Accuracy Improvement of Neuron Concept Discovery Using CLIP with Grad-CAM-Based Attention Regions

Takahiro Sannomiya[a] and Kazuhiro Hotta[b]
Meijo University, Nagoya, Japan

Keywords: Explainable AI, CLIP, Concept of Neurons.

Abstract: WWW is a method that computes the similarity between image and text features using CLIP and assigns a concept to each neuron of the target model whose behavior is to be determined. However, because this method calculates similarity using center crop for images, it may include features that are not related to the original class of the image and may not correctly reflect the similarity between the image and text. Additionally, WWW uses cosine similarity to calculate the similarity between images and text. Cosine similarity can sometimes result in a broad similarity distribution, which may not accurately capture the similarity between vectors. To address them, we propose a method that leverages Grad-CAM to crop the model's attention region, filtering out the features unrelated to the original characteristics of the image. By using t-vMF to measure the similarity between the image and text, we achieved a more accurate discovery of neuron concepts.

1 INTRODUCTION

In recent years, image recognition models have been used in a variety of fields, but the problem is that it is unclear how the models are making decisions. To address this issue, visualization methods such as Class Activation Maps (CAM)(Wang et al., 2020; Zhou et al., 2016) have been proposed, but they only show the regions of interest and cannot explain what concepts and features the model is learning. A method was proposed to identify the concepts of the model's neurons using CLIP(Radford et al., 2021), which can measure the similarity between images and text. This method allows us to explain in concrete terms that humans can understand what concepts the model is basing its decisions on, and to deepen our understanding of the model's decision-making process and internal behavior.

WWW(Ahn et al., 2024) is a method for identifying neuron concepts. Since this method calculates similarity using center crop for images, it includes features that are not related to the original class of the image and may not correctly reflect the similarity between the image and the text. The WWW uses cosine similarity in calculating the similarity between images and text. Cosine similarity may not accurately reflect the similarity between vectors due to the wide similarity distribution. To address these issues, we propose a method to more accurately discover neuron concepts by using t-vMF similarity between images and text, while using Grad-CAM to crop the regions of interest in the model and eliminating features that are not related to the original features of the image.

Experiments were conducted on ImageNet validation datasets consisting of 1000 classes, such as animals and vehicles, and text datasets such as Broaden and WordNet. The results showed that the accuracy of some evaluation metrics, such as CLIP cos, mpnet cos and F1-score, exceeded that of conventional method.

The paper is organized as follows. Section 2 describes related works. Section 3 details the proposed method. Section 4 presents experimental results. Section 5 discusses the Ablation Study. Finally, Section 6 concludes our paper.

2 RELATED WORKS

We explain WWW, a method for identifying neuron concepts. Figure 1 illustrates WWW. Let the i-th neuron in layer l of the target model be denoted as (l,i). We denote the number of text samples as j. First, we crop a center region in each image in the probing dataset (evaluation data) D_{probe}, and feed them into the target model. We then select images where

[a] https://orcid.org/0009-0005-3644-381X
[b] https://orcid.org/0000-0002-5675-8713

Sannomiya, T. and Hotta, K.
Accuracy Improvement of Neuron Concept Discovery Using CLIP with Grad-CAM-Based Attention Regions.
DOI: 10.5220/0013247500003912

In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 2: VISAPP*, pages 497-502
ISBN: 978-989-758-728-3; ISSN: 2184-4321

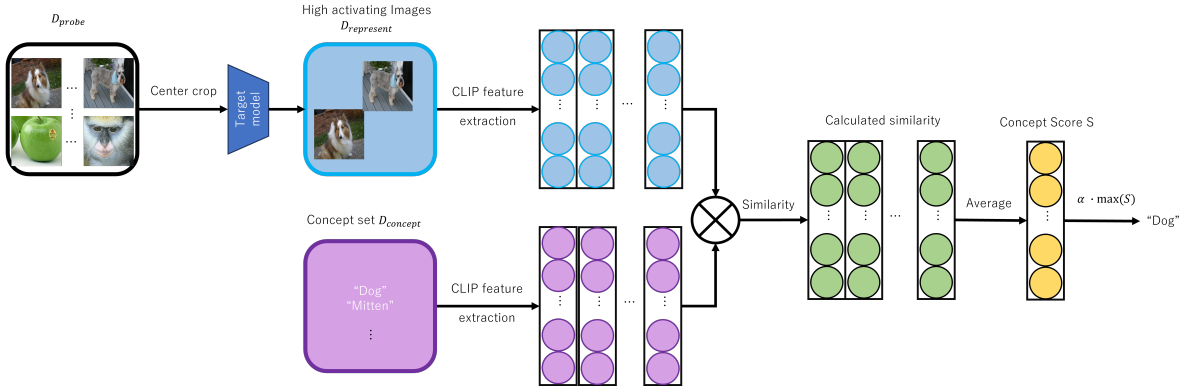

Figure 1: Overview of WWW.

the neuron exhibits a strong response (high activation) and denote this set as $D_{represent}$. By feeding both $D_{represent}$ and the text dataset $D_{concept}$ into CLIP, we compute the concept score $S_j^{(l,i)}$ using the following equation.

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

where $v_o^{(l,i)}$ is the CLIP image feature vector, t_j represents the CLIP text feature vector, and t_{tem} is the CLIP text feature vector of a base template word, such as "a photo of a," which has similarity with any image. In Equation (1), subtracting the cosine similarity between $v_o^{(l,i)}$ and t_{tem} from the cosine similarity between $v_o^{(l,i)}$ and t_j removes the influence of the base template, allowing us to focus solely on the similarity between the image and the text itself. After calculating the concept score $S_j^{l,i}$, the texts corresponding to scores above the threshold $\delta^{(l,i)}$ are considered as concepts for the neuron (l,i). $\delta^{(l,i)}$ is defined by the following equation.

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

where α is a hyper parameter representing concept sensitivity.

This method reduces the influence of the base template t_{tem}, enabling the measurement of similarity between the image and text itself. However, since images in D_{probe} are center-cropped and then passed through the model to select high-activation images, which are used as $D_{represent}$ for calculating similarity with text, the unrelated features to the original class of the image, such as background details, may be included. This can prevent an accurate reflection of the similarity between the image and text, potentially affecting the identification of neuron concepts. Additionally, although WWW uses cosine similarity for similarity calculations, this similarity measure has a wide distribution, which may be insufficient for measuring precise similarity.

3 PROPOSED METHOD

To solve this issue, we propose a method that computes Grad-CAM on D_{probe}, crops the attention region, and calculates the similarity using images that retain class-relevant features. Additionally, to measure the similarity more accurately, we replace cosine similarity with t-vMF similarity (Kobayashi, 2021), which narrows the similarity distribution for more precise measurement. As shown in Figure 2, the proposed method feeds D_{probe} into the target model and computes Grad-CAM. The image is then cropped the area centered on the highest Grad-CAM value. The cropped image is fed into the target model, and images in which the i-th neuron in layer l (i.e., (l,i)) shows high activation are selected as $D_{represent}$. The concept score $S_j^{(l,i)}$ is then computed using the following equation.

$$S_j^{(l,i)} = \frac{1}{n}\sum_{o=1}^{n}\left\{\text{t-vMF}\left(v_o'^{(l,i)}, t_j\right) - \text{t-vMF}\left(v_o'^{(l,i)}, t_{tem}\right)\right\} \tag{3}$$

where $v_o'^{(l,i)}$ is the image feature vector obtained when high-activation images, determined using Grad-CAM, are fed into CLIP. By calculating $S_j^{l,i}$ and selecting texts corresponding to scores above the threshold, as done in WWW, we can define these texts as the concepts for the neuron (l,i). This approach allows us to retain only the essential features of the image, enabling a more accurate calculation of similarity between features related to the image class and the text.

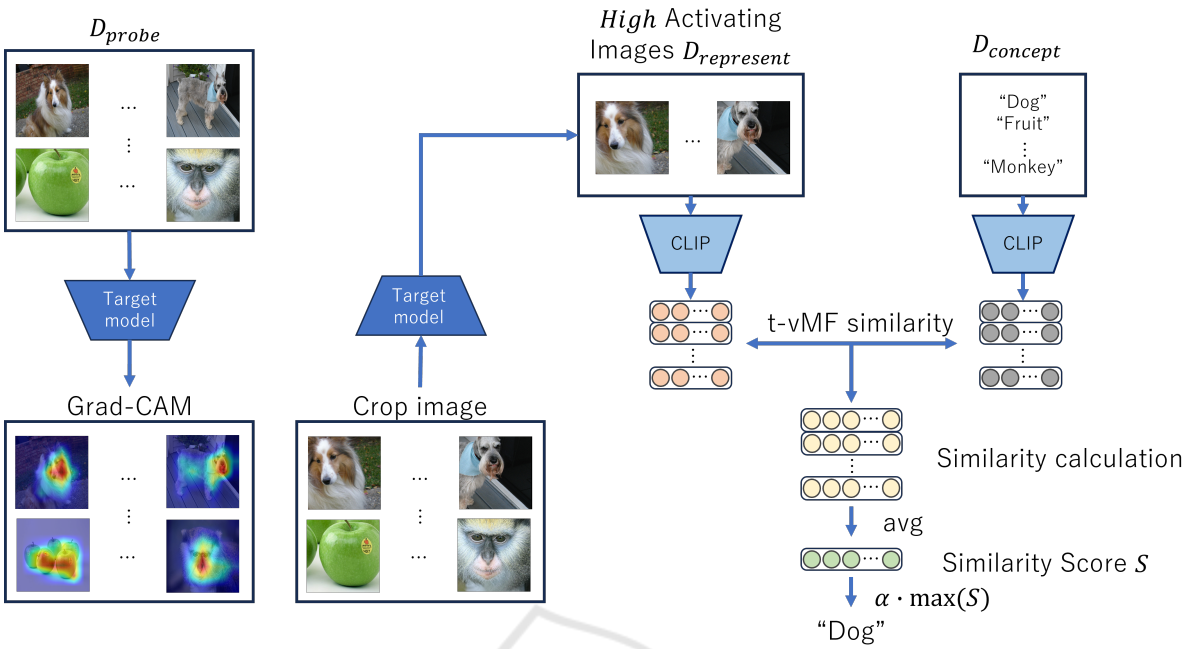

Figure 2: Overview of the proposed method.

4 EXPERIMENTS

In this section, we describe the experimental setup and results. Section 4.1 describes the experimental setup. Section 4.2 presents the results of the qualitative evaluation experiments. Section 4.3 shows the results of the quantitative evaluation experiments.

4.1 Experimental Settings

Following the previous research, we use a ResNet-50 model pre-trained on ImageNet-1k (Russakovsky et al., 2015) as the target model. For D_{probe}, we utilize the validation images in the ImageNet-1k dataset, while $D_{concept}$ includes ImageNet-1k, WordNet (Fellbaum, 2005), and Broaden (Bau et al., 2017). We selected 40 images from D_{probe} that exhibited high activation in the neurons to form $D_{represent}$. Evaluation metrics include CLIP cos, mpnet cos, F1-Score, and Hit Rate, using CLIP and mpnet (Song et al., 2020). CLIP cos and mpnet cos are metrics that measure cosine similarity by feeding class labels and selected concepts into CLIP and mpnet, respectively. The F1-score is an evaluation metric that measures the accuracy and flexibility of the discovered concepts, while the Hit Rate is calculated based on the proportion of selected concepts that match the class labels. Higher values for any of these evaluation metrics indicate better performance. The concept sensitivity α is set to 0.95 for both methods.

4.2 Qualitative Result

The results of the qualitative evaluation are shown in Figure 3. WordNet is used for $D_{concept}$, and images with each neuron in the final layer of ResNet-50 highly activated are compared. Below each image, the proposed method and the concept identified in WWW are shown. Figure 3 confirms the superiority of the proposed method. For example, in Neuron 0, Neuron 10, Neuron 446, and Neuron 479, the concept identified by the proposed method matches the ground truth of the image, while the concept identified by WWW is similar to the ground truth but different. For Neuron 296, Neuron 460, Neuron 671,Neuron 742, and Neuron 850, the proposed method identifies concepts that are consistent with ground truth, while WWW identifies significantly different concepts. This may be due to the fact that WWW uses center crop, which includes unnecessary features in addition to the original image features, making it easier to identify unrelated or similar concepts. This can be seen from Figure 4. Figure 4 shows that if center crop is simply used, the left side of the image contains many objects unrelated to the ground truth, which may result in incorrect similarity calculations. On the other hand, if the image is cropped based on Grad-CAM, it is possible to remove the areas unrelated to the ground truth

Neuron 0
Ground truth : tench

Ours : tench
WWW : crucian carp

Neuron 10
Ground truth : ostrich

Ours : ostrich
WWW : emu

Neuron 296
Ground truth : ice bear

Ours : ice bear
WWW : Spitsbergen

Neuron 446
Ground truth : binder

Ours : binder
WWW : stationery, message pad

Neuron 460
Ground truth :breakwater

Ours : breakwater
WWW : Bristol Channel, Solway Firth, pollutant

Neuron 479
Ground truth : car wheel

Ours : car wheel
WWW : drum brake, roundel

Neuron 671
Ground truth : mountain bike

Ours : mountain bike
WWW : pack riding, Wrangell-St, Elias National Park

Neuron 742
Ground truth : prayer rug

Ours : prayer rug
WWW : sarong

Neuron 850
Ground truth : teddy

Ours : teddy
WWW : hamper

Figure 3: Qualitative evaluation of each method.

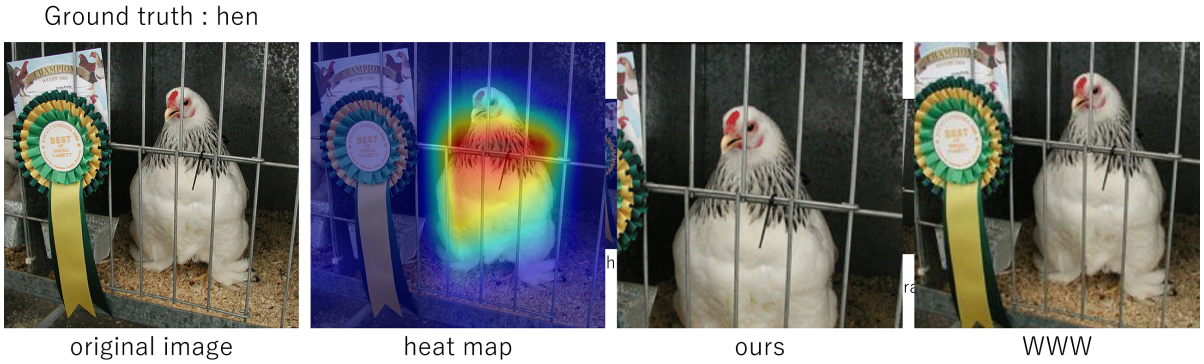

Figure 4: Comparison of the proposed method and WWW crop images.

Table 1: Comparison results with Resnet-50 as target model.

Method	D_{probe}	$D_{concept}$	CLIP cos	mpnet cos	F1-score	Hit Rate
WWW	ImageNet val	ImageNet(1k)	93.31	83.47	**77.53**	**95.5**
	ImageNet val	Broaden(1.2k)	77.79	44.47	6.6	**9.3**
	ImageNet val	Wordnet(80k)	88.76	70.05	42.62	66.2
Ours	ImageNet val	ImageNet(1k)	**93.46**	**83.90**	76.98	**95.5**
	ImageNet val	Broaden(1.2k)	**78.31**	**45.45**	**6.73**	**9.3**
	ImageNet val	Wordnet(80k)	**89.49**	**72.58**	**45.77**	**69.8**

and keep many of the original features of the image. we believe that the proposed method could more accurately measure the relationship between image features and text features by t-vMF similarity while excluding areas that are irrelevant to the original image features using Grad-CAM.

4.3 Quantitative Results

From Table 1, it can be observed that the accuracy has improved for almost all evaluation metrics. The maximum improvements are 0.73% for CLIP cos, 2.53% for mpnet cos, 3.15% for F1-score, and 3.6% for Hit Rate. This improvement is believed to result from the use of Grad-CAM to remove non-essential features of the images while accurately calculating the similarity between image features and text features using t-vMF similarity. Additionally, regarding Broaden and WordNet, both methods show an increase in accuracy as the size of $D_{concept}$ increases. This improvement is considered to arise from the enhanced expressiveness of the concepts that can be assigned to the neurons as $D_{concept}$ grows. The proposed method can reflect the similarity between image features and text features more accurately, which likely leads to a greater increase in the accuracy of evaluation metrics when $D_{concept}$ is changed from Broaden to WordNet compared to existing methods.

Table 2: Results of Ablation Study.

Grad-CAM	t-vMF	CLIP cos	mpnet cos	F1-score
		88.76	70.05	42.62
	✓	88.89	70.45	42.42
✓		89.43	72.42	**45.83**
✓	✓	**89.49**	**72.58**	45.77

5 ABLATION STUDY

In this section, we discuss the contributions of Grad-CAM and t-vMF to the improvements in accuracy. The experimental results are presented in Table 2. From Table 2, it can be inferred that WWW uses center-cropped images, which contain a significant amount of redundant features, leading to relatively low accuracy. Furthermore, when we use t-vMF, a slight improvement in accuracy is observed in CLIP cos and mpnet cos. In contrast, when Grad-CAM is used, improvements in accuracy are confirmed across all evaluation metrics. Additionally, when both Grad-CAM and t-vMF are used together, the maximum accuracy is achieved for two evaluation metrics: CLIP cos and mpnet cos. This is believed to be due to the elimination of redundant features by Grad-CAM while allowing for accurate similarity calculations through t-vMF.

6 CONCLUSION

In this paper, we proposed a method that utilizes Grad-CAM and t-vMF similarity to accurately measure the similarity between the intrinsic features of images and text for improving the discovery accuracy of neuron concepts. As a result, we achieved more accurate identification of neuron concepts across various datasets.

REFERENCES

Ahn, Y. H. et al. (2024). WWW: A unified framework for explaining what, where and why of neural networks by interpretation of neuron concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10968–10977.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549.

Fellbaum, C. (2005). Wordnet and wordnets. encyclopedia of language and linguistics.

Kobayashi, T. (2021). T-vMF similarity for regularizing intra-class feature distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6616–6625.

Radford, A. et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. (2020). Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., and Hu, X. (2020). Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929.