Consistency Explanation for Vision Transformer

Seokhyeon Jeong
sh102201@snu.ac.kr

Wonkyun Kim
wonkyunkim.sj@snu.ac.kr

Juneau Jung
sean2ie@snu.ac.kr

Abstract

Explainable Artificial Intelligence (XAI) has become increasingly important in fields where understanding the decision-making process of a model is critical. Various XAI methods have been proposed to interpret the black box of deep learning. However, these methods have been challenged for their lack of consistency. Several studies have been conducted to enhance the consistency of XAI, utilizing a convolutional neural network (CNN) model and Grad-CAM as the target XAI technique. These studies showed that improving consistency improves not only image classification accuracy but also classification accuracy on fine-grained datasets and in limited-label data environments. However, traditional methods to enhance consistency in large pre-trained models such as ViT and CLIP are computationally expensive, and fine-tuning does not produce significant results. This paper introduces a combined fine-tuning approach utilizing ***VPT: Visual Prompt Tuning*** *and a regularization term designed for application with large pre-trained models and extensive datasets. Additionally, we present the evaluation metrics to assess consistency more precisely.*

1. Introduction

The field of computer vision has achieved outstanding performance in the era of deep learning. Although it has been achieved state-of-the-art on the vision task, the neural network is a black box since we don't understand how the model makes a decision making. For instance, when diagnosing cancer in the medical field, it was a critical problem not knowing how the model diagnose it with which variables [5]. Hence, it's essential to understand how the model processes decision-making [17]. It led to how we can interpret and know how the model works. Then, various XAI methods have been proposed to interpret the Deep Neural Network. Recent methods for image classification tasks have been proposed and can be applied to classification activation maps or attention maps. Since the model can explain decision-making by XAI methods, it is utilized in various applications. However, XAI methods have been challenged for the lack of consistency [3]. For instance, when an image is classified into a class, it does not work as intended by a person. Despite the position transformation of the image, the decision-making of the model is carried out based on some areas of the image that are completely different from the original image. It is related to the reliability and whether the model is working properly. Furthermore, consistency affects the performance of the deep learning model. A model that does not guarantee consistency has not been generalized and is biased toward a specific dataset [10]. In particular, consistency tends to be worse on fewer datasets, affecting model performance [11]. Several studies have been conducted to enhance the consistency of explainable artificial intelligence (XAI) [11, 12]. These studies utilized a convolutional neural network (CNN) model as a foundation and Grad-CAM, a commonly used method in vision, as the target XAI technique. To improve the consistency of Grad-CAM, a regularization term was added to the loss function to ensure that Grad-CAM results were consistent with one another. The model was then retrained with a new regularization term. The results of these studies demonstrate that enhancing consistency not only improves image classification accuracy but also enhances performance in fine-grained datasets and limited-label data environments.

Large pre-trained models such as ViT [6] and CLIP [13] are becoming increasingly prominent in vision tasks. As shown in Figure 1, after transforming the image, the XAI method focuses on different parts of the image compared with the original image in the ImageNet1000 dataset in the ViT as same as the CNN network. This trend is evident in both attention and Grad-CAM methods. However, the computational cost of using traditional methods to enhance consistency in such large models is prohibitively high. As a result, we introduce a cost effective, fine-tuning approach utilizing **VPT: Visual Prompt Tuning** combined with a regularization term. This method has been specifically designed for applications involving large pre-trained models and extensive datasets. Moreover, we have introduced an evaluation metric to measure consistency, aiming to demonstrate the improvements our approach and measure the consistency of explanations precisely.

Our contributions are as follows:

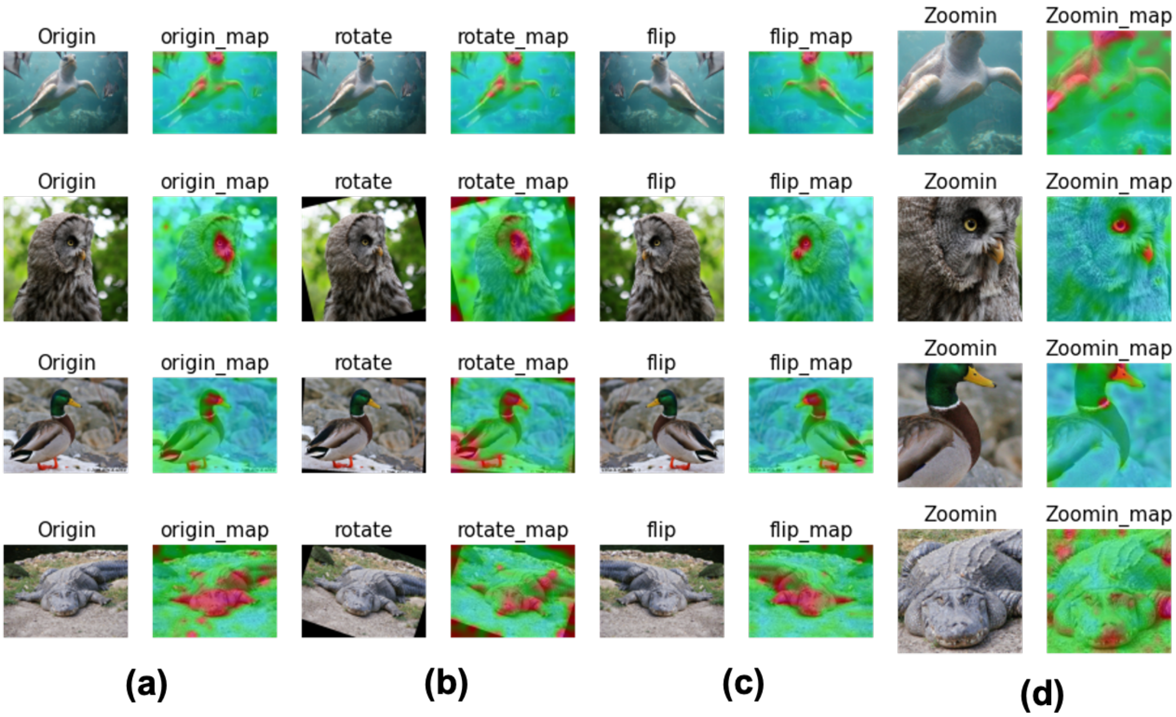


Figure 1. Example of consistency problem in Vision transformer, (a): origin, (b): rotation (c): flipped, (d): zoom-in. Each transformed image is pointing to a different object.

- We confirmed that both attention-based and Grad-cam-based methods have consistency issues in large models such as vision transformers.
- Instead of Conventional fine-tuning, we chose to use a method that combines VPT with the existing regularization term approach. This approach is more effective as it reduces the computational cost and allows us to utilize prompt information.
- We improved the evaluation method used in previous studies.This method allows for a more detailed analysis of consistency than previous methods, thereby providing more objective and intuitive results.

2. Related works

This section provides an overview of relevant existing works on consistency in explanations, with a specific focus on the Vision Transformer model, which is the primary subject of our research.

2.1. Explanations

Deep learning models are often referred to as ”blackbox” due to their lack of interpretability. Various methods have been proposed to address it. Class Activation Mapping-based(CAM-based) techniques offer intuitive interpretability and several benefits. It’s designed in such a way that you can generate a heatmap that highlights the areas in the input image that contributed most to the final decision of the model. This can be used to understand why a model made a certain prediction. Grad-Cam is one of the methods. It use the gradients of the class scores with respect to input image during the backpropagation process to compute the importance of each location. However, deep learning model does not always produces consistent explanations. Slight transformation that are not change the semantics to input images change the explanations frequently.

2.2. Consistency

Several methods have been proposed to achieve consistency in explanations. One approach, presented in [14], incorporates domain knowledge to align explanations with prior knowledge. Another method, discussed in [8], utilizes adversarial perturbations to ensure explanation consistency. Additionally, [18] employs causal masking to generate contrastive images, aiming to improve interpretability, and [7] introduces a perceptual consistency prior to attention heatmaps in the context of multi-label image classification. This approach is based on the notion that the CAM attention heatmap should undergo the same transformation as the image if it is transformed.

In the pursuit of reducing spurious correlations in inter-

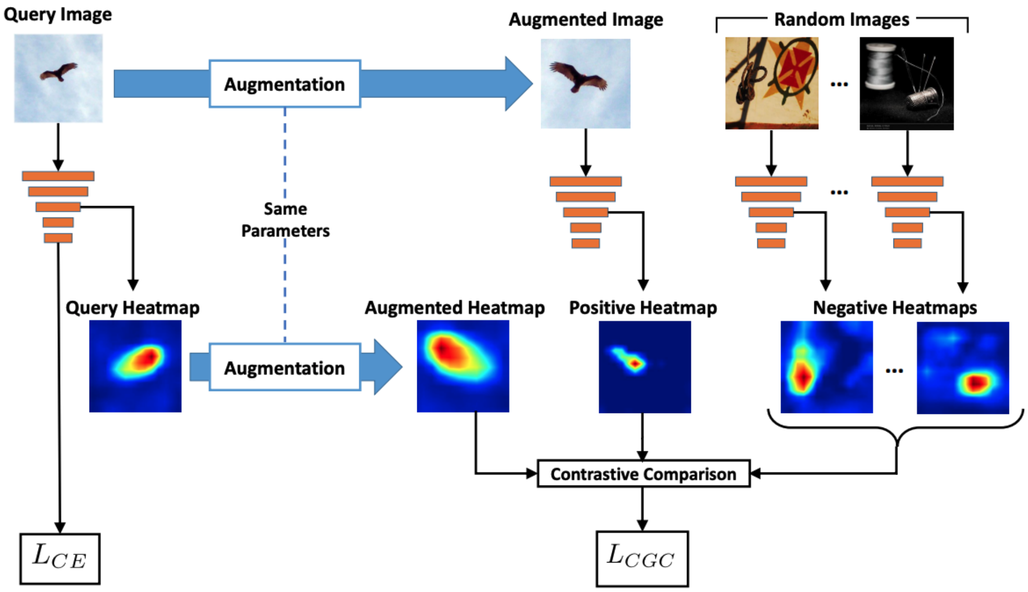


Figure 2. An illustration of the method of CGC loss. We extract the picture in [11]

pretation heatmaps, [11, 12] focus on contrastive learning. These approaches aim to mitigate undesired correlations and enhance the consistency of interpretation heatmaps. The CGC method in Figure 2 is designed to make the model produce more consistent explanations. The authors adopt ideas from contrastive self-supervised learning and apply them to the interpretations of the model rather than its embeddings. The CGC method works by encouraging the Grad-CAM of an image to be close to the Grad-CAM of an augmented version of the same image while being far from the Grad-CAM of other random images. This is achieved by designing a loss function that takes into account these factors. The CGC method acts as a regularizer and improves the accuracy of limited-data, fine-grained classification settings. Through the use of contrastive learning, they try to generate explanations that are both meaningful and consistent. The same author has proposed a similar method. This method involves generating a larger composite image using a 2x2 grid, where four images are randomly placed within the grid's cells [12]. The model is then trained to minimize the difference between the interpretations of the original image and the composite image. However, applying this method to Vision Transformers can be challenging. We adopt the CGC method [11] as a baseline.

2.3. Vision Transformer

Vision Transformer [6] is significant in that the architecture of the Transformer is used for image processing. When attention techniques are used in the existing computer vision field, they are primarily used with CNN or used to replace only specific components while attracting the entire CNN structure. However, the vision transformer showed better performance than the existing CNN-based model by applying a transformer that uses a sequence of image patches as an input value without relying on CNN.

Vision Transformer (ViT) constructed a model in the form of directly putting the image itself into the standard Transformer. To this end, the image was divided into patch units. By applying linear embedding to the patch, it can be transmitted as a Transformer input value in the form of a sequence. When training on a medium-sized data set, it does not show good performance compared to the existing ResNets model. This reason can be confirmed that the Transformer structure itself lacks inductive bias compared to the CNN structure, and generalization is not achieved without a large amount of data. However, when training a large data set of 14 million to 300 million pages, it was shown to overcome the structural limitation. To improve the computationally intensive vision transformer, an efficient vision transformer model called the DeiT [16] model has also been studied. We used both ViT and DeiT in this study.

3. Proposed methods

In existing CNN-based methods, a regularization term is added to consider the consistency of Grad-CAM during learning. To apply these methods to large models such as vision transformers, fine-tuning is necessary. In general, there are two types of fine-tuning methods that utilize a large pre-trained model to learn data, (a) linear probing and (b) end-

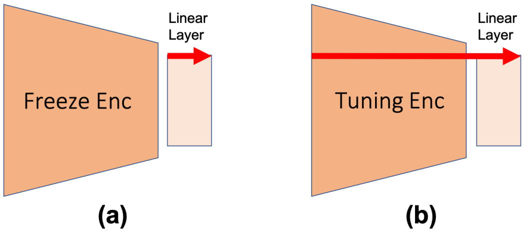


Figure 3. Example of the fine-tuning case, (a): linear probing (b): end-to-end fine-tuining

to-end fine-tuning as depicted in figure 3.

End-to-end fine-tuning involves learning at a high rate on a new dataset, resulting in a good performance but high training costs due to the need to update new parameters. Linear probing, on the other hand, involves freezing the pre-trained model parameters and adding a linear layer for training. Sometimes, it is good to add a few MLP layers and a head layer. This approach is often used to analyze the performance of existing multimodal encoders.

In the case of linear probing, the performance is inferior to full fine-tuning because only the linear layer is trained, and end-to-end fine-tuning takes a long time. Specifically, applying the method used in our previous method paper [11] to end-to-end fine-tuning would consume a considerable amount of time, thus diminishing its effectiveness.

Therefore, we will adopt a new tuning method, which will be introduced in the following section.

3.1. VPT: Visual Prompting Tuning

Prompt learning has recently been applied to various fields for in-context learning of large language models. In [9], the pre-trained vision transformer model was enhanced by adding learnable prompts to the input path, resulting in significant improvements in downstream task performance. This method can be applied to a range of tasks, including classification, segmentation, and object detection. Figure 4 from the original paper illustrates how visual prompt tuning is performed(We took this figure in the original paper). In (a), learnable parameters are prepended to the input of each Transformer encoder layer (VPT-Deep), while in (b), prompt parameters are only inserted into the first layer's input (VPT-shallow). During downstream task training, only the prompt and linear head parameters are updated while the entire Transformer encoder remains frozen.

We will investigate the extent to which visual prompting aids in achieving consistency. This will involve breaking the process down into steps and examining each method sequentially to assess its individual usefulness.

3.1.1 Visual Prompting Tuning: Shallow and Deep

In this approach, the prompt will be a vector of the same size as the image token, which will undergo training. The initial prompt is generated from a random vector that follows a Gaussian distribution and is subsequently updated using a loss function tailored to the downstream task. For instance, in classification tasks, the prompt vector can be updated with a cross-entropy loss.

As previously discussed, there are two strategies for visual-prompt tuning: VPT-Deep, which learns the visual prompt and head by feeding the learnable visual prompt through the transformer encoder at each step, and VPT-Shallow, which only learns the visual prompt and head, allowing the visual prompt to be learnable solely for the first input, thus having the fewest parameters.

Even though VPT-Deep requires learning more parameters than the Shallow method, it still involves learning less than 1% of the parameters compared to the end-to-end fine-tuning of the entire model. Consequently, we will initially experiment with visual prompting tuning rather than end-to-end fine-tuning on pre-trained ViT and DeiT models.

3.1.2 Visual Prompting with Contrastive Grad-CAM Loss

In this approach, we will train the pre-trained model using contrastive Grad-CAM loss as presented in the paper [14]. Instead of employing CGC loss solely for fine-tuning, we combine CGC loss with cross-entropy loss for visual prompting tuning. Regrettably, the resulting accuracy is subpar, and the process is quite time-consuming due to the requirement of multiple gradient calculations for Grad-CAM. Given the low accuracy, we are not confident in the evaluation metric, so these findings are excluded from the results.

3.1.3 Interpreting Image Tokens with Visual Prompt

During visual prompting tuning, the prompt token learns and adapts. We have observed that new information is gained in this process, suggesting that if we can better interpret the visual prompting and align it with the image data, we may enhance the consistency of Grad-CAM. A simple approach to achieve this would be to average the visual prompt tokens and incorporate them into each image.

To explore a more sophisticated technique, we are searching for research papers utilizing the application of clips to visual prompting, aiming to improve alignment with images and overall consistency.

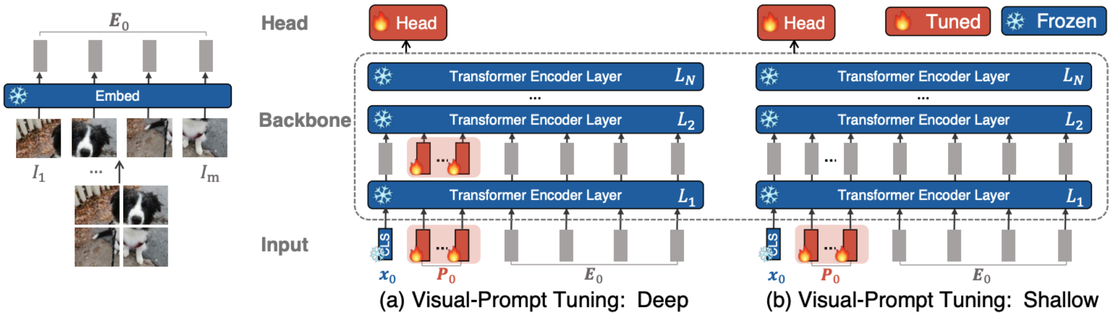
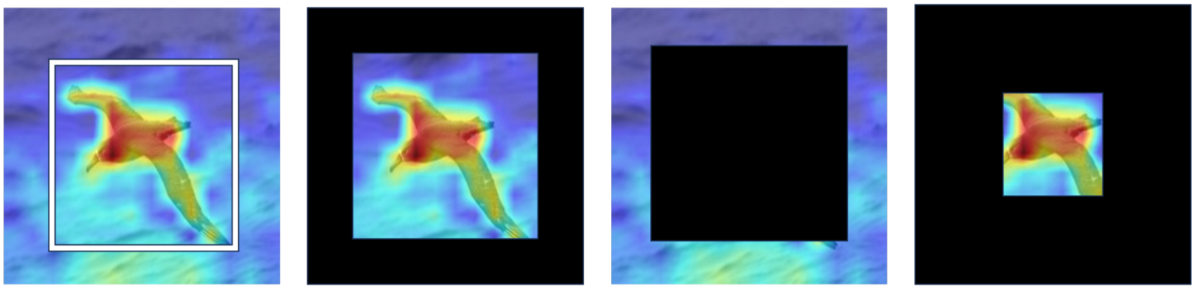

E_0
Embed
I_1
...
I_m
Head
Backbone
Input
Head
Transformer Encoder Layer L_N
Transformer Encoder Layer L_2
Transformer Encoder Layer L_1
CLS
x_0
P_0
E_0
Head
Tuned
Frozen
Transformer Encoder Layer L_N
Transformer Encoder Layer L_2
Transformer Encoder Layer L_1
CLS
x_0
P_0
E_0
(a) Visual-Prompt Tuning: Deep
(b) Visual-Prompt Tuning: Shallow

Figure 4. Visual Prompting tuning framework, (a): Learnable parameters are prepended to the input of each Transformer encoder layer (VPT-Deep), (b): Prompt parameters are only inserted into the first layer's input (VPT-shallow). We extract the picture in [9]



(A) bounding box

(B) Base content heatmap

(C) Reverse content heatmap

(D) Zoomin content heatmap

Figure 5. An illustration for content heatmaps

4. Experiments

4.1. XAI-methods

We used Grad-CAM, Grad-CAM++ [15], Score-CAM [19], and RollOut [2] as XAI Method for evaluation. The first three methods are class-activation map-based, while the last is attention-based. Our investigation of consistency trends according to various XAI methods will focus solely on Grad-CAM++, Score-CAM, and RollOut. For evaluation against the baseline, however, we will exclusively employ the Grad-CAM technique. This decision is justified by the desire for a fair comparison, as [11] also utilizes Grad-CAM in its analysis. Techniques that perform effectively for Grad-CAM are expected to demonstrate similar effectiveness for attention-based methods in the future.

4.2. Implementation details

In our work, we utilized the pre-trained ViT("vit-base-patch16-224") trained on JFT-300M dataset and DeiT("deit-tiny-patch16-224") models trained on ImageNet1000 data. We are implementing the pre-trained model while our results thus far have been obtained using Torch Hub [1]. Both models are also used in our baseline and proposed method. For the fine-grained dataset employed in our study, overtraining could lead to overfitting due to its limited size. Consequently, during end-to-end fine-tuning on the CUB200 dataset, we utilized a batch size of 32 for 10 epochs with DeiT and a batch size of 64 for 7 epochs with ViT. The AdamW optimizer was employed for training using a learning rate of 1e-4, resulting in successful learning. To address potential underfitting in the DeiT visual prompting tuning, we increased the epochs to 20, consistently implementing this adjustment for both VPT-deep and VPT-shallow. During the evaluation process of VPT-deep(mean) and VPT-deep(shallow), the VPT model simply integrated the average of the prompt values into the image tensor.

Our exploration of CGC loss encompassed various experimental approaches. Initially, the existing CGC loss method applies Grad-CAM at the onset of contrastive learning. In contrast to CNN models, applying Grad-CAM to transformers produces noisy results due to multiple objects being captured in underfitting situations. This interference hinders achieving accurate learning. To overcome this challenge, we trained models using two techniques and selected the top-performing model. We experimented with modifying the λ value of the CGC loss regularization term and utilizing a two-phase approach, wherein initial training was

	DEIT				VIT			
	Acc	Base CH(%) ↑	Reverse CH(%)↓	Zoomin CH(%)↑	Acc	Base CH(%)↑	Reverse CH(%)↓	Zoomin CH(%)↑
Base	72.25	67.97	59.14	62.68	79.55	**85.86**	75.79	**82.52**
CGC	72.90	69.06	60.15	63.75	81.62	83.14	72.46	76.96
VPT-deep	63.79	**79.53**	71.48	**75.13**	81.03	82.09	76.54	81.47
VPT-deep(mean)		79.45	71.56	75.19		80.14	75.60	80.04
VPT-shallow	64.61	72.62	58.55	64.74	77.06	67.50	56.47	64.10
VPT-shallow(mean)		72.45	58.43	64.47		61.52	**50.43**	57.60
VPT-shallow+CGC	58.77	70.70	55.50	61.33	-	-	-	-
VPT-shallow(mean)+CGC		70.35	**55.20**	61.30		-	-	-

Table 1. Accuracy and Content Heatmap for CUB200-2011 test set

conducted using cross-entropy loss followed by additional training with the CGC loss term. Our findings indicate that the CGC method, though ineffective in the transformer context, warrants further investigation for application in CNN models.

4.3. Baseline

We present the baseline methods outlined in Table 2.

1. **Base** The pre-trained model undergoes end-to-end fine-tuning with cross-entropy loss.

2. **CGC** The pre-trained model is end-to-end tuned using CGC loss.

3. **VPT-deep** Deep visual prompt tuning is applied to the pre-trained model.

4. **VPT-shallow** Shallow visual prompt tuning is implemented on the pre-trained model.

5. **VPT-deep(mean), VPT-shallow(mean)** Same as VPT-deep, VPT-shallow, but When calculating Grad-CAM for evaluation, the mean of the prompt token is added to the image token.

6. **VPT-shallow+CGC** Deep visual prompt tuning with CGC loss is applied to the pre-trained model. We only trained VPT-shallow with CGC because of computation resources.

It is important to note that when computing Grad-CAM for VPT-deep and shallow, only image tokens are considered, excluding learned prompts. In contrast, for VPT-deep(mean) and VPT-shallow(mean), the image token calculation incorporates the average of the prompt tokens.

4.4. Dataset

We will make use of small datasets with fine-grained and bounding box annotations, such as the CUB-200 [20]. For the dataset, we have implemented a function to assess the degree of overlap between the bounding box and the Grad-CAM.

4.5. Evaluation

4.5.1 Base Content Heatmap(BCH)

This metric, introduced in [12], measures the sum of the ℓ_1-normalized heatmap within the annotated bounding box of an object. If the model interpretation aligns with human annotations of the object's location, it can be assumed that the percentage of the heatmap within the object annotation mask should be close to 100%. As a result, a high value for this metric is expected. Refer to the figure 5 (B).

A higher degree of overlap between the bounding box and model interpretation signifies enhanced localization of the model interpretation. This can be perceived as increased **consistency**, as the model focuses on sparse areas even when the image undergoes transformations.

4.5.2 Reverse Content Heatmap(RCH)

This metric could represent the sum of the ℓ_1 normalized heatmaps outside the bounding boxes annotated with objects. In this case, if the model interpretation matches human annotations concerning object locations, the heatmap ratio outside the object annotation mask should be low. In figure 5 (C), one can observe the percentage of model interpretations that correspond to the background outside the bounding box in the entire figure. Lower values indicate superior model localization, whereas higher values suggest noisy model interpretations.

4.5.3 ZOOMin Content Heatmap(ZCH)

This metric might measure the sum of the ℓ_1 normalized heatmaps within the bounding boxes annotated with objects, focusing on more specific or localized regions rather than the entire bounding box. Figure 5 (D) illustrates the distribution of model interpretation within the localized space, with the bounding box further constricted around the center. For this metric, a value closer to 100% is deemed more desirable.”

4.6. Experiment Result

4.6.1 Content Heatmap in ImageNet1000

Prior to conducting our full-scale experiment, we compared the Grad-CAM-based and Attention-based methods for the Content heatmap metric on the ImageNet1000 validation set. While the accuracy was lower than that of ResNet18, the CH metric was also lower for the transformer series. For the RollOut method, the CH score was low due to poor localization on the vision transformer. The following experiments were performed on Grad-CAM.

	DEIT		VIT		ResNet18	
	Acc	Base CH(%)	Acc	Base CH(%)	Acc	Base CH(%)
Grad-CAM	72.1	58.4	78.1	52.5	69.7	54.47
Grad-CAM++		47.8		40.2		-
RollOut		34.4		36.6		-
Score-CAM		54.3		49.7		-

Table 2. Accuracy and Content Heatmap for ImageNet1000 validation set

4.6.2 GradCAM Content Heatmap in CUB200

Before examining the experimental results, we would like to clarify our prior understanding: higher values of Base Content Heatmap and Zoom-in Content Heatmap, alongside lower values of Reverse Content Heatmap, indicate better model interpretation alignment to the object rather than the background. This consistency leads to more robust model interpretation and improved model prediction.

We conducted experiments in Table 1 with end-to-end fine-tuning, CGC loss, interpretation of models trained by visual prompt-tuning (a variant of VPT), and a combination of VPT and CGC loss, as outlined in the prior section. Initially, in DeiT, a relatively small model, we observed high BCH and ZCH values when training solely with VPT and assessing model interpretation. It can be inferred that it detects more localized areas effectively; however, the RCH value is concurrently high, rendering the overall interpretation rather noisy. It is remarkable that incorporating VPT and CGC in DeiT resulted in higher BCH and lower RCH compared to end-to-end fine-tuning outcomes. In terms of BOR (Base content heatmap over Reverse), VPT with CGC loss demonstrates higher effectiveness, indicating that a larger BOR value signifies less noise when localized. Conversely, we noticed that merely combining CGC with end-to-end fine-tuning had minimal impact. Integrating CGC with standalone training didn't yield substantial benefits for the transformer family of models.

We also discovered that employing prompts in VPT, which was our initial expectation, did not prove efficacious for DeiT and ViT. VPT-shallow was even less effective on ViT. But we believe visual prompting remains a promising approach. Techniques like CGC continued to exhibit strong performance when applied in conjunction with visual prompting. Moreover, there exists research investigating the concrete implications of learned visual prompting, often employing clips. [4] We plan to leverage these studies in the future.

5. Conclusion

In conclusion, we have presented a combined method that addresses the challenge of consistency in Explainable Artificial Intelligence (XAI) for deep learning models. Our approach combines Visual Prompt Tuning (VPT) and a regularization term, making it suitable for large pretrained models such as ViT and CLIP, as well as extensive datasets. The proposed method alleviates the computational expense associated with traditional techniques and yields improved results compared to fine-tuning, especially in the DeiT model. Furthermore, we have introduced an evaluation metric to assess consistency. In future work, we plan to investigate the effective utilization of the tuned prompts in visual prompt tuning to further enhance interpretability and performance.

References

[1] Torch hub. https://pytorch.org/hub. Accessed: 2023-05-11. 5

[2] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020. 5

[3] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018. 1

[4] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 1(3):4, 2022. 7

[5] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015. 1

[6] 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. 1, 3

[7] Hao Guo, Kang Zheng, Xiaochuan Fan, Hongkai Yu, and Song Wang. Visual attention consistency under image transforms for multi-label image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 729–739, 2019. 2

[8] Tao Han, Wei-Wei Tu, and Yu-Feng Li. Explanation consistency training: Facilitating consistency-based semi-supervised learning with interpretability. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 7639–7646, 2021. 2

[9] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 709–727. Springer, 2022. 4, 5

[10] Ali Mirzazadeh, Florian Dubost, Maxwell Pike, Krish Maniar, Max Zuo, Christopher Lee-Messer, and Daniel Rubin. Atcon: Attention consistency for vision models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1880–1889, 2023. 1

[11] Vipin Pillai, Soroush Abbasi Koohpayegani, Ashley Ouligian, Dennis Fong, and Hamed Pirsiavash. Consistent explanations by contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10213–10222, 2022. 1, 3, 4, 5

[12] Vipin Pillai and Hamed Pirsiavash. Explainable models with consistent interpretations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2431–2439, 2021. 1, 3, 6

[13] 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. 1

[14] Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International conference on machine learning*, pages 8116–8126. PMLR, 2020. 2

[15] 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. 5

[16] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 3

[17] Alfredo Vellido. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications*, 32(24):18069–18083, 2020. 1

[18] Dong Wang, Yuewei Yang, Chenyang Tao, Zhe Gan, Liqun Chen, Fanjie Kong, Ricardo Henao, and Lawrence Carin. Proactive pseudo-intervention: Causally informed contrastive learning for interpretable vision models. *arXiv preprint arXiv:2012.03369*, 2020. 2

[19] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. 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, 2020. 5

[20] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010. 6