Agree to Disagree: Diversity through Disagreement for Better Transferability

Guiv Farmanfarmaian
Mentor: Frédéric Berdoz
Seminar in Deep Neural Networks
19.03.2024, ETH Zurich
Motivation – Shortcomings of DNN

• Out of Distribution (OOD) setting: training and test data differ

From Beery et al. [2]
Motivation – Spurious vs Transferable Features

- Spurious Features (Correlation without Causation): Grass, mountains
- Transferable Features (Causation): Eyes, Ears, Body

From Beery et al. [2]
Shortcut Learning – Simplicity Bias

Learns Colors not Shape
Motivation - Objectives

Main Objectives

Avoid Shortcut Learning
- Generalize to OOD Distributions

Improve Uncertainty Estimation
Previous Work - Ensembles

• Solutions to increase diversity of ensemble:
  1. Train on different subsets of dataset
  2. Add orthogonality constraints on predictor’s gradient

From Breiman [3]
From Ross et al. [4]
Previous Work – OOD Generalization

| Methods to Increase Generalization |
|-----------------------------------|
| **Robust Learning**                |
| • Set of plausible test distributions U |
| • Minimize over worst distribution in U |
| **Invariant Learning**             |
| • Define a set of Environments     |
| • Output Indistinguishable among them |
Previous Work – Weakness of Invariant Learning

• Invariance $\not\Rightarrow$ Correctness

From Pagliardini et al. [1]
Previous Work – OOD generalization

Spurious Feature (i.e. Color) fully predictive
Previous work – Uncertainty Estimation

• Monte-Carlo Dropout, Bayesian Neural Networks, etc. improve uncertainty estimation

• Problem: Fail on OOD samples away from decision boundary

From van Amersfoort et al. [5]
From Liu et al. [6]
Previous work – Seminal Work (1)

| Simplicity Bias                  |
|----------------------------------|
| Teney et al. (2021)              |
| • Gradient orthogonality constraints at an intermediary level |
| • Problem: Reliance on pre-trained encoder; Large # of models needed |

From Teney et al. [7]
### Previous work – Seminal Work (2)

| OOD generalization                  |
|-------------------------------------|
| Lee et al. (2022)                   |

- Use mutual information
- Problem: don’t investigate uncertainty estimation; MI on entire dataset is costly

From Lee et al. [8]
Agree to Disagree – Diversity-By-disAgreement Training (D-BAT)

Core Idea

“Diverse hypotheses should agree on the source distribution $D$ while disagreeing on the OOD distribution $D_{ood}$”

From Pagliardini et al. [1]
D-BAT Intuition – Maximize Disagreement on White Space

Training Data

Model 1

Model 2

Ensemble

Code from Pagliardini et al. [1]
D-BAT - Metrics

\( \mathcal{X} \) input space \quad h : \mathcal{X} \to \mathcal{Y} \) labelling function

\( \mathcal{Y} \) output space \quad (\mathcal{D}, h) \) domain

\( \mathcal{D} \) distribution over \( \mathcal{X} \) \quad L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+ \) loss function

Expected Loss

\[
\mathcal{L}_\mathcal{D}(h_1, h_2) = \mathbb{E}_{x \sim \mathcal{D}} [L(h_1(x), h_2(x))]
\]
D-BAT – OOD Generalization

$(\mathcal{D}_t, h_t)$ training domain
$(\mathcal{D}_{ood}, h_{ood})$ unlabelled OOD domain

$\mathcal{H}$ set of all labelling functions
$\mathcal{H}_t^* := \arg\min_{h \in \mathcal{H}} \mathcal{L}_{\mathcal{D}_t}(h, h_t)$
$\mathcal{H}_{ood}^* := \arg\min_{h \in \mathcal{H}} \mathcal{L}_{\mathcal{D}_{ood}}(h, h_{ood})$

Key Assumption

$\mathcal{H}_t^* \cap \mathcal{H}_{ood}^* \neq \emptyset$
D-BAT - Objective

No OOD labels ⇒ Minimize a proxy

$$\mathcal{L}_{D_{ood}}(h_1, h_{ood}) = \max_{h_2 \in H_t^* \cap H_{ood}^*} \mathcal{L}_{D_{ood}}(h_1, h_2) \leq \max_{h_2 \in H_t^*} \mathcal{L}_{D_{ood}}(h_1, h_2) \approx \mathcal{L}_{D_{ood}}(h_1, h_{D-BAT})$$

Objective

$$h_{D-BAT} \in \min_{h_2 \in H} \left[ \mathcal{L}_{D_t}(h_2, h_t) + \alpha \mathcal{A}_{D_{ood}}(h_1, h_2) \right]$$
D-BAT Algorithm for 2 predictors

1. Train $h_1$ by minimizing the training data loss
2. Train $h_2$ by also considering the agreement with $h_1$ on the OOD data

$$h_2^* \in \arg\min_{h_2 \in \mathcal{H}} \frac{1}{N} \left( \sum_{(x,y) \in \mathcal{D}} \mathcal{L}(h_2(x), y) + \alpha \sum_{\tilde{x} \in \mathcal{D}_{ood}} A_{\tilde{x}}(h_1, h_2) \right)$$
D-BAT – Ensemble of predictors

Inspired by Pagliardini et al. [1]
D-BAT Theorem: Assumptions

Color, Shape and Label Combinations

Training Data $D$

- **A**
- **B**

Probability 1/2

Uniform OOD Distribution $D_{ood}$

- **A**, **B**, **A**, **B**, **A**, **B**

Probability 1/8
D-BAT Theorem: Assumptions

Training Data D

Model 1: Learns Colors to Predict Labels

P(Label = ‘A’ | Color = Blue) = 1
P(Label = ‘A’ | Color = Red) = 0
D-BAT Theorem: Predict Labels

Training Data D

Model 1: Learns Colors
Model 2: Learns Shapes

P(Label = ‘A’ | Shape = ■) = 1
P(Label = ‘A’ | Shape = ●) = 0
Assumptions for D-BAT

• Existence of a transferable function: \( h^* \in \mathcal{H}_t^* \cap \mathcal{H}_{ood}^* \)

• Counterfactual correlations essential for OOD distribution

OOD data
Colored MNIST Dataset
Experimental Results: Performance Comparison

Camelyon17 dataset

From Pagliardini et al. [1]
Experimental Results - Uncertainty Estimation

CIFAR-10 Dataset
3 Models with Similar Performance ->
D-BAT Better at Uncertainty Estimation on OOD samples

From Pagliardini et al. [1]
## Experimental Results - Key Takeaways

### D-BAT Achievements

| Better Generalization:                                                                 | Improves Uncertainty Estimation |
|--------------------------------------------------------------------------------------|---------------------------------|
| • On Natural Domains                                                                 |                                 |
| • With Ensemble                                                                      |                                 |
| • When OOD test data (i.e. new domains)                                              |                                 |
Personal Opinion

• Approach beautifully self-evident
• Training ensemble of models computationally expensive
• No control over OOD distribution -> hard to know whether features have counterfactual correlations
Questions / Your Opinions
Sources

[1]: Pagliardini, M., Jaggi, M., Fleuret, F., and Karimireddy, S. P. Agree to disagree: Diversity through disagreement for better transferability. arXiv preprint arXiv:2202.04414, 2022.

[2]: Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In ECCV (16), volume 11220 of Lecture Notes in Computer Science, pp. 472–489. Springer, 2018.

[3]: Leo Breiman. Bagging predictors. Mach. Learn., 24(2):123–140, 1996.

[4]: Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In ECCV (16), volume 11220 of Lecture Notes in Computer Science, pp. 472–489. Springer, 2018.

[5]: Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In ICML, volume 119 of Proceedings of Machine Learning Research, pp. 9690–9700. PMLR, 2020.

[6]: Yehao Liu, Matteo Pagliardini, Tatjana Chavdarova, and Sebastian U. Stich. The peril of popular deep learning uncertainty estimation methods. 2021b.

[7]: Damien Teney, Ehsan Abbasnejad, Simon Lucey, and Anton van den Hengel. Evading the simplicity bias: Training a diverse set of models discovers solutions with superior OOD generalization. CoRR, abs/2105.05612, 2021.

[8]: Yoonho Lee, Huaxiu Yao, and Chelsea Finn. Diversify and disambiguate: Learning from underspecified data. CoRR, abs/2202.03418, 2022.
Appendix: Experimental Results – Artificial Datasets

From Pagliardini et al. [1]

| Dataset $\mathcal{D}$ | Single Model |  |
|-----------------------|--------------|---|
|                       | ERM          | D-BAT        |
| C-MNIST               | 12.3 ± 0.7   | 90.2 ± 3.7   |
| M/F-D                 | 52.9 ± 0.1   | 94.8 ± 0.3   |
| M/C-D                 | 50.0 ± 0.0   | 73.3 ± 1.2   |

Case where OOD data = test data
## Appendix: Experimental Results – Natural Datasets (1)

| Dataset $\mathcal{D}$ | Single Model | | | Ensemble | | |
|------------------------|--------------|-------------|--------------|--------------|-------------|
|                        | ERM          | D-BAT       | ERM          | D-BAT        |
| Waterbirds             | 86.0 ± 0.5   | **88.7 ± 0.2** | 85.8 ± 0.4   | **87.5 ± 0.0** |
| Office-Home            | **50.4 ± 1.0** | **51.1 ± 0.7** | 52.0 ± 0.5   | **52.7 ± 0.2** |
| Camelyon17             | 80.3 ± 0.4   | **93.1 ± 0.3** | 80.9 ± 1.5   | **91.9 ± 0.4** |

Case where OOD data = test data

From Pagliardini et al. [1]
Appendix: Experimental Results – Natural Datasets (2)

| Case where OOD data ≠ test data | \( D_{ood} \neq \text{test data} \) |
|---------------------------------|----------------------------------------|
|                                 | Single Model | Ensemble |
|                                 | ERM | D-BAT | ERM | D-BAT |
| Office-Home                    | 51.7 ± 0.6 | 51.7 ± 0.3 | 53.9 ± 0.4 | 54.5 ± 0.5 |
| Camelyon17                     | 80.3 ± 0.4 | 88.8 ± 1.4 | 80.9 ± 1.5 | 85.9 ± 0.9 |

Case where OOD data ≠ test data

From Pagliardini et al. [1]
Appendix: Experimental Results – Ensemble on Natural Datasets

Waterbirds Dataset

Office-Home Dataset

From Pagliardini et al. [1]
Appendix: Choice of the Hyperparameter $\alpha$