Imperfect Information Learning Team

Masashi Sugiyama

不完全情報学習チーム　杉山 将

RIKEN Center for Advanced Intelligence Project

Team's Vision and Social Impact:

- Develop reliable and robust **machine learning** methods/algorithms that can cope with imperfect information such as **weak/noisy/zero supervision** and **adversarial attacks**.
- Enable machine learning in applications under imperfect, limited and adversarial data, such as health care, finance and natural disasters.

Members

- Masashi Sugiyama (Team Leader)
- Gang Niu (Research Scientist)
- Shuo Chen (Postdoctoral Researcher)
- Jingfeng Zhang (Postdoctoral Researcher)
- Jiaqi Lv (Postdoctoral Researcher)

Self-Supervised Learning

Data annotation in the real world is expensive to obtain. Self-supervised learning considers a scenario without human annotation.

- Many self-supervised contrastive learning methods have been proposed, but they cannot explicitly discriminate the inter-cluster and intra-cluster, failing to capture semantic information hidden in the data:
  - We proposed a new regularizer to constrain the distance region, obtaining a margin to discriminate the inter-cluster and intra-cluster (Chen et al, ICML 2021).
- Current self-attention modules use different tokens for different tasks, but they ignore the mutual information between two tokens so that the intrinsic correlations among different tasks cannot be utilized:
  - We designed new self-attention modules based on biological-evolution, making different tokens interactive (Tao et al., NeurIPS 2021).
- Existing self-supervised subspace clustering needs iteration, which is time-consuming:
  - We overcame this by building a new linearity-aware metric that directly predicts the point-to-point similarity without iteration in the test phase (Xu et al., AAAI 2022, oral).

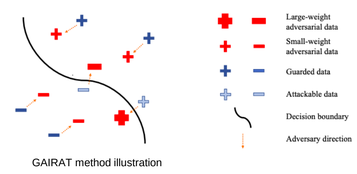
Adversarial Robustness

It was shown that classification results (by standard training) can be easily changed by adding small (adversarial) noise to test inputs.

- Many adversarial training methods have been proposed, but they are:
  - Too pessimistic due to equal treatment for all training data: We proposed weighing training data according to their hardness for classification (Zhang et al., ICLR 2021 oral; Wang et al., NeurIPS 2021).
  - Computationally too demanding in architecture search: We proposed effective atomic blocks for efficient search (Du et al., ICML 2021).
  - Less robust due to some irregular neural activations: We proposed a module to correct the irregularity for ConvNets (Yan et al., ICML 2021).
- Existing works showed that two-sample tests cannot detect adversarial data:
  - We overcame this based on semantic-awareness, allowing abstention from making predictions (Gao et al. ICML 2021).

Geometry-aware Instance-reweighted Adversarial Training

- Two facts: In adversarial training, 1) over-parametrized networks still encounter insufficient model capacity; 2) data points inherently have different degree of robustness.
- Idea: Given limited model capacity, data should have unequal importance for learning the robust decision boundary.
- Methodology: For updating the model, we explicitly give larger weights on the losses of adversarial data, whose natural counterparts are closer to the decision boundary and vice versa.
- Impact: shed new light on instance-dependent views on adversarial robustness.

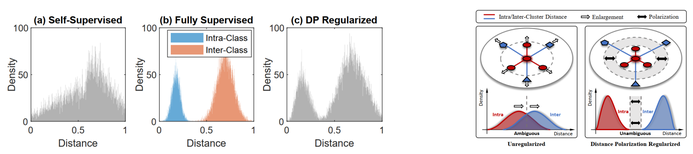


GAIRAT method illustration

Zhang, Zhu, Niu, Han, Sugiyama & Kankanhalli (ICLR 2021, oral)

Large-Margin Contrastive Learning

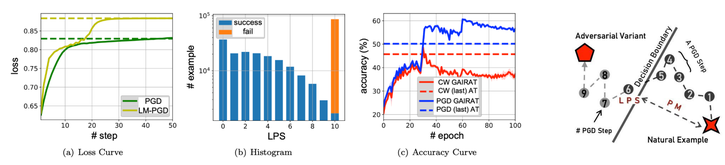
- Contrastive Learning: Learning representation by pushing away each pair of instances in the training data.
- Motivation: Conventional contrastive learning cannot produce a margin between inter-cluster and intra-cluster for discrimination.
- Methodology: We propose large-margin contrastive learning (LMCL) with distance polarization regularizer, motivated by the distribution characteristic of metric learning. In LMCL, we can only push away inter-cluster pairs.



Chen, Niu, Gong, Li, Yang & Sugiyama (ICML 2021)

Probabilistic Margins for Instance Reweighting in Adversarial learning

- Motivation: In GAIRAT (above), the closeness measure of data points to decision boundary is not reliable, i.e., discrete and path-dependent.
- Idea: We propose probabilistic margin (PM), which are continuous and path-independent, for measuring the closeness measure and reweighting adversarial data.



(a) Loss Curve (b) Histogram (c) Accuracy Curve

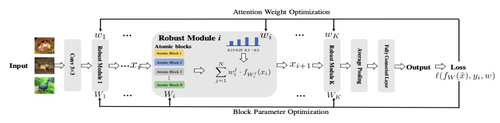
Wang, Liu, Han, Liu, Gong, Niu, Zhou & Sugiyama (NeurIPS 2021)

Learning Diverse-structure Networks

- Motivation: The optimal network structures in standard training would be no longer optimal in adversarial training (AT).

Model	Standard Acc.	Ranking	Robustness	Ranking
WRN-28-10	0.9646	1	0.4872	3
ResNet-62	0.9596	2	0.4855	4
DenseNet-121	0.9504	3	0.4993	2
MobileNetV2	0.9443	4	0.4732	6
AdaRKNet-62	0.9403	5	0.5016	1
ResNet-50	0.9362	6	0.4807	5

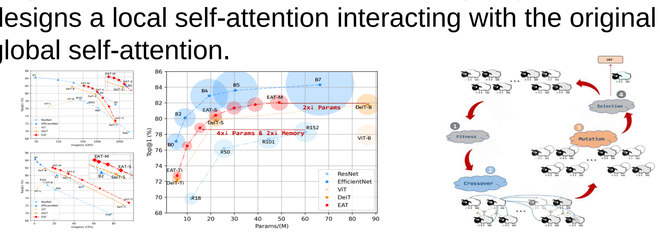
- Challenge: AT (time-consuming) * network architecture search (time-consuming) = computational explosion
- Solution: Search only predefined atomic blocks, where an atomic block is a time-tested building block such as the residual block, etc.



Du, Zhang, Han, Liu, Rong, Niu, Huang & Sugiyama (ICML 2021)

Evolutionary Algorithm based Transformer

- Vision Transformer: Learning representation by building upon the feed forward network with self-attention blocks.
- Motivation: Current models initialize different tokens for different tasks, and they participate in every level of calculation that is incompatible with other tokens for internal operations.
- Methodology : Inspired by the biological evolution, we propose a novel EA based transformer (EAT) that intuitively designs a local self-attention interacting with the original global self-attention.

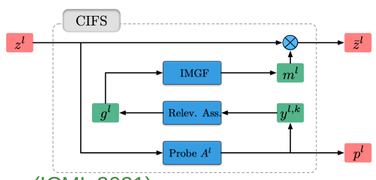


Tao, Ji, Wang, Chen, Lin, Cao, Lu, Luo & Tai (NeurIPS 2021)

Channel-wise Importance-based Feature Selections

- Observations: Adversarial data as input to convolutional neural networks (CNN) has some irregular activations compared with their natural counterparts.
- Methodology: Activation alignments.

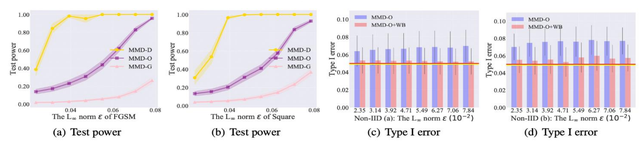
1) Probe Network A^l first makes a raw prediction p^l for z^l.
2) Channels' relevances g^l are assessed based on the gradients of the top-k prediction results $y^{l,k}$.
3) The IMGF module generates an importance mask m^l from g^l for channel adjustment.



Yan, Zhang, Niu, Feng, Tan & Sugiyama (ICML 2021)

Maximum Mean Discrepancy is Aware of Adversarial Attacks

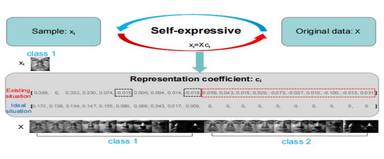
- An interesting question: Are natural data and adversarial data from the same distribution?
- Challenge: Existing statistics tests cannot distinguish adversarial data from natural data.
- Our contributions: We propose *semantic-aware maximum mean discrepancy* (SAMMD) that can indeed detect adversarial data.
  - We uncover the previous use of the MMD test on the purpose missed three key factors: Gaussian kernel has limited representation power; adversarial data are non-independent.

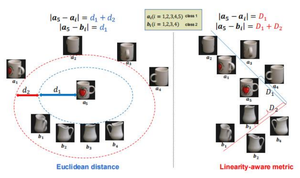


Gao, Liu, Zhang, Han, Liu, Niu & Sugiyama (ICML 2021)

Linearity-Aware Self-Expression

- Subspace Clustering: Learning and clustering from unlabeled data that lies in a low-dimensional subspace.
- Motivation: Existing methods adopt a small part of instances as basis to represent others. It is difficult to describe the linear relation between pairwise instances.
- Methodology: We propose a new linearity-aware metric to directly measure the degree of linear correlation between pairwise instances, and thus providing reliable point-to-point similarities.





Xu, Chen, Li & Qian (AAAI 2022, oral)

Selected Publications in FY2021

- J. Zhang, J. Zhu, G. Niu, B. Han, M. Sugiyama, and M. Kankanhalli, Geometry-aware Instance-reweighted Adversarial Training. ICLR 2021, oral.
- Z. Xie, I. Sato, and M. Sugiyama, A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. ICLR 2021.
- S. Chen, G. Niu, C. Gong, J. Li, J. Yang, and M. Sugiyama, Large-Margin Contrastive Learning with Distance Polarization Regularizer. ICML 2021.
- H. Yan, J. Zhang, G. Niu, J. Feng, V. Y. F. Tan, and M. Sugiyama, CIFS: Improving Adversarial Robustness of CNNs via Channel-wise Importance-based Feature Selection. ICML 2021.
- R. Gao, F. Liu, J. Zhang, B. Han, T. Liu, G. Niu, and M. Sugiyama, Maximum Mean Discrepancy is Aware of Adversarial Attacks. ICML 2021.
- X. Du, J. Zhang, B. Han, T. Liu, Y. Rong, G. Niu, J. Huang and M. Sugiyama, Learning Diverse-Structured Networks for Adversarial Robustness. ICML 2021.
- X. Li, T. Liu, B. Han, G. Niu, and M. Sugiyama. Provably end-to-end label-noise learning without anchor points. ICML 2021.
- A. Berthon, B. Han, G. Niu, T. Liu, and M. Sugiyama. Confidence scores make instance-dependent label-noise learning possible. ICML 2021.
- Y. Zhang, G. Niu, and M. Sugiyama. Learning noise transition matrix from only noisy labels via total variation regularization. ICML 2021.
- L. Feng, S. Shu, N. Lu, B. Han, M. Xu, G. Niu, B. An, and M. Sugiyama. Pointwise binary classification with pairwise confidence comparisons. ICML 2021.
- N. Lu, S. Lei, G. Niu, I. Sato, and M. Sugiyama.Binary classification from multiple unlabeled datasets via surrogate set classification. ICML 2021.
- Y. Cao, L. Feng, Y. Xu, B. An, G. Niu, and M. Sugiyama. Learning from similarity-confidence data. ICML 2021.
- S. Wu, X. Xia, T. Liu, B. Han, M. Gong, N. Wang, H. Liu, and G. Niu. Class2Simi: A noise reduction perspective on learning with noisy labels. ICML 2021.
- N. Charoenphakdee, Z. Cui, Y. Zhang, and M. Sugiyama, Classification with rejection based on cost-sensitive classification. ICML 2021.
- Z. Xie, L. Yuan, Z. Zhu, and M. Sugiyama, Positive-negative momentum: Manipulating stochastic gradient noise to improve generalization. ICML 2021.
- I. Yamane, J. Honda, F. Yger, and M. Sugiyama, Mediated uncoupled learning: Learning functions without direct input-output correspondences. ICML 2021.
- S. M. Yoshida, T. Takenouchi, and M. Sugiyama, Lower-bounded proper losses for weakly supervised classification. ICML 2021.
- J. Zhang, C. Xu, J. Li, W. Chen, Y. Wang, Y. Tai, S. Chen, C. Wang, F. Huang, and Y. Liu, Analogous to Evolutionary Algorithm: Designing a Unified Sequence Model. NeurIPS 2021.
- F. Futami, T. Iwata, N. Ueda, I. Sato, and M. Sugiyama, Loss function based second-order Jensen inequality and its application to particle variational inference. NeurIPS 2021.
- G. Tao, X. Ji, W. Wang, S. Chen, C. Lin, Y. Cao, T. Lu, D. Luo, Y. Tai, Spectrum-to-Kernel Translation for Accurate Blind Image Super-Resolution. NeurIPS 2021.
- Q. Wang, F. Liu, B. Han, T. Liu, C. Gong, G. Niu, M. Zhou, and M. Sugiyama, Probabilistic Margins for Instance Reweighting in Adversarial Training. NeurIPS 2021.
- Y. Xu, S. Chen, J. Li, and J. Qian, Linearity-Aware Subspace Clustering. AAAI 2022 oral.