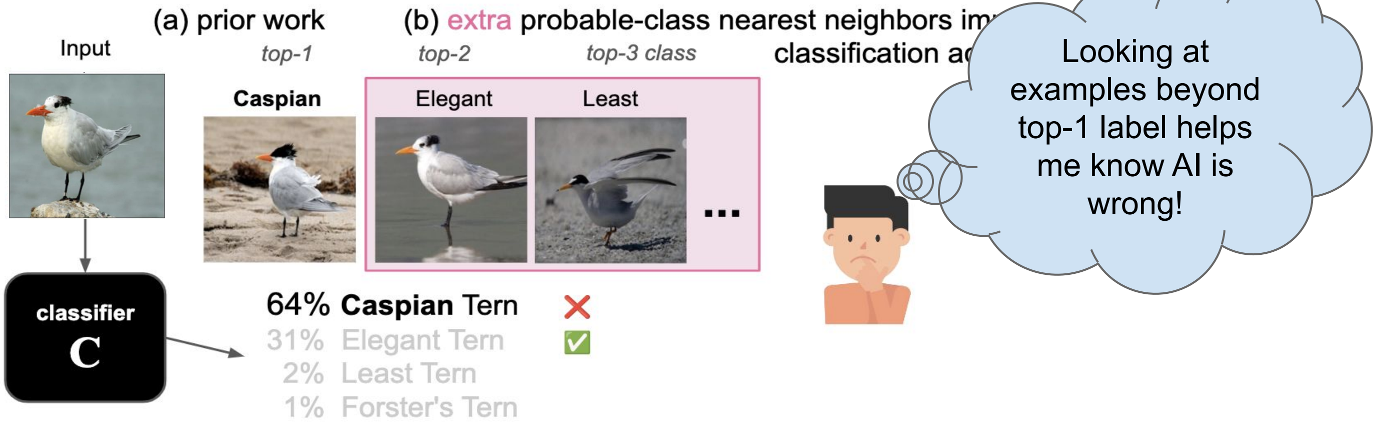
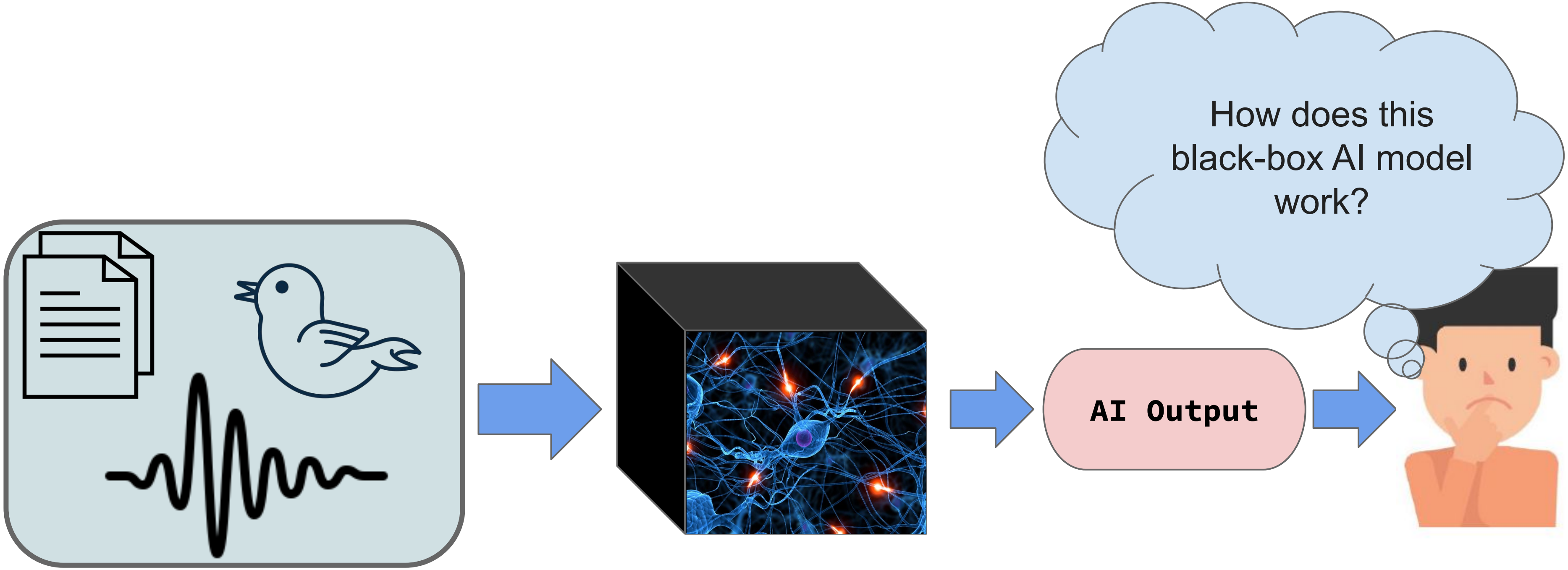
Probable-class Nearest-neighbor Explanations Improve AI & Human Accuracy

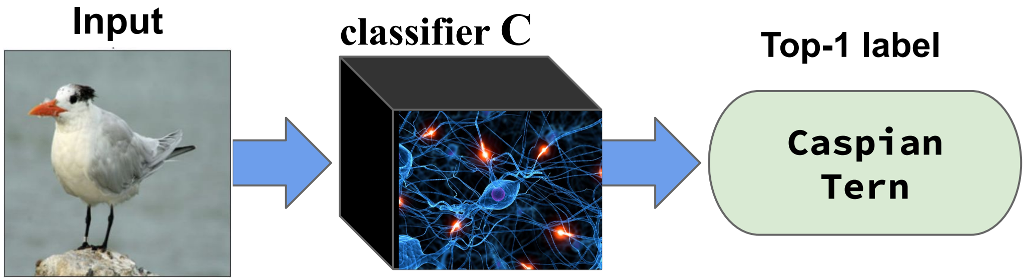
Giang Nguyen, Valerie Chen, Mohammad Taesiri, Anh Nguyen



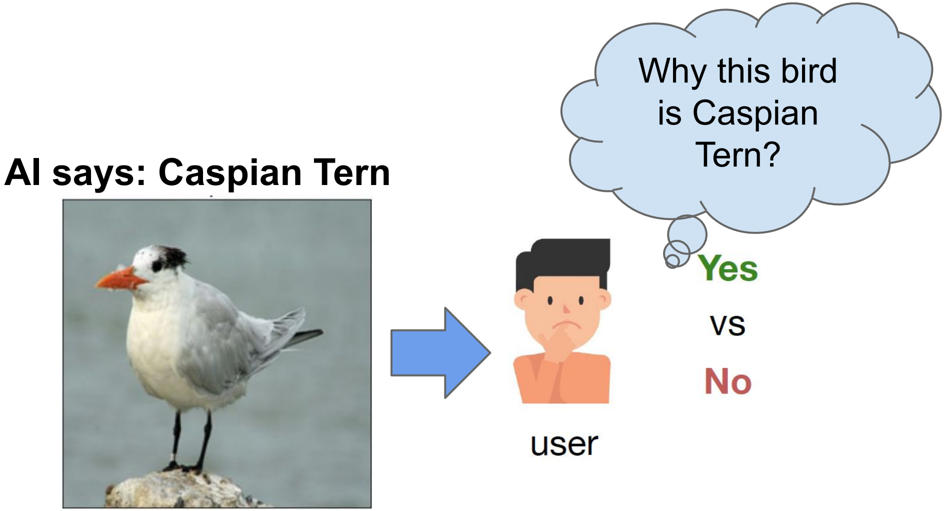
Motivation



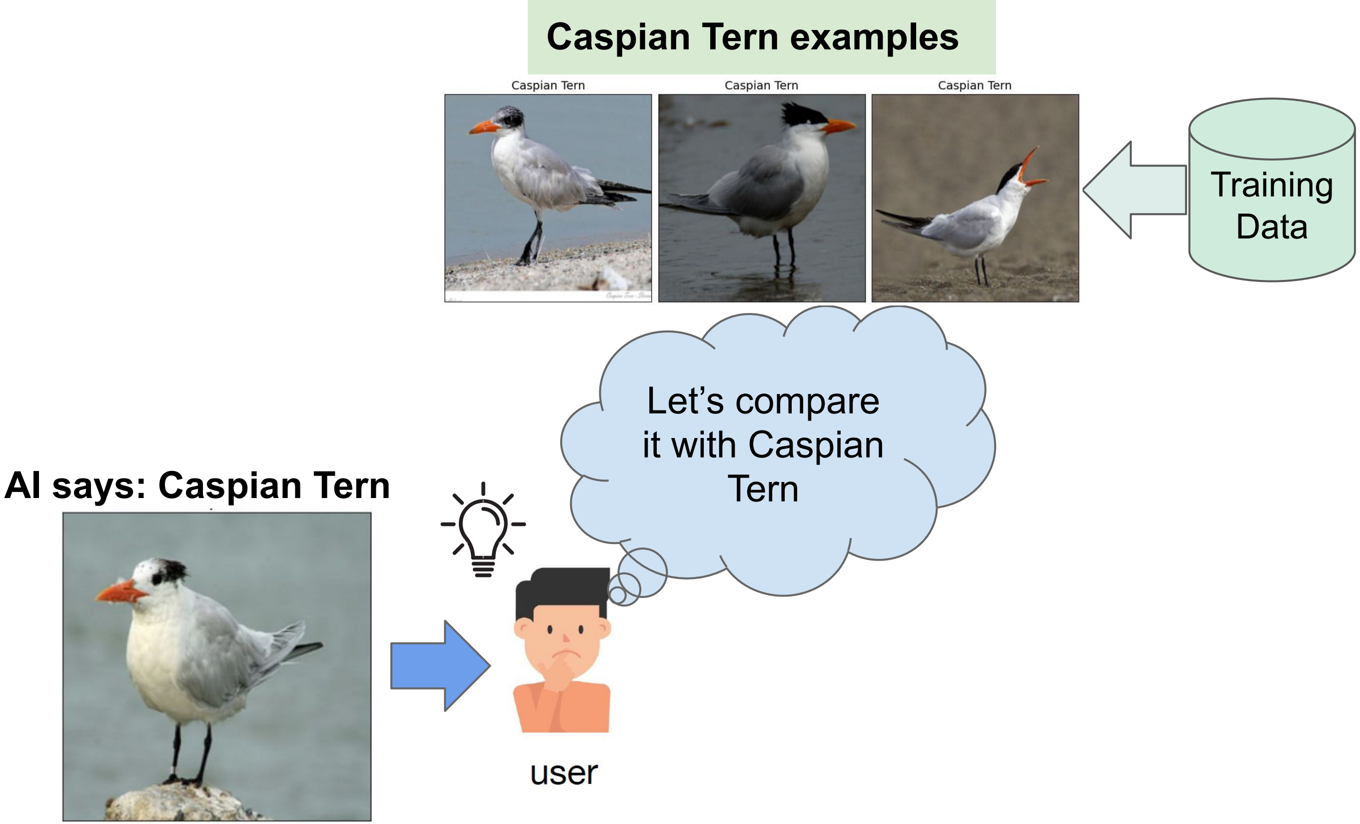
Background



Background

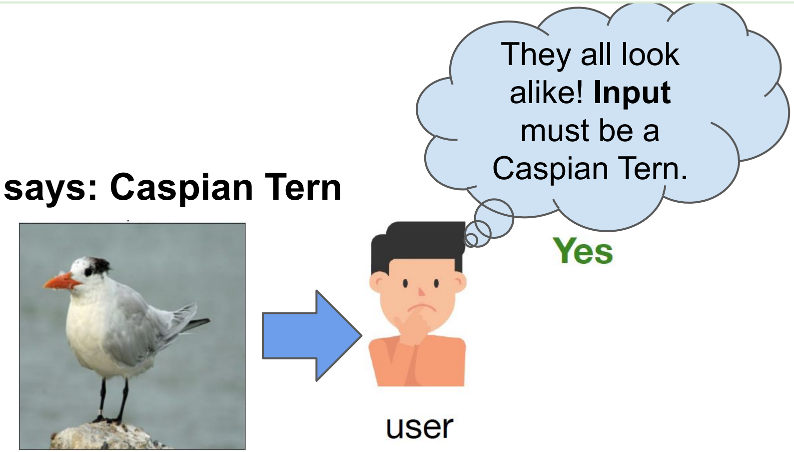

AI says: Caspian Tern
Why this bird is Caspian Tern?
Yes
vs
No
user

Background


Caspian Tern examples
Caspian Tern
Caspian Tern
Caspian Tern
Training Data
Let's compare it with Caspian Tern
AI says: Caspian Tern
user

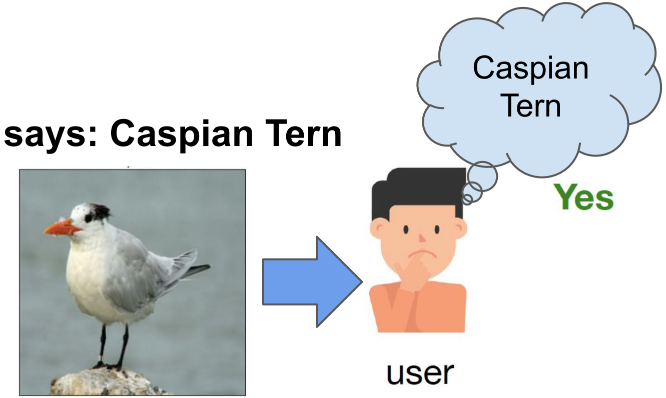
(Motivation) The deceptiveness of nearest neighbors

AI says: Caspian Tern



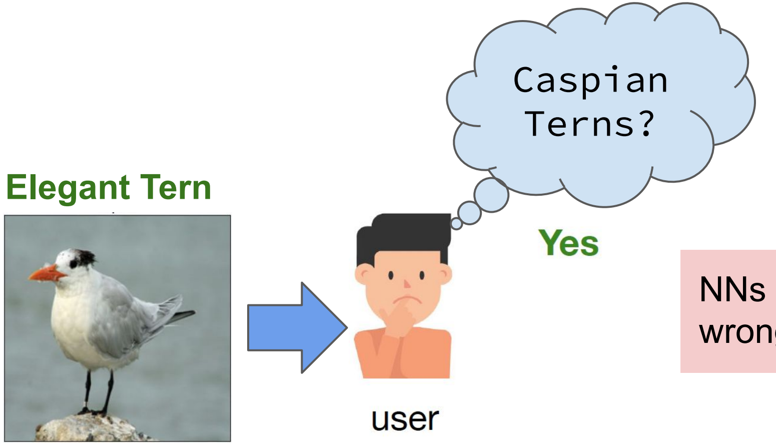
(Motivation) The deceptiveness of nearest neighbors

AI says: Caspian Tern



Input is actually Elegant Tern

(Motivation) The deceptiveness of nearest neighbors

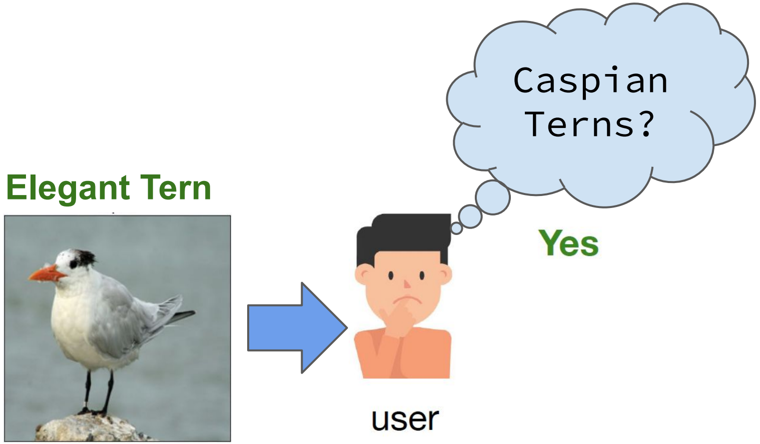


NNs are **deceptively** convincing, leading users to accept (even wrong) AI recommendations most of the time

(Motivation) The deceptiveness of nearest neighbors

Deceptively convincing nearest neighbors

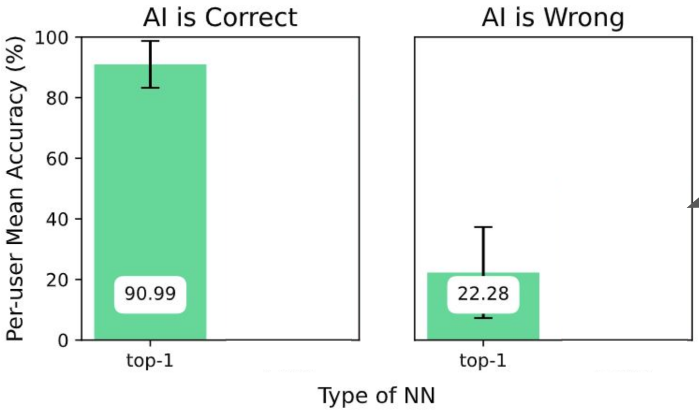
Elegant Tern



Yes

user

Consequently, leading to low accuracy on identifying AI errors

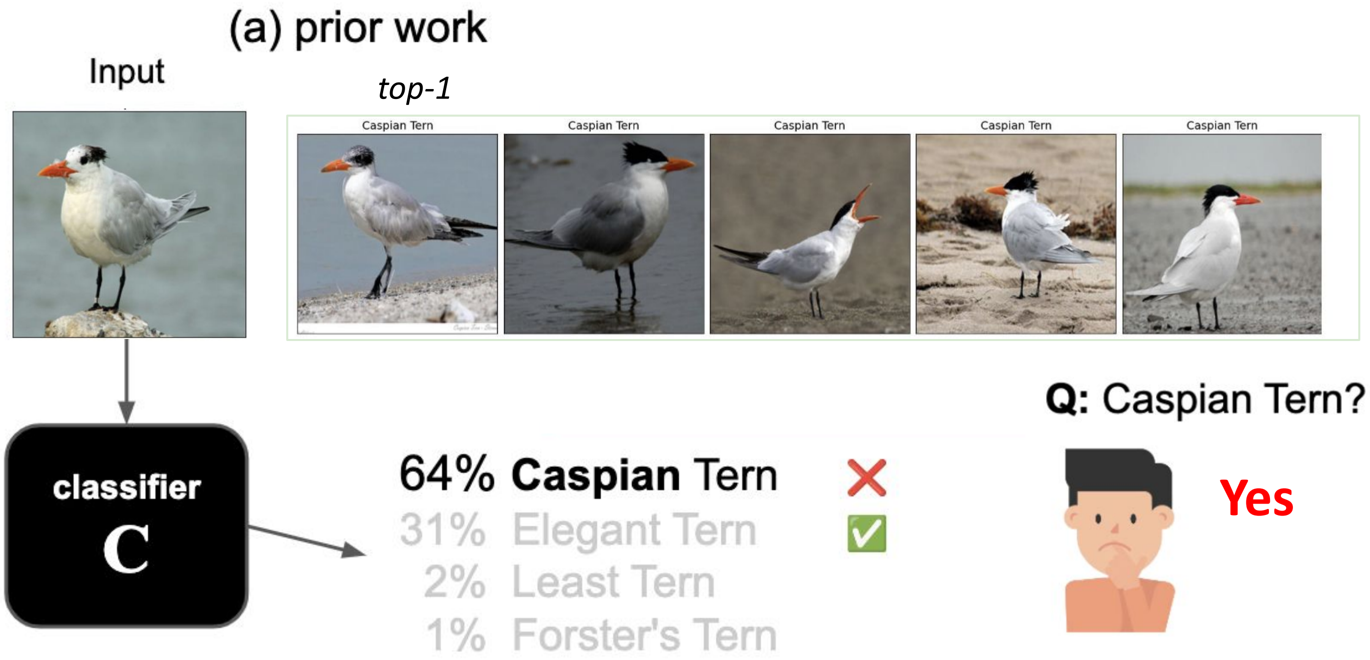


Research question

Research question:

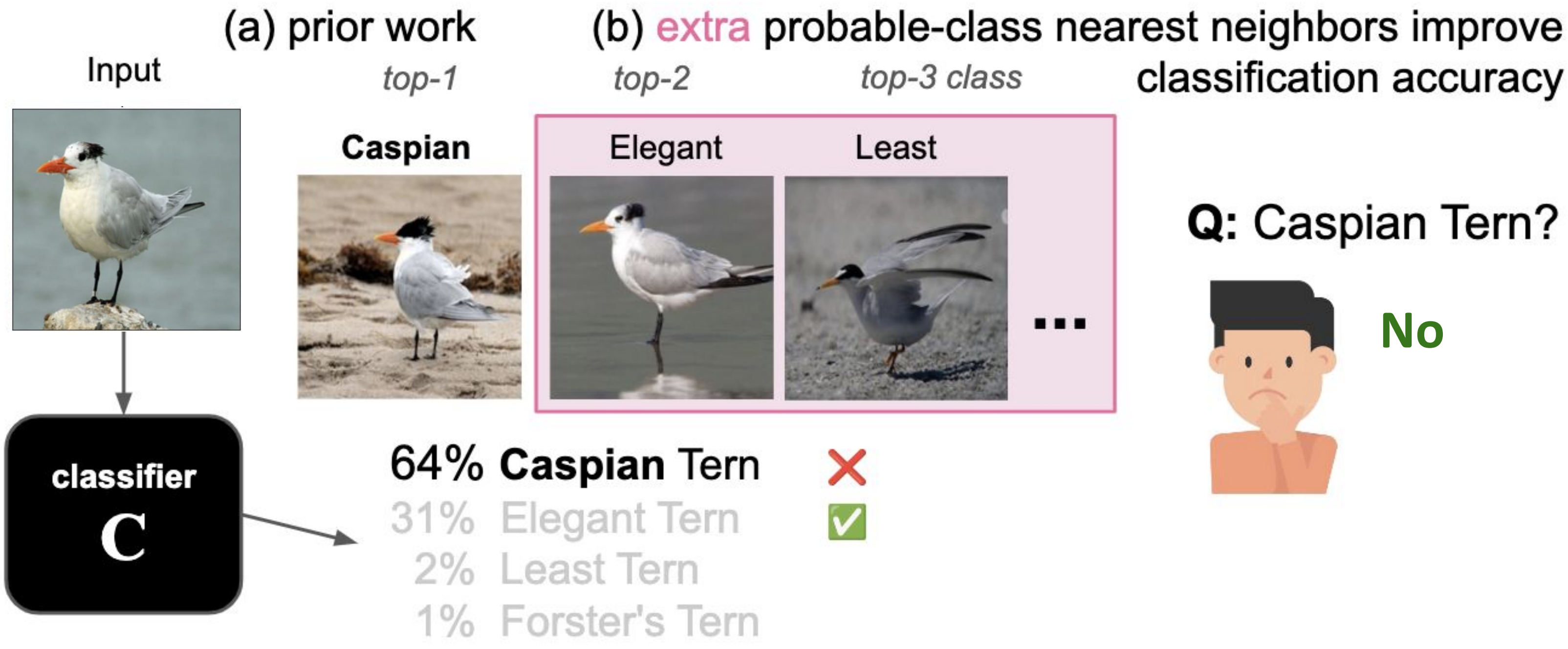
- ❑ 1. (Interpretability) Can we leverage the rich information from nearest-neighbor explanations to mitigate the deception?

Probable-class nearest neighbors (PCNN)



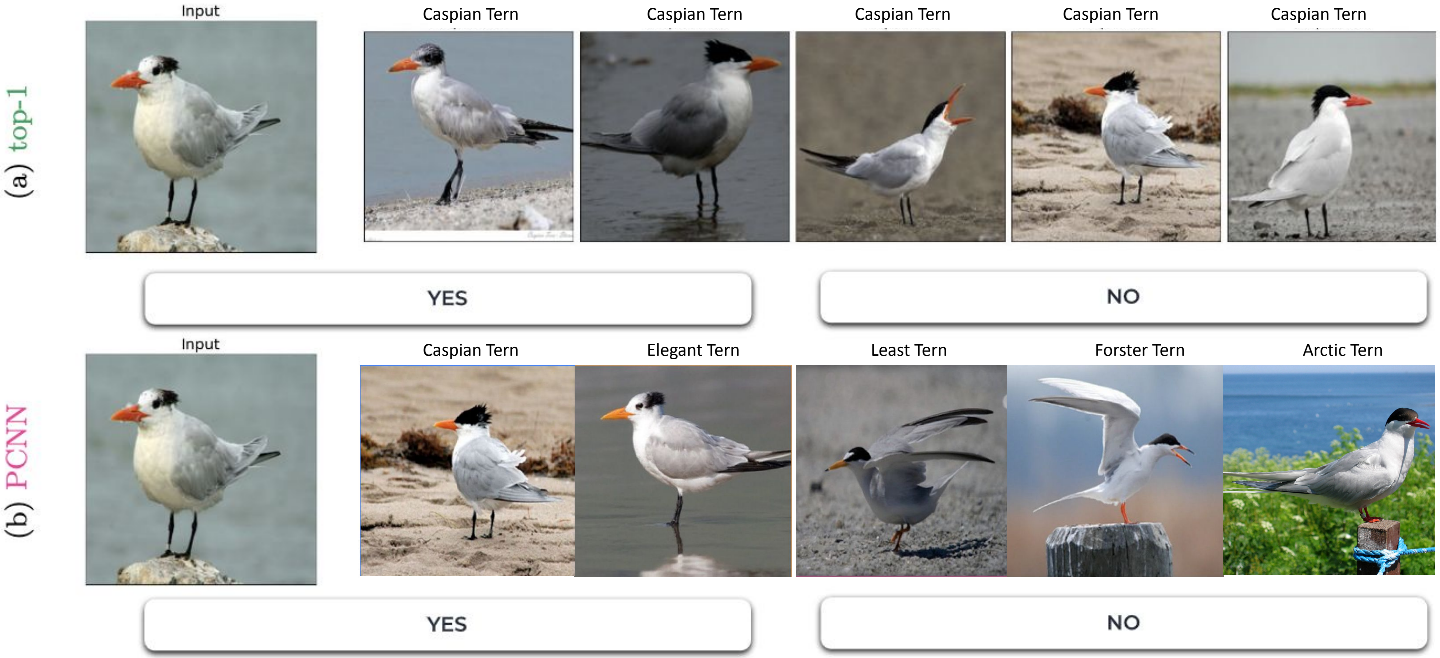
Prior work (a) often shows only the nearest neighbors from the top-1 predicted class as explanations for the decision, which often fools humans into accepting wrong decisions (here, Caspian Tern) due to the similarity between the input and top-1 class examples

Probable-class nearest neighbors (PCNN)



Instead, we present extra nearest neighbors (b) from top-2 to top-K classes that improves human accuracy when AI is wrong via providing contrastive evidence.

Human evaluation on PCNN

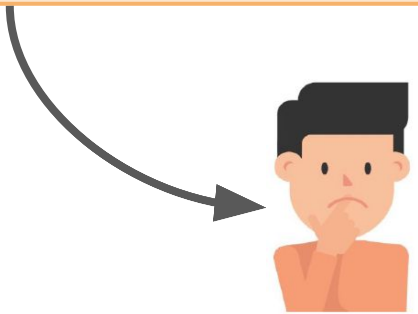


Human evaluation settings where users assess whether the top-1 predicted label is correct or incorrect

Human evaluation on PCNN

Elegant Tern | Caspian Tern | Elegant Tern | Least Tern | Forster Tern | Arctic Tern

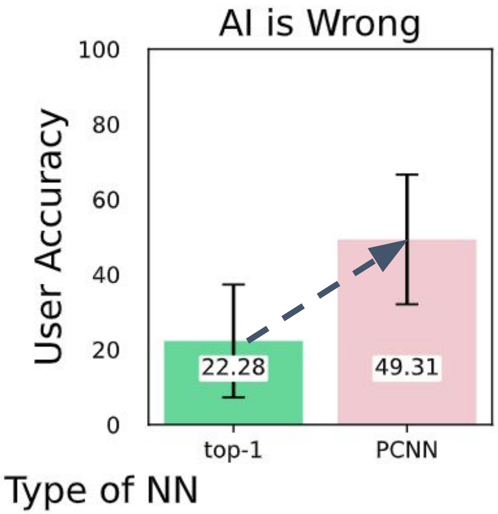
Sam guessed the Input image is **Caspian Tern** with 64% confidence.
Is this bird a **Caspian Tern**?



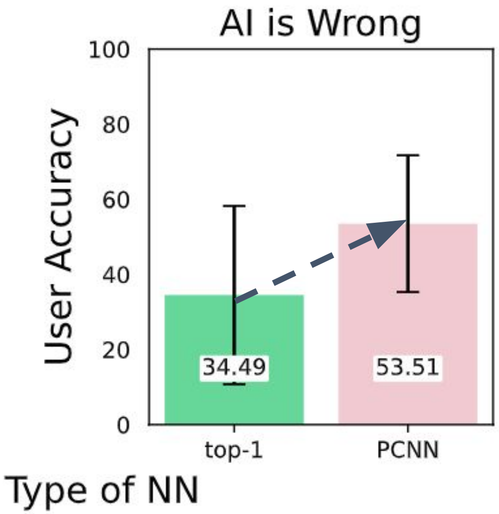
Caspian Tern?
Yes!

Caspian Tern?
No!

Human evaluation on PCNN



(a) CUB-200.



(b) Dogs-120.

Finding 1: In both settings (CUB-200 and Dogs-120), humans show significantly improved accuracy in identifying AI errors.

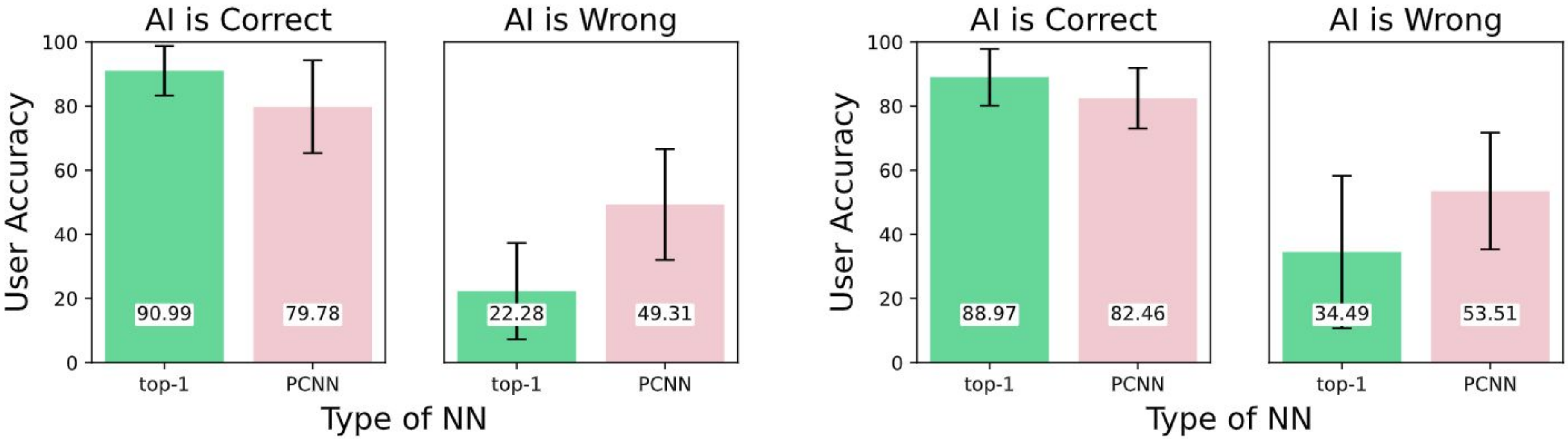
Human evaluation on PCNN

Limited information, no alternative labels

Diverse & contrastive information

PCNN's richer information helps users distinguish similar species, while top-1 predictions provide little context and no alternative labels, leading to easier acceptance of errors.

Human evaluation on PCNN

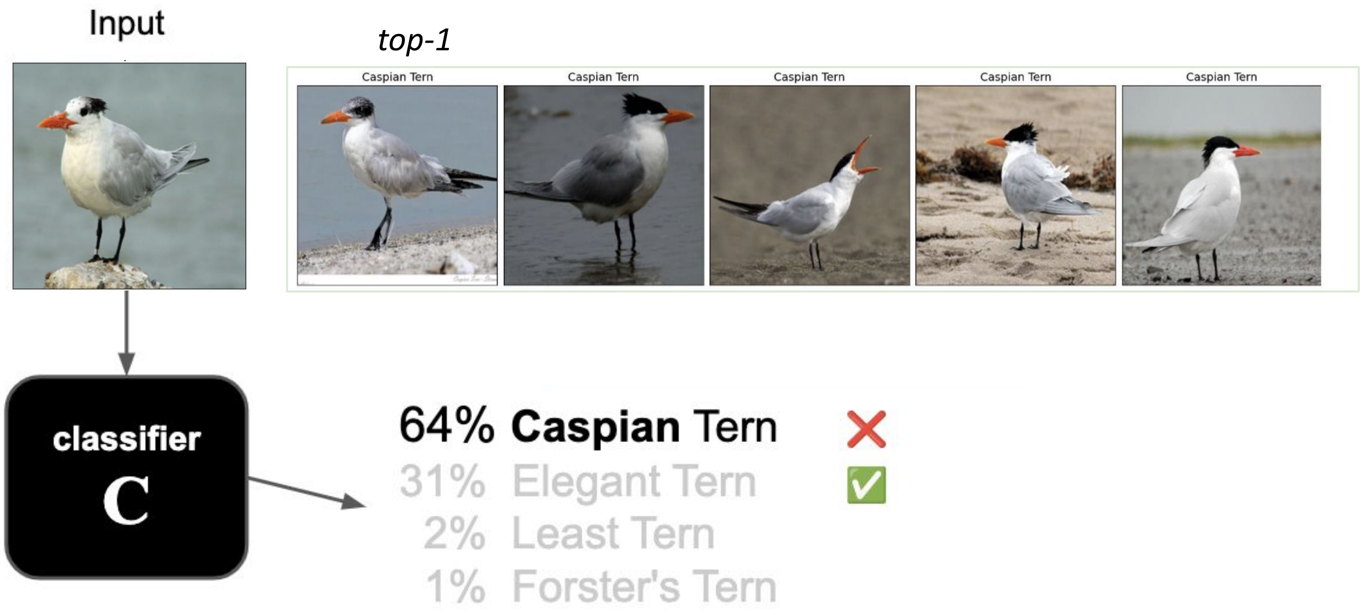


(a) CUB-200.

(b) Dogs-120.

Finding 2: On all (correct & wrong) samples, PCNN improves user accuracy by 10 points on CUB-200 (54.55% → 64.58%) and over 5 points on Dogs-120 (63.55% → 69.21%).

Pretrained image classifiers struggle with close species



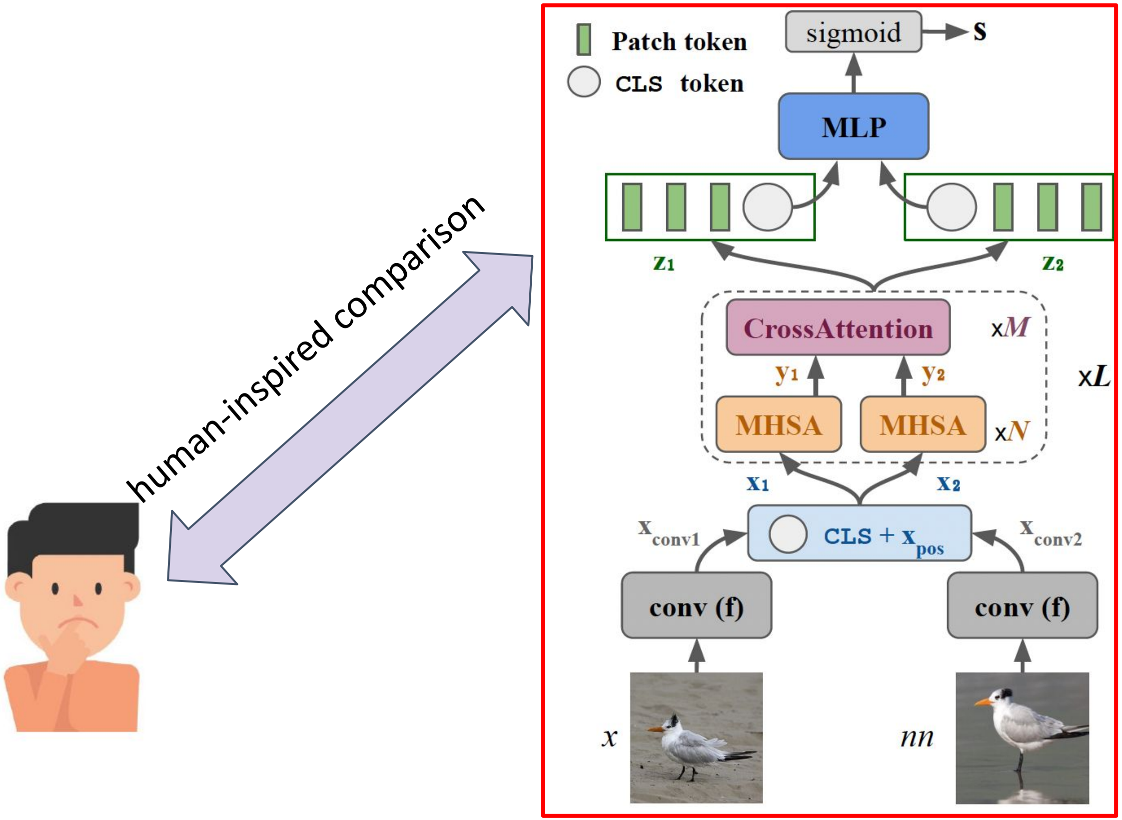
Pretrained image classifiers often struggle to distinguish with close species

Research question

Research question:

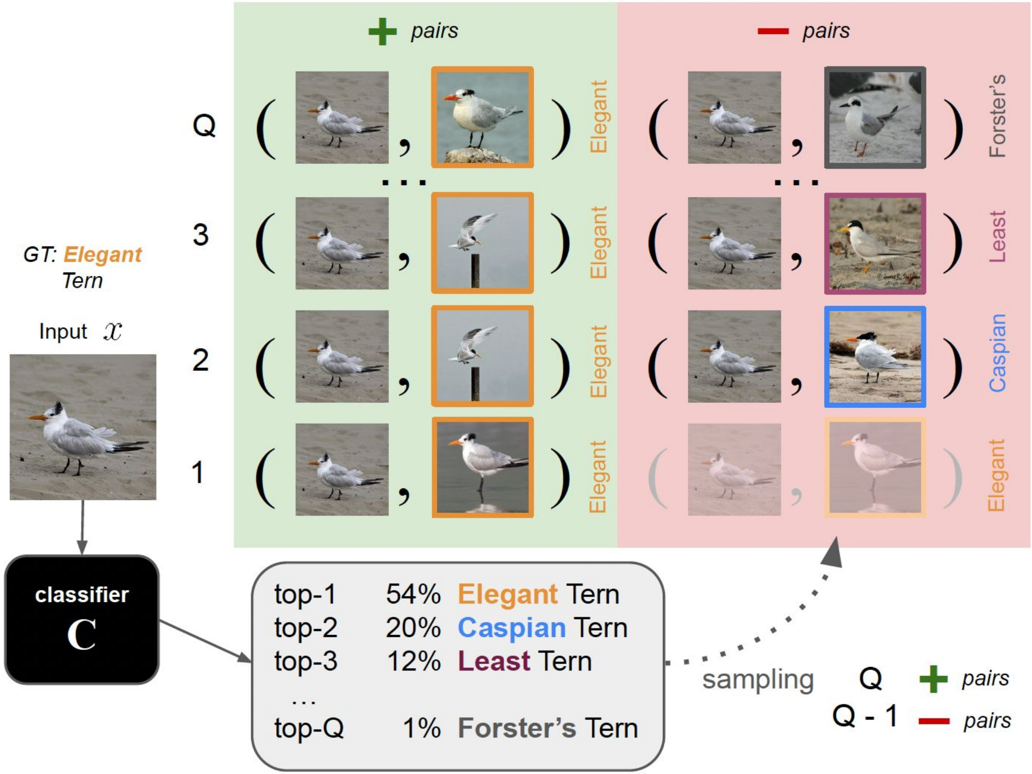
- ❑ 1. (Interpretability) Can we leverage the rich information from nearest-neighbor explanations to mitigate the deception?
- ❑ 2. (Accuracy) Can pretrained models benefit from the rich information of probable-class nearest neighbors (PCNN)?

Image comparator network S



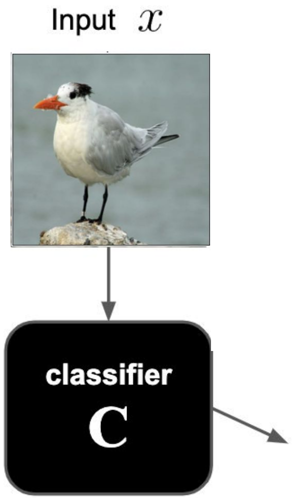
A novel image comparator network S that takes two images as input and outputs a probability score [0→1] indicating the likelihood that they belong to the same class

Sampling algorithm for training S



A sampling algorithm for selecting positive and negative image pairs to train the image comparator $\mathbf{S}$. The classes (top1 $\rightarrow$ Q) are determined based on the likelihood scores of classifier $\mathbf{C}$ for the input $\boldsymbol{x}$

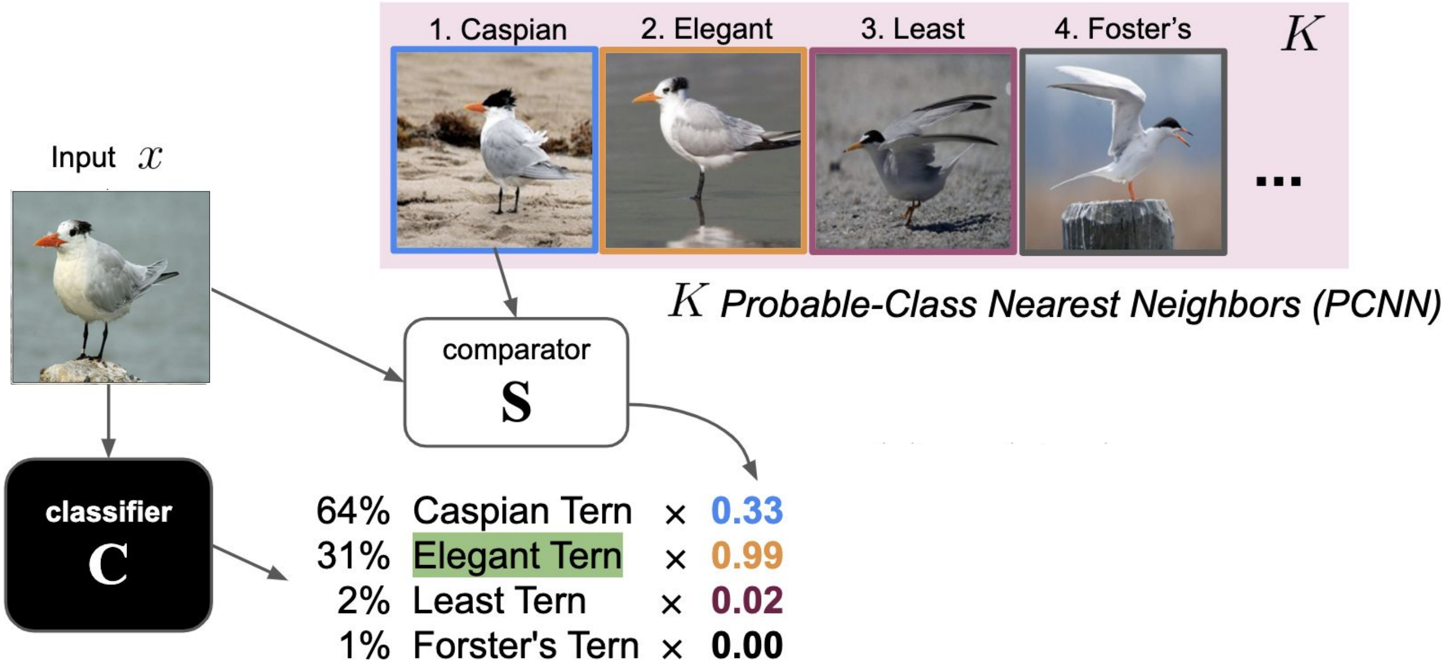
Reranking with image comparator S



64% Caspian Tern
31% Elegant Tern
2% Least Tern
1% Forster's Tern

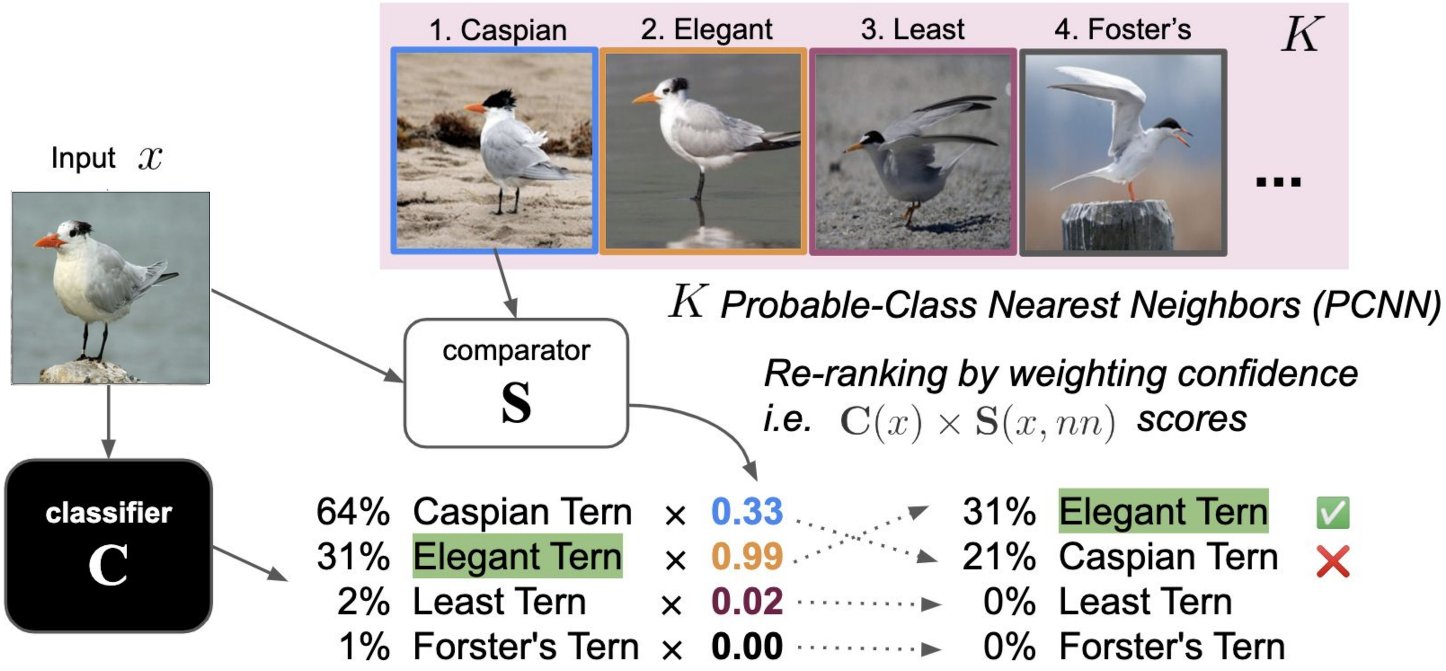
First, pretrained classifier $\mathbf{C}$ predicts the label for input $\boldsymbol{x}$

Reranking with image comparator S



Then, from each class among the top-K predicted classes by **C**, we find the nearest neighbor ***nn*** to the input ***x*** and compute a sigmoid similarity score ***S(x, nn)***

Reranking with image comparator $\mathbf{S}$



These scores weight the original $\mathbf{C}(x)$ probabilities, re-ranking the labels (here **Elegant Tern** was pushed into top-1 label).

Reranking with image comparator S

C: Original pretrained image classifier

C → S: Reranking with similarity scores from S only

C x S: Reranking with weighted probabilities C*S

Classifier architecture		ResNet-18 (a)			ResNet-34 (b)			ResNet-50 (c)		
Dataset	Pretraining	**C**	**C → S**	**C × S**	**C**	**C → S**	**C × S**	**C**	**C → S**	**C × S**
CUB-200	iNaturalist	n/a	n/a	n/a	n/a	n/a	n/a	85.83	87.72	88.59 (+2.76)
	ImageNet	60.22	66.78	71.09 (+10.87)	62.81	71.92	74.59 (+11.78)	62.98	71.63	74.46 (+11.48)
Cars-196	ImageNet	86.17	85.70	88.27 (+2.10)	82.99	83.57	86.02 (+3.03)	89.73	89.90	91.06 (+1.33)
Dogs-120	ImageNet	78.75	75.34	79.58 (+0.83)	82.58	80.82	83.62 (+1.04)	85.82	83.39	86.31 (+0.49)

Reranking with image comparator S consistently improves over the pretrained classifier C

Reranking for CUB-200

Initial class ranking by pretrained classifier C

Query: Elegant Tern | Top 1: Caspian Tern | Top 2: Elegant Tern | Top 3: Least Tern | Top 4: Forster Tern | Top 5: Arctic Tern

RN50: 64% | S: 0.33 | RN50: 31% | S: 0.99 | RN50: 2% | S: 0.02 | RN50: 1% | S: 0.00 | RN50: 1% | S: 0.00

Refined class ranking by Product of Experts C x S

Top 1: Elegant Tern | Top 2: Caspian Tern | Top 3: Forster Tern | Top 4: Least Tern | Top 5: Arctic Tern

RN50 x S: 31% | RN50 x S: 21% | RN50 x S: 0% | RN50 x S: 0% | RN50 x S: 0%

CUB-200 re-ranking: Caspian Tern → Elegant Tern

Reranking for Cars-196

Cars-196 re-ranking: BMW M6 → Jaguar XK

Reranking for Dogs-120

Dogs-120 re-ranking: Irish Terrier → Otterhound

Achieving state-of-the-art accuracy on FG classification

Classifier	Ex	Img	Patch	R	Acc
k-NN + cosine Taesiri et al. (2022)	✓	✓	-	-	85.46
k-NN + **S**	✓	✓	-	-	86.88
ProtoPNet Chen et al. (2019)	-	-	✓	-	81.10†
PIPNet Nauta et al. (2023)	-	-	✓	-	82.00
ProtoTree Nauta et al. (2021)	-	-	✓	-	82.20
ProtoPool Rymarczyk et al. (2022)	-	-	✓	-	85.50
Def-ProtoPNet Donnelly et al. (2021)	-	-	✓	-	86.40
TesNet Wang et al. (2021)	-	-	✓	-	86.50†
ST-ProtoPNet Wang et al. (2023b)	-	-	✓	-	86.60
ProtoKNN Ukai et al. (2023)	✓	-	✓	-	87.00
CHM-Corr Taesiri et al. (2022)	✓	✓	✓	✓	83.27
EMD-Corr Taesiri et al. (2022)	✓	✓	✓	✓	84.98
C × **S** (ours)	✓	✓	-	✓	**88.59** ± 0.17

(a) CUB-200

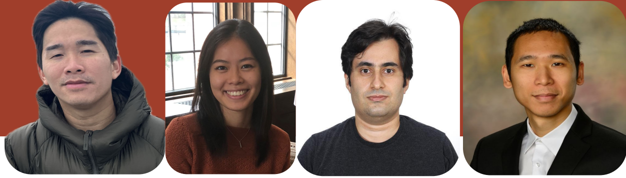
Classifier	Ex	Img	Patch	R	Acc
k-NN + cosine	✓	✓	-	-	87.48
k-NN + **S**	✓	✓	-	-	88.90
ProtoPNet	-	-	✓	-	85.31†
ProtoPShare	-	-	✓	-	86.40*
PIPNet	-	-	✓	-	86.50
ProtoTree	-	-	✓	-	86.60
ProtoPool	-	-	✓	-	88.90
ProtoKNN	✓	-	✓	-	90.20
CHM-Corr	✓	✓	✓	✓	85.03
EMD-Corr	✓	✓	✓	✓	87.40
C × **S** (ours)	✓	✓	-	✓	**91.06** ± 0.15

(b) Cars-196

Classifier	Ex	Img	Patch	R	Acc
k-NN + cosine	✓	✓	-	-	85.56
k-NN + **S**	✓	✓	-	-	82.33
ProtoPNet	-	-	✓	-	76.40†
TesNet	-	-	✓	-	82.40†
Def-ProtoPNet	-	-	✓	-	82.20†
ST-ProtoPNet	-	-	✓	-	84.00
MGProto	-	-	✓	-	85.40
CHM-Corr	✓	✓	✓	✓	85.59
EMD-Corr	✓	✓	✓	✓	85.57
C × **S** (ours)	✓	✓	-	✓	**86.31** ± 0.03

(c) Dogs-120

Summary



Giang Valerie Mohammad Anh

Take-away messages:

1. *Probable-class nearest neighbors* improve human accuracy on fine-grained images by providing contrastive evidence rather than solely supportive (top–1) evidence.
2. *Probable-class nearest neighbors* aid in training an image comparator, significantly improving the image classification accuracy of pretrained models.

Future Works:

1. Improving interpretability of LLMs/VLMs by displaying multimodal contrastive evidence behind answers.
2. Improving accuracy of LLMs/VLMs via first answer, then deep think via comparing with supportive/contrastive evidence (2-stage generation).

Acknowledgements:

1. Transactions on Machine Learning Research for inviting us to present at ICLR.
2. Son Nguyen and Hung Dao from KAIST; Peijie Chen, Thang Pham, and Pooyan R. for valuable.
3. NaphCare Foundation, Adobe Research, and Auburn University for financial support.