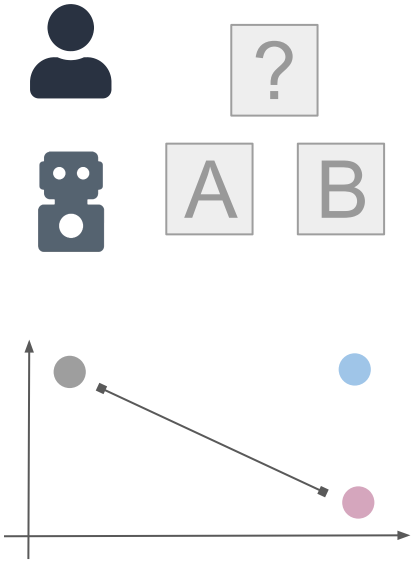
Learning Human-Compatible Representations for Case-Based Decision Support

Han Liu, Yizhou Tian, Chacha Chen, Shi Feng, Yuxin Chen, Chenhao Tan
University of Chicago



Examples explain model predictions

Examples in the training set can serve as **Justification** for <u>model predictions</u>.

Image Classification Task:
Butterfly v.s. Moth

Predicted:
Butterfly

Justification:
A similar-looking Butterfly in the training set

Examples support decision making

Examples in the training set can serve as **Decision Support** for decision makers.

Image Classification Task:
Butterfly v.s. Moth

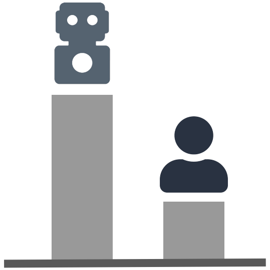
To be Predicted:
?

Decision Support:
A Butterfly and a Moth
in the training set

Two desirable properties of ML models

For effective **decision support** and **justification**, two properties are often desired:

- High autonomous performance
  - An ML model should have satisfying performance on trained tasks

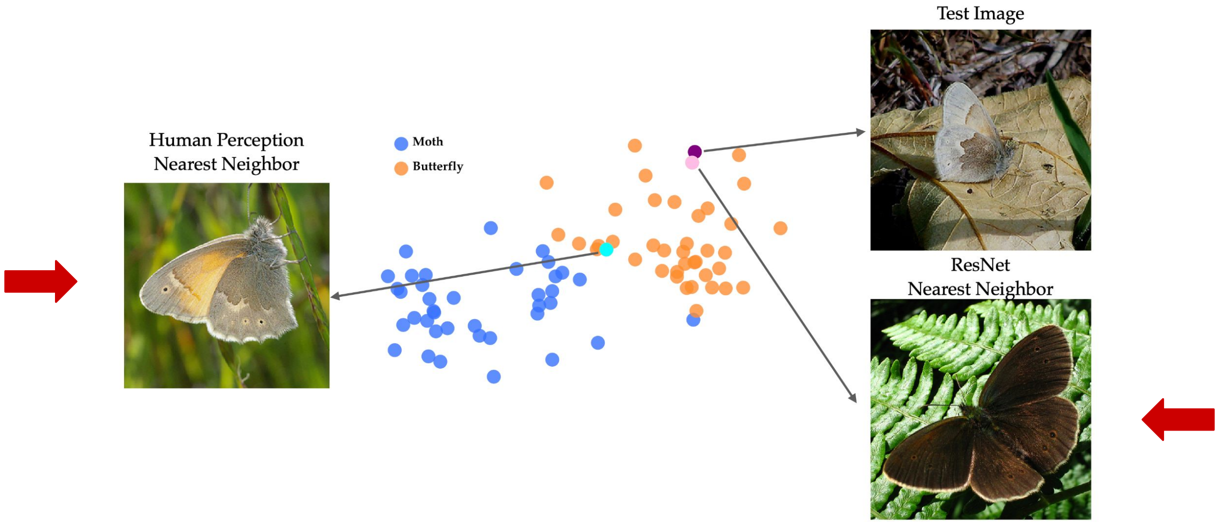


- Transparency and explainability
  - An ML model should provide or be able to derive comprehensible explanations for its predictions

Traditional ML may NOT align with human perception

Projected ImageNet Representations from ResNet (He et al., 2016):

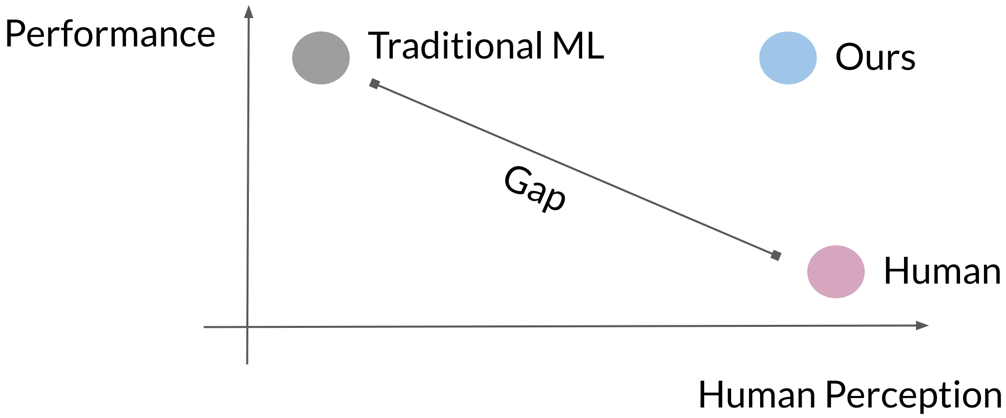
Nearby examples may look very different for humans.



Human-Compatible (HC) Representations

A machine learning model with **two objectives:**

- Achieving high autonomous performance
- Aligned with human perceptions and intuitions



Human-Compatible (HC) Representations

Multi-task learning framework with **two objectives:**

- Task 1: Natural or medical image classification
- Task 2: Human visual similarity judgment prediction
  - Triplet prediction: Two-alternative forced choice (2AFC)

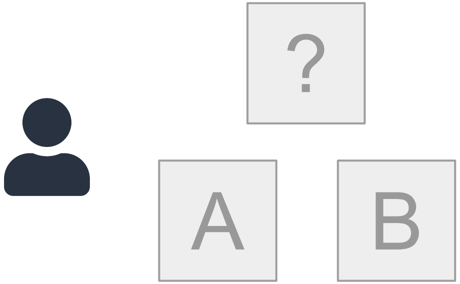
(X, y)

Butterfly

Pneumonia

Moth

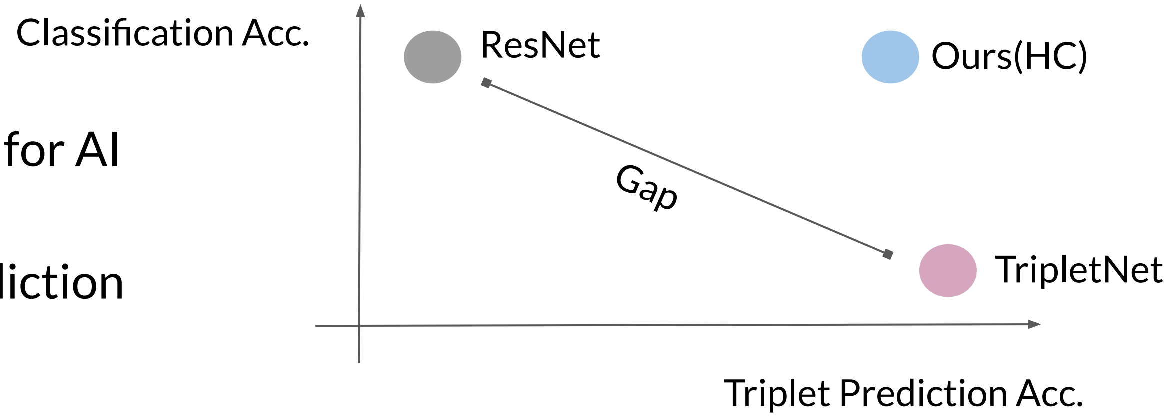
Normal



Human-Compatible (HC) Representations

Multi-task learning

- **Two** decision making tasks for AI
  - Image classification
  - Human judgment prediction



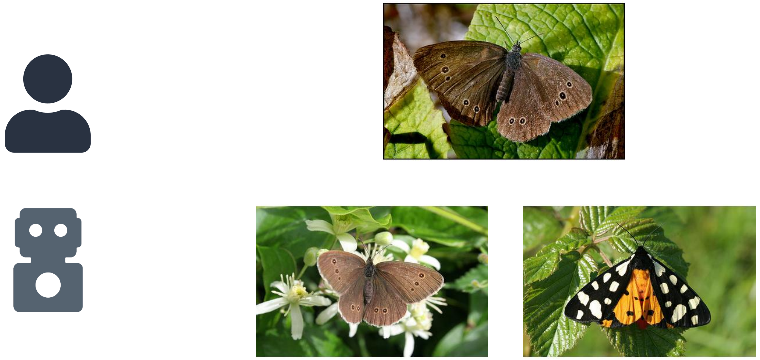
Loss function

$$\lambda \underbrace{\left[- \sum_{(x,y)\sim D} \log\left(p_\theta(y|x)\right) \right]}_{\text{Cross-entropy loss}} + (1-\lambda) \underbrace{\left[\sum_{(x^r, x^+, x^-)\sim T} \max\left(d_\theta(x^r, x^+) - d(x^r, x^-) + \mu, 0\right) \right]}_{\text{Triplet margin loss}}$$

Case-Based Decision Support

Human makes decisions.

Machine provides examples as decision support.

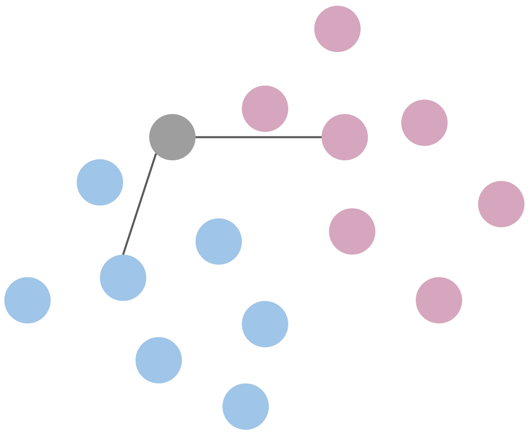


Question: How do we select the decision support examples?

Decision Support Policies

Three types of support policies:

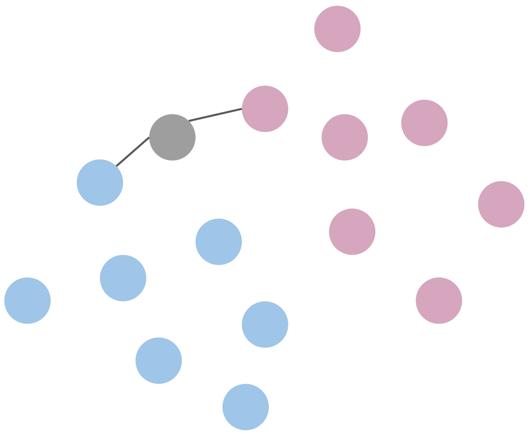
- Random decision support:
  - Random example from each class



Decision Support Policies

Three types of support policies:

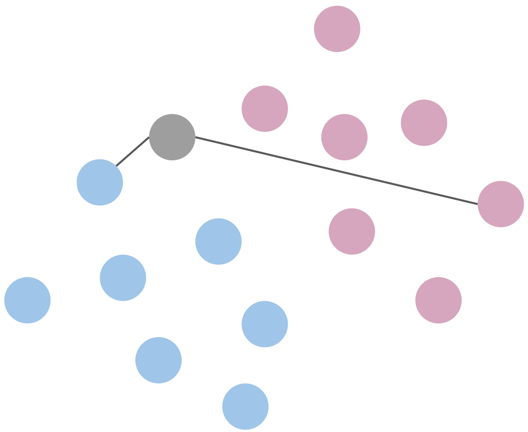
- Random decision support:
  - Random example from each class
- Neutral decision support:
  - Nearest neighbors from each class



Decision Support Policies

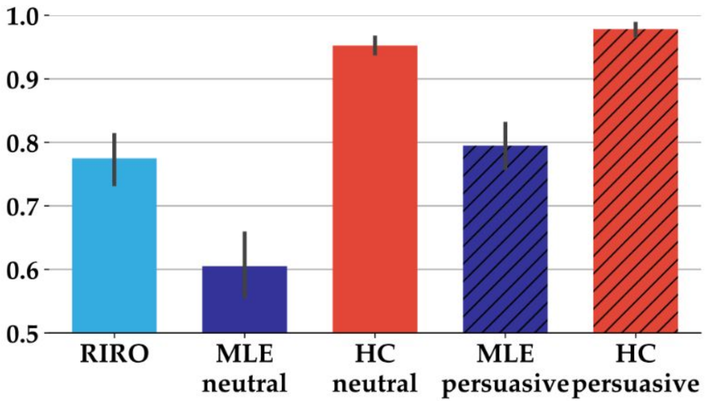
Three types of support policies:

- Random decision support:
  - Random example from each class
- Neutral decision support:
  - Nearest neighbors from each class
- Persuasive decision support:
  - Nearest neighbor from the predicted class
  - Furthest example from the other class(es)

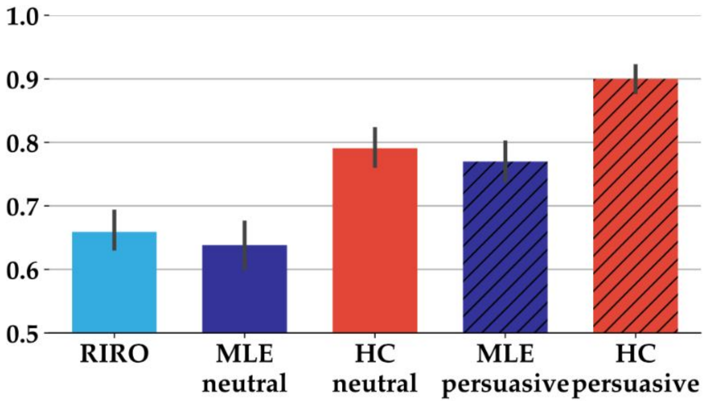


Human subject study (N=50 x 5=250)

Ours (Red) outperforms Random (Blue) and ResNet (Navy).



(a) Butterfly vs. Moth.



(b) Pneumonia classification.

Takeaways

- We highlight the importance of alignment in representation learning towards effective human-machine collaboration.
- We propose a multi-task learning framework that combines supervised learning and metric learning to simultaneously learn classification and human visual similarity.
- We design a novel evaluation framework for comparing representations in case-based decision support.
- Empirical results with synthetic data and human subject experiments demonstrate the effectiveness of our approach.

Thank you!

Email: hanliu@uchicago.edu
Data & Code: https://github.com/ChicagoHAI/learning-human-compatible-representations

📢 Shout-out to my awesome collaborators! 👏





Chicago Human+AI