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

Yong Hyun Ahn, [1], Hyeon Bae Kim, [1], Seong Tae Kim[2*]
[1]Department of Artificial Intelligence, Kyung Hee University, Republic of Korea
[2]Department of Computer Science and Engineering, Kyung Hee University, Republic of Korea
{yhahn, hyeonbae.kim, st.kim}@khu.ac.kr

Abstract

Recent advancements in neural networks have showcased their remarkable capabilities across various domains. Despite these successes, the "black box" problem still remains. To address this, we propose a novel framework, WWW, that offers the 'what', 'where', and 'why' of the neural network decisions in human-understandable terms. Specifically, WWW utilizes adaptive selection for concept discovery, employing adaptive cosine similarity and thresholding techniques to effectively explain 'what'. To address the 'where' and 'why', we proposed a novel combination of neuron activation maps (NAMs) with Shapley values, generating localized concept maps and heatmaps for individual inputs. Furthermore, WWW introduces a method for predicting uncertainty, leveraging heatmap similarities to estimate the prediction's reliability. Experimental evaluations of WWW demonstrate superior performance in both quantitative and qualitative metrics, outperforming existing methods in interpretability. WWW provides a unified solution for explaining 'what', 'where', and 'why', introducing a method for localized explanations from global interpretations and offering a plug-and-play solution adaptable to various architectures. Code is available at: https://github.com/ailab-kyunghee/WWW

1. Introduction

Neural networks have demonstrated impressive performance in various fields in recent years. Despite these successes, their widespread adoption in more diverse areas is slowed by several challenges. A fundamental issue is the "black box" problem, referring to the often hidden and unclear decision-making processes of neural network models. This lack of clarity raises concerns about the stability and reliability of these models, leading to a growing consensus that artificial intelligence should be reliable, robust, and safe [4, 13]. As a response to this need for trustworthy AI, there has been an emergence of laws and regulations [15, 18] that require neural networks to base their decisions on principles that are understandable to humans.

Table 1. **Illustration of What, Where, and Why of recent concept-based neural network interpretation methods.** Green, yellow, and red marks illustrate that the method is able to interpret well, is partially interpretable, and has limitations for interpretation, respectively. Recent methods are able to interpret one or two 'w's but have limitations regarding interpreting three 'w's at once.

Method	What	Where	Why
CLIP-Dissect [20] (ICLR'23)	✔	✗	✗
FALCON [14] (ICML'23)	✔	✗	▲
CRAFT [7] (CVPR'23)	✗	✔	▲
WWW (Ours)	✔	✔	✔

According to Doshi-Velez *et al.* [4], interpretability is defined as the capability to provide explanations in terms that are understandable to humans. Zhang *et al.* [29] expand on this definition, emphasizing that interpretability involves providing explanations in understandable terms to humans. This requirement for explanations in human-understandable terms is a consistent theme in the literature on interpretable methods. Furthermore, this concept aligns with a fundamental principle in journalism and problem-solving: the importance of clear and understandable communication. This principle is often encapsulated in the five Ws (Who, What, When, Where, Why) [25]. Building on this, we propose that practical explanations or interpretations should include three critical elements: 'what' (the nature of the decision or outcome), 'where' (the context or specific region of the input that affects the decision-making process), and 'why' (the reasoning or factors behind a decision). This approach aims to make the interpretations more accessible and relevant to human understanding.

From this point of view, former works can be re-grouped into several groups which those works aim to solve. For ex-

*Corresponding author.

ample, a series of works in feature attribution [10, 23, 28] can be classified as works mainly focusing on 'where'. These methods explain model decisions by identifying highly contributed input regions. However, identifying highly related input regions primarily focuses on 'where' only, not enough explanation for 'what' and 'why'.

On the other hand, concept-based explanations mainly focus on 'why'. Concept-based explanations [6, 7, 10, 17] explain model decisions by decomposing model responses into smaller units called concepts. Kim *et al.* proposed TCAV [17], understanding and interpreting model decisions into a form of concept activation vectors. Recently, Fel *et al.* proposed CRAFT [7], which utilizes non-negative matrix factorization for identifying concept vectors. The works mentioned above mainly decompose activations into the form of vectors, which gives a great advantage for finding out 'why' but often fails to match or annotate the human-understandable term (i.e., name) that is needed for 'what'.

Other streams of work, called Neuron-concept association, aim to name what each neuron (e.g. convolution filters, layer outputs) represents, which answers 'what' the model cares about. Bau *et al.* proposed Network Dissection [3], which identifies each neuron's representing concept from Broaden dataset [3]. Recently, CLIP-Dissect [20] and FALCON [14] have shown that representing the concept of neurons can be matched automatically by using a pre-trained CLIP model [21]. The aforementioned methods are suitable for explaining 'what' but partially explain or have limitations for explaining 'where' and 'why'. Furthermore, these methods generate global explanations focusing on each neuron's role in the network, not the local explanations for each sample input.

To address these issues, we propose a novel framework that can explain '**W**hat', '**W**here', and '**W**hy' (WWW) at once. WWW introduces adaptive selection for discovering each neuron's concept and interpreting each neuron to explain 'what'. By leveraging adaptive cosine similarity (ACS) and adaptive selection techniques, we achieve advanced performance compared to competitive methods. Moreover, we combined neuron activation map (NAM) and Shapley value [2, 16, 24] to generate class and sample concept maps and heatmaps to explain 'where' and 'why'. We also conducted various objective evaluations to assess the performance of the proposed method. In the experiments, WWW achieves better results in both quantitative and qualitative evaluations. Our key contributions are summarized as follows:

- We introduce a novel and effective way to generate high-quality explanations that explain 'what', 'where', and 'why' at once. Due to the powerful performance of the adaptive selection for concept discovery, WWW is able to achieve higher quantitative and better qualitative results in various metrics.
- We introduce a way to generate localized explanations from global neural network interpretation. By the novel combination of Shapley value and neuron activation maps, WWW is able to generate localized explanations with concept annotations for sample input.
- WWW can be attached to various target models with different architectures, from conventional convolution neural networks to the recent attention-based Vision transformers, in a plug-and-play manner.

2. Related Works

2.1. Neuron-Concept Assosiation

Bau *et al.* introduced Network Dissection [3], using the Broden dataset to identify which concepts individual neurons in a network represent. They use overlap between segmentation masks and feature maps to annotate concepts for neurons. Fong and Vedaldi expanded on this with Net2Vec [9], which looks at individual neurons and their combinations. Mu and Andreas further extended these ideas with Compositional Explanation [19], aiming to generate more complex and detailed explanations. These methods primarily focused on understanding 'what' a neuron represents. However, the aforementioned methods fell short in explaining the 'where' and 'why' of neuron representations. Additionally, they relied heavily on image-concept-matched datasets like Broden, which are often costly and hard to collect due to the need for pixel-wise labels. To address these limitations, recent approaches like HINT [27] and MILAN [12] have been developed. These methods train concept classifiers or models that reduce the dependency on image-concept-matched datasets. Moreover, approaches like CLIP-Dissect [20] and FALCON [14] leverage the pre-trained CLIP model to use separate sets of image datasets and concept datasets. Despite these advancements in understanding 'what' neurons represent, there remains a gap in fully explaining the 'where' and 'why' of the neuron representations.

2.2. Vector-based Explanation

The field of interpretability has significantly advanced, particularly in understanding and interpreting model decisions using Concept Activation Vectors (CAVs). Kim *et al.* introduced TCAV [17], understanding and interpreting model decisions into the form of CAVs. Recently, CRAFT [7] was introduced, leveraging non-negative matrix factorization to identify CAVs and localize the most relevant input regions for each CAV. Also, Achtibat *et al.* [1] and Fel *et al.* [6] proposed methods to interpret each CAV's role in decision-making. Despite these developments, a key challenge remains: translating CAVs into human-understandable terms.

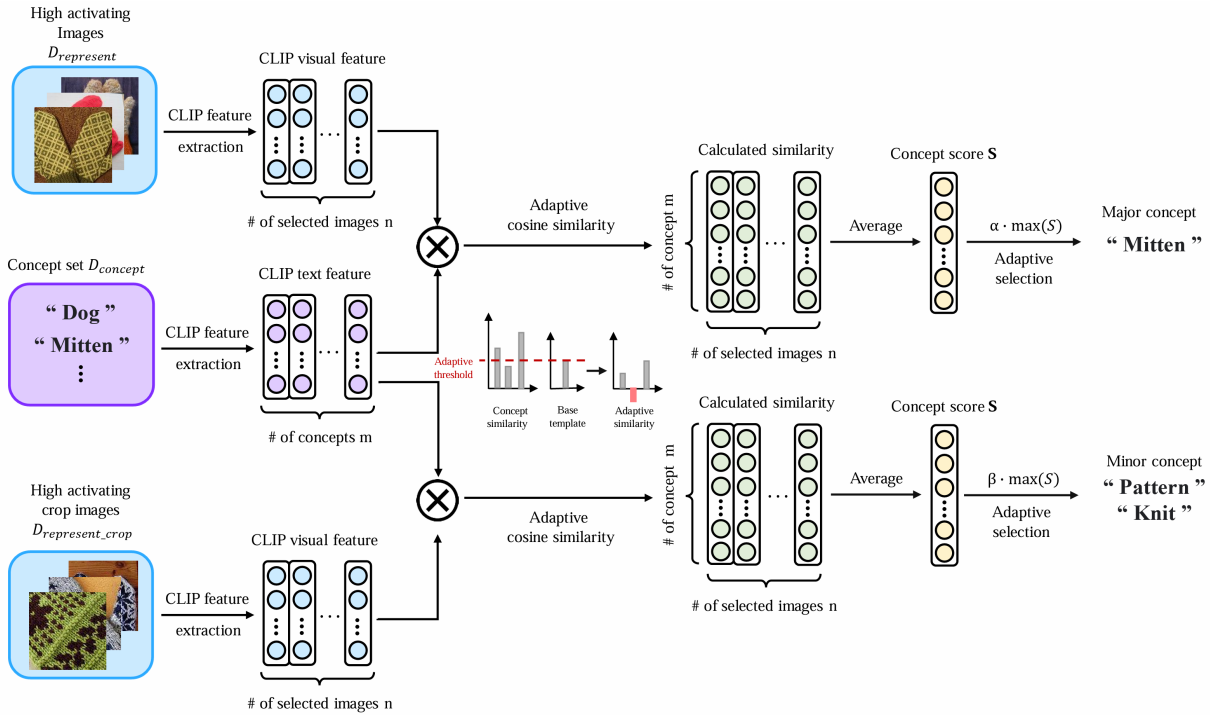

Figure 1. **Overall flow of Concept Discovery module identifying concepts for a single neuron.** We first calculate the cosine similarity of CLIP features between images and concepts with the template from the selected high-activating images. Then, we subtract the cosine similarity of CLIP features between images and the base template by only considering the similarity between the concept and image. From calculated adaptive cosine similarity (ACS), we generate concept score S by the average similarity of images. Note that concept score $S = \{s_1, s_2, ..., s_m\}$ are a group of scores, not a single scalar. From the calculated concept scores S, we select major concepts by adaptive selection. We also discover minor concepts using the same process but with crop images.

3. Method

3.1. Method Overview

WWW consists of three modules: Concept discovery, Localization, and Reasoning. The concept discovery module identifies each neuron's concept, which explains part 'what'. It selects neuron concepts from the concept set ($D_{concept}$) by leveraging adaptive cosine similarity and adaptive selection. The localization module is to identify highly contributed input regions of the test sample, which concept is present at 'where'. Also, the combination of the neuron activation map leveraging Shapley value helps identify 'where' the concept is and tells the 'why' of the individual predictions. The reasoning module identifies important neurons of the test sample and predicted class. By comparing the differences between the sample and class explanations, users can understand the 'why' of the model prediction and, even more, whether the prediction is reliable or not.

3.2. Concept Discovery Module

The concept discovery module aims to identify proper major and minor concepts from $D_{concept}$ that match an example-based representation of each neuron. Let target model f d, and let (l, i) as i-th neuron in layer l of the target model. From the images in probing dataset D_{probe}, we select high activating images for neuron (l, i) as $D_{rep}^{(l,i)}$. With selected images in $D_{rep}^{(l,i)}$ and concepts in $D_{concept}$, we calculate CLIP visual feature of $D_{rep}^{(l,i)}$ as $V^{(l,i)} = \left[v_1^{(l,i)}, v_2^{(l,i)}, \cdots, v_n^{(l,i)}\right]$ and CLIP text feature of $D_{concept}$ as $T = [t_1, t_2, \cdots, t_m]$. n denotes number of example images for neuron (l, i) and m stands for the number of concept in concept set $D_{concept}$.

We calculate concept score $s^{(l,i)}$ by calculating adaptive cosine similarity (ACS), which allows us to reduce the effect of the base template and only consider the similarity between the image and the concept itself. Concept score $s_j^{(l,i)}$ for j-th concept t_j of neuron (l, i) is calculated as follows:

$$s_j^{(l,i)} = \frac{1}{n} \sum_{o=1}^{n} \left\{ cos(v_o^{(l,i)}, t_j) - cos(v_o^{(l,i)}, t_{tem}) \right\}, \quad (1)$$

where $1 \leq j \leq m$, t_{tem} denotes CLIP text feature of base templete(e.g. 'a photo of.') and $cos(x, y)$ denotes cosine similarity between x and y.

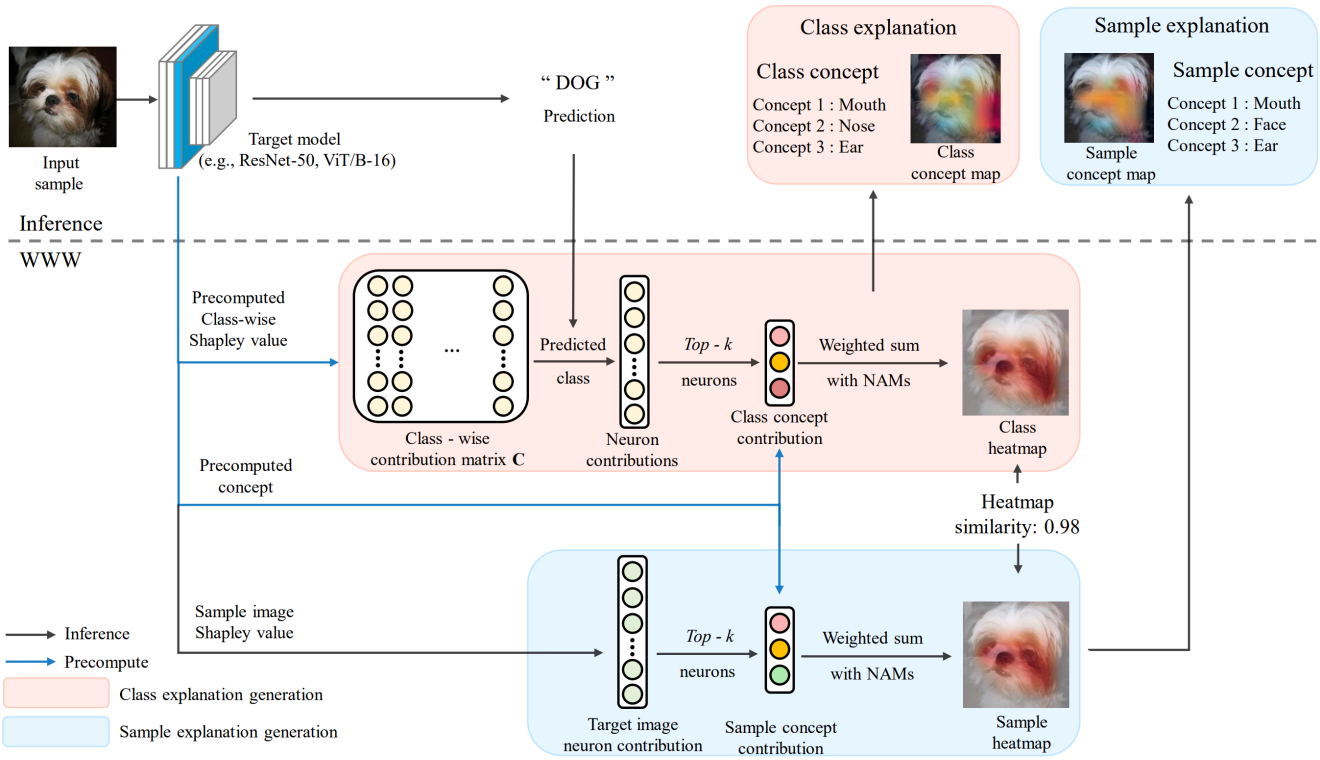

Figure 2. **Illustration of overall test time flow of WWW.** In the test time (i.e., inference), class explanation selects important neurons with pre-computed Shapley value of the predicted class. On the other hand, the sample explanation selects important neurons with a Shapley value for the input sample. With selected important neurons, pre-computed concepts are annotated. After the concept annotation, a class heatmap is generated with the pre-computed Shapley value of the predicted class. On the other hand, a sample heatmap is generated with the Shapley value of the sample input.

From the calculated concept scores $s^{(l,i)} = \left[s_1^{(l,i)}, \cdots, s_m^{(l,i)}\right]$, we select concepts where $s_j^{(l,i)} > \delta^{(l,i)}$. Adaptive selection threshold $\delta^{(l,i)}$ for neuron (l,i) is calculated as follows:

$$\delta^{(l,i)} = \alpha \times max(s^{(l,i)}), \tag{2}$$

where α denotes concept sensitivity for major and minor concepts. For discovering minor concepts, we select $D_{rep}^{(l,i)}$ with cropped images of probing dataset D_{probe}.

3.3. Localization Module

The localization module aims to identify highly contributed input regions of each concept and generate a concept map and concept heatmap. We select important neurons with Taylor approximation of Shapley value introduced in [16] for generating a concept region map. For input image sample x, Neuron contribution $w^{(l,i)}$ of neuron (l,i) is calculated as follows:

$$w^{(l,i)}(x) = \left| f(x) - f(x; a^{(l,i)} \leftarrow 0) \right| = \left| a^{(l,i)} \nabla_{a^{(l,i)}} f(x) \right|. \tag{3}$$

Where $a^{(l,i)}$ denotes activation of neuron (l,i). After calculating each layer's Neuron contribution, we rank neurons by calculated contribution (i.e., Shapley value) and select the top-k important neurons for the sample. The concept heatmap M is calculated with the weighted sum of important neurons' neuron activation maps (NAMs). Concept heatmap $M^l(x)$ of important neurons is calculated as follows:

$$M^l(x) = \sum^{u} w^{(l,u)}(x) A^{(l,u)}(x) \tag{4}$$

where u denotes the index of important neurons, w denotes the neuron contribution, and A denotes the neuron activation map of the neuron. Note that M^l shows only related regions of a single or combination of important neurons, not the whole network.

3.4. Reasoning Module

The reasoning module is designed to help users understand the 'why' of the model output. This not only explains the result but also includes 'why' this prediction is reliable or not. By leveraging the class-wise Shapley value introduced in [2], we can understand each neuron's class-wise contri-

bution and identify important neurons for each class. From Ahn *et al.* [2], the class-wise contribution of a neuron can be calculated by the average contribution of a neuron in class samples. From the class-wise contribution, we rank top-k important neurons for each class. We can also generate a class-wise concept region map and concept heatmap with the index of important neurons. Class-wise maps can be used as a guideline for understanding the general case of the prediction. By comparing Class-wise maps and sample maps, users can identify which concept and region of the sample differs from general cases. This naturally helps users understand the 'why' of the model output and also 'why' the results are reliable or not. Figure 2 for further understanding the WWW flow for generating class and sample explanations. Figure 5 shows the generated explanation example.

3.5. Overall flow and Output of WWW

As described above, WWW has three main modules explaining three 'w' of 'what', 'where', and 'why', respectively. In Section 3.5, we are going to follow the overall flow of our method by time sequence. Before starting inference, WWW needs pre-computing for discovering major and minor concepts with the concept discovery module (Sec 3.2) and also a class-wise contribution for class-wise analysis in the Reasoning Module (Sec 3.4). After pre-computing concepts and contributions, WWW is ready to generate test sample explanations. Figure 2 shows the overall flow of explaining the generation of WWW in the inference time. In the test time (i.e., inference), WWW generates a class-wise explanation of the predicted class and sample explanation into two tracks. WWW leverages the predicted class and class-wise contribution matrix to generate a class-wise explanation to select the most critical neurons for the predicted class. After identifying important neurons, WWW leverages the localization module (Sec 3.3) to generate a class concept region map and class concept combination heatmap. On the other hand, for the sample explanation, WWW calculates the Taylor-approximated Shapley value of the sample for the predicted class and identifies critical neurons for the prediction of the sample. After identifying important neurons for the sample, WWW leverages the sample Shapley value to generate the sample concept region map and sample concept combination heat map.

4. Experiment

Comprehensive experiments have been conducted to evaluate our method. In section 4.1, we evaluate the performance of the concept discovery module both qualitatively (Sec. 4.1.1) and quantitatively (Sec. 4.1.2). Section 4.2 is an ablation study for the concept discovery module. In Section 4.3, we analyze generated explanations of both correct and wrong predictions with examples.

Figure 3. **Qualitative comparison of WWW with other baselines.** We compared WWW with three competitive baselines (CLIP-Dissect [20], MILAN [12], FALCON [14]) in two final layer neurons and four penultimate layer (i.e., layer 4) neurons with each neuron's highly activating images. layer-4 neurons are top-2 important neurons of the final layer class. 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.

4.1. Performance Evaluation for Concept Module

In this experiment, we evaluate the concept matching performance of the concept discovery module with four other baselines, Network Dissection [3], MILAN [12], CLIP-Dissect [20] and FALCON [14]. We evaluate the performance of methods on the various models (e.g., ResNet-18 [11], ResNet-50 [11], and ViT-B/16 [5]), various probing datasets (e.g., Imagenet, Places365) and various concept sets (e.g., Wordnet nouns, and labels of Places365, Broaden, and ImageNet)

4.1.1 Qualitative Results

Settings. We compared WWW with the three most comparable methods (CLIP-Dissect [20], MILAN [12], FALCON [14]) in the penultimate layer and final layer of the model. We do not compare with Network Dissection [3] due to the limitation that probe image data D_{probe} and concept set $D_{concept}$ is fixed to Broaden. We used a ResNet-50 [11] model pre-trained in the ImageNet-1k [22] dataset. For probe image data D_{probe}, we used the ImageNet-1k validation set, and we extracted all nouns in Wordnet[8] (about 80k) dataset for a concept set $D_{concept}$.

Results. Figure 3 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 same major concept as the final layer neuron.

Table 2. **Quantitative comparison on final layer concept matching performance of ResNet-50 trained on ImageNet.** We compared predicted neuron concepts with ground truth labels of ImageNet. We used the Imagenet-1k validation set for D_{probe}. **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	D_{probe}	$D_{concept}$	CLIP cos	mpnet cos	F1-score	Hit Rate
Network Dissection [3]	Broden	Broden(1.2k)	0.7229 ± 0.003	0.2989 ± 0.005	0.0010 ± 0.001	0.001
MILAN(b) [12]	ImageNet val	-	0.7300 ± 0.003	0.2485 ± 0.005	0.0005 ± 0.000	0.001
FALCON [14]	ImageNet val	LAION-400m	0.7065 ± 0.003	0.1790 ± 0.001	0.0002 ± 0.000	0.001
CLIP-Dissect [20]	ImageNet val	ImageNet (1k)	$\mathbf{0.9340 \pm 0.003}$	$\mathbf{0.8376 \pm 0.006}$	0.7286 ± 0.009	0.933
	ImageNet val	Broden (1.2k)	0.7369 ± 0.003	0.3432 ± 0.004	0.0328 ± 0.003	**0.108**
	ImageNet val	Wordnet (80k)	0.8689 ± 0.004	0.6846 ± 0.008	0.3647 ± 0.014	0.456
WWW (Ours)	ImageNet val	ImageNet (1k)	0.9325 ± 0.003	0.8327 ± 0.006	$\mathbf{0.7719 \pm 0.009}$	**0.955**
	ImageNet val	Broden (1.2k)	$\mathbf{0.7758 \pm 0.004}$	$\mathbf{0.4414 \pm 0.007}$	$\mathbf{0.0645 \pm 0.007}$	0.091
	ImageNet val	Wordnet (80k)	$\mathbf{0.8858 \pm 0.003}$	$\mathbf{0.6945 \pm 0.008}$	$\mathbf{0.4197 \pm 0.012}$	**0.645**

Table 3. **Quantitative comparison on final layer concept matching performance of ViT-B/16 trained on ImageNet.** We compared predicted neuron concepts with ground truth labels of ImageNet. We used the Imagenet-1k validation set for D_{probe}. **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	D_{probe}	$D_{concept}$	CLIP cos	mpnet cos	F1-score	Hit Rate
CLIP-Dissect [20]	ImageNet val	ImageNet (1k)	$\mathbf{0.9337 \pm 0.003}$	$\mathbf{0.8375 \pm 0.006}$	0.7289 ± 0.009	0.933
	ImageNet val	Broden (1.2k)	0.7365 ± 0.003	0.3416 ± 0.004	0.0319 ± 0.003	**0.106**
	ImageNet val	Wordnet (80k)	0.8700 ± 0.004	0.6886 ± 0.008	0.3679 ± 0.014	0.460
WWW (Ours)	ImageNet val	ImageNet (1k)	0.9331 ± 0.003	0.8347 ± 0.006	$\mathbf{0.7718 \pm 0.009}$	**0.955**
	ImageNet val	Broden (1.2k)	$\mathbf{0.7754 \pm 0.004}$	$\mathbf{0.4407 \pm 0.007}$	$\mathbf{0.0651 \pm 0.007}$	0.091
	ImageNet val	Wordnet (80k)	$\mathbf{0.8857 \pm 0.003}$	$\mathbf{0.6970 \pm 0.008}$	$\mathbf{0.4166 \pm 0.012}$	**0.634**

4.1.2 Quantitative Results

In this section, we compare the performance of our methods with baselines. As introduced in [20], we evaluate final layer neuron concepts with the class labels with various metrics. By comparing generated explanations with the class labels, we can objectively evaluate the quality of the generated neuron labels with what each neuron is trained to represent.

Metrics. CLIP cos and mpnet cos are measured as cosine similarities between the encoded feature of the class label and selected concepts with CLIP [21] model and mpnet [26] model, respectively. F1-score is measured to evaluate the discovered concept's balance of exactness and flexibility. Also, the hit rate is calculated as a rate of selected concepts that exactly match the class label.

Quantitative comparison on ResNet-50. Table 2 compares WWW with Network Dissection [3], MILAN(b) [12], FALCON [14], and CLIP-Dissect [20]. We evaluate the final layer neuron concepts of ResNet-50 with the class labels of ImageNet-1k. In table 2, WWW showed better performance as the concept set $D_{concept}$ gets larger and outperformed all other baselines when concept set $D_{concept}$ is set to Broden and Wordnet nouns. In the results of Wordnet nouns ($D_{concept}$), comparison between WWW and CLIP-Dissect [20] are statistically significant across all metrics ($p < 0.05$).

Quantitative comparison on ViT-B/16. We compared predicted concepts of the final layer to ground truth labels of ViT-B/16 pre-trained on ImageNet. In Table 3, WWW showed better performance as the concept set $D_{concept}$ gets larger and outperformed other baselines when concept set $D_{concept}$ is set to Broden and Wordnet nouns.

Quantitative comparison on ResNet-18 pre-trained in Places365. We compared predicted labels to ground truth labels in final layer neurons of ResNet-18 pre-trained on Places365. In table 4, WWW showed better performance as the concept set $D_{concept}$ gets larger and outperformed other baselines when concept set $D_{concept}$ is set to Broden and Wordnet nouns.

4.2. Ablation study

Evaluation on the effect of leveraging base template and ACS. Table 5 shows an ablation study over various parts used in WWW. Without using a base template, WWW showed the lowest performance overall. With the use of a base template, WWW shows slightly increased performance in CLIP cos and F1-score but not much advance overall. However, performance was significantly increased, and the highest scores in all three metrics were achieved when leveraging ACS. The findings suggest that while us-

Table 4. **Quantitative comparison on final layer concept matching performance of ResNet-18 trained on Places365.** We compared predicted neuron concepts with ground truth labels of Places365. We used the Places365 test set for D_{probe}. **Bold** numbers represent the best scores between the same settings. The average score and standard errors of the 365 final layer neurons are reported.

Method	D_{probe}	$D_{concept}$	CLIP cos	mpnet cos	F1-score	Hit Rate
CLIP-Dissect [20]	Places365 test	Places365 (0.4k)	$\mathbf{0.9562 \pm 0.004}$	$\mathbf{0.8687 \pm 0.012}$	$\mathbf{0.7233 \pm 0.023}$	**0.723**
	Places365 test	Broden (1.2k)	0.8304 ± 0.003	0.4678 ± 0.007	0.1096 ± 0.008	**0.329**
	Places365 test	Wordnet (80k)	0.8378 ± 0.006	$\mathbf{0.5107 \pm 0.014}$	0.1041 ± 0.016	0.104
WWW (Ours)	Places365 test	Places365 (0.4k)	0.9402 ± 0.004	0.8204 ± 0.013	0.6361 ± 0.024	0.660
	Places365 test	Broden (1.2k)	$\mathbf{0.8925 \pm 0.005}$	$\mathbf{0.6415 \pm 0.013}$	$\mathbf{0.2242 \pm 0.021}$	0.255
	Places365 test	Wordnet (80k)	$\mathbf{0.8492 \pm 0.004}$	0.5000 ± 0.013	$\mathbf{0.1106 \pm 0.015}$	**0.142**

Table 5. **Ablation study on the use of the template and ACS.** '✔' in the Template represents WWW using a template (e.g., 'a photo of *word*') '✔' in the ACS represents WWW using adaptive cosine similarity.

Template	ACS	CLIP cos	mpnet cos	F1-score
		0.8499 ± 0.003	0.6123 ± 0.007	0.3265 ± 0.009
✔		0.8547 ± 0.003	0.6075 ± 0.007	0.3361 ± 0.009
✔	✔	$\mathbf{0.8858 \pm 0.003}$	$\mathbf{0.6945 \pm 0.008}$	$\mathbf{0.4197 \pm 0.012}$

ing a base template in the CLIP model offers some benefits, it also has its limitations, particularly in terms of embedding concepts within a similar CLIP feature space. In contrast, ACS appears to reduce the uniformity between concepts, allowing the concept discovery module to discover more distinct and accurate concepts for each neuron representation. This highlights the effectiveness of ACS in enhancing the overall performance of the concept discovery module.

Ablation study on Concept Sensitivity (α). The ablation study depicted on the left side of Figure 4 examined the impact on the F1-score by varying levels of major concept sensitivity (denoted as α). We found that as concept sensitivity decreases, the F1-score initially increases($\alpha > 0.95$) and then decreases. This pattern is due to a trade-off. Higher concept sensitivity leads to more precise concept identification but at the cost of identifying fewer concepts. Conversely, lower sensitivity results in more concepts being identified that may be less similar to the target concept. The point at which the F1-score is maximized can be seen as the optimal balance in this trade-off, providing a guideline for setting the most effective level of concept sensitivity for the major concept discovery.

4.3. Discussion

4.3.1 Overall Explanation Generated by WWW

Figure 5 illustrates an explanation example generated by WWW, where the left side of the figure provides a class explanation for a test sample by highlighting important neurons for the predicted class, and the right side depicts a sample explanation showing the important neurons identified for the test sample image prediction. This explanation target neurons from the penultimate layer of a ResNet-50 model pre-trained on ImageNet, using the ImageNet validation set (D_{probe}) and Wordnet nouns ($D_{concept}$) for the

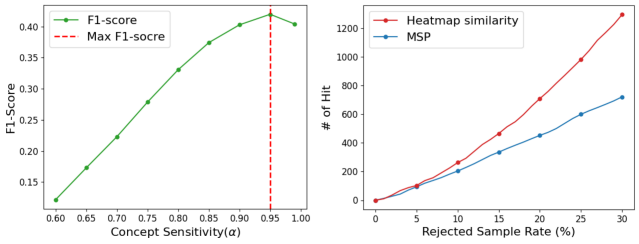

Figure 4. **Ablation of concept sensitivity and heatmap similarity feasibility result.** Left figure illustrates the F1 score with respect to Concept Sensitivity (α) changes. Concept sensitivity (α) that maximizes the F1-score is illustrated as the red line. The right figure illustrates the rejection test result of heatmap similarity and maximum softmax probability (MSP). # of hit denotes the number of correctly detected samples as a misprediction.

Figure 5. **Example of generated explanation by WWW.** From top to bottom, important neurons are displayed in the order of importance. Images in each neuron are examples of a major and minor concept, respectively. Colors in the top localization image show highly related regions for each concept. The bottom localization image is a weighted sum of important neuron activation maps displayed as a heatmap.

explanation. Notably, neuron 30 is consistently identified as the most crucial in both explanations. It represents the 'tench' concept, which matches the prediction and the test sample's ground truth. Interestingly, while neuron 1246 is the third most important for the concept, it is the second most significant for the test sample's prediction, and neuron 1480, not highlighted in the class explanation, emerges as the third important neuron in the sample explanation. Despite these neuron selection and ranking variations, the model accurately predicts the correct class. The overall heatmap of the class and sample explanations can explain this. The heatmap explanation reveals a high cosine similarity between the two heatmaps, indicating that the calculated weighted sum of NAMs is similar. Despite the difference in individual neuron importance, the highly related input region is remarkably similar in both cases.

4.3.2 Anlaysis of Explanation of Failure Case

Figure 6 provides an example of a mispredicted sample explanation generated by the WWW. In this particular case of failure, 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, measuring at 0.19. Interestingly, 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 of 0.47. From this observation, we explore the potential utility of heatmap similarity to predict uncertainty. On the right side of Figure 4, we present the results of a rejection test that evaluates heatmap similarity and the maximum softmax probability (MSP). When the rejection method detects the mispredicted sample correctly, we consider that as a hit. The number of hit is measured with respect to the rejected sample rate. As the rejection rate increases based on their uncertainty level, it becomes evident that heatmap similarity outperforms MSP regarding the detection of misprediction samples. This suggests that heatmap similarity can serve as a more effective measure of uncertainty compared to MSP. These findings indicate that heatmap similarity can be used as a tool for predicting uncertainty.

5. Conclusion

We proposed WWW, a unified framework that provides comprehensive explanations for the 'what', 'where', and 'why' of neural network decisions. WWW demonstrates superior performance in both quantitative and qualitative measures, offering a deeper and more detailed understanding of neural network behavior. This is achieved through a novel integration of adaptive selection for concept discovery, neuron activation maps, and Shapley values. WWW's adaptability is also shown across various neural network architectures, including convolutional networks and attention-based Vision Transformers. Additionally, our approach to predicting uncertainty through heatmap similarity analysis introduces a new way to obtain the reliability of their predictions. By offering localized explanations with concept annotations for individual inputs, WWW enhances the transparency of the model's decision-making process, contributing to the broader goal of making AI more reliable and trustworthy.

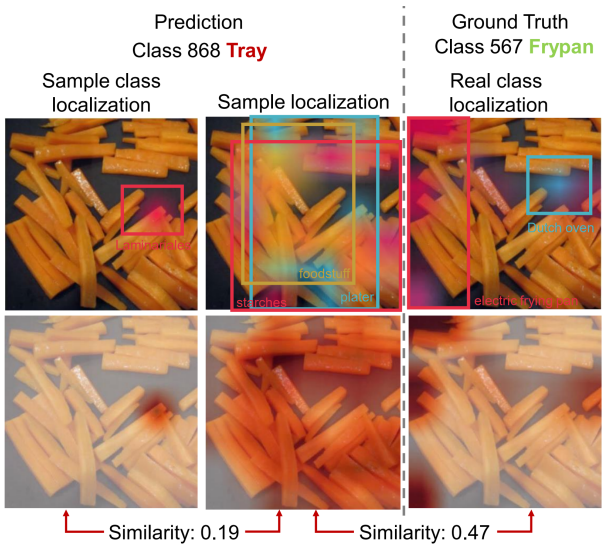

Figure 6. **Example of failure case explanation by WWW.** The explanations of the predicted label are presented on the left side. On the right side, the explanation of the ground-truth label is shown. In the upper half, we displayed concept attribution maps, which show important concepts and their respective regions. In the bottom half, we showed the overall heatmap which shows important regions for the model decision.

Acknowledgements

This work was supported in part by the National Research Foundation of Korea (NRF) Grant funded by the Korea Government (MSIT) under Grant 2021R1G1A1094990, by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korea Government (MSIT) under Grant 2022-0-00078 (Explainable Logical Reasoning for Medical Knowledge Generation), Grant IITP-2024-RS-2023-00258649 (ITRC(Information Technology Research Center) support program), Grant 2021-0-02068 (Artificial Intelligence Innovation Hub), Grant RS-2022-00155911 (Artificial Intelligence Convergence Innovation Human Resources Development (Kyung Hee University)), and by Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD (UD230017TD).

References

[1] Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence*, 5(9):1006–1019, 2023. 2

[2] Yong Hyun Ahn, Gyeong-Moon Park, and Seong Tae Kim. Line: Out-of-distribution detection by leveraging important neurons. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19852–19862, 2023. 2, 4, 5

[3] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017. 2, 5, 6

[4] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. 1

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[6] Thomas Fel, Victor Boutin, Mazda Moayeri, Rémi Cadène, Louis Bethune, Léo andéol, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. *arXiv preprint arXiv:2306.07304*, 2023. 2

[7] Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2711–2721, 2023. 1, 2

[8] Christiane Fellbaum. Wordnet and wordnets. encyclopedia of language and linguistics, 2005. 5

[9] Ruth Fong and Andrea Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8730–8738, 2018. 2

[10] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pages 3429–3437, 2017. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Proceedings of the European conference on computer vision*, pages 630–645. Springer, 2016. 5

[12] Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep visual features. In *International Conference on Learning Representations*, 2021. 2, 5, 6, 1

[13] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 624–635, 2021. 1

[14] Neha Kalibhat, Shweta Bhardwaj, C Bayan Bruss, Hamed Firooz, Maziar Sanjabi, and Soheil Feizi. Identifying interpretable subspaces in image representations. *International Conference on Machine Learning*, 2023. 1, 2, 5, 6

[15] Margot E Kaminski and Jennifer M Urban. The right to contest ai. *Columbia Law Review*, 121(7):1957–2048, 2021. 1

[16] Ashkan Khakzar, Soroosh Baselizadeh, Saurabh Khanduja, Christian Rupprecht, Seong Tae Kim, and Nassir Navab. Neural response interpretation through the lens of critical pathways. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13528–13538, 2021. 2, 4

[17] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018. 2

[18] Mauritz Kop. Eu artificial intelligence act: The european approach to ai. Stanford-Vienna Transatlantic Technology Law Forum, Transatlantic Antitrust ..., 2021. 1

[19] Jesse Mu and Jacob Andreas. Compositional explanations of neurons. *Advances in Neural Information Processing Systems*, 33:17153–17163, 2020. 2

[20] Tuomas Oikarinen and Tsui-Wei Weng. CLIP-dissect: Automatic description of neuron representations in deep vision networks. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 5, 6, 7

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 6

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 5

[23] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2

[24] Lloyd S Shapley. A value for n-person games. *Classics in game theory*, 69, 1997. 2

[25] Michael C Sloan. Aristotle's nicomachean ethics as the original locus for the septem circumstantiae. *Classical Philology*, 105(3):236–251, 2010. 1

[26] K. Song et al. Mpnet: Masked and permuted pre-training for language understanding. *NeurIPS*, 33:16857–16867, 2020. 6

[27] Andong Wang, Wei-Ning Lee, and Xiaojuan Qi. Hint: Hierarchical neuron concept explainer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10254–10264, 2022. 2

[28] Yang Zhang, Ashkan Khakzar, Yawei Li, Azade Farshad, Seong Tae Kim, and Nassir Navab. Fine-grained neural network explanation by identifying input features with predictive information. *Advances in Neural Information Processing Systems*, 34:20040–20051, 2021. 2

[29] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021. 1