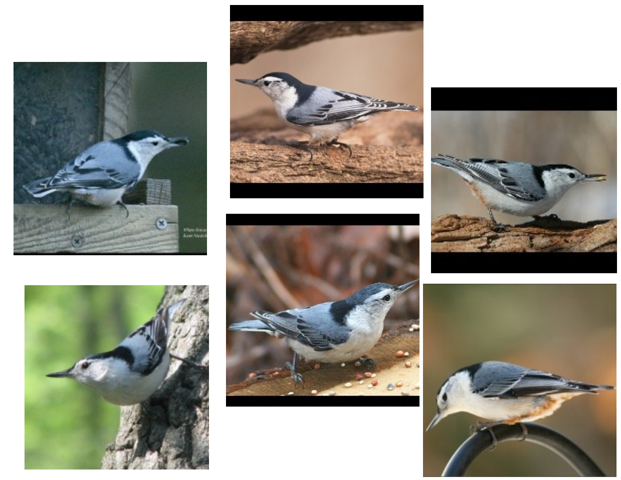
Explanation-based Data Augmentation for Image Classification

Sandareka Wickramanayake, Mong Li Lee, Wynne Hsu
{sandaw, leeml, whsu}@comp.nus.edu.sg
School of Computing, National University of Singapore

Training Dataset must be Representative.

Black Tern

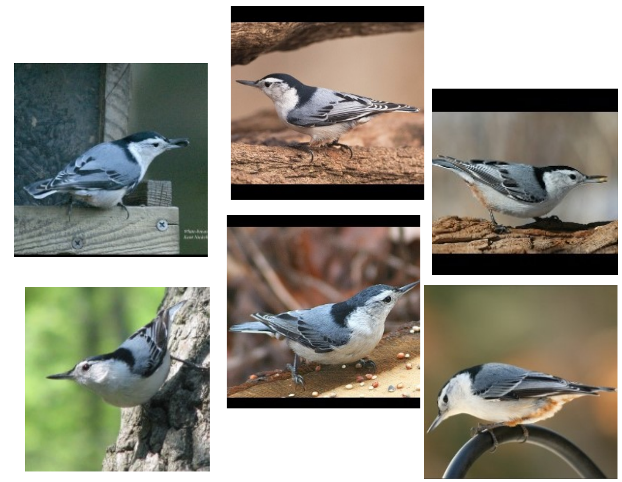


White Breasted Nuthatch

Training Dataset must be Representative.

Juvenile Black Terns

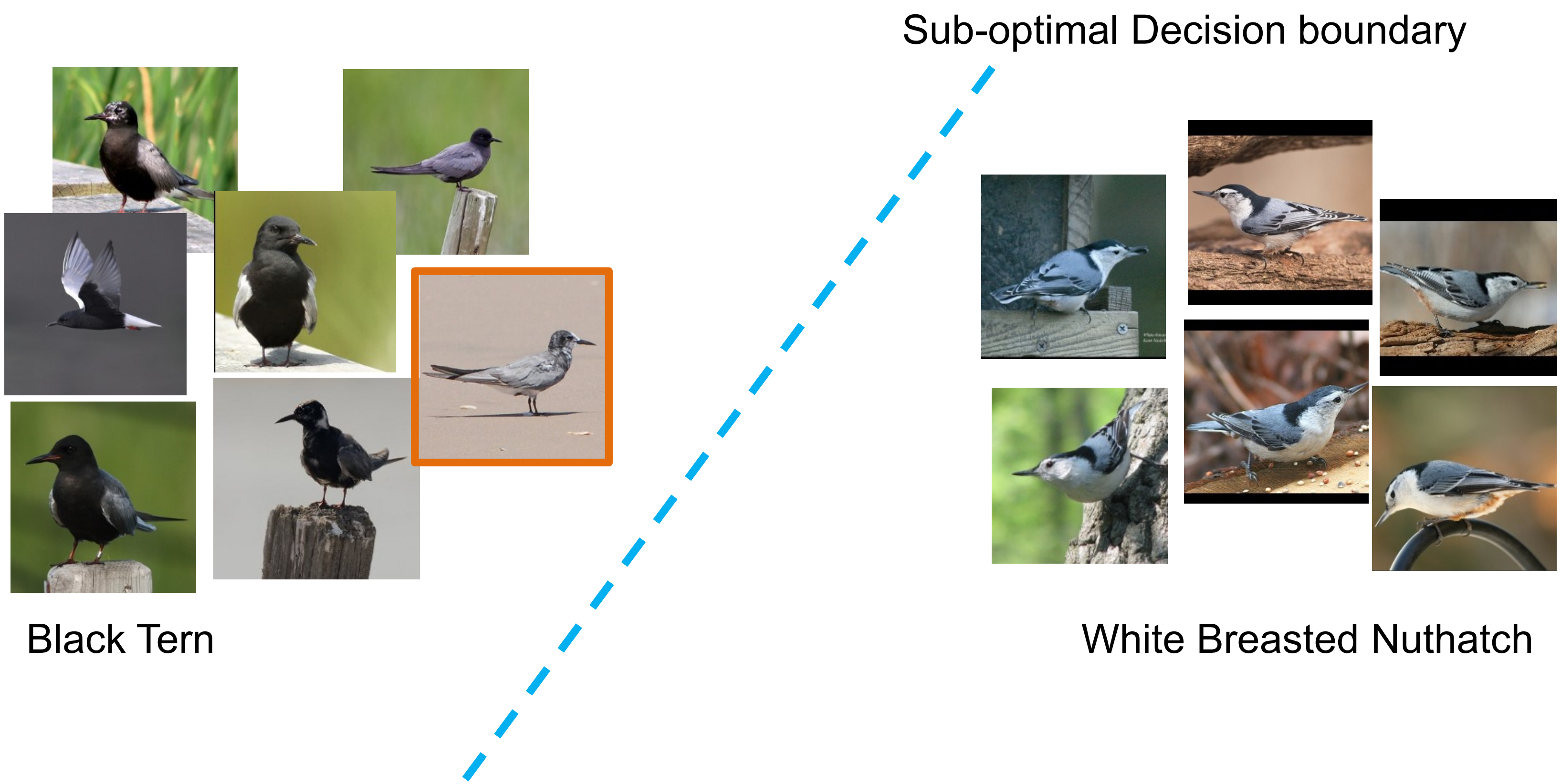
Black Tern



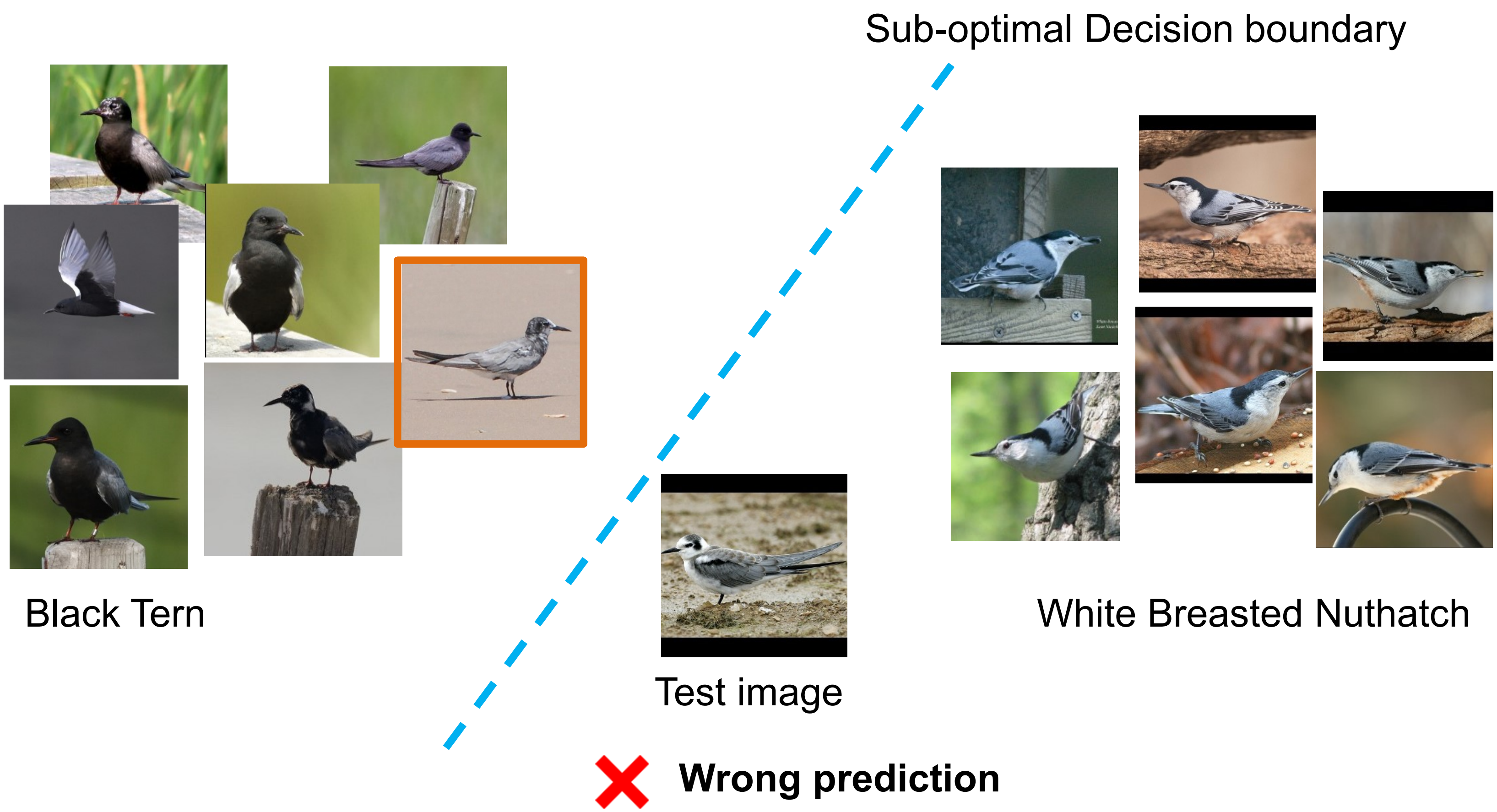
White Breasted Nuthatch

- Train dataset contains very few images of Juvenile Black Terns.
- Juvenile Black Terns are under-represented.

Training Dataset must be Representative.

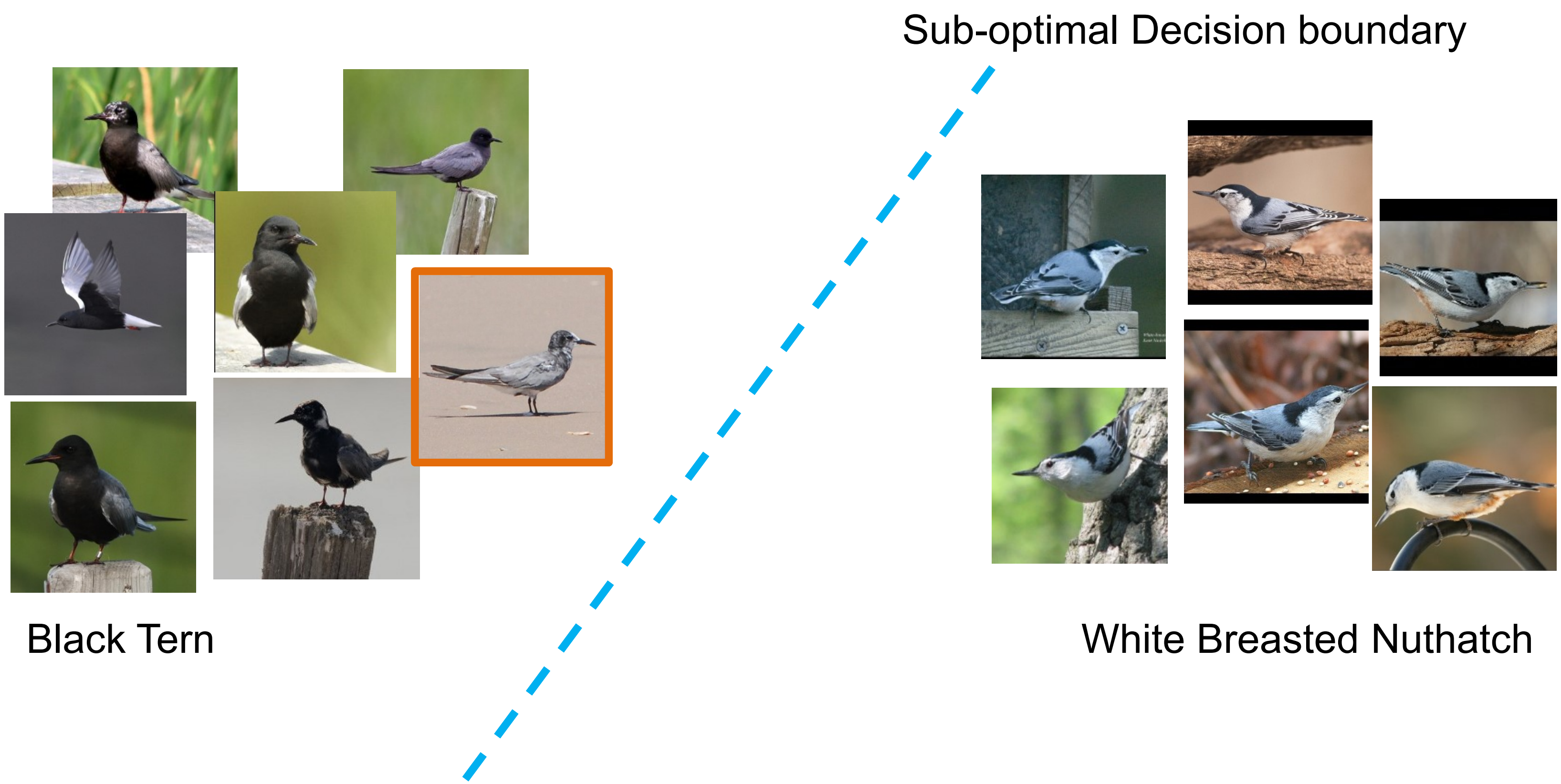


Training Dataset must be Representative.



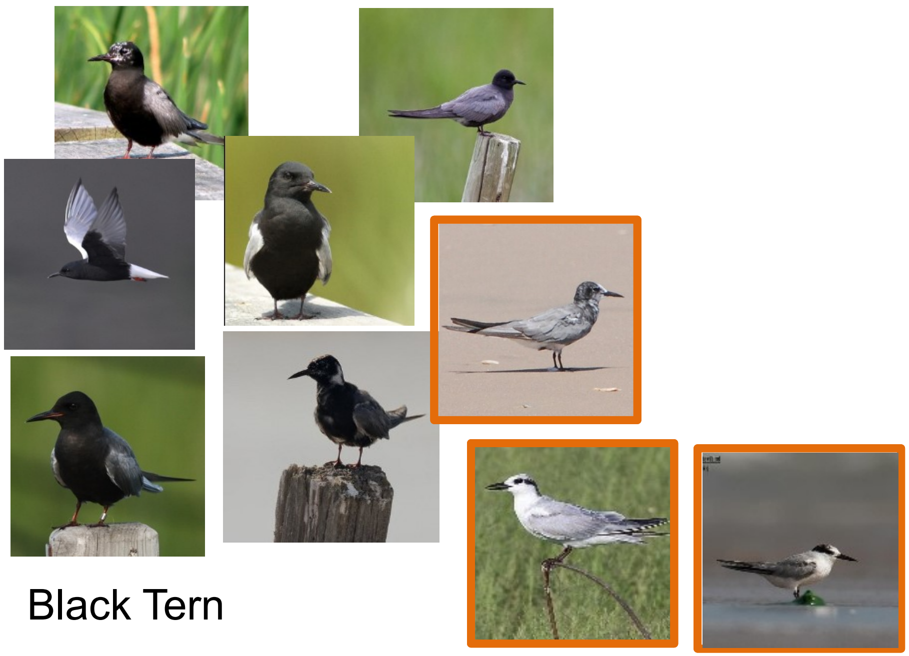
BRACE - BetteR Accuracy from Concept-based Explanation

Transformation-based Data Augmentation?

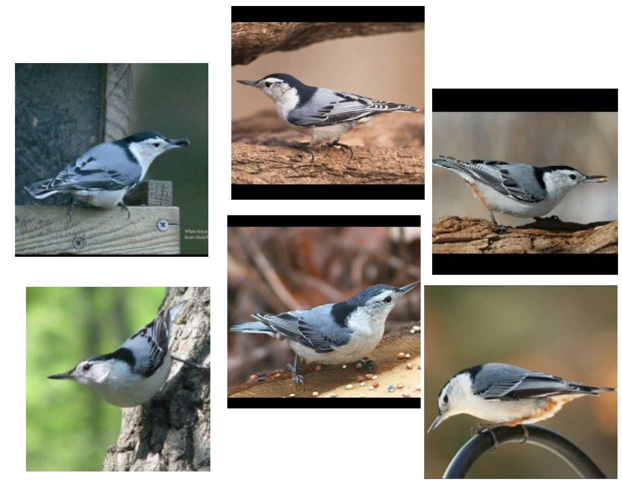


Only explore the neighborhood of existing samples and may not cover the under-represented regions.

Obtain Images from Image Repositories?



Black Tern

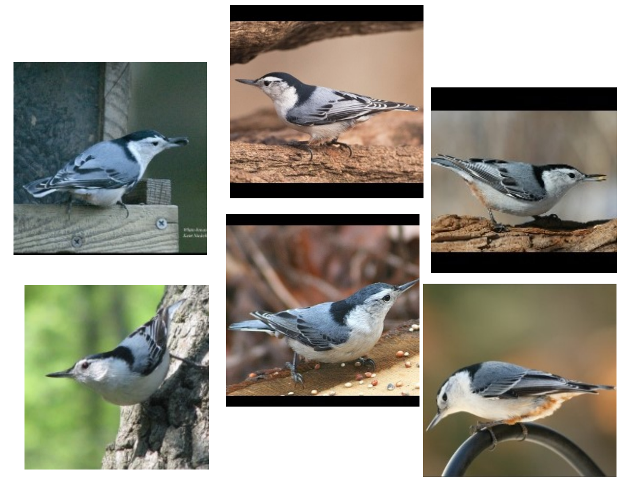


White Breasted Nuthatch

Obtain Images from Image Repositories?

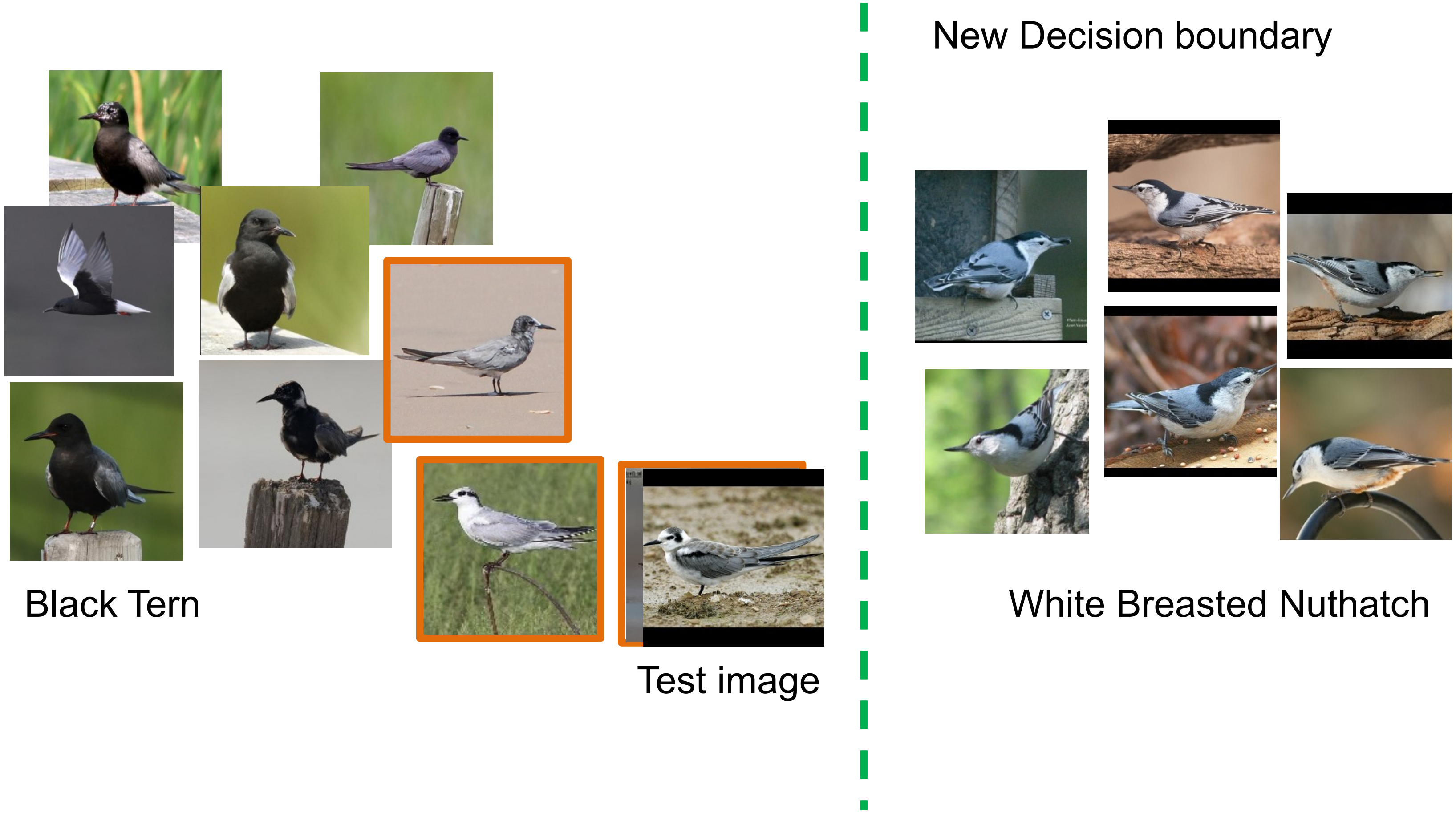
New Decision boundary

Black Tern



White Breasted Nuthatch

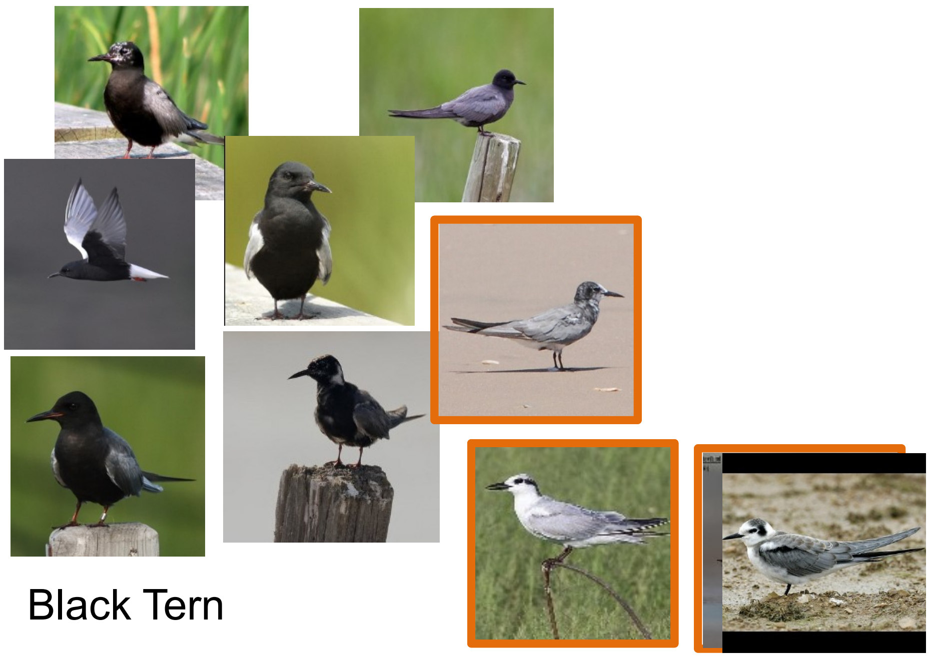
Obtain Images from Image Repositories?



BRACE - BetteR Accuracy from Concept-based Explanation

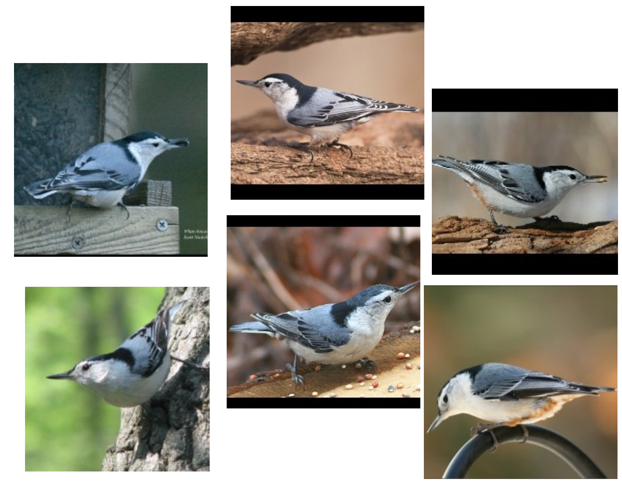
Obtain Images from Image Repositories?

New Decision boundary



Black Tern

Test image



White Breasted Nuthatch

✓ **Correct prediction**

Obtain Images from Image Repositories?

- Should not introduce noise
  - e.g., out-of-distribution images.

(a) Images obtained for Black Tern class

- Added images should be informative.

Explanations Provide Useful Insights!

- E.g., **CCNN**[1] misclassifies juvenile Black Tern in Fig (a) as White Breasted Nuthatch.
- **Concepts caused misclassification :** *White breast, White belly, Black crown*

(a) Juvenile Black Tern

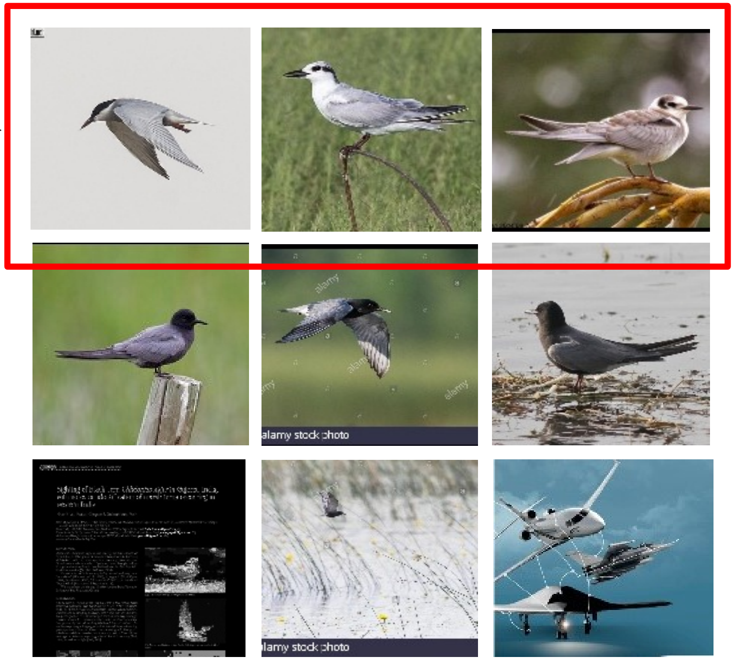
(b) Images of adult Black Tern

(c) Images of White Breasted Nuthatch

[1] Comprehensible Convolutional Neural Networks via Guided Concept Learning, IJCNN, 2021

Use Explanations to Identify Informative Images?

Informative images



(a) Images obtained for Black Tern class

- BRACE - **B**ette**R** **A**ccuracy from **C**oncept-based **E**xplanation

Assess Utility of New Images

- Suppose an image of c is misclassified into $\bar{c}$.
- **Is the new image from the under-represented region?**
  - Similar to the existing images of c.
  - Model's confidence that the new image belongs to a $\bar{c}$ is high.

$$\boldsymbol{\beta(x, c, \bar{c}) = \frac{f_x \cdot f_c}{\|f_x\| \, \|f_c\|} \times e^{P(\bar{c}|x)}}$$

x = new image f_x = visual features of x f_c = avg visual features of class c

Visual features are extracted using the classifier trained with original train dataset.

Assess Utility of New Images

- Does the new image contain concepts caused misclassifications?
  - Derive concepts caused misclassifications - $\boldsymbol{S_{c \rightarrow \bar{c}}}$
  - Computer degree of match between visual features in the new image and $\boldsymbol{S_{c \rightarrow \bar{c}}}$ - $\boldsymbol{\Delta(S_{c \rightarrow \bar{c}}, x)}$

- $\boldsymbol{utility} = \sum_{\bar{c} \in \bar{C}} [\boldsymbol{\beta(x, c, \bar{c})} \times \boldsymbol{\Delta(S_{c \rightarrow \bar{c}}, x)}]$ where $\bar{C} = \cup\, \bar{c}$

Which Explanation Methods?

- Concept-based explanation methods
- Post-hoc explanations or explanations from inherently interpretable models
- Post-hoc methods – GradCAM [1], ACE [2], IBD [3]
- Inherently interpretable models – CCNN [4], ProtoPNet [5]

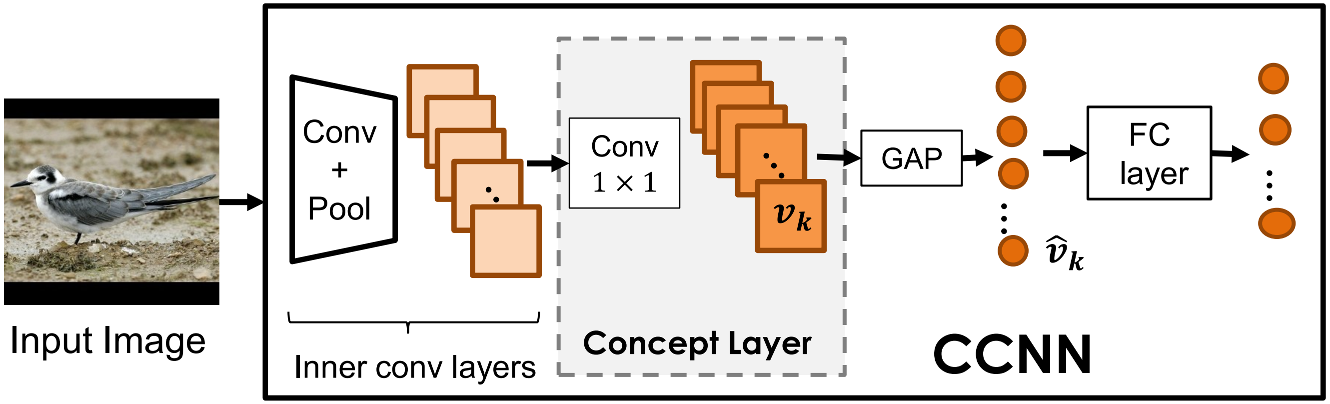
[1] Grad-cam: Visual explanations from deep networks via gradient-based localization, ICCV, 2017.
[2] Towards automatic concept-based explanations, NeurIPS, 2019.
[3] Interpretable basis decomposition for visual explanation, ECCV, 2018.
[4] Comprehensible Convolutional Neural Networks via Guided Concept Learning, IJCNN, 2021.
[5] This looks like that: deep learning for interpretable image recognition, NeurIPS, 2019.

Which Explanation Methods?

- Concept-based explanation methods
- Post-hoc explanations or explanations from inherently interpretable models
- Post-hoc methods – **GradCAM [1]**, ACE [2], IBD [3]
- Inherently interpretable models – **CCNN [4]**, ProtoPNet [5]

[1] Grad-cam: Visual explanations from deep networks via gradient-based localization, ICCV, 2017.
[2] Towards automatic concept-based explanations, NeurIPS, 2019.
[3] Interpretable basis decomposition for visual explanation, ECCV, 2018.
[4] Comprehensible Convolutional Neural Networks via Guided Concept Learning, IJCNN, 2021.
[5] This looks like that: deep learning for interpretable image recognition, NeurIPS, 2019.

BRACE – CCNN

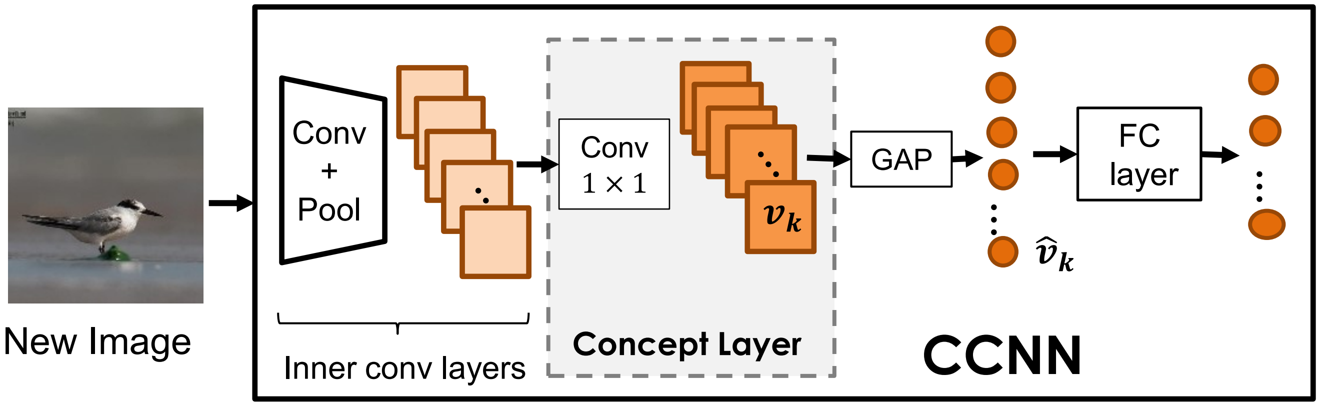


Decision: White Breasted Nuthatch
Contributed concepts:
White breast (0.50)
Black crown (0.30)
White belly (0.20)

- For each misclassified image obtained top-contributed concepts.
- Select top-r concepts that have contributed for highest number of misclassifications of class c - $S_{c \to \bar{c}}$

Comprehensible Convolutional Neural Networks via Guided Concept Learning, IJCNN, 2021.

BRACE – CCNN



- For each concept i in $S_{c \to \bar{c}}$, calculate degree of matching with a new image using the corresponding activation value of GAP layer in CCNN - a_i
- $\Delta(S_{c \to \bar{c}}, x) = \sum_{i=1}^{r} a_i$

Comprehensible Convolutional Neural Networks via Guided Concept Learning, IJCNN, 2021.

Performance Evaluation

- Source code - ***https://github.com/sandareka/BRACE***
- Datasets – CUB, CUB-Families, Tiny Imagenet
- Comparative study
  - Data augmentation methods
    - Cut -mix - Cutmix: Regularization strategy to train strong classifiers with localizable features, ICCV, 2019.
    - Snap-mix - Snapmix: Semantically proportional mixing for augmenting fine-grained data, AAAI, 2021.
    - WS-DAN - See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification, arXiv, 2019.
    - Metaset-based - Data-driven meta-set based fine-grained visual recognition, ACM-MM, 2020.
  - Sample selection methods
    - Random – samples are selected randomly.
    - Confidence – the most confident samples are selected.
    - Core-set – Active learning for convolutional neural networks: A core-set approach, ICLR, 2018.
    - L-loss - Learning loss for active learning, CVPR, 2019.

Comparison of Data Augmentation Methods

Performance of fully interpretable CCNN classifier based on ResNet-34.

Method	CUB	CUB-Families
Original dataset	84.3	83.8
Cut-mix	80.6	79.0
Snap-mix	82.4	79.9
WS-DAN	81.6	81.8
Metaset-based	85.1	88.1
BRACE	**86.0**	**88.7**

BRACE-augmented datasets achieve the highest accuracy.
The improvement is bigger in CUB-Families, where there are more under-represented regions.

Comparison of Sample Selection Methods

Performance of fully interpretable CCNN classifier based on ResNet-101.

Method	CUB	CUB-Families
Original dataset	86.6	85.7
Core-set	84.6	85.4
L-loss		
Random	87.0	88.0
Confidence	86.7	85.8
BRACE	**88.4**	**92.2**

BRACE-augmented datasets consistently achieves the highest accuracy.

Comparison of Samples Selected by Different Methods

Sample original images | Sample images of removed sub-categories | BRACE | Random | Confidence

Icteridae class of Cub-Families

- **BRACE selects images similar to those are in the removed subcategories.**
- **Random may select out-of-distribution images.**
- **Confidence selects images similar to those are in the original dataset.**

Performance Evaluation – Generalizability with BRACE

Comparison on generalizability of ResNet-34 trained with different data augmentation methods.

Method	NAbirds-Sub	ImageNet-V2-Sub
Original dataset	81.5	54.4
Cut-mix	72.1	37.2
Snap-mix	76.0	42.6
WS-DAN	67.0	56.1
Metaset-based	83.5	42.4
BRACE	**84.9**	**70.0**

BRACE enables the classifier to learn features that are generalizable to handle more diverse images.

Thank You