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

2022-12-14, Novo Nordisk Research Centre Oxford

Outline

Outline

1. Work and experience: a quick glance, more in the CV
2. Technical novelty and principle: Example weighting is universal
3. Deep distance metric learning
   - Deep distance metric: Definition & Overview
   - Deep distance metric: Vision applications
4. Robust deep learning (scope: transparent and interpretable ML)
   - Understanding real-world data and why do we need it
   - Learning objectives
   - Project: Robust video-based person ReID
   - Project: To uncover how a deep model learns under noise
5. Omics AI
   - Bioinformatics: Sequence alignment and distance
   - Robust protein understanding
   - Deep distance metric: EDA + Active Learning
6. Industrial R&D experience
   - Industrial R&D overview
   - Industrial AI Research = model-centric AI + data-centric AI

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

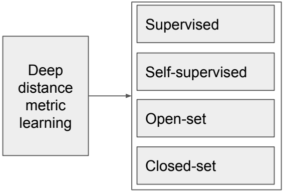
We define our interpretation of example weighting [5]:

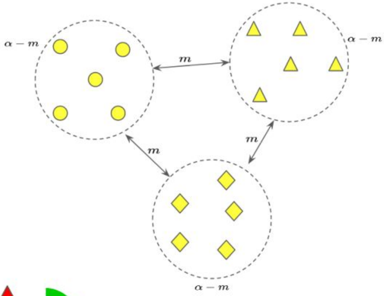
Definition (Example Weighting). *In gradient-based optimisation, the derivative of an example can be interpreted as its effect on the update of a model. Therefore, a derivative's magnitude function equals to a weighting scheme.*

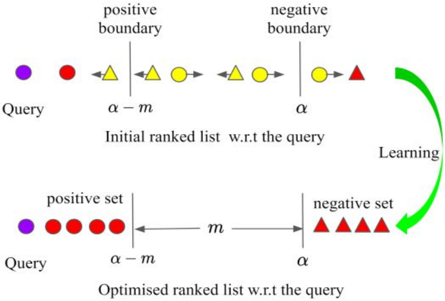
Accordingly, a change of the derivative magnitude function, is implicitly equivalent to, modifying an example weighting scheme.

Outline

Deep distance metric: Overview

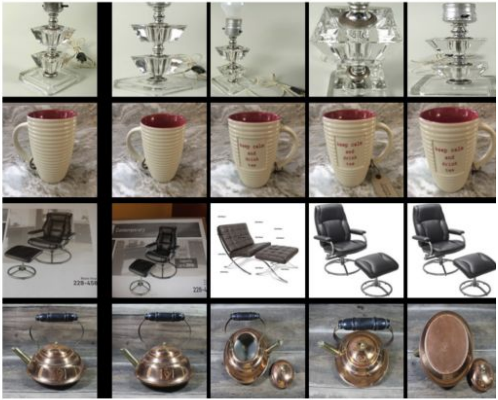






Deep distance metric: Vision applications

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

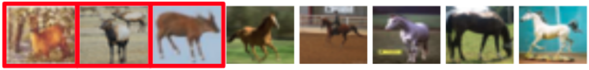


Query The top 4 images in the ranked list of each query

Outline

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.

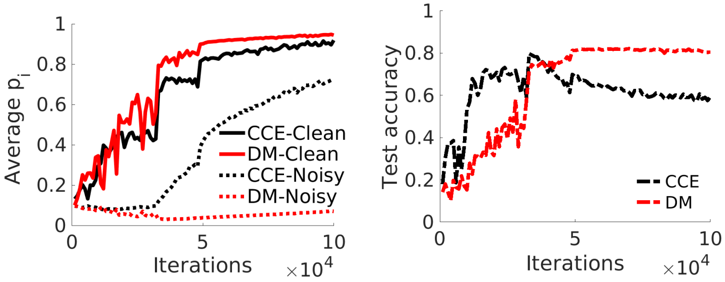
In-distribution anomalies: Detect => Ignore or Correct

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**.

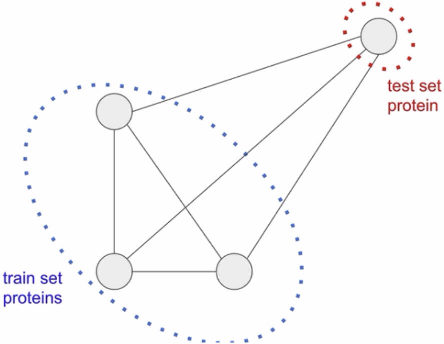


$p_i = p(y_i|\mathrm{x}_i)$: probability of predicting x_i to its oracle y_i.

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).

Robust deep learning projects

Video-based person re-identification (ReID)

(a) with/without blur

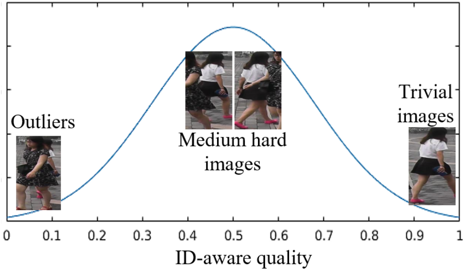
(b) complete/incomplete body

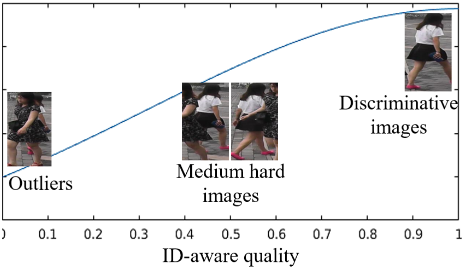
(c) with/without occlusion

(d) single/multiple people in one image

If the cost is expensive to improve the data quality, e.g., **improving super-resolution or detection** in set/video-based person ReID,

Example weighting and robust deep learning can help!





Robust deep learning projects

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

DNN has strong fitting capability, *but we find*:

Deep models easily fit random noise.

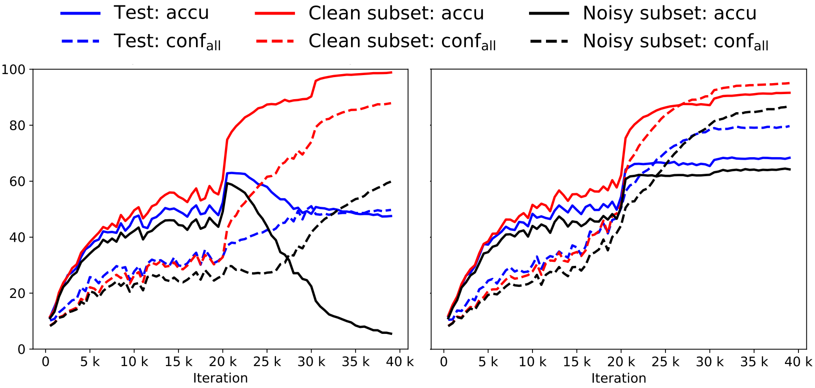
Deep networks learn simple semantic patterns before fitting noise.

Modern deep neural works tend to be over-confident.

(Ours) Deep neural networks become less confident of learning semantic patterns before fitting noise when the label noise rises.

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

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



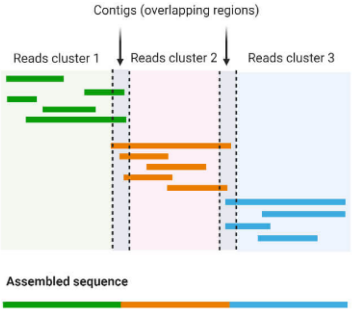
(a) Standard CCE. (b) ProSelfLC with an AT (ours).

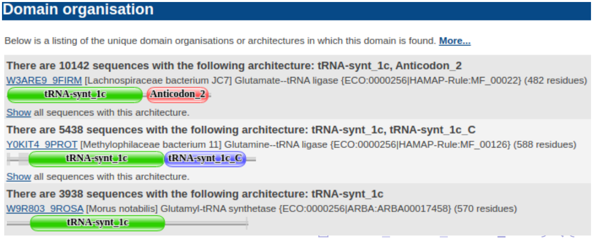
Figure: accu and $\mathrm{conf}_{\mathrm{all}}$ along with the iteration when training ResNet18 on CIFAR-100. The symmetric noise rate is $r = 40\%$.

Outline

1. Work and experience: a quick glance, more in the CV
2. Technical novelty and principle: Example weighting is universal
3. Deep distance metric learning
   - Deep distance metric: Definition & Overview
   - Deep distance metric: Vision applications
4. Robust deep learning (scope: transparent and interpretable ML)
   - Understanding real-world data and why do we need it
   - Learning objectives
   - Project: Robust video-based person ReID
   - Project: To uncover how a deep model learns under noise
5. Omics AI
   - Bioinformatics: Sequence alignment and distance
   - Robust protein understanding
   - Deep distance metric: EDA + Active Learning
6. Industrial R&D experience
   - Industrial R&D overview
   - Industrial AI Research = model-centric AI + data-centric AI

1. NGS Sequencing and sequence analysis
   - With reference: Bowtie + Samtools
   - No reference: overlap-based de novo assembly, then comparing to known sequences using BLAST
2. Sequence alignment and distance calculation
   - MMSeq2
   - Pfam domain database + HMMer-based distance

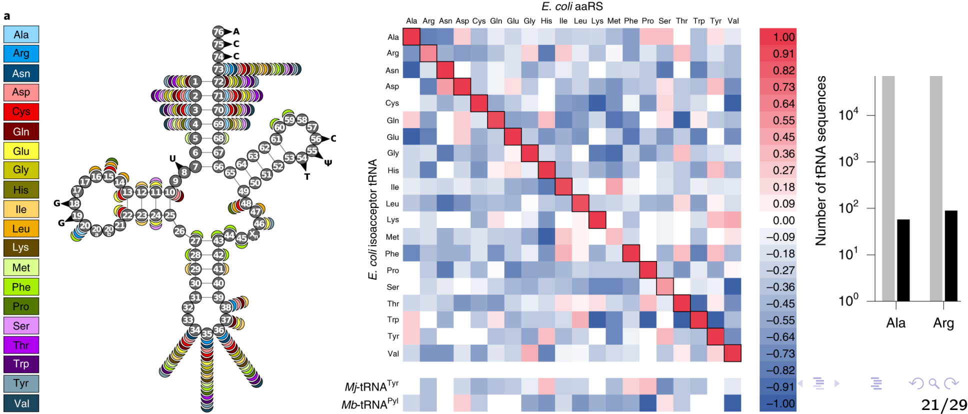




Bioinformatics

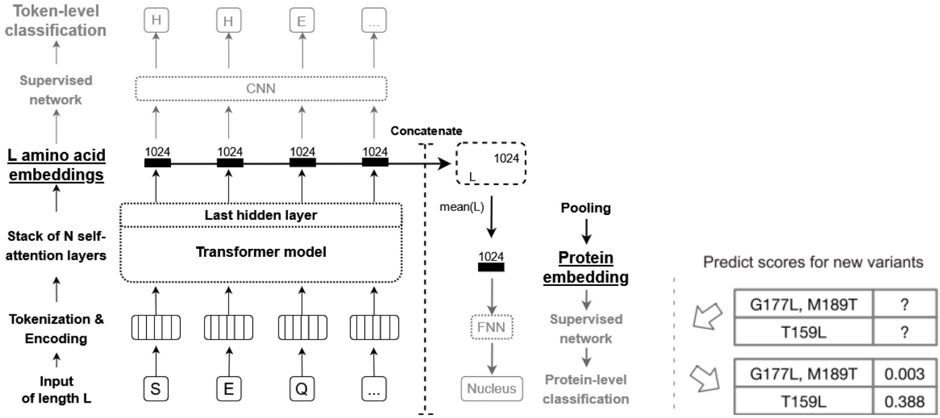
Iterative set-based validation -> improve the algorithm/model

1. Biological wetlab annotation/validation: enrichment/pathway analysis via Null-Hypothesis Statistical Test.
2. tRNA distance calculation [2]
   - Secondary structure and the canonical numbering scheme.
   - Identity elements, responsible for recognising cognate aaRS.



Robust protein understanding

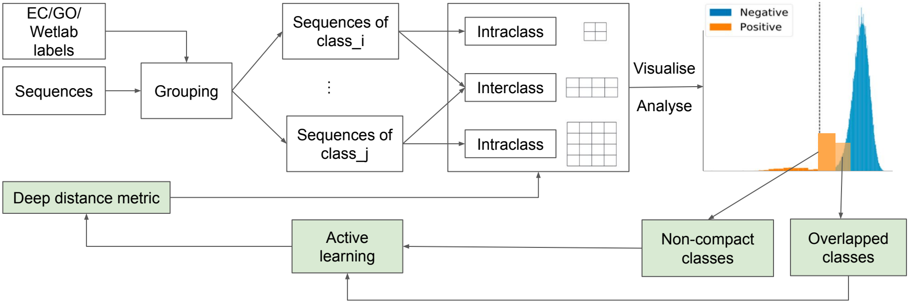
Use the pre-trained transformers to predict or design



[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

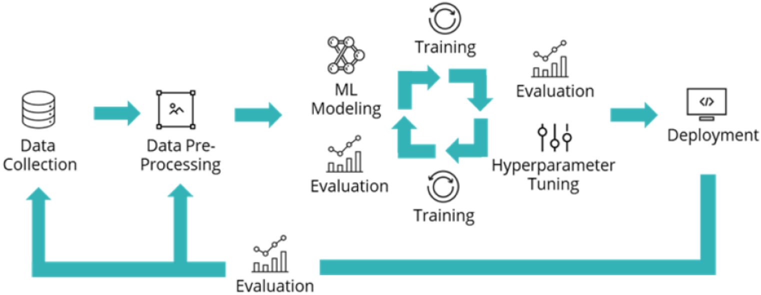


Outline

Industrial R&D overview

Focus: solve & productise real-world impactful problems

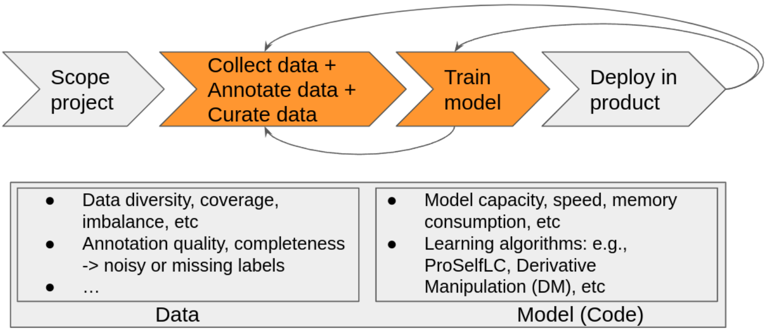
- Publication to share research is a plus and encouraged.
- Collaboratively build AI toolboxes, end-to-end AI service pipelines, etc.



Source: https://dida.do/blog/data-centric-machine-learning

Industrial AI Research

Methods = model-centric AI + data-centric AI



AI System = Data + Model (Code)

1. Suggest research **directions** and write **proposals** to the board.
2. Lead research via task **breakdown** and an estimated/agreed **timeline** using Jira and Confluence.
3. Collaborate, control quality, and maintain conventions using github code **peer review**.

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.