Breaking Inter-Layer Co-Adaptation by Classifier Anonymization

Ikuro Sato\textsuperscript{1}  
Kohta Ishikawa\textsuperscript{1}  
Guoqing Liu\textsuperscript{1}  
Masayuki Tanaka\textsuperscript{2}  

\textsuperscript{1} Denso IT Laboratory. Inc., Japan  
\textsuperscript{2} National Institute of Advanced Industrial Science and Technology, Japan
Summary first

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.
E2E optimization scheme flourishes. Is it always good?

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

Feature extractor \( F_{\phi^*} \) adapts to a particular classifier \( C_\theta \).

Toy ex.) 2-class regression

Features may form excessively complex distribution.
- Disjointed
- Split 😞
FOCA: Feature-extractor Optimization through **Classifier Anonymization**

\[ \phi^* = \arg \min_{\phi} \frac{1}{\|D\|_0} \sum_{(x,t) \in D} \mathbb{E}_{\theta \sim \Theta_{\phi}} L \left( C_{\theta} \left( F_{\phi}(x) \right), t \right) \]

**Random weak classifier:** \( \theta \sim \Theta_{\phi} \)

Feature extractor \( F_{\phi^*} \) adapts to a set of weak classifiers \( \{C_{\theta}\} \).

Features form simple point-like distribution per class under some conditions.

---

**Additional note:**

```
Want to know more about \( \theta_{\phi} \)?
Please come to the poster!
```
Proposition about the **point-like** property

In words,

*If feature extractor has an enough representation ability, all input data of the same class are projected to a single point in the feature space in a class-separable way under certain conditions.*

Please see the paper for the proof.
Toy problem demonstration

Small-batch classifier works as a weak classifier to the entire dataset.

Small perturbations lead to point-like distribution.

I. Sato, et al., *Breaking Inter-Layer Co-Adaptation by Classifier Anonymization*, ICML 2019
Experiment #1: partial-dataset training

Thing we wish to confirm:

Do they perform similarly for given $F_{\phi^*}$?
Experiment #1: partial-dataset training

CIFAR10 test error rates

- FOCA (ours)
- Plain
- Noisy
- Dropout
- Batch Norm

Performance gap

large for other methods

much smaller for FOCA

One indication of point-like property

classifier trained with large dataset

classifier trained with small dataset

(The same, fixed feature extractor is used within each method.)
More experiments ...

including:

- Approximate geodesic distance measurements between large- and small-dataset solutions
- Low-dimensional analyses

to further study the point-like property.
**Poster #28 tonight**

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

**How?** By classifier anonymization.

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

**Reality?** Point-like property largely confirmed on real datasets.