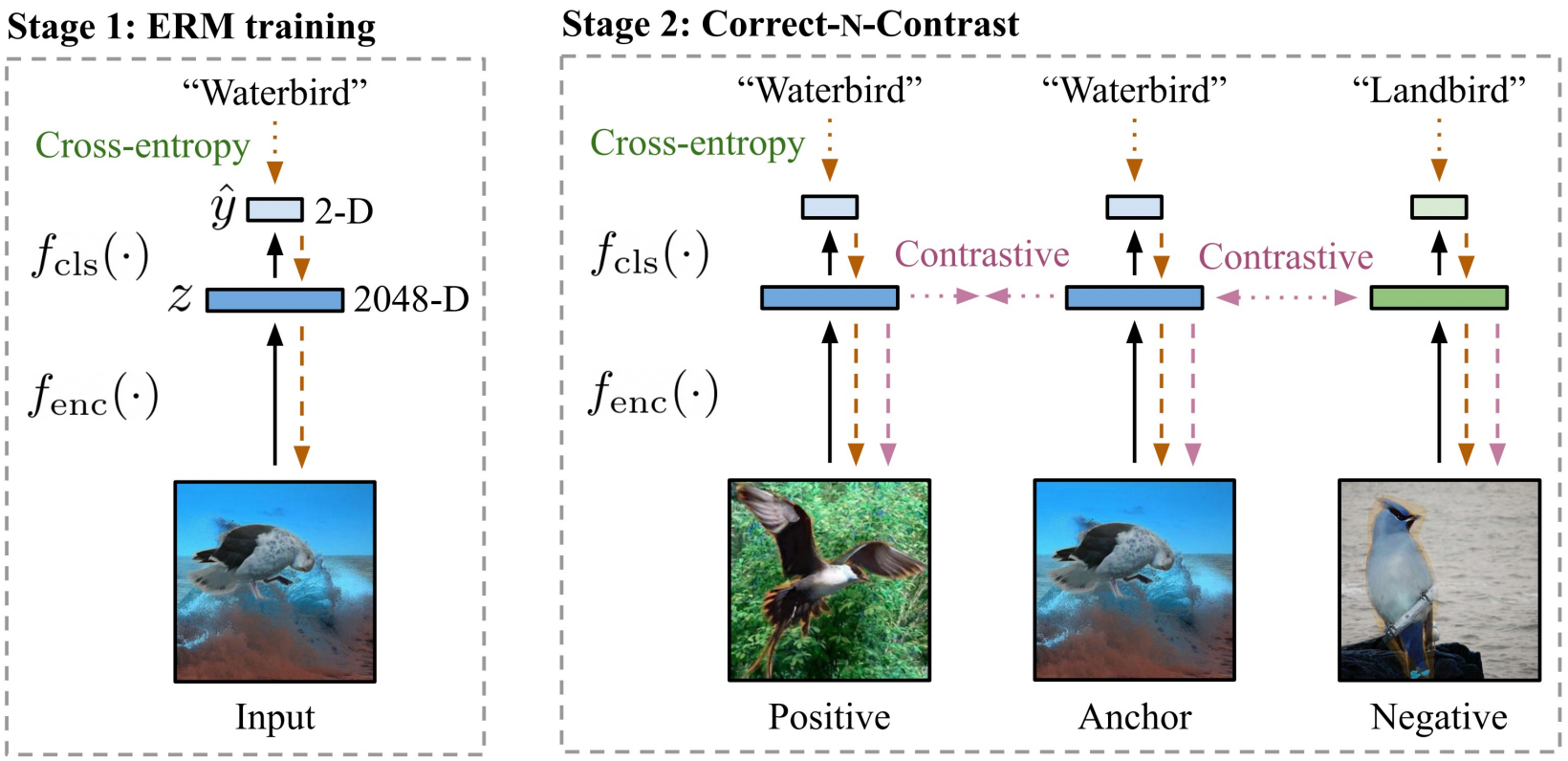
$$\mathcal{L}_{\text{con}}^{\text{sup}}(x; f_{\text{enc}}) = \mathop{\mathbb{E}}_{z,\{z_m^+\}_{m=1}^M,\{z_n^-\}_{n=1}^N} \left[-\log \frac{\exp(z^\top z_m^+/\tau)}{\sum_{m=1}^M \exp(z^\top z_m^+/\tau) + \sum_{n=1}^N \exp(z^\top z_n^-/\tau)} \right]$$

Loss внутри батча

$$\hat{\mathcal{L}}_{\text{con}}^{\text{sup}}(f_{\text{enc}}) = \hat{\mathcal{L}}_{\text{con}}^{\text{sup}}\left(x_1, \{x_m^+\}_{m=1}^M, \{x_n^-\}_{n=1}^N; f_{\text{enc}}\right) + \hat{\mathcal{L}}_{\text{con}}^{\text{sup}}\left(x_1^+, \{x_i\}_{i=1}^M, \{x_n'^-\}_{n=1}^N; f_{\text{enc}}\right)$$

$$\hat{\mathcal{L}}(f_\theta; x, y) = \lambda \hat{\mathcal{L}}_{\text{con}}^{\text{sup}}(f_{\text{enc}}; x, y) + (1-\lambda)\hat{\mathcal{L}}_{\text{cross}}(f_\theta; x, y)$$

CnC


Stage 1: ERM training
"Waterbird"
Cross-entropy
$\hat{y}$
2-D
$f_{\text{cls}}(\cdot)$
z
2048-D
$f_{\text{enc}}(\cdot)$
Input
Stage 2: Correct-N-Contrast
"Waterbird"
"Waterbird"
"Landbird"
Cross-entropy
$f_{\text{cls}}(\cdot)$
Contrastive
Contrastive
$f_{\text{enc}}(\cdot)$
Positive
Anchor
Negative

Simple data balancing

- **SUBY** – subsampling large classes
- **SUBG** - subsampling large groups
- **RWY** - reweighting the sampling probability of each example, mini-batches are class-balanced in expectation
- **RWG** - reweighting the sampling probability of each example, mini-batches are group-balanced in expectation

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy.

Simple data balancing

Method	#HP	Groups	Worst Acc				Average
			CelebA	Waterbirds	MultiNLI	CivilComments	
ERM	4	No	79.7±3.7	85.5±1.0	67.6±1.2	61.3±2.0	73.5
JTT	6	No	75.6±7.7	85.6±0.2	67.5±1.9	67.8±1.6	74.1
RWY	4	No	82.9±2.2	86.1±0.7	68.0±1.9	67.5±0.6	76.2
SUBY	4	No	79.9±3.3	82.4±1.7	64.9±1.4	51.2±3.0	69.6
RWG	4	Yes	84.3±1.8	87.6±1.6	69.6±1.0	72.0±1.9	78.4
SUBG	4	Yes	85.6±2.3	89.1±1.1	68.9±0.8	71.8±1.4	78.8
gDRO	5	Yes	86.9±1.1	87.1±3.4	78.0±0.7	69.9±1.2	80.5