Evading the Simplicity Bias: Training a Diverse Set of Models Discovers Solutions with Superior OOD Generalization[Teney et al., 2022]

Presenter: Yang Zhang

SDS, Fudan University

February 21, 2024

Introduction

- Evading the Simplicity Bias: Training a Diverse Set of Models Discovers Solutions with Superior OOD Generalization
- Publication: *CVPR 2022*
- Abstract:
  - Neural networks trained with SGD are shown to have **simplicity bias** which can explain their lack of robustness out of distribution (OOD).
  - They **train a set of similar models to fit the data in different ways using a penalty on the alignment of their input gradients**. They show theoretically and empirically that this induces the learning of more complex predictive patterns.
  - OOD generalization fundamentally requires information beyond i.i.d. examples. Their approach shows that we can **defer this requirement from training stage to an independent model selection stage**.

Inductive Bias and OOD Generalization

- At the core of every learning algorithm are a set of inductive biases. They define the learned function outside of training examples and they allow extrapolation to novel test points.

- shape(bird) or background(sky)? This is where a learning algorithm's inductive biases come into play.
- OOD generalization is not achievable only through regularizers, network architectures, or unsupervised control of inductive biases.

Simplicity Bias

- Simplicity is defined corresponding to the **feature** which induce minimal linear decision boundary.
- **Not a property of neural networks themselves**. [Shah et al., 2020] showed that neural networks trained with SGD are biased to learn the simplest predictive features in the data while ignoring others.
- **Pros**: by promoting simpler decision boundary, can act as an implicit regularizer and improves generalization.
- **Cons**: mechanisms to learn are more likely to be overshadowed by **simpler spurious patterns**. This will lead to **shortcut learning** or **poor OOD generalization**.
  - CV: use the background rather than the shape of the object to do image recognition.
  - NLP: use the presence of certain words rather than the overall meaning of a sentence to do atural language understanding.

Method Overview

The regularizer is required because trivial options such as training models with **different initial weights, hyperparameters, architectures, or shuffling of the data** do not prevent converging to very similar solutions affected by the simplicity bias.

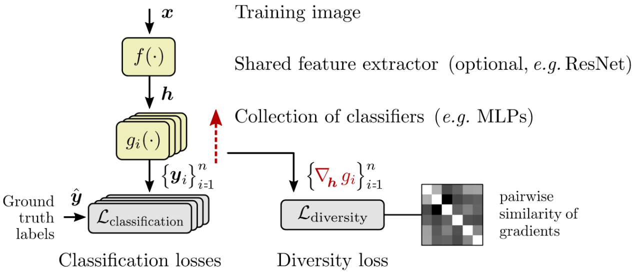

Figure: A **diversity loss** penalizes pairwise similarities between models, using each classifier's input gradient at training points.

Setup

- Dataset: $T = \{(x^k, y^k)\}_{k=1}^K$
- Model: $F : \mathrm{supp}(x) \to \mathrm{supp}(y)$, suppose $F = g \circ f$ where f_θ is a feature extractor and g_ϕ is a classifier. $h = f_\theta(x)$ is hidden representation of input data.
- Train:

$$\min_{\theta,\phi} \quad \mathcal{R}(F_{\theta,\phi})$$

where $\mathcal{R}(F_{\theta,\phi}) = \sum_{k=1}^K \mathcal{L}_{\mathrm{cls}}(\hat{y}^k, y^k)$ and $\hat{y}^k = F_{\theta,\phi}(x^k)$.

- Diversity Loss: we compare the functions implemented classifiers using their input gradients

$$\delta_{g_{\phi_1}, g_{\phi_2}}(h) = \nabla_h g^*_{\phi_1}(h) \cdot \nabla_h g^*_{\phi_2}(h)$$

where ∇g^* is the gradient of its largest component (top predicted score).

- Complete Method:

$$\min_{(\theta,\{\phi_i\})} \quad {\color{blue}\sum_i \mathcal{R}(F_{\theta,\phi_i})} + \lambda {\color{red}\sum_{i\neq j} \sum_k \delta_{g_{\phi_i}, g_{\phi_j}}(h^k)}$$

FAQ

- How diversity can induce complexity?
  - By assumption of the simplicity bias the model learned by default lies at one end of the space of solutions.
- Why use input gradients to quantify diversity?
  - [Selvaraju et al., 2017] show input gradients are indicative of the **features** used by the model.
  - Furthermore $g(h) = g(h') + (h - h')\nabla_h g(h')$ where h is a test point and h' is a nearby training point.
- See more in Appendix A:
  - Where to split a model into "feature extractor" and "classifier"?
  - Why not design the diversity regularizer on the activations of the models but on the input gradients?
  - Is the introduction of more diversity just a fancy random search?

Multi-dataset collages

- This experiment try to figure out: *Can we learn predictive patterns otherwise ignored by standard SGD and existing regularizers?*

Collages dataset (accuracy in %)	Best model on MNIST	SVHN	Fashion-M.	CIFAR-10	Average
Upper bounds: one predictive block in tr.	99.7	89.7	77.4	68.7	83.9
Baseline, 32 models with different seeds	99.7 ±0.0	50.0 ±0.1	51.2 ±0.3	50.1 ±0.1	62.7
With dropout (best rate: 0.5)	98.7	54.8	52.9	54.9	65.3
With penalty on L1 norm of gradients	98.9	49.8	50.7	49.9	62.3
With Jacobian regularization [28]	98.8	49.8	50.7	49.9	62.3
With spectral decoupling [57]	99.1	49.8	50.7	49.9	62.4
Proposed, 8 models	**97.3** ±0.5	82.1 ±6.0	59.6 ±4.0	55.8 ±1.9	73.7
Proposed, 16 models	96.6 ±1.2	72.1 ±10.3	64.6 ±4.0	58.4 ±1.4	72.9
Proposed, 32 models	95.6 ±0.3	81.8 ±5.3	69.2 ±2.8	61.1 ±1.0	76.9
Proposed, 64 models	95.5 ±0.1	80.9 ±5.8	70.7 ±1.5	60.8 ±0.9	77.0
Proposed, 96 models	95.8 ±0.8	**84.7** ±4.0	**71.7** ±1.1	**61.7** ±1.2	**78.5**

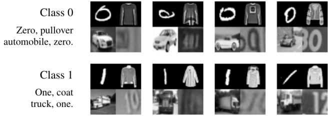

Figure 3. Training examples of collages. Each block features one of two pre-selected classes from MNIST, Fashion-MNIST, CIFAR-10, SVHN. All four blocks are predictive of training labels. Because of the simplicity bias, a standard classifier latches on MNIST and ignores others.

Biased activity recognition

- This experiment try to figure out: *Are these patterns relevant for OOD generalization in computer vision tasks?*

	Biased activity recognition (BAR) dataset		
Training collection of 64 models, reporting performance of:	**Single model** (average accuracy in the collection)	**Ensemble** (whole collection)	**Best single model** (oracle selection)
Baseline in [48]	51.9 ± 5.92	N/A	N/A
Learning from failure [48]	63.0 ± 2.76	N/A	N/A
Our strong baseline: frozen ResNet-50, 2-layer MLP	62.0 ± 0.3	63.1 ± 0.2	64.9 ± 0.7
Penalty on sq. L2 norm of gradient (Jacobian reg. [28])	63.7 ± 0.4	64.5 ± 0.7	67.0 ± 0.9
Penalty on sq. L2 norm of feature-gradient product (App. E)	62.8 ± 0.1	63.9 ± 0.6	65.9 ± 0.5
Penalty on L1 norm of feature-gradient product (App. E)	63.9 ± 0.3	64.6 ± 0.4	66.1 ± 0.3
Penalty on sq. L2 norm of logits (spectral decoupling [57])	64.3 ± 0.2	65.2 ± 0.5	67.0 ± 0.4
Proposed, 8 models	**64.9** ± 0.8	65.9 ± 0.4	66.8 ± 0.5
Proposed, 64 models	64.4 ± 0.2	**66.1** ± 0.3	**67.1** ± 0.3

Domain generalization

- PACS dataset is a standard benchmark for visual domain generalization (DG). PACS contains 4 domain(Art, Cartoon, Photo and Sketch) and each domain contains 7 categories.
- VLCS is included for an additional cross-dataset evaluation i.e. zero-shot transfer.

Training set	PACS (cartoon, photo, sketch)			
Test set	PACS (art)			VLCS (horse/person AUC)
Model evaluated	Single	Ensemble	Best	Best model on PACS
Baseline, 64 models, no regularizer	84.48 ± 0.23	84.62	85.71	74.57
Penalty on sq. L2 norm of grad. (Jacobian reg. [28])	85.12 ± 0.33	**85.06**	85.84	74.10
Penalty on sq. L2 norm of ReLU of grad. (App. E)	84.62 ± 0.19	84.62	85.16	75.84
Penalty on sq. L2 norm of feature-grad. prod. (App. E)	84.61 ± 0.26	84.77	85.45	73.51
Penalty on L1 norm of grad. (App. E)	84.66 ± 0.45	84.67	86.13	76.29
Penalty on sq. L2 norm of logits (spectral dec. [57])	84.46 ± 0.32	84.81	85.16	74.60
Combination: proposed + spectral dec. [57]	84.31 ± 0.83	84.72	86.08	74.51
Proposed, 64 models	**85.14** ± 0.59	84.62	**86.80**	**79.66**

Domain generalization

Proposed method compared with existing methods on PACS.

PACS Dataset					
Test style (leave-one-out)	Art	Cartoon	Photo	Sketch	Avg.
D-SAM baseline [15]	77.9	75.9	95.2	69.3	79.6
D-SAM*	77.3	72.4	95.3	77.8	80.7
Epi-FCR baseline [37]	77.6	73.9	94.4	74.3	79.1
Epi-FCR*	82.1	77.0	93.9	73.0	81.5
DMG baseline [9]	72.6	78.5	93.2	65.2	77.4
DMG*	76.9	80.4	93.4	75.2	81.5
DecAug baseline [4]	78.4	78.3	94.2	72.1	80.8
DecAug*	79.0	79.6	95.3	75.6	82.4
JiGen baseline [8]	77.9	74.9	95.7	67.7	79.1
JiGen	79.4	75.3	96.0	71.4	80.5
Latent domains baseline [44]	78.3	75.0	96.2	65.2	78.7
Latent domains	81.3	77.2	96.1	72.3	81.8
Our baseline, 64 independent models					
Random single model	84.4 ±0.3	77.8 ±0.3	95.8 ±0.1	69.8 ±0.7	82.0
Ensemble of all models	84.6 ±0.1	77.9 ±0.1	96.0 ±0.1	69.6 ±0.2	82.0
Best single model	85.1 ±0.1	78.7 ±0.3	**96.2** ±0.1	71.7 ±0.5	82.9
Proposed, 64 models					
Random single model	85.2 ±0.6	79.6 ±0.7	95.9 ±0.1	70.8 ±0.8	82.9
Ensemble of all models	84.8 ±0.1	79.0 ±0.1	96.0 ±0.1	70.3 ±0.1	82.5
Best single model	**86.5** ±0.1	**81.1** ±0.4	**96.2** ±0.4	**72.8** ±0.2	**84.2**

Discussion

- Limitations of the method
  - The main hyperparameters(the regularizer strength and the number of models learned) setting give no guarantees.
- Model fitting and model selection are equally hard?
  - In this approach the two steps can be completely decoupled.
- Universality of inductive biases
  - The inductive biases of any learning algorithm cannot be universally superior to another's.
  - This method does not affect inductive biases in a directed way. It only increases the variety of the learned models, so it could be seen as a "meta-regularizer".
  - Experiments also show that intuitive notions behind classical regularizers like smoothness (Jacobian regularization), sparsity (L1 norm), or simplicity (L2 norm) are sometimes detrimental.

References I

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017).
Grad-cam: Visual explanations from deep networks via gradient-based localization.
In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Shah, H., Tamuly, K., Raghunathan, A., Jain, P., and Netrapalli, P. (2020).
The pitfalls of simplicity bias in neural networks.
Advances in Neural Information Processing Systems, 33:9573–9585.

Teney, D., Abbasnejad, E., Lucey, S., and Van den Hengel, A. (2022).
Evading the simplicity bias: Training a diverse set of models discovers solutions with superior ood generalization.
In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16761–16772.

Thank you!