WWW: A Unified Framework for Explaining What, Where and Why of Neural Networks by Interpretation of Neuron Concepts

Supplementary Material

A. Implementation Details

In this section, we discuss implementation details for the experiment.

A.1. ResNet-50 trained on ImageNet

We use ResNet-50 of ImageNet-1k pre-trained weight on the torchvision 0.14.0 version.

WWW (Ours). We select 40 high-activating images and 40 high-activating crop images for major and minor concepts for each neuron, respectively. Adaptive threshold α is set to 0.95 for major concept selection and α is set to 0.90 for minor concept selection.

CLIP-Dissect [20]. We implemented from the official GitHub code (https://github.com/Trustworthy-ML-Lab/CLIP-dissect). We only change pre-trained model weight.

MILAN(b) [12]. We implemented from the official GitHub code (https://github.com/evandez/neuron-descriptions). We only change pre-trained model weight.

FALCON [14]. When matching neuron concepts with official FALCON threshold, the concepts are not matched to all final layer neurons. So, we modified the threshold to 0.35 and implemented it to match concepts to most neurons in the ResNet-50 final layer. Additionally, to evaluate concept matching performance for each neuron, the official FALCON of matching concepts to a group of neurons was modified and implemented to match concepts only to each neuron.

A.2. ViT-B/16 trained on ImageNet

We use the pre-trained weight of ViT B/16 on the ImageNet dataset in the timm 0.9.7 version.

WWW (Ours). We select 40 high-activating images and 40 high-activating crop images for major and minor concepts for each neuron, respectively. Adaptive threshold α is set to 0.95 for major concept selection and α is set to 0.90 for minor concept selection.

CLIP-Dissect [20]. We implemented from the official GitHub code (https://github.com/Trustworthy-ML-Lab/CLIP-dissect). We only change pre-trained model weight.

A.3. ResNet-18 trained on Places365

We used the same ResNet-18 pre-trained in places365 dataset weight used in CLIP-Dissect official implementation (https://github.com/Trustworthy-ML-Lab/CLIP-dissect).

WWW (Ours). We select 40 high-activating images and 40 high-activating crop images for major and minor concepts for each neuron, respectively. Adaptive threshold α is set to 0.95 for major concept selection and α is set to 0.90 for minor concept selection.

CLIP-Dissect [20]. We implemented from the official GitHub code (https://github.com/Trustworthy-ML-Lab/CLIP-dissect).

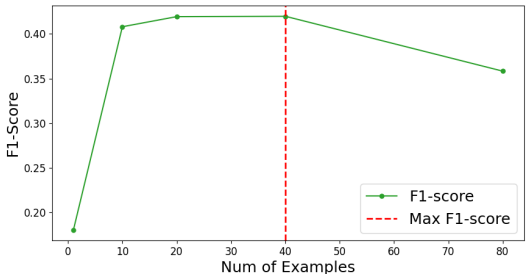

Figure S.1. **Ablation study on the number of selected examples.** We displayed F1-score change regarding the number of selected example images for each neuron. The red line refers to the point that maximizes F1 score.

B. Ablation Studies

B.1. Ablation study on the Number of High Activating Images

In this section, we are going to discuss the effect of changing the numbers of selected example-based representations(i.e., high-activating samples).

Implementation Details. We used ImageNet-1k pre-trained ResNet-50, ImageNet Validation set as D_{probe}, and Wordnet nouns as $D_{concept}$. For hyperparameter settings, We used the same settings in A.1

Results. In figure S.1, we show the relation between the number of selected images and performance. With the small number of examples (i.e., $k = 1$), WWW shows low performance. That is because, in one single example image, there are dozens of different concepts which is not related to the neuron representation. But with a sufficient number of examples, the F1-score showed the best performance. At a large number of examples (i.e., $k = 80$), WWW showed decreased performance due to the less similar concept images.

Table S.1. **Ablation on concept selection methods.** We compared Cosine similarity, L1, L2, and WWW (ACS + AT) and their performance on four different metrics. We used ResNet-50, ImageNet validation(D_{probe}), Wordnet nouns($D_{concept}$). **Bold** numbers represent the best scores between the same settings. The average score and standard errors of the 1000 final layer neurons are reported.

Method	CLIP cos	mpnet cos	F1-score
Cosine	0.8499 ± 0.003	0.6123 ± 0.007	0.3265 ± 0.009
L1	0.8497 ± 0.003	0.6033 ± 0.008	0.2644 ± 0.013
L2	0.6331 ± 0.006	0.3861 ± 0.008	0.1112 ± 0.009
WWW (Ours)	$\mathbf{0.8858 \pm 0.003}$	$\mathbf{0.6945 \pm 0.008}$	$\mathbf{0.4197 \pm 0.012}$

B.2. Ablation Study on the ACS to Other Baselines

In this section, we compared four different methods for selecting concepts: Cosine similarity, L1, L2, and WWW (ACS + AT). Cosine similarity refers to the concept that has the highest cosine similarity with images selected as a concept. L1 and L2 refer to selecting a concept that has minimal L1 and L2 errors with images, respectively. For WWW, we used Adaptive Cosine Similarity (ACS) and Adaptive Thresholding (AT) for concept selection.

Implementation Details. We used ImageNet-1k pre-trained ResNet-50, ImageNet Validation set as D_{probe}, and Wordnet nouns as $D_{concept}$. For hyperparameter settings, We used the same settings in A.1

Results. In table S.1, WWW shows the best performance, except for the Hit rate, compared to other baselines.

C. Explanation Case Analysis

In this section, we displayed sample cases for explanation generated by WWW. We showed two additional failure cases in figure S.2 and figure S.3. In both cases, not only do the selected important neurons differ between the class and sample explanations, but the cosine similarity between their respective heatmaps is relatively low. Even though the ground-truth class explanations and the sample explanations highlight different important neurons, they both localize to similar regions in the heatmap, showing a relatively high similarity score in both cases.

Implemenetation Details. We used ImageNet-1k pre-trained ResNet-50, ImageNet Validation set as D_{probe}, and Wordnet nouns as $D_{concept}$. For hyperparameter settings, We used the same settings in A.1

D. Qualitative Results

In section D.1, we displayed concept selection results on different layers of ResNet-50. In section D.2, we displayed concept selection results on the final layer of ViT/B-16 model.

Implemenetation Details. For the ResNet-50 qualitative result, we used the same settings as Section A.1. We used ImageNet-1k pre-trained ResNet-50, ImageNet Validation set as D_{probe}, and Wordnet nouns as $D_{concept}$. For the ViT/b-16 qualitative result, we used the same settings as Section A.2. We also used the ImageNet pre-trained ViT/B-16 model, ImageNet Validation set as D_{probe}, and Wordnet nouns as $D_{concept}$.

D.1. ResNet-50

In this section, we displayed the qualitative results of WWW on ResNet-50. In figure S.4 and figure S.5 shows examples of descriptions for hidden neurons in the penultimate and final layers. Neurons in the penultimate layer are top-2 important neurons of the final layer neuron's ground truth label class. We observed that WWW not only interpreted each neuron well but also showed robust interpretation that the most important neuron of the class in the penultimate layer represents the exact same major concept as the final layer neuron.

In figure S.6 and figure S.7, we showed each neuron's 5 highest activating examples with the ground truth label. We have colored the descriptions green if they match the images, yellow if they match but are too generic or similar, and red if they do not match.

D.2. ViT/b-16

In this section, we displayed the qualitative results of WWW on Vision transformers. We showed each neuron's five highest activating examples with the ground truth label. We have colored the descriptions green if they match the images, yellow if they match but are too generic or similar, and red if they do not match. In figure S.8, WWW exactly matches the ground truth label, outperforming the other baseline.

E. t-tests in tables

We conducted paired t-tests to measure the statistical significance of the differences. Over Tables 2 and 3, the performance gap between WWW and the most comparable method (i.e., CLIP-Dissect) is statistically significant for both metrics of CLIP cos ($p < 0.001$) and F1-score ($p < 0.001$) in settings of D_{probe} = ImageNet validation with $D_{concept}$ = Wordnet (80k). In Table 4, WWW significantly improves CLIP cos compared to CLIP-Dissect in the setting of D_{probe} = ImageNet validation with $D_{concept}$ = Wordnet 80k ($p = 0.018$).

F. Heatmap similarity and misprediction detection based on uncertainty measure

In misprediction detection, our approach is to detect uncertain samples (i.e., samples with low heatmap similarity), which can be a highly potential failure case. We conducted

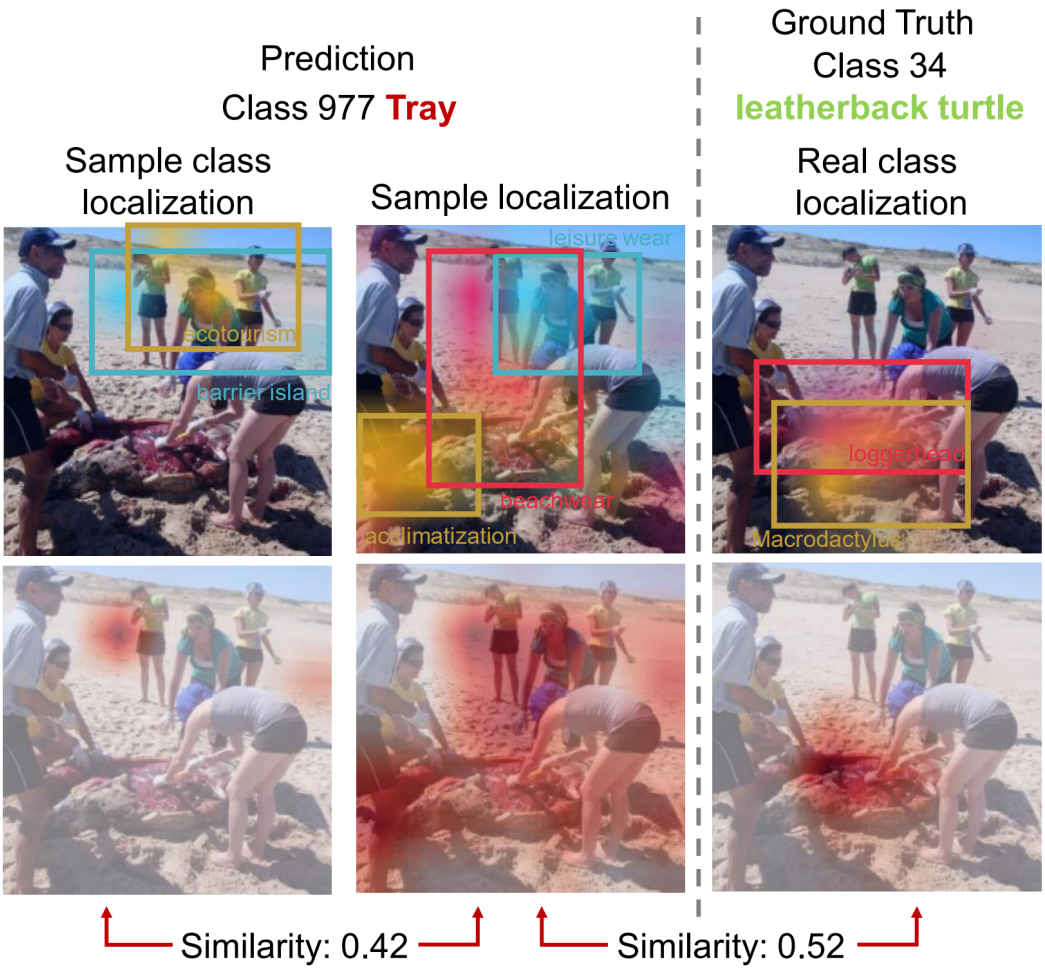

Figure S.2. **Example of failure case explanation by WWW.** The explanations of the predicted label are presented on the left side. On the right side, the explanation of the ground-truth label is shown. We displayed related regions of the concept as bounding boxes in the order of blue, red, and yellow boxes. (blue represents the most important concept) At the bottom, we showed the similarity between the two heatmaps.

Table S.2. **Experiment on AUROC of MSP and our method for mis-prediction detection.**

Method	AUROC
MSP	0.808
WWW (Ours)	**0.903**

a 'large-scale' experiment to quantify the quality of the Reasoning module. To show the distribution of the similarities across correct predictions and mispredictions, we calculated AUROC on both MSP and WWW (i.e., heatmap similarity) for the binary classification task of misprediction based on estimated uncertainty. We used the ImageNet pre-trained ResNet-50 model and ImageNet validation set. As in Table S.2, WWW shows outstanding performance compared to the MSP. This result indicates that WWW can be used as a misprediction detector.