#28
Wed. June 12
ICML 2019

Breaking Inter-Layer Co-Adaptation by Classifier Anonymization

Ikuro Sato[1], Kohta Ishikawa[1], Guoqing Liu[1], and Masayuki Tanaka[2]

[1]Denso IT Laboratory, Janan
[2]AIST, Japan

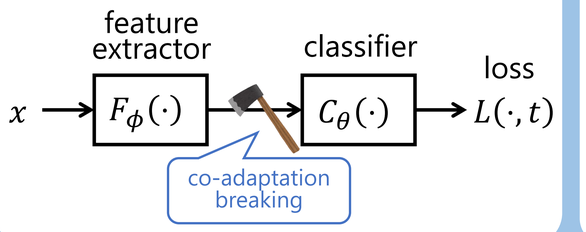
1. Summary

About what? **Breaking co-adaptation** between feature extractor and classifier.

How? By **classifier anonymization** technique.

Theory? Proved: Features form simple **point-like distribution**.

In reality? Point-like property largely confirmed on real datasets.



2. Possible problem with co-adaptation

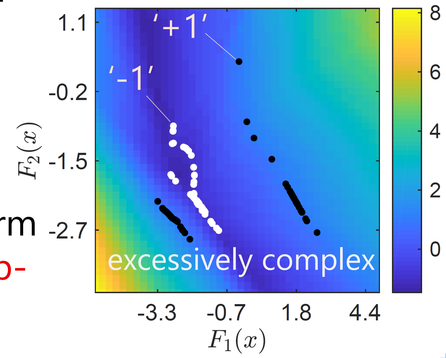
E2E opt.

$$(\phi^\star, \theta^\star) = \arg\min_{\phi,\theta} \frac{1}{\|\mathcal{D}\|_0} \sum_{(x,t)\in\mathcal{D}} L\left(C_\theta\left(F_\phi(x)\right), t\right) \quad (1)$$

Feature extractor $F_{\phi^\star}$ adapts to a particular classifier C_θ.

Features may form excessively complex distribution.

Toy) 2-class regression



3. FOCA: Feature-extractor Optimization through Classifier Anonymization

FOCA $\phi^\star = \arg\min_{\phi} \frac{1}{\|\mathcal{D}\|_0} \sum_{(x,t)\in\mathcal{D}} \mathbb{E}_{\theta\sim\Theta_\phi} L\left(C_\theta\left(F_\phi(x)\right), t\right)$ (2)

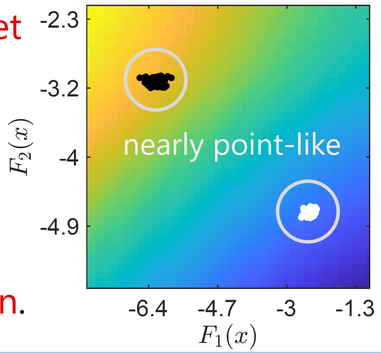
Classifier of small data-subset b (works as weak classifier to the entire dataset)

$\Theta_\phi = \mathcal{U}(\{\theta_{\phi,b}; b = b_1, b_2, \cdots\})$, $\mathcal{U}$: discrete uniform dist. (3)

$\theta_{\phi,b} = \arg\min_{\theta} \frac{1}{\|b\|_0} \sum_{(x,t)\in b} L\left(C_\theta\left(F_\phi(x)\right), t\right) + \lambda\|\theta\|_2^2$ (4)

$F_{\phi^\star}$ adapts to a set of random, weak classifiers $\{C_\theta\}$.

Features are expected to form simple distribution.



Algorithm 1 Approximate minimization in Eq. (2)

Input: total number of iterations T; number of classes C; number of class-c samples $n_c (c = 1, \cdots, C)$; number of samples per class for θ-update k; total number of samples $n_\mathcal{D}$; minibatch size for ϕ-update m; learning rate η

Begin
  Initialize ϕ by random variables.
  for $t = 1 : T$ **do**
    $I_c = [\text{randi}(n_1, k), \cdots, \text{randi}(n_C, k)]$
    $\theta = \arg\min_{\theta'} \sum_{i\in I_c} L\left(C_{\theta'}(F_\phi(x_i)), t_i\right) + \lambda\|\theta'\|_2^2$
    $I_f = \text{randi}(n_\mathcal{D}, m)$
    $\phi \leftarrow \phi - \frac{\eta}{m} \sum_{i\in I_f} \partial L\left(C_\theta(F_\phi(x_i)), t_i\right)/\partial\phi$
  end for
End

Output: feature-extractor parameters $\phi^\star = \phi$

4. Proposition of **point-like** property

Proposition 3.2. *Suppose that $\phi^\star$ simultaneously minimizes the classifier-anonymized, sample-wise losses $\mathbb{E}_{\theta\sim\Theta_\phi}\tilde{L}_{\phi,\theta}(x,t)$ in a class-separable fashion for all $(x,t) \in \mathcal{D}$. Then, samples from the same class share the same features; i.e., $F_{\phi^\star}(x) = F_{\phi^\star}(x'), \forall x, x' \in \mathcal{X}_c$, but samples from different classes do not; i.e., $F_{\phi^\star}(x) \neq F_{\phi^\star}(x'), \forall x \in \mathcal{X}_c, \forall x' \in \mathcal{X}_{c'\neq c}$.*

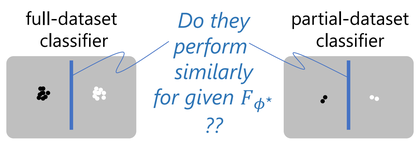
($\mathcal{X}_c$ is the set of input var with class c.)

Assumption
$$\begin{cases} \tilde{L}_{\phi,\theta}(x,t) = \left(C_\theta\left(F_\phi(x)\right) - t\right)^2 \\ C_\theta\left(F_\phi(x)\right) = \bar{\theta}^\top F_\phi(x) + \theta^0 \\ \Theta_\phi \text{ given by Eq.(3, 4)} \end{cases}$$

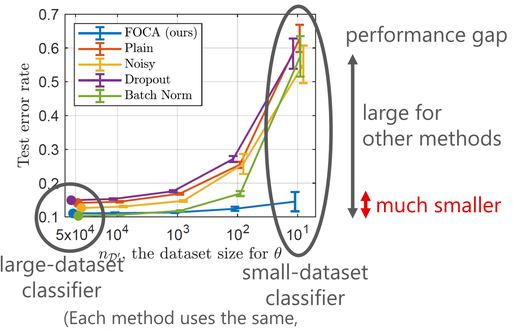
(Proof is given in the main text.)

5. Exp. #1: Partial-dataset opt.

Q. If features form point-like distribution, the decision boundary should be robust against the dataset size. *Is the classification performance robust?*



Result Yes, FOCA exhibits much smaller performance gap.



(Each method uses the same, fixed feature extractor is used.)

Exp. #2: Approx. geodesic dist.

Q. Let: θ^{LD}: large-dataset classifier param.
θ^{SD}: small-dataset classifier param.
Are they close?

Procedure
1) Partitions straight line connecting θ^{LD} and θ^{SD} into P line segments of equal lengths.
2) Comp. approx. geodesic distance:

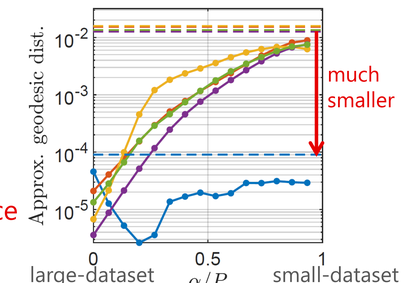
$$d(\theta^{LD}, \theta^{SD}) = \left[\sum_{\alpha=0}^{P-1} d(\theta^\alpha, \theta^{\alpha+1})^2\right]^{\frac{1}{2}}. \quad (15)$$

where $d(\theta^\alpha, \theta^{\alpha+1})^2 = (\theta^{\alpha+1} - \theta^\alpha)^\top \mathcal{I}^\alpha (\theta^{\alpha+1} - \theta^\alpha)$, (16)

$$\mathcal{I}^\alpha = \mathbb{E}_{(x,t)\in\tilde{\mathcal{D}}} \left(\frac{\partial L_{\phi^\star,\theta}(x,t)}{\partial\theta}\right) \left(\frac{\partial L_{\phi^\star,\theta}(x,t)}{\partial\theta}\right)^\top \Big|_{\theta=\theta^\alpha},$$

Result

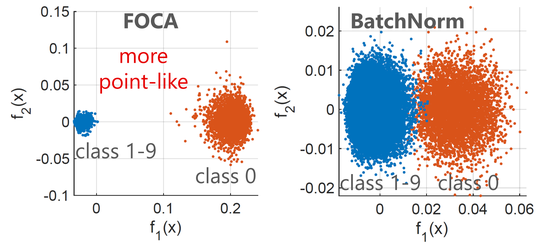
Yes, FOCA exhibits orders-of-magnitude smaller approx. geodesic distance (dashed lines).



Exp. #3: Low-dim. analyses

Q. *Qualitatively, is the low-dimensional structure point-like?*

Result Yes, FOCA (left) looks more point-like than BatchNorm (right) after LDA on normalized features.



Further, FOCA has the highest 2-class linear separability along principle axis:

Method	Eigenvalue	Test error rate
FOCA (ours)	**247.28**	**2.01%**
Plain	5.74	2.71%
Noisy	7.49	2.86%
Dropout	5.81	2.78%
Bach Norm	7.28	2.43%