FY2022/2022年度

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 robust image classification, single image denoising, and unsupervised text recognition.

Members

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

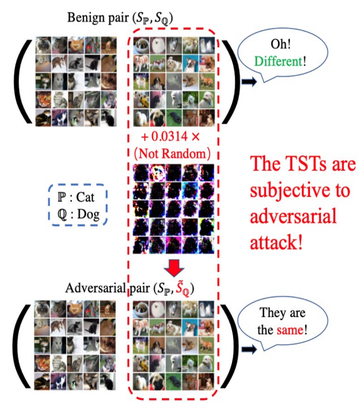
Adversarial Robustness

It was shown that classification results (by standard training) can be easily changed by adding small (adversarial) noise to test inputs. We explored the adversarial robustness of existing machine learning approaches from the following two aspects:

- Adversarial attacks added tiny perturbations on natural data that can effectively mislead artificial intelligence (AI) to give wrong predictions.
  - We focused on a specific task called two sample test and uncovered its failure mode of through adversarial attacks. Then we proposed corresponding defense strategies (Xu et al., ICML 2022).
- Adversarial training enhanced AI robustness against adversarial attacks. However, the practical performances of some specific learning tasks were still not satisfactory.
  - Existing deep image denoisers (DID) were vulnerable to the adversarial noise, so we proposed a new learning framework, jointly training DIDs with adversarial and non-adversarial noisy data (Yan et al., IJCAI 2022).
  - We investigated noisy labels (NLs) injection into AT's inner maximization & outer minimization. Then we proposed a new method NoiLIn that randomly injects NLs into training data at each training epoch and dynamically increases the NL injection rate (Zhang et al., TMLR 2022).

Adversarial Attack and Defense For Non-Parametric Two-Sample Tests

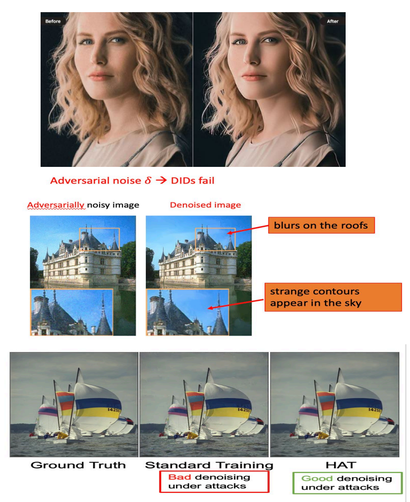
- **Two-Sample Test (TST):** A statistical hypothesis testing, judging whether the two sets of samples are drawn from the same distributions
- **Motivation:** Powered by recent AI methods, Non-parametric TST greatly enhance its testing abilities on differentiate complex distributions, and an interesting question is that where is the vulnerability of non-parametric TST?
- **Methodology:** We conduct adversarial attack to uncover the failure mode of non-parametric TST. Then we implement the adversarial training, so that can obtain a robust non-parametric TST that has excellent testing ability and withstands adversarial attack



Xu, Zhang, Liu, Sugiyama & Kankanhalli (ICML 2022)

Towards Adversarially Robust Image Denoising

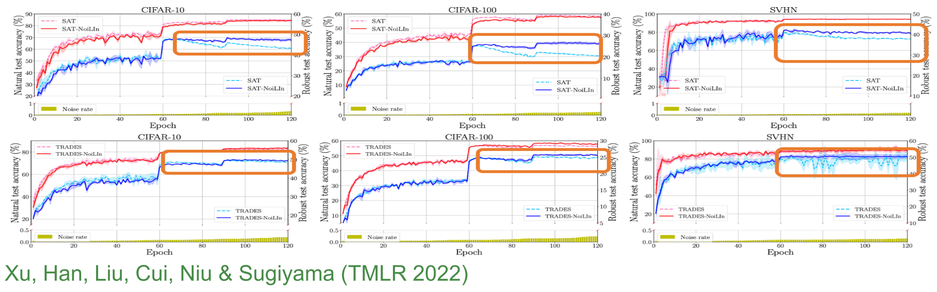
- **Image Denoising**: Reconstructing clean images from their noisy counterparts
- **Motivation:** AI methods powered by deep neural networks have achieved excellent performance, yet an interesting question is that where is the vulnerability of DID?
- **Methodology**: We design a zero-mean adversarial attack to uncover the failure mode of DID: hybrid adversarial training (HAT), and we can obtain a robust DID that performs well on both benign and adversarial noise



Yan, Zhang, Feng, Sugiyama & Tan (IJCAI 2022)

NoiLIn: Improving Adversarial Training and Correcting Noisy Labels

- **Adversarial Training (AT):** AT is a basic learning paradigm to enhance model's adversarial robustness against attacks
- **Motivation:** AT has the issues of robust overfitting, i.e., over the training process the robust accuracy first climbs and then drops
- **Methodology:** We propose NoiLIn, which is plug-in any AT methods that fix AT's drawback



Zhang, Xu, Han, Liu, Cui, Niu & Sugiyama (TMLR 2022)

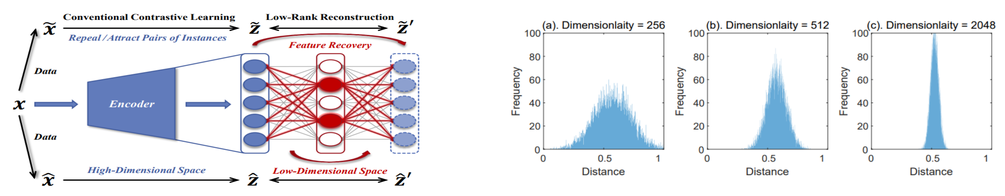
Self-Supervised Learning

Data annotation in the real world is expensive to obtain. Self-supervised contrastive learning considers a scenario without human annotation. We have obtained the following three achievements that focus on the methodology research and application tasks of contrastive learning.

- Existing contrastive learning (CL) methods used high-dimensional embeddings to push instances as far as possible, yet such high-dimensional features may fail to capture underlying pairwise similarities:
  - We proposed a new CL approach by low-dimensional reconstruction to capture similarities in low-dimensional space, meanwhile maintaining the instance discrimination (Chen et al., NeurIPS 2022).
- Existing frameworks of generalized zero-shot (GZSL) learning synthesized the missing visual features of unseen classes, yet it lacked semantic information for unseen classes:
  - We designed a new contrastive learning based GZSL framework to consider the pairwise similarity, and capturing the critical semantic information hidden within data (Han et al., IJCV 2022).
- The SOTA cross-modal recognition built the contrastive set by random sampling between audio and visual instances, which may result in faulty negatives:
  - We proposed a new contrastive set mining technique that explored the data pairs with informative and diverse negatives for more robust cross-modal recognition (Xuan et al., IJCAI 2022).

Learning Contrastive Embedding in Low-Dimensional Space

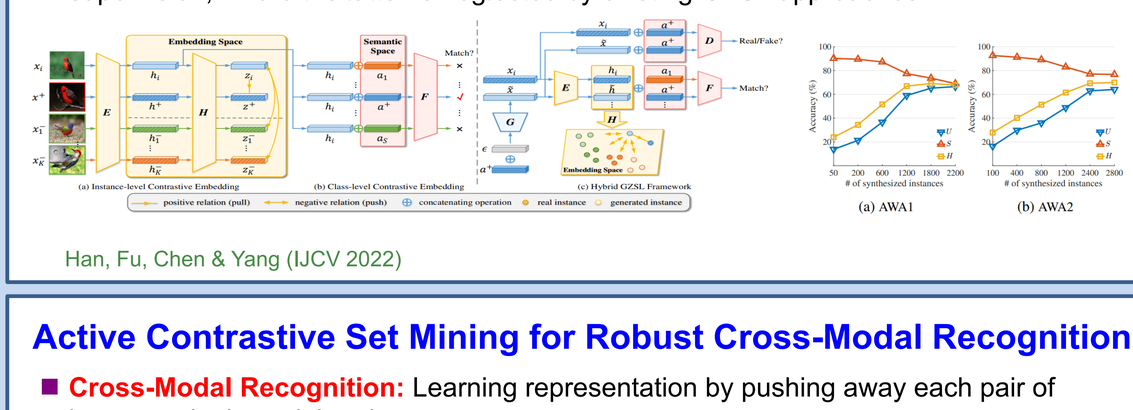
- **Contrastive Learning:** Learning representation by pushing away each pair of instances in the training data
- **Motivation:** Conventional approaches encourage using high-dimensional features to scatter instances. Yet it may lead to that data points sparsely distributed in the feature space, making it hardly captures the underlying similarity
- **Methodology:** We propose a new reconstruction layer to recover low-dimensional features, and meanwhile we maintain the original objective of instance discrimination



Chen, Gong, Li, Yang, Niu & Sugiyama (NeurIPS 2022)

Semantic Contrastive Embedding for Zero-Shot Recognition

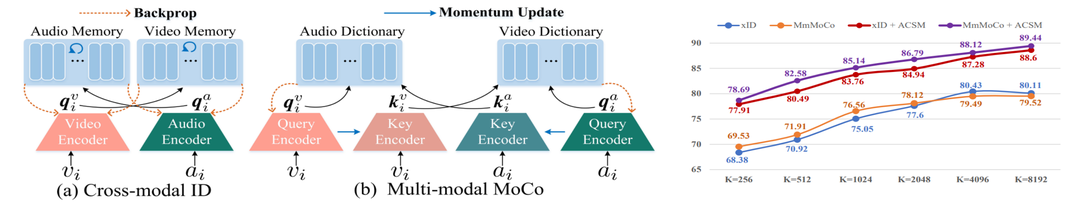
- **(Generalized) Zero-Shot Recognition:** Classifying both the seen and unseen class instances without retraining the model
- **Motivation:** Conventional GZSL does not consider the semantic relationship between seen class and unseen class data
- **Methodology:** We proposed a contrastive embedding (CE) paradigm for the GZSL task. CE can leverage not only the class-wise supervision but also the instance-wise supervision, where the latter is neglected by existing GZSL approaches



Han, Fu, Chen & Yang (IJCV 2022)

Active Contrastive Set Mining for Robust Cross-Modal Recognition

- **Cross-Modal Recognition:** Learning representation by pushing away each pair of instances in the training data
- **Motivation:** Existing methods construct the contrastive set by random sampling based on the assumption that the audio and visual clips from videos are not semantically related
- **Methodology:** We argue that this assumption is rough, as the resulting contrastive sets have a large number of faulty negatives. We overcome this limitation by actively exploring the informative and diverse negatives for building an active contrastive set



Xuan, Xu, Chen, Wu, Yang, Yan, Alameda-Pineda (IJCAI 2022)

Selected Publications in FY2022

- S. Cui, J. Zhang, J. Liang, B. Han, M. Sugiyama, C. Zhang. Synergy-of-Experts: Collaborate to Improve Adversarial Robustness, NeurIPS 2022.
- J. Zhou, J. Zhou, J. Zhang, T. Liu, G. Niu, B. Han, M. Sugiyama. Adversarial Training with Complementary Labels: On the Benefit of Gradually Informative Attacks, NeurIPS 2022.
- Y. Cao, T. Cai, L. Feng, L. Gu, J. Gu, B. An, G. Niu, and M. Sugiyama. Generalizing consistent multi-class classification with rejection to be compatible with arbitrary losses, NeurIPS 2022.
- S. Chen, C. Gong, J. Li, J. Yang, G. Niu, M. Sugiyama. Learning Contrastive Embedding in Low-Dimensional Space, NeurIPS 2022.
- J. Wei, H. Liu, T. Liu, G. Niu, M. Sugiyama, and Y. Liu. To smooth or not? When label smoothing meets noisy labels, ICML 2022.
- R. Gao, J. Wang, K. Zhou, F. Liu, B. Xie, G. Niu, B. Han, and J. Cheng. Fast and reliable evaluation of adversarial robustness with minimum-margin attack, ICML 2022.
- X. Xu, J. Zhang, F. Liu, M. Sugiyama, and M. Kankanhalli. Adversarial Attacks and Defense For Non-parametric Two Sample Tests, ICML 2022.
- J. Zhang, X. Xu, B. Han, T. Liu, N. Gang, L. Cui, M. Sugiyama. NoiLin: Improving Adversarial Training and Correcting Stereotype of Noisy Labels, TMLR 2022.
- X. Peng, F. Liu, J. Zhang, L. Lan, J. Ye, T. Liu, B. Han. Bilateral Dependency Optimization: Defending Against Model-inversion Attacks, KDD 2022.
- J. Zhu, J. Yao, B. Han, J. Zhang, T. Liu, G. Niu, J. Zhou, J. Xu, H. Yang. Reliable Adversarial Distillation with Unreliable Teachers, ICLR 2022.
- H. Chi*, F. Liu*, W. Yang, L. Lan, T. Liu, B. Han, G. Niu, M. Zhou, and M. Sugiyama. Meta discovery: Learning to discover novel classes given very limited data, ICLR 2022 (spotlight).
- H. Wang, R. Xiao, Y. Li, L. Feng, G. Niu, G. Chen, and J. Zhao. PiCO: Contrastive label disambiguation for partial label learning, ICLR 2022 (Outstanding Paper Honorable Mention).
- Z. Han, Z. Fu, S. Chen, J. Yang. Semantic Contrastive Embedding for Generalized Zero-Shot Learning, IJCV 2022.
- H. Xuan, Y. Xu, S. Chen, Z. Wu, J. Yang, Y. Yan, Xavier Alameda-Pineda. Active Contrastive Set Mining for Robust Audio-Visual Instance Discrimination, IJCAI 2022.
- H. Yan, J. Zhang, J. Feng, M. Sugiyama, and V. Y. F. Tan. Towards Adversarially Robust Image Denoising, IJCAI 2022.
- Y. Yang, F. Guo, S. Chen, J. Li, J. Yang. Industrial Style Transfer with Large-scale Geometric Warping and Content Preservation, CVPR 2022.