Unveiling AI's Blind Spots: An Oracle for In-Domain, Out-of-Domain, and Adversarial Errors

Shuangpeng Han[1,2], Mengmi Zhang[1,2]

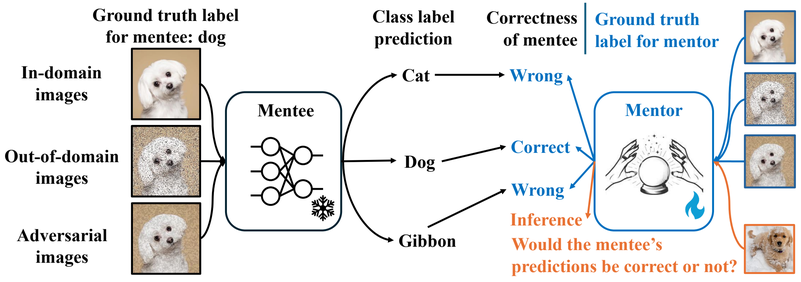
[1] Deep NeuroCognition Lab, College of Computing and Data Science, NTU, Singapore [2]I2R and CFAR, Agency for Science, Technology and Research, Singapore

Introduction

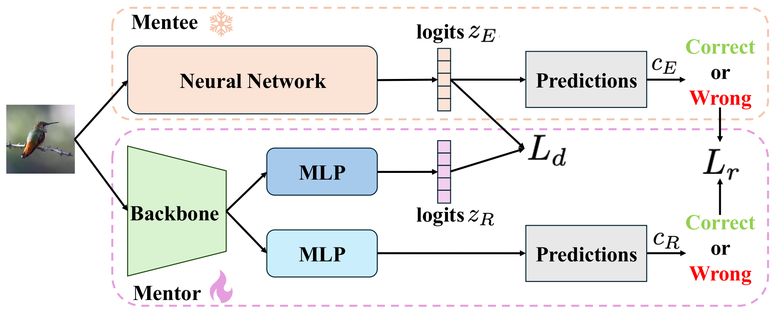
Motivation: AI models make mistakes when recognizing images—whether in-domain, out-of-domain, or adversarial. Understand what mistakes AI models make, why they occur, and how to predict them remains an open challenge.

Use a deep neural network designed to predict another neural network's errors.

- **Mentor**: AI model that predicts errors.
- **Mentee**: AI model being evaluated for performance.



Mentor Architecture



Experimental Setups

Mentor/Mentee architectures:

- ResNet50 (R)
- ViT (V)

Datasets:

- CIFAR-10 (C10)
- CIFAR-100 (C100)
- ImageNet-1K (IN)

Experimental Setups

Error types of mentee:

- In-Domain (ID) Errors.
- Out-of-domain (OOD) Errors: speckle noise, Gaussian blur, spatter and saturate.
- Adversarial Attack (AA) Errors: PGD, CW, Jitter and PIFGSM.

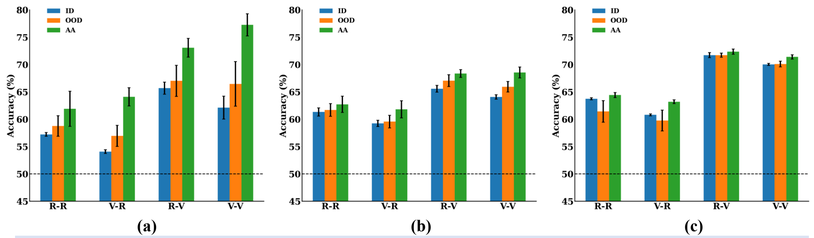
Seven Baselines:

- Self Error Rate (SER)
- Maximum Class Probability (MCP)
- Class Probability Entropy (CPE)
- Distance To Centroid (DTC)
- ConfidNet
- TrustScore
- Steep Slope Loss (SSL)

Evaluation metric: The average accuracy of a mentor across all test sets, including one ID error, four OOD errors, and four AA errors.

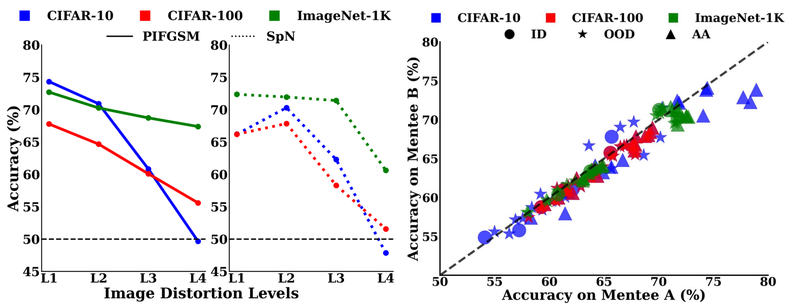
Results

- Training on specific errors of mentees impacts the performance of mentors.
- Mentor architectures matter in error predictions.



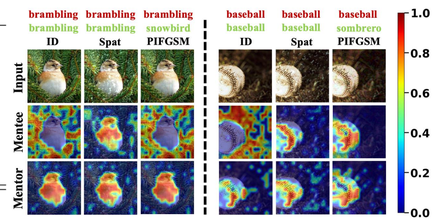
- Training on images with smaller perturbations helps error predictions.
- Mentors generalize across mentees.

Results



- Analysis on our SuperMentor reveals key insights.
- SuperMentor: Adopt ViT as the backbone architecture and train on adversarial images with small perturbations of a mentee.

	Accuracy
SER	50.1±0.0
MCP ($\gamma = 0.7$)	65.1±0.0
CPE ($\alpha = 0.1$)	63.0±0.0
DTC ($d = 10$)	51.8±0.0
ConfidNet (Corbière et al., 2019)	59.3±0.2
TrustScore (Jiang et al., 2018)	61.6±0.0
SSL (Luo et al., 2021)	66.7±0.3
SuperMentor (ours)	**70.4±0.1**



Conclusion & Discussion

- Mentor excels at learning from a mentee's errors on adversarial images with minimal perturbations.
- Mentor generalizes well to both in-domain and out-of-domain predictions of the same mentee.
- Transformer-based mentors generalize better than 2D-CNN-based ones.
- Introduce the SuperMentor which outperforms all existing mentor baselines.

We support Open Science!
Scan QR code for
Papers, code, data

PhD students and postdocs wanted! Join us!

NATIONAL RESEARCH FOUNDATION