Dr. (Amos) Xinshao Wang: Data + Modelling

Deep distance metric learning;
Robust deep learning;
Diverse applications (CV, NLP, Health Care, Omics, etc)

Homepage: https://xinshaoamoswang.github.io/about/
Blogs: https://xinshaoamoswang.github.io/blogs/
LinkedIn: https://www.linkedin.com/in/xinshaowang/
Github: https://github.com/XinshaoAmosWang
Google Scholar: yOBhB7UAAAAJ

Email: xinshaowang@gmail.com
Phone: +44 (0) 7712 114316

2022-12-23, Shanghai DianJi University

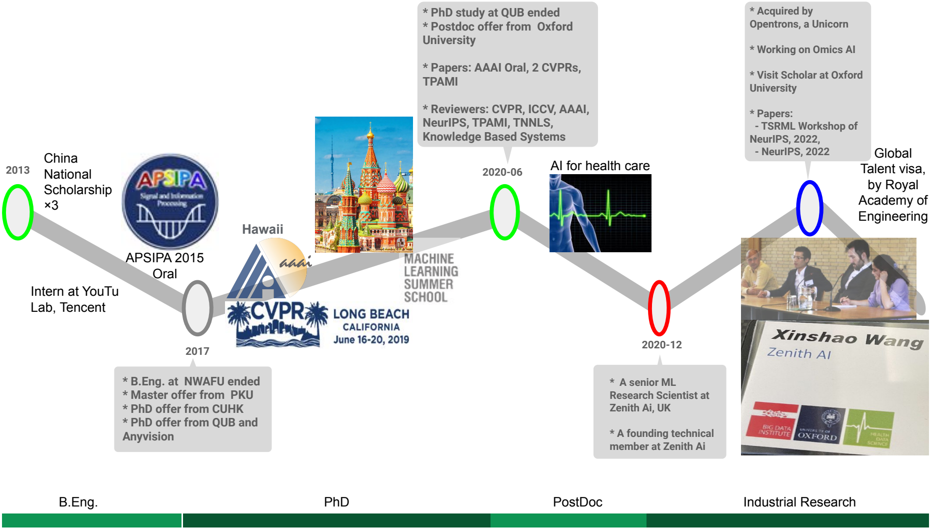
Outline

1. Self introduction
2. Highlights of work and experience
3. Technical novelty and principle: Example weighting is universal
4. Deep distance metric learning
   - Definition & Overview
   - Projects: Vision applications
5. Robust deep learning (scope: transparent and interpretable ML)
   - Understanding real-world data and why do we need it
   - Learning objectives
   - Project: ProSelfLC: Progressive Self Label Correction
6. Omics AI
   - Robust protein understanding
   - Deep distance metric: EDA + Active Learning

Outline

1. Self introduction
2. Highlights of work and experience
3. Technical novelty and principle: Example weighting is universal
4. Deep distance metric learning
   - Definition & Overview
   - Projects: Vision applications
5. Robust deep learning (scope: transparent and interpretable ML)
   - Understanding real-world data and why do we need it
   - Learning objectives
   - Project: ProSelfLC: Progressive Self Label Correction
6. Omics AI
   - Robust protein understanding
   - Deep distance metric: EDA + Active Learning

Self introduction



2013

China National Scholarship ×3

Intern at YouTu Lab, Tencent

APSIPA 2015 Oral

2017

* B.Eng. at NWAFU ended
* Master offer from PKU
* PhD offer from CUHK
* PhD offer from QUB and Anyvision

Hawaii

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

MACHINE LEARNING SUMMER SCHOOL

2020-06

* PhD study at QUB ended
* Postdoc offer from Oxford University

* Papers: AAAI Oral, 2 CVPRs, TPAMI

* Reviewers: CVPR, ICCV, AAAI, NeurIPS, TPAMI, TNNLS, Knowledge Based Systems

AI for health care

2020-12

* A senior ML Research Scientist at Zenith Ai, UK

* A founding technical member at Zenith Ai

* Acquired by Opentrons, a Unicorn

* Working on Omics AI

* Visit Scholar at Oxford University

* Papers:
  - TSRML Workshop of NeurIPS, 2022,
  - NeurIPS, 2022

Global Talent visa, by Royal Academy of Engineering

B.Eng. | PhD | PostDoc | Industrial Research

Outline

1. Self introduction
2. Highlights of work and experience
3. Technical novelty and principle: Example weighting is universal
4. Deep distance metric learning
   - Definition & Overview
   - Projects: Vision applications
5. Robust deep learning (scope: transparent and interpretable ML)
   - Understanding real-world data and why do we need it
   - Learning objectives
   - Project: ProSelfLC: Progressive Self Label Correction
6. Omics AI
   - Robust protein understanding
   - Deep distance metric: EDA + Active Learning

Work and experience ($\approx$ 90% leading)

1. Deep distance metric learning
    - CVPR 2019 and TPAMI 2021
    - AAAI 2019 Oral
2. Robust deep learning, model calibration and uncertainties:
    - **A PhD student as the 1st author**: Trustworthy and Socially Responsible ML, NeurIPS 2022
    - CVPR 2021
    - **Preprint (56 citations)**: "IMAE for Noise-Robust Learning: Mean Absolute Error Does Not Treat Examples Equally and Gradient Magnitude's Variance Matters"
3. (Oxford) Postdoc, Visit Scholar on AI health care (e.g., ECG)
4. (Zenith Ai) Sr. Researcher on Omics AI (e.g., DNA, tRNA, protein, amino acid, ribosome, **molecule docking** / **AlphaFold2** / **biochemistry**)

Outline

1. Self introduction
2. Highlights of work and experience
3. Technical novelty and principle: Example weighting is universal
4. Deep distance metric learning
   - Definition & Overview
   - Projects: Vision applications
5. Robust deep learning (scope: transparent and interpretable ML)
   - Understanding real-world data and why do we need it
   - Learning objectives
   - Project: ProSelfLC: Progressive Self Label Correction
6. Omics AI
   - Robust protein understanding
   - Deep distance metric: EDA + Active Learning

We define our interpretation of example weighting [5]:

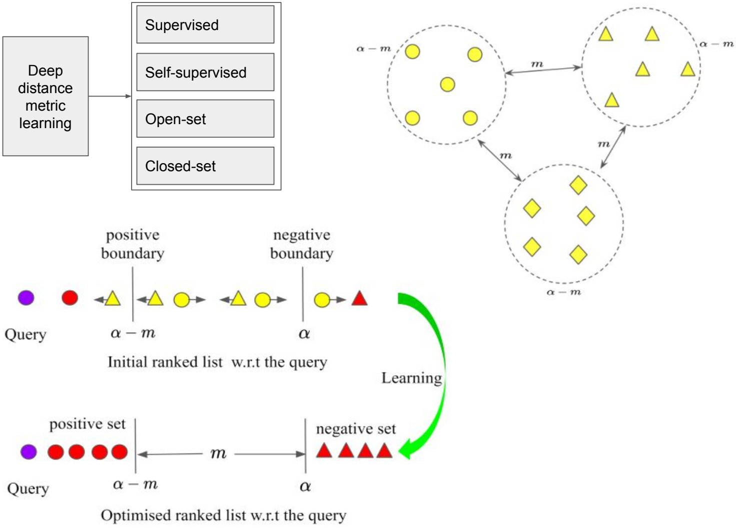
Definition (Example Weighting). *In gradient-based optimisation, the derivative of an example can be interpreted as its effect on the update of a model. Therefore, a derivative's magnitude function equals to a weighting scheme.*

Accordingly, a change of the derivative magnitude function, is implicitly equivalent to, modifying an example weighting scheme.

Outline

1. Self introduction
2. Highlights of work and experience
3. Technical novelty and principle: Example weighting is universal
4. Deep distance metric learning
   - Definition & Overview
   - Projects: Vision applications
5. Robust deep learning (scope: transparent and interpretable ML)
   - Understanding real-world data and why do we need it
   - Learning objectives
   - Project: ProSelfLC: Progressive Self Label Correction
6. Omics AI
   - Robust protein understanding
   - Deep distance metric: EDA + Active Learning

Deep distance metric: Overview

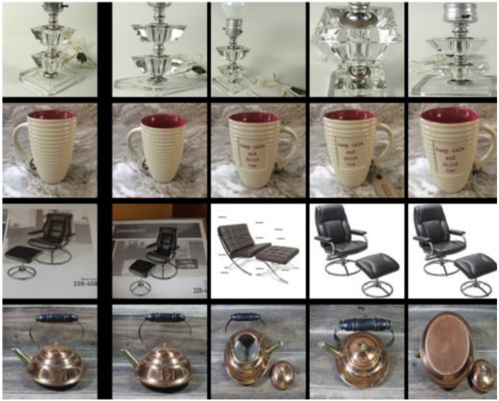


Deep distance metric learning

- Supervised
- Self-supervised
- Open-set
- Closed-set

Deep distance metric: Vision applications

1. **General-purpose image/video** clustering / retrieval
2. **Body/Face image/video** re-identification (i.e., retrieval)



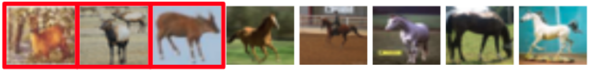
Query The top 4 images in the ranked list of each query

Outline

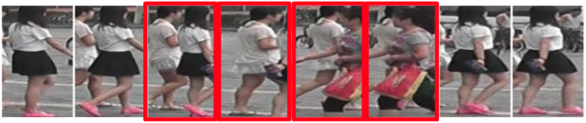
1. Self introduction
2. Highlights of work and experience
3. Technical novelty and principle: Example weighting is universal
4. Deep distance metric learning
   - Definition & Overview
   - Projects: Vision applications
5. Robust deep learning (scope: transparent and interpretable ML)
   - Understanding real-world data and why do we need it
   - Learning objectives
   - Project: ProSelfLC: Progressive Self Label Correction
6. Omics AI
   - Robust protein understanding
   - Deep distance metric: EDA + Active Learning

Why do we need robust deep learning?

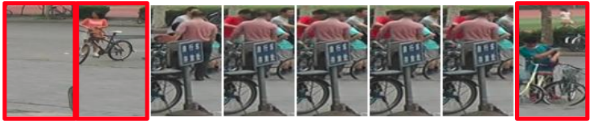
Understanding the real-world data with adverse cases



Horse class: The first three images are deer semantically.



This video is labelled as the person wearing black skirt.



This video is labelled as the person wearing green shirt.

Adverse cases in real-world data

Out-of-distribution anomalies: Know the unknown

1. The inputs contain only background: no semantic information.
2. The labels do not belong to any class in the training set.

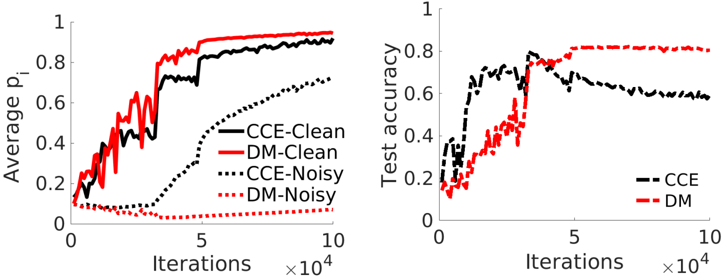
In-distribution anomalies: Detect => Ignore or Correct

1. Single-label noise: also common in annotating molecules
   - Noisy annotations.
   - Missing annotations.
2. Multi-label noise: to be solved by multi-label training. E.g., one molecule may have multiple biological functions.

Learning objectives of robust deep learning

What is the meaning of robustness here?

1. **To learn meaningful patterns** on semantically clean data.
2. **Without fitting errors/bias**.

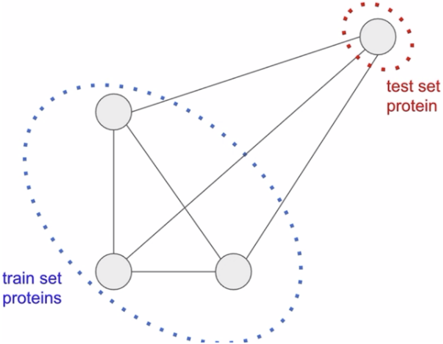


$p_i = p(y_i|\mathrm{x}_i)$: probability of predicting x_i to its oracle y_i.

3. **Generalisation** to **unseen** data.

Generalise to the unseen (e.g., remote homology)

Build train-validation datasets properly

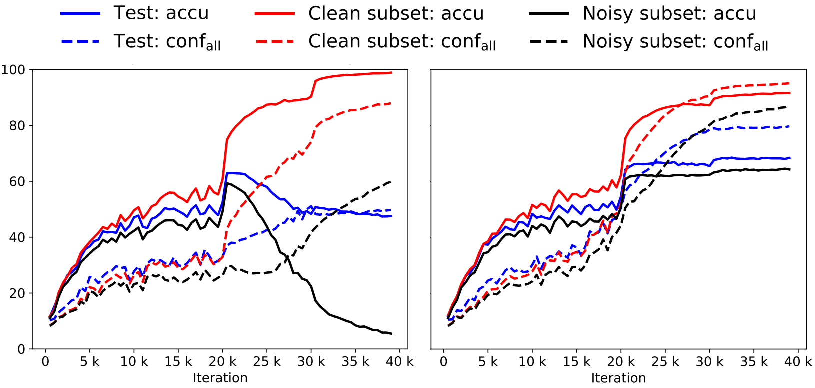


[1] "Using deep learning to annotate the protein universe." Nature Biotechnology (2022).

Project: Progressive Self Label Correction

To uncover how a deep model learns under noise [6]

1. To reward a low-entropy status other than penalise.
2. Model calibration.



(a) Standard CCE.

(b) ProSelfLC with an AT (ours).

Figure: accu and $\mathrm{conf}_{\mathrm{all}}$ along with the iteration when training ResNet18 on CIFAR-100. The symmetric noise rate is $r = 40\%$.

Outline

1. Self introduction

2. Highlights of work and experience

3. Technical novelty and principle: Example weighting is universal

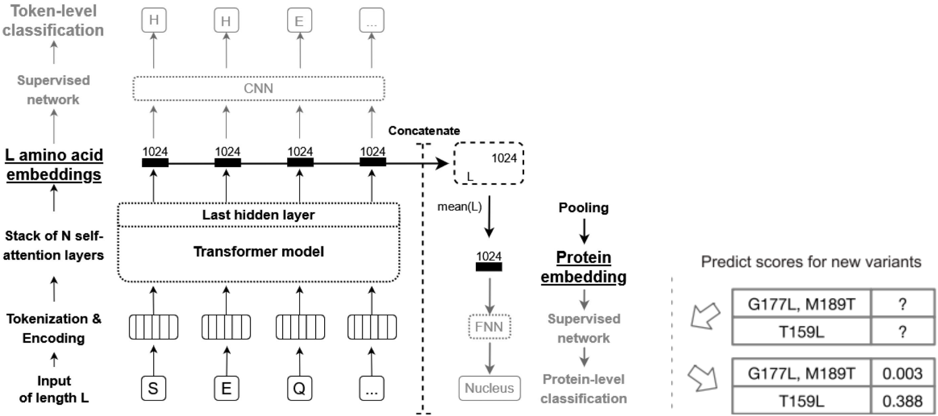
4. Deep distance metric learning
   - Definition & Overview
   - Projects: Vision applications

5. Robust deep learning (scope: transparent and interpretable ML)
   - Understanding real-world data and why do we need it
   - Learning objectives
   - Project: ProSelfLC: Progressive Self Label Correction

6. Omics AI
   - Robust protein understanding
   - Deep distance metric: EDA + Active Learning

Robust protein understanding

Use the pre-trained transformers to predict or design



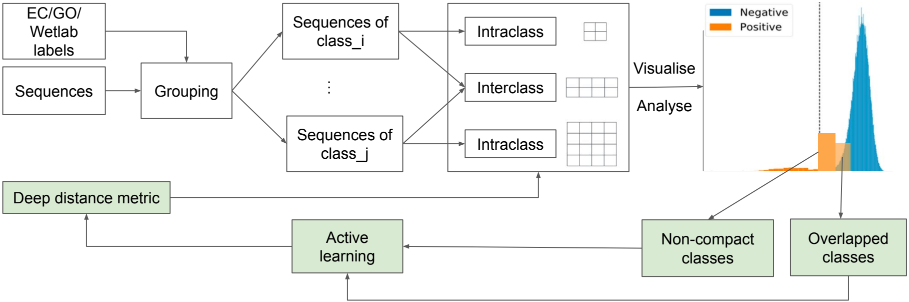
G177L, M189T	?
T159L	?

G177L, M189T	0.003
T159L	0.388

[3] "ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning." IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).
[4] "Neural networks to learn protein sequence–function relationships from deep mutational scanning data." Proceedings of the National Academy of Sciences (2021).
[6] "ProSelfLC: Progressive Self Label Correction Towards A Low-Temperature Entropy State." Ours under peer review.

Deep distance metric for Omics AI

1. Exploratory data analysis (EDA): data variances.
   - Intraclass: remote homology/evolutionary information
   - Interclass: how to discriminate biological molecules/sequences.
2. Iterative active learning for the efficiency of time, data and labels



Thanks for your attention.
Questions and discussions are very welcome.

Research topics of Dr. (Amos) Xinshao Wang:

- Deep distance metric learning
- Robust deep learning
- Omics AI + Bioinformatics
- Active learning
- EDA + Visualisation

Homepage: https://xinshaoamoswang.github.io/about/
Blogs: https://xinshaoamoswang.github.io/blogs/
LinkedIn: https://www.linkedin.com/in/xinshaowang/
Github: https://github.com/XinshaoAmosWang
Google Scholar: yOBhB7UAAAAJ

References

[1] Bileschi, M. L., Belanger, D., Bryant, D. H., Sanderson, T., Carter, B., Sculley, D., Bateman, A., DePristo, M. A., and Colwell, L. J. Using deep learning to annotate the protein universe. *Nature Biotechnology*, pp. 1–6, 2022.

[2] Cervettini, D., Tang, S., Fried, S. D., Willis, J. C., Funke, L. F., Colwell, L. J., and Chin, J. W. Rapid discovery and evolution of orthogonal aminoacyl-trna synthetase–trna pairs. *Nature biotechnology*, pp. 989–999, 2020.

[3] Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Yu, W., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 7112–7112.

[4] Gelman, S., Fahlberg, S. A., Heinzelman, P., Romero, P. A., and Gitter, A. Neural networks to learn protein sequence–function relationships from deep mutational scanning data. *Proceedings of the National Academy of Sciences*, 2021.

[5] Wang, X., Kodirov, E., Hua, Y., and Robertson, N. M. Derivative manipulation for general example weighting. *arXiv preprint arXiv:1905.11233*, 2019.

[6] Wang, X., Hua, Y., Kodirov, E., Mukherjee, S. S., Clifton, D. A., and Robertson, N. M. Proselflc: Progressive self label correction towards a low-temperature entropy state. *arXiv preprint arXiv:2207.00118*, 2022.