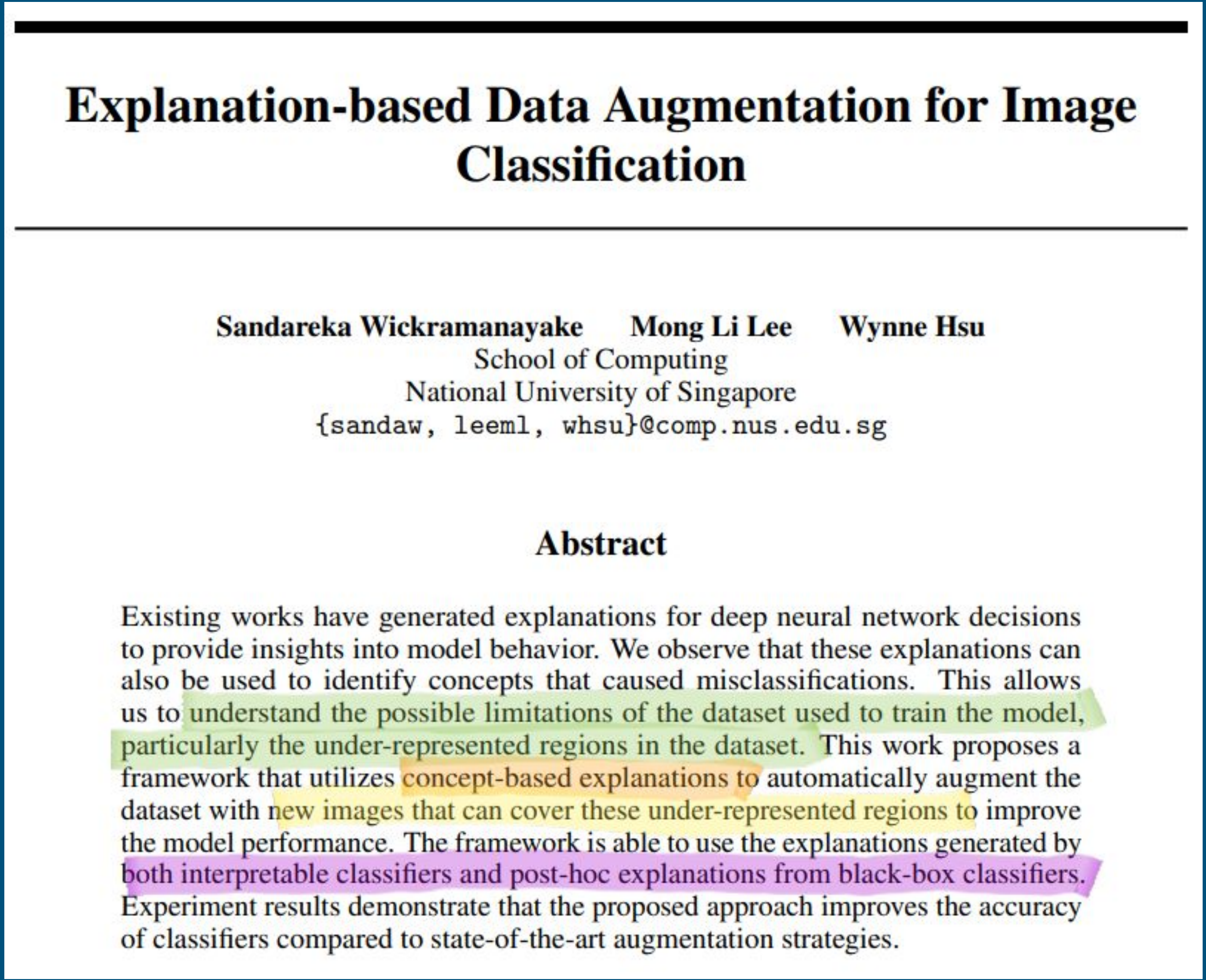
Explanation-based Data Augmentation for Image Classification

Beata Baczyńska

Table of contents

- Abstract
- Data augmentation
- Paper’s solution - BRACE
- Utility function
- BRACE on CCNN
- BRACE on GradCAM
- Datasets
- Evaluation - data augmentation methods
- Evaluation - sample selection methods
- Conclusion and Critic



Explanation-based Data Augmentation for Image Classification

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

Abstract

Existing works have generated explanations for deep neural network decisions to provide insights into model behavior. We observe that these explanations can also be used to identify concepts that caused misclassifications. This allows us to understand the possible limitations of the dataset used to train the model, particularly the under-represented regions in the dataset. This work proposes a framework that utilizes concept-based explanations to automatically augment the dataset with new images that can cover these under-represented regions to improve the model performance. The framework is able to use the explanations generated by both interpretable classifiers and post-hoc explanations from black-box classifiers. Experiment results demonstrate that the proposed approach improves the accuracy of classifiers compared to state-of-the-art augmentation strategies.

Data augmentation

Common approaches and limitations:

I. Approach: to apply transformations to training samples or mix existing samples [1] [2]

   Limitation: may not cover the under-represented regions

II. Approach: to obtain new random images from image resources based on label

   Limitation: may introduce out-of-distribution images

[1] X. W. Shaoli Huang and D. Tao, "Snapmix: Semantically proportional mixing for augmenting fine-grained data," in AAAI, 2021.
[2] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in ICCV, 2019.

Paper’s solution

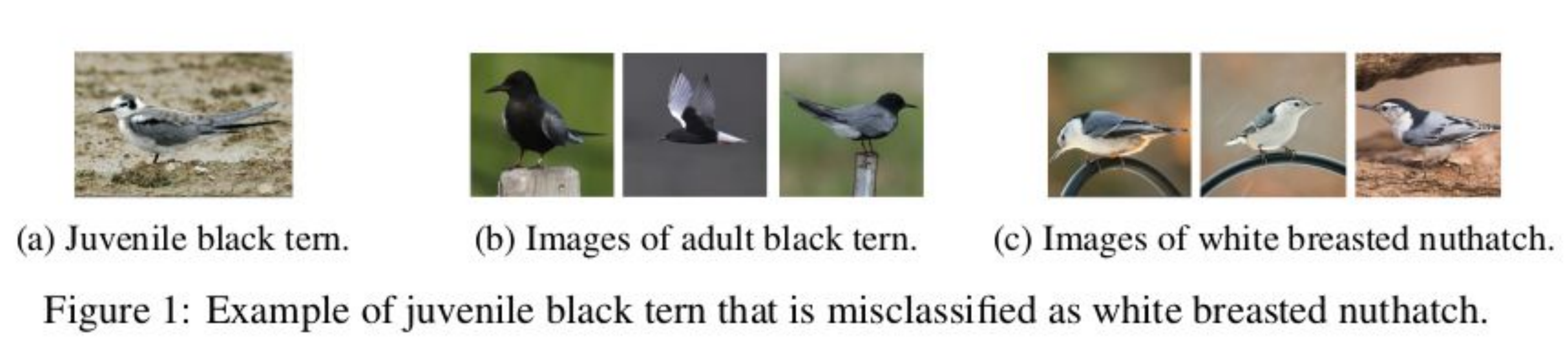
Use model decision explanations to facilitate the search for new samples from under-represented regions

BRACE
(BetteR Accuracy from Concept-based Explanation)

Key idea:

to identify concepts that have led to misclassifications and augment the dataset with images that contain these concepts

Example



(a) Juvenile black tern.

(b) Images of adult black tern.

(c) Images of white breasted nuthatch.

Figure 1: Example of juvenile black tern that is misclassified as white breasted nuthatch.

Concepts caused misclassification:

white breast, white belly, black crown

Utility function

We say that x is an under-represented sample if it has visual features of class c, but the model's confidence that x belongs to a different class is high.

Under-representation score ->

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

fx - feature vector of x
fc - the average feature vector of all the images in class c
P(c̄|x) - the predicted probability of x belonging to class c̄

A high β value indicates that x is similar to some images in class c, but the model has classified it as class c̄

Utility function

The utility score of a sample x ->

$$utility(x) = \sum_{\bar{c}\in\bar{C}} [\beta(x, c, \bar{c}) \times \Delta(\mathcal{S}_{c\to\bar{c}}, x))]$$

Sc→c̄ - set of concepts that caused M to misclassify images of class c into c̄

Δ(Sc→c̄ , x) - the function that computes the degree of match between the visual features in an image x and the concepts in S c→c̄

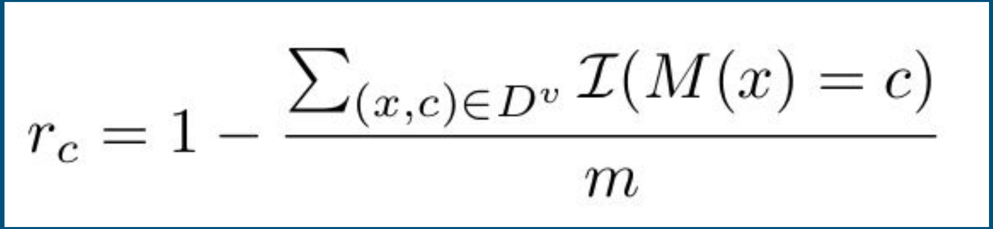
The appropriate number of samples with the highest utility scores are added to the original train dataset. This process is repeated for all the classes.

The appropriate number of samples that are added

The |D_c| × r_c samples are added.

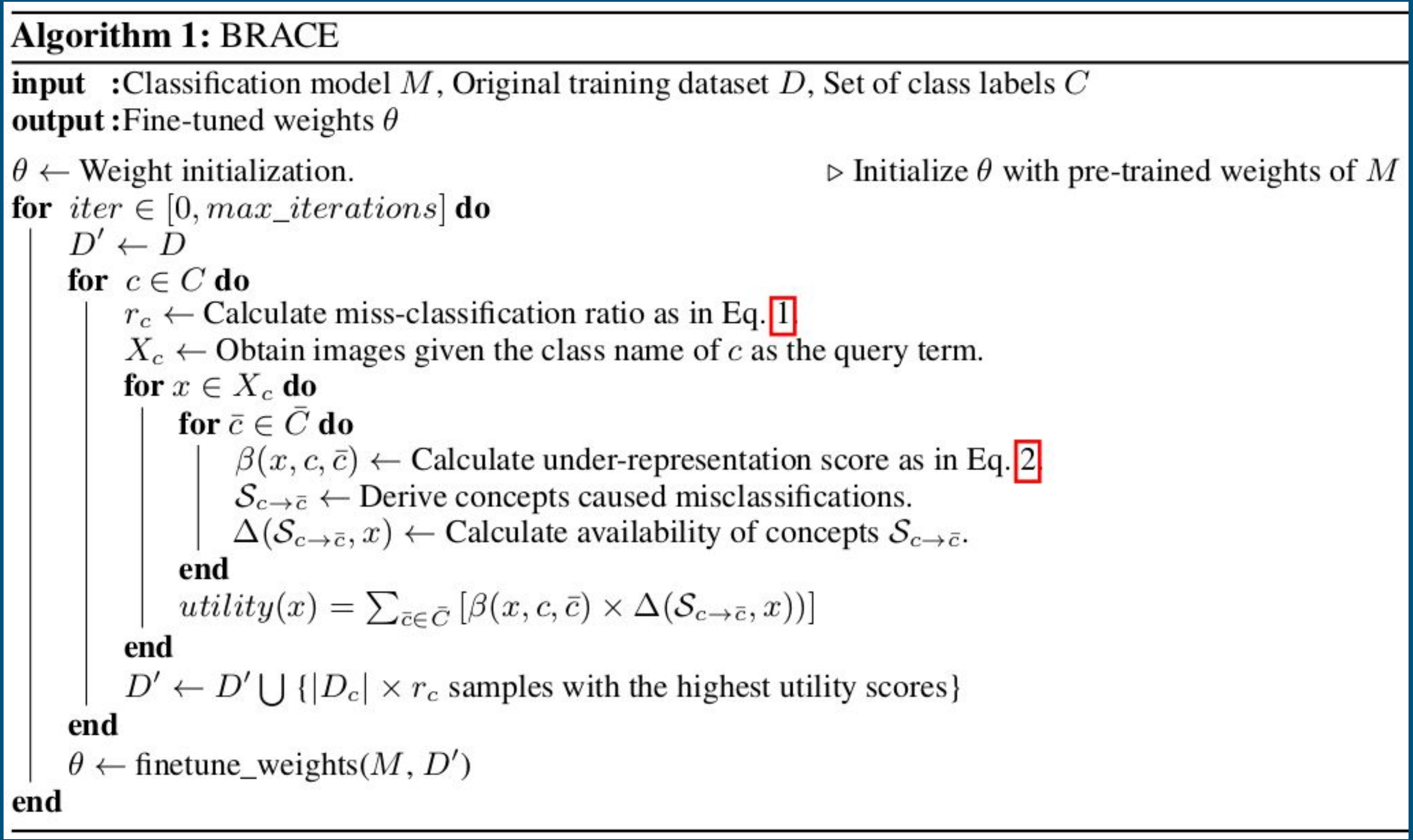
Where D_c is the number of images with label c.

The ratio of misclassifications->



$$r_c = 1 - \frac{\sum_{(x,c)\in D^v} \mathcal{I}(M(x) = c)}{m}$$

m - the number of validation samples with class label c
I - a function that returns 1 when its argument is true and 0 otherwise
Dv - validation dataset



Algorithm 1: BRACE

```
input  :Classification model $M$, Original training dataset $D$, Set of class labels $C$
output :Fine-tuned weights $\theta$

$\theta \leftarrow$ Weight initialization.        ▷ Initialize $\theta$ with pre-trained weights of $M$
for $iter \in [0, max\_iterations]$ do
    $D' \leftarrow D$
    for $c \in C$ do
        $r_c \leftarrow$ Calculate miss-classification ratio as in Eq. 1.
        $X_c \leftarrow$ Obtain images given the class name of $c$ as the query term.
        for $x \in X_c$ do
            for $\bar{c} \in \bar{C}$ do
                $\beta(x, c, \bar{c}) \leftarrow$ Calculate under-representation score as in Eq. 2.
                $\mathcal{S}_{c \rightarrow \bar{c}} \leftarrow$ Derive concepts caused misclassifications.
                $\Delta(\mathcal{S}_{c \rightarrow \bar{c}}, x) \leftarrow$ Calculate availability of concepts $\mathcal{S}_{c \rightarrow \bar{c}}$.
            end
            $utility(x) = \sum_{\bar{c} \in \bar{C}} [\beta(x, c, \bar{c}) \times \Delta(\mathcal{S}_{c \rightarrow \bar{c}}, x))]$
        end
        $D' \leftarrow D' \bigcup \{|D_c| \times r_c$ samples with the highest utility scores$\}$
    end
    $\theta \leftarrow$ finetune_weights($M$, $D'$)
end
```

About $S_{c \to \bar{c}}$ and $\Delta(S_{c \to \bar{c}}, x)$

The BRACE framework can be used for concept-based explanation methods such as:

- the fully interpretable classifier with linguistic explanation (eg. CCNN)
- the black-box classifier with post-hoc explanation using heat maps

(eg. GradCAM)

CCNN (Comprehensible Convolutional Neural Network)

The Comprehensible Convolutional Neural Network (CCNN) introduces an additional concept layer to the CNN-based architecture to guide the learning of the associations between visual features and word phrases extracted from image descriptions.

The CCNN explains its decisions in word phrases corresponding to different visual concepts.

S. Wickramanayake, W. Hsu, and M. L. Lee, “Comprehensible convolutional neural networks via guided concept learning,” in IJCNN, 2021.

CCNN (Comprehensible Convolutional Neural Network)

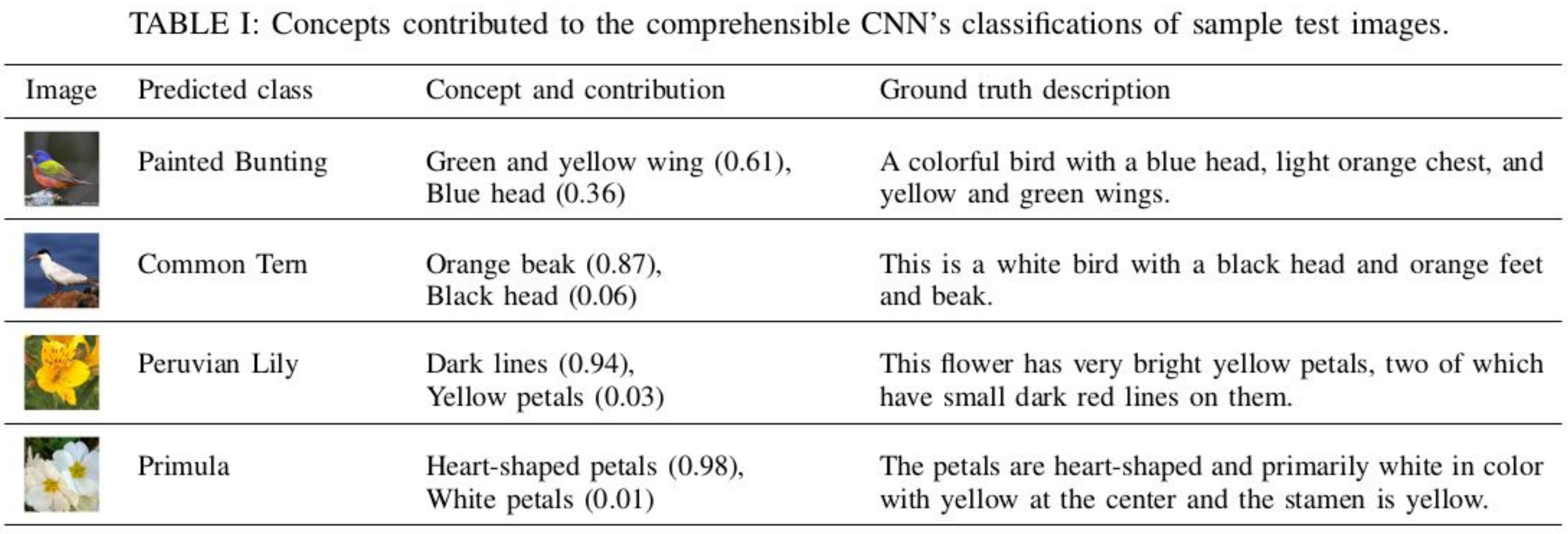


TABLE I: Concepts contributed to the comprehensible CNN's classifications of sample test images.

Image	Predicted class	Concept and contribution	Ground truth description
	Painted Bunting	Green and yellow wing (0.61), Blue head (0.36)	A colorful bird with a blue head, light orange chest, and yellow and green wings.
	Common Tern	Orange beak (0.87), Black head (0.06)	This is a white bird with a black head and orange feet and beak.
	Peruvian Lily	Dark lines (0.94), Yellow petals (0.03)	This flower has very bright yellow petals, two of which have small dark red lines on them.
	Primula	Heart-shaped petals (0.98), White petals (0.01)	The petals are heart-shaped and primarily white in color with yellow at the center and the stamen is yellow.

Formulas for CCNN

Given X c→c̄ , the set of images with class label c that have been misclassified as c̄, we compute the number of images in X c→c̄ that contains the concept i. The top k concepts with the highest number form the set S c→c̄

Given the set of concepts S c→c̄ , we utilize their corresponding visual feature activations to determine the degree of match of a concept i in an image x.

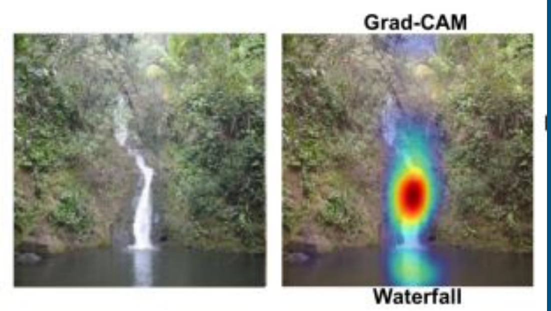
$$\alpha_i = -\log\left[1 - \frac{1}{1 + e^{-\bar{o}_i}} + \epsilon\right]$$

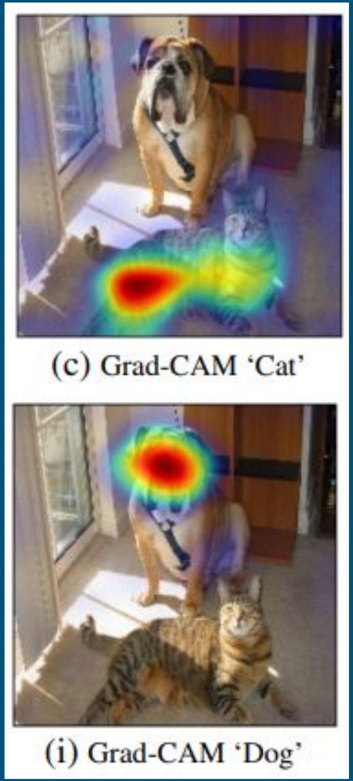
With this, the degree of match between the set S c→c̄ and image x is given by:

$$\Delta(\mathcal{S}_{c \to \bar{c}}, x) = \sum_{i=1}^{k} \alpha_i$$

GradCAM (Gradient-weighted Class Activation Mapping)

“GradCAM is an image-level post-hoc explanation mechanism. GradCAM uses the gradients of any target concept flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept.”




(c) Grad-CAM ‘Cat’

(i) Grad-CAM ‘Dog’

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in ICCV, 2017

Formulas for GradCAM

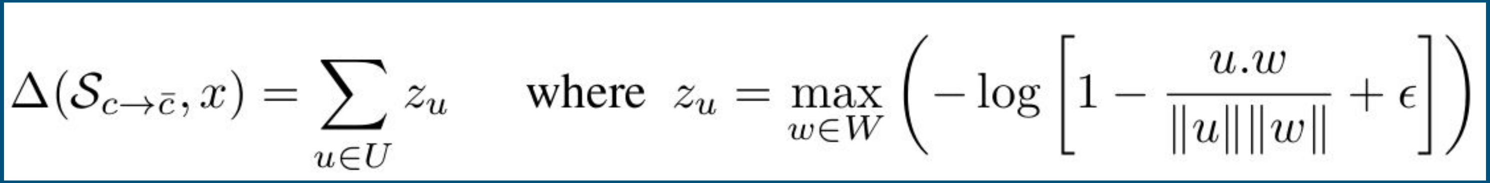
BRACE normalize the pixels in the saliency map to have values between 0 and 1, and regions whose values are greater than the threshold of 0.5 correspond to semantic concepts that are responsible for the model decision.

Sc→c̄ is the set of regions in Xc→c̄ obtained from the saliency maps of Grad-CAM.

The regions in S c→c̄ are passed through M , and the output of the last layer before the classification layer form their visual features, denoted as U .

Given an image x, we use a state-of-the-art RCNN model [1] to derive salient region proposals that correspond to an object or object part. These region proposals are passed through M to obtain the corresponding visual features denoted as W .

The degree of match between Sc→c̄ , and x is given by the similarity between visual features in U and W .



$$\Delta(\mathcal{S}_{c\rightarrow\bar{c}}, x) = \sum_{u\in U} z_u \qquad \text{where} \;\; z_u = \max_{w\in W}\left(-\log\left[1 - \frac{u.w}{\|u\|\,\|w\|} + \epsilon\right]\right)$$

S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” NeurIPS, 2015.

Evaluation

Datasets:

- CUB

  (11,788 images of birds belonging to 200 classes)

- CUB-Families

  (This dataset groups the 200 species of birds in CUB into 37 families comprised of multiple bird species)

- Tiny ImageNet

  (Tiny ImageNet contains 110,000 images with 200 classes)

Evaluation

Datasets:

- CUB

  (11,788 images of birds belonging to 200 classes. The dataset is divided into a train set of 3994 images, a validation set of 2000 images, and a test set of 5794 images. Each image has ten sentences describing the bird)

- CUB-Families

  (This dataset groups the 200 species of birds in CUB into 37 families comprised of multiple bird species. We create under-represented regions by removing some species in the test split of [30] from each family in the training dataset. No species are removed in the validation and test sets. The resultant dataset contains 4585 training images, 2343 validation images, and 4860 test images)

- Tiny ImageNet

  (Tiny ImageNet contains 110,000 images with 200 classes. In the standard dataset split, each class has 500 training images (400 for training the model and 100 for tuning the hyper-parameters) and 50 validation images. We report our results on the validation set)

Evaluation

Comparative studies: data augmentation methods:

- Cut-mix [1]

  This mixed-based data augmentation method cuts out one image patch, pastes it on another image, and mixes their labels according to the area proportion

- Snap-mix [2]

  This is another mixed-based data augmentation method. It combines images similar to CutMix but uses the semantic composition of the resultant image to derive the label

- WS-DAN [3]

  WS-DAN is an attention-based approach that uses attention-cropping and attention-dropping. Attention-cropping refers to cropping the region of the image attended by the model to create a new sample. Attention-dropping refers to erasing the attended region to create a new sample

[1] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, “Cutmix: Regularization strategy to train strong classifiers with localizable features,” in ICCV, 2019.
[2] X. W. Shaoli Huang and D. Tao, “Snapmix: Semantically proportional mixing for augmenting fine-grained data,” in AAAI, 2021.
[3] T. Hu, H. Qi, Q. Huang, and Y. Lu, “See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification,” arXiv preprint arXiv:1901.09891, 2019.

Evaluation

Comparative studies: data augmentation methods:

- Part-based [1]

  This method obtains samples from image repositories by using object bounding boxes and part landmark annotations

- Metaset-based [2]

  This method also obtains samples from image repositories. It trains two models to ensure the samples are in-distribution and to correct noisy labels

[1] Z. Xu, S. Huang, Y. Zhang, and D. Tao, "Webly-supervised fine-grained visual categorization via deep domain adaptation," IEEE transactions on pattern analysis and machine intelligence, 2016.
[2] C. Zhang, Y. Yao, X. Shu, Z. Li, Z. Tang, and Q. Wu, "Data-driven meta-set based fine-grained visual recognition," in ACM-MM, 2020.

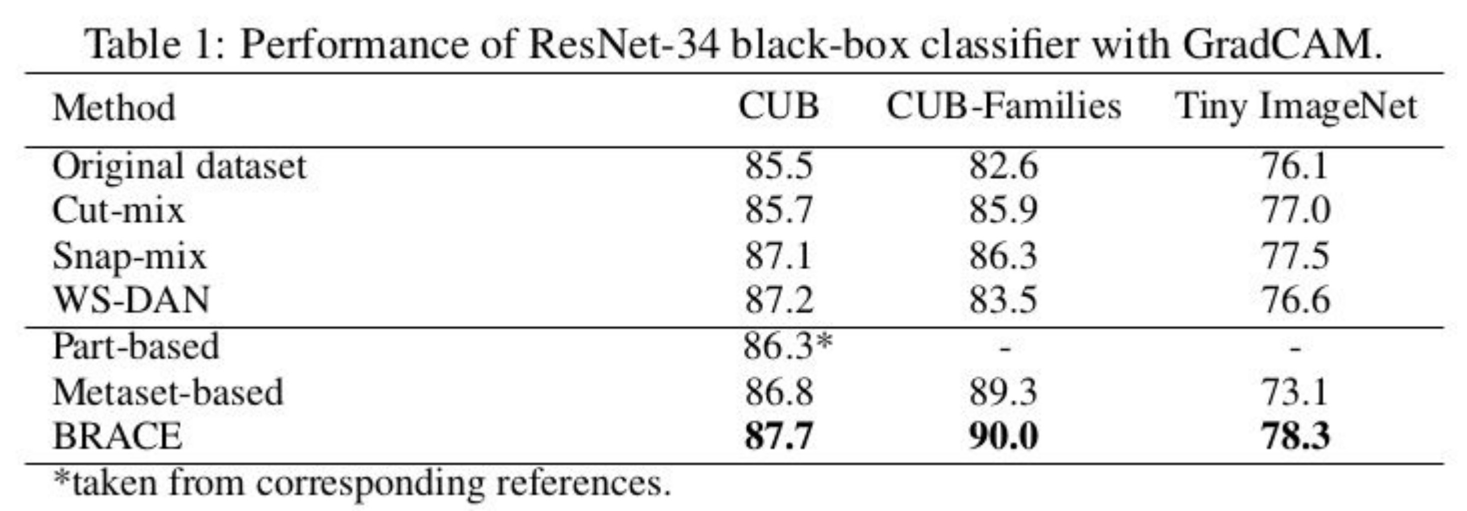


Table 1: Performance of ResNet-34 black-box classifier with GradCAM.

Method	CUB	CUB-Families	Tiny ImageNet
Original dataset	85.5	82.6	76.1
Cut-mix	85.7	85.9	77.0
Snap-mix	87.1	86.3	77.5
WS-DAN	87.2	83.5	76.6
Part-based	86.3*	-	-
Metaset-based	86.8	89.3	73.1
BRACE	**87.7**	**90.0**	**78.3**

*taken from corresponding references.

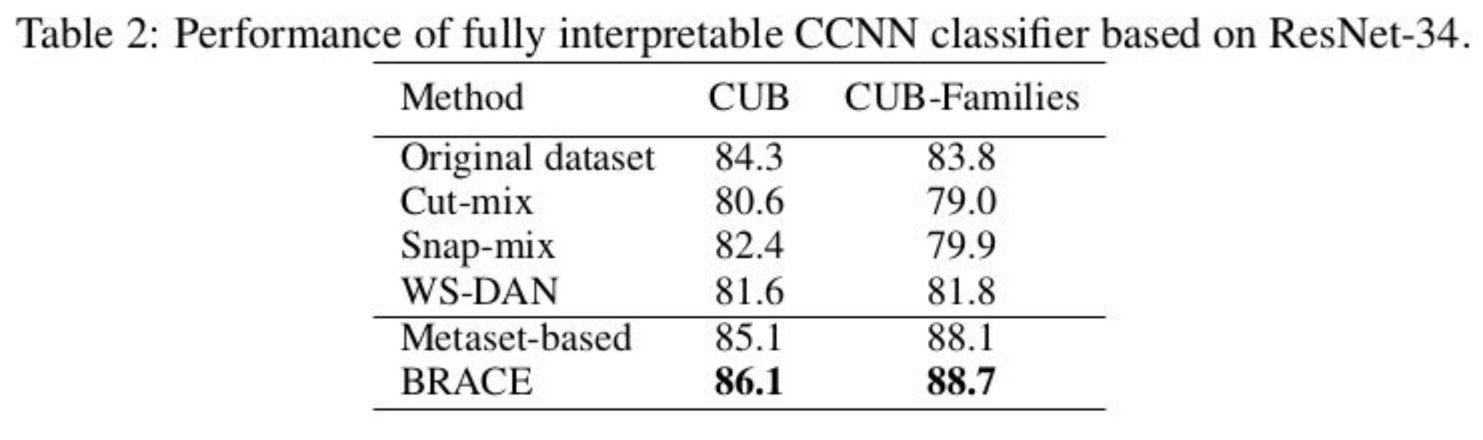


Table 2: Performance of fully interpretable CCNN classifier based on ResNet-34.

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	**86.1**	**88.7**

Evaluation

Comparative studies: sample selection methods:

- Random
- Confidence

  The samples obtained for each class are passed through the classifier. We rank the samples based on the classifier’s scores, and the top scored samples are selected

- Core-set

  k representative samples are selected using the K-Center-Greedy algorithm

- L-loss

  The classification loss of a sample is predicted, and samples with the top-k losses are returned

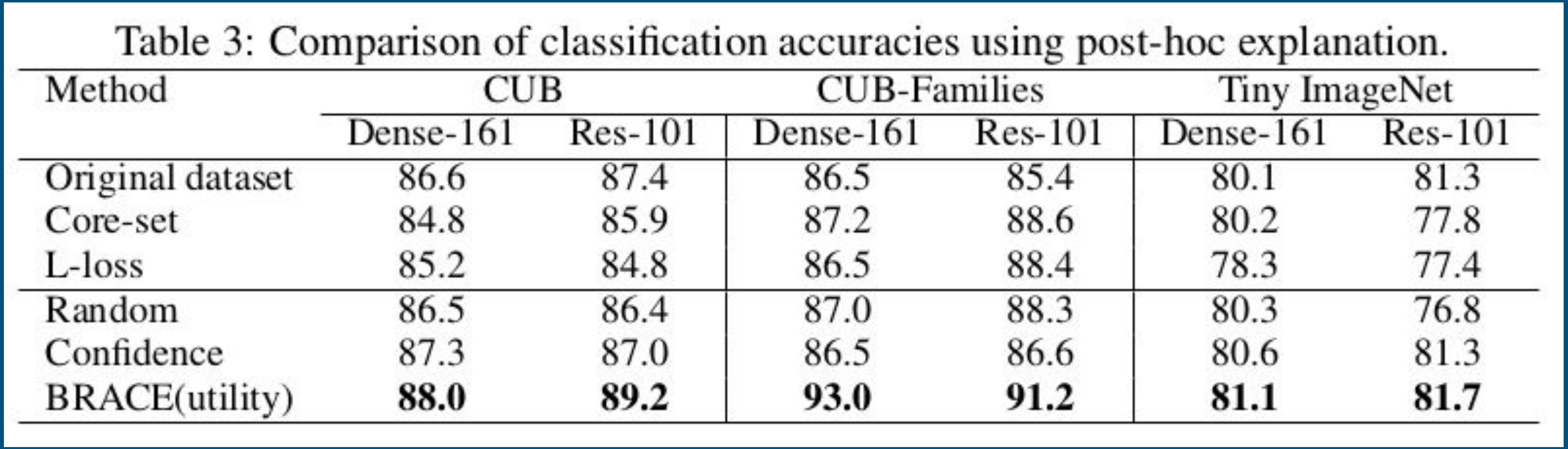


Table 3: Comparison of classification accuracies using post-hoc explanation.

Method	CUB		CUB-Families		Tiny ImageNet	
	Dense-161	Res-101	Dense-161	Res-101	Dense-161	Res-101
Original dataset	86.6	87.4	86.5	85.4	80.1	81.3
Core-set	84.8	85.9	87.2	88.6	80.2	77.8
L-loss	85.2	84.8	86.5	88.4	78.3	77.4
Random	86.5	86.4	87.0	88.3	80.3	76.8
Confidence	87.3	87.0	86.5	86.6	80.6	81.3
BRACE(utility)	**88.0**	**89.2**	**93.0**	**91.2**	**81.1**	**81.7**

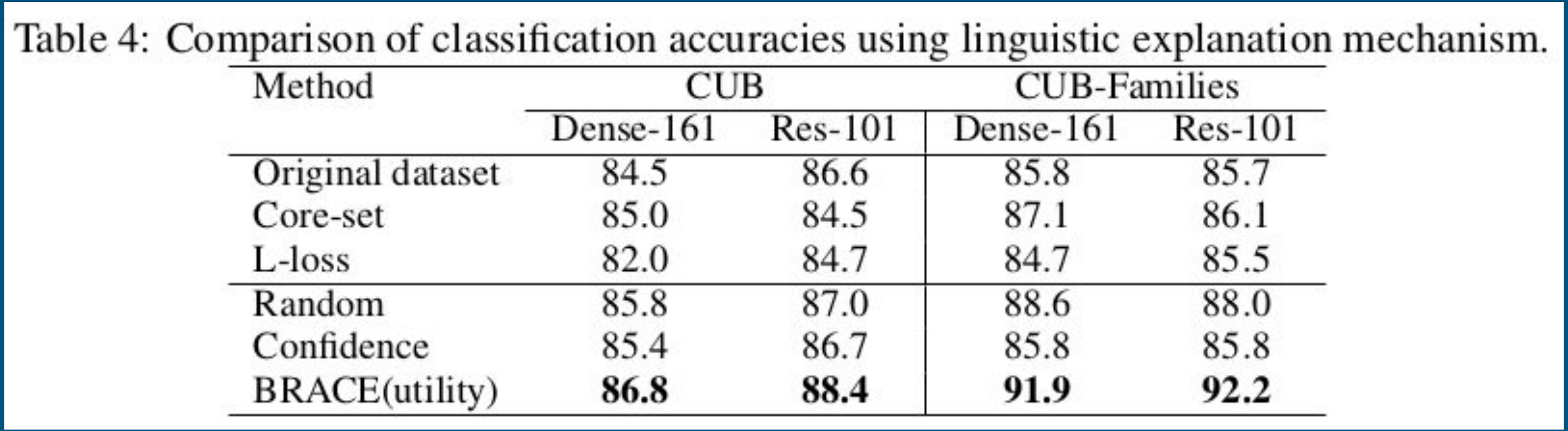


Table 4: Comparison of classification accuracies using linguistic explanation mechanism.

Method	CUB		CUB-Families	
	Dense-161	Res-101	Dense-161	Res-101
Original dataset	84.5	86.6	85.8	85.7
Core-set	85.0	84.5	87.1	86.1
L-loss	82.0	84.7	84.7	85.5
Random	85.8	87.0	88.6	88.0
Confidence	85.4	86.7	85.8	85.8
BRACE(utility)	**86.8**	**88.4**	**91.9**	**92.2**

Conclusion and Critic

Advantages:

- The augmentation method seems to cure the root cause of a weak performance of the model
- This method can be also used with GradCAM and other post-hoc explanations with heat maps

Disadvantages:

- The utility function must be adapt depending on the explanation method
- For approach with post-hoc explanations using heat maps the second model (RCNN) must be able to detect object we are interested in