ViT-Calibrator: Decision Stream Calibration for Vision Transformer

Lin Chen[1], Zhijie Jia[1], Tian Qiu[1], Lechao Cheng[2,†],
Jie Lei[3], Zunlei Feng[1], Mingli Song[1]
[1]Zhejiang University, [2]Zhejiang Lab, [3]Zhejiang University of Technology
{lin_chen,tqiu,zunleifeng,brooksong}@zju.edu.cn,
zhijiejia1998@outlook.com, chenglc@zhejianglab.com, jasonlei@zjut.edu.cn

Abstract

A surge of interest has emerged in utilizing Transformers in diverse vision tasks owing to its formidable performance. However, existing approaches primarily focus on optimizing internal model architecture designs that often entail significant trial and error with high burdens. In this work, we propose a new paradigm dubbed Decision Stream Calibration that boosts the performance of general Vision Transformers. To achieve this, we shed light on the information propagation mechanism in the learning procedure by exploring the correlation between different tokens and the relevance coefficient of multiple dimensions. Upon further analysis, it was discovered that 1) the final decision is associated with tokens of foreground targets, while token features of foreground target will be transmitted into the next layer as much as possible, and the useless token features of background area will be eliminated gradually in the forward propagation. 2) Each category is solely associated with specific sparse dimensions in the tokens. Based on the discoveries mentioned above, we designed a two-stage calibration scheme, namely ViT-Calibrator, including token propagation calibration stage and dimension propagation calibration stage. Extensive experiments on commonly used datasets show that the proposed approach can achieve promising results. The source codes are given in the supplements.

1. Introduction

Image classification is an important research area in computer vision that involves quantitatively analyzing digital images and categorizing them into different classes. The associated approaches have been widely applied in practical scenarios such as medical image diagnosis [21], security monitoring [31], and autonomous driving [5]. The remarkable performance of deep neural networks has propelled them to the forefront of image classification, where they are now regarded as a mainstream approach. As an indispensable part of deep neural networks, Convolutional Neural Networks (CNNs) have exhibited exceptional processing capabilities for translation invariance and local structures [29], leading to outstanding classification performance. The emergence of Transformers has recently opened up new possibilities for visual feature learning. The self-attention mechanism in Transformer models captures global semantic information [38], allowing Transformers to be more proficient in handling long sequential data. Consequently, deep models based on Transformers have achieved comparable or even superior performance to CNNs in computer vision, as demonstrated by [17, 32, 36, 41]. Although the Transformer model demonstrates satisfactory classification performance, its opaque internal transformations and learning processes impede a profound understanding and analysis of its internal mechanisms, making it challenging to improve its performance through modification and adjustment.

Currently, there have been some model diagnosis works, mainly focusing on traditional model repair and detection [27]. The latest work [20] on optimizing deep models exploit gradient constraint strategies to diagnose and repair convolutional neural network models. In addition, some works [3, 30] apply interpretable decision tree methods to approximate deep learning models for analysis and detection. Furthermore, Other works, such as, visualization techniques [25] have also explored to explain and diagnose [24, 44] the predictions of deep learning models to boost the performance. Despite these related works for optimizing deep learning models, there are no effective methods yet for diagnosing and repairing Transformer models based on the attention mechanism.

In this work, we introduce a novel Decision Stream Calibration paradigm that boosts performance by explicating the information propagation mechanism based on the correlation among tokens and the relevance coefficient across dimensions. We have derived two insightful discoveries from

†Corresponding author.

empirical experiments, one of which is derived from the biological neural feedback principle [14], and we incorporate this idea by developing a dynamic feedback loop mechanism that enables the interaction between high-level and shallow-level semantic information. While the remaining one manifests, that specific sparse dimensions of the deep features of the Transformer are highly correlated with the target category. In contrast, other irrelevant dimensions can negatively affect classification performance. Consequently, we devise a two-stage information propagation mechanism to address the defects at both the token and dimension levels.

Specifically, we first introduce an elaborated network with a feedback module, as illustrated in Figure 3 (Token Feedback Stage). Inside the module, we define a feedback input layer capable of capturing and providing more pertinent semantic information for various categories. Besides, we also present a shallow network that extracts basic visual information as the target layer of the feedback mechanism. Next, the output features of the feedback layer and the target layer are harmonized. The proportion of deep feature feedback to the shallow network is determined by measuring the similarity between the deep and shallow networks. Moreover, the feedback mechanism selectively feeds back advantageous deep features to the shallow network using the similarity measure. To provide feedback on deep semantics, fusion tuning is performed for both attention and token features in the feedback mechanism.

In the second phase, the anticipated class assignment vector for particular layer categorization tokens is utilized as a singular relevance gauge, reflecting the extent of interrelation between the target category and the corresponding dimension. For all training samples of the same category, we aggregate these relevance metrics in the dimensions of the transformer-specific classification tokens, which serve as the criteria for calibrating erroneous samples. Subsequently, a distillation technique is employed to reformulate and constrain the flawed association between specific dimensions and target categories.

Therefore, our contribution is to propose a Decision Stream Calibration framework called ViT-Calibrator. We provide two new perspectives for optimizing the Transformer model: token feedback and dimension constraints. Experiments show that the proposed method can effectively calibrate wrong feature stream in the forward propagation and further improve the performance of the Transformer model. Apart from this, the ViT-Calibrator is based on the original network and only requires fine-tuning, avoiding the huge time cost of retraining the Transformer.

2. Related Work

2.1. Transformer-based Classification

Inspired by the tremendous success of Transformer in natural language processing [16, 35], it has also been widely used in computer vision. To enhance the model's receptive field and global dependency, ViT [17] was first proposed for image classification, and it outperformed many traditional convolutional neural networks. Afterward, various models based on ViT were proposed for image classification. Some works enhanced the Transformer with the spatial inductive bias of CNN. Hugo Touvron et al. [36] proposed a model Deit based on knowledge distillation, data augmentation, and small-batch training to address the poor performance of ViT on small datasets. ConVit [18] combined CNN and Transformer to improve computational efficiency and classification performance through an adaptive feature importance weighting mechanism. Chun-Fu Richard Chen et al. [10] proposed a Transformer model CPVT with a variable receptive field and variable precision and designed a more flexible local and global information interaction mechanism. Zihang Dai et al. [13] designed a lightweight visual Transformer network CoAtNet based on multi-resolution input and grouped convolution, achieving efficient multi-resolution feature extraction and information interaction.

In addition to convolution, many researchers have proposed a local attention mechanism to focus on adjacent elements and enhance local feature extraction dynamically. One representative method is Swin Transformer [32]. Swin used a moving window along the spatial dimension to model global and boundary features. On the other hand, ViT ignored fine-grained features and brought high computational costs due to the fixed-resolution pillar structure used throughout the Transformer layer. Li Yuan et al. [41] proposed a model T2T-ViT that introduced the paradigm of hierarchical Transformers and used overlapping unfold operations for downsampling. However, this operation brings heavy memory and computational costs. Therefore, Wenhai Wang et al. [39] used non-overlapping patch partitioning to reduce feature size in model PVT. In general, increasing the depth of the model can enhance its learning ability. Hugo Touvron et al. [37] proposed a cross-scale attention mechanism, Cait, that simultaneously considers global and local information to improve performance in image classification tasks. Daquan Zho et al. [47] aggregates cross-head attention maps and increases cross-layer feature diversity in model DeepViT by regenerating new attention maps using linear layers . Additionally, some other research attempts to design various self-supervised learning schemes [2, 8] for ViT in a generative and discriminative manner.

2.2. Model Diagnosis

The operational mechanism of machine learning models is often very complex and lacks interpretability and transparency, which makes it difficult for researchers to debug the models. In order to help researchers debug and analyze models, some model interpretation techniques have been developed to improve their comprehensibility and reliability. Cadamuro et al. [7] proposed a machine learning model debugging method based on optimization techniques, which can be used to identify training items that are most likely to cause model bias. In addition, there are works [6, 25, 28] devoted to interactive visualization analysis, supporting users in visually inspecting predictions of black box machine learning models to understand the internal logic of the model's predictions. The above works focus on analyzing and debugging traditional machine learning models, which require human-machine interaction and cannot fully explain model problems.

For deep models, Bastani et al. [3] proposed a method of using symbolic regression to explain deep learning models, and Jose Gustavo S Paiva et al. [34] proposed an incremental visualization data classification method to debug and improve models as training data increases. In addition, using interpretable random forests [30] to approximate black box models is also a deep model interpretation method, and debugging black box models by checking interpretable models. Model-independent explanation and diagnostic methods [44] are also a way to use visualization analysis technology to support the explanation, debugging, and comparison of machine learning models interactively. Recently, Feng et al. [20] proposed a gradient-constrained convolutional neural network model optimization method that uses gradient constraints to optimize convolutional neural networks automatically. Unlike the above methods, we focus on automatically processing Transformer models based on diagnostic results.

2.3. Model Interpretability in Computer Vision

The interpretability of computer vision typically refers to explaining why a model makes specific predictions and which features are crucial in predictions, usually by generating a heatmap that describes the correlation between image locations and prediction results. Currently, there are various interpretability methods, including perturbation-based [22, 23], backpropagation-based [43], saliency map-based [12, 42, 45, 46], and Shapley value-based methods [11, 33]. Among these, LRP [3] is an outstanding interpretability model that recursively allocates relevance from deep layers to earlier ones while ensuring the total sum of relevance across all layers remains constant.

The interpretability research of Transformer models mainly focuses on attention mechanisms. Abnar et al. [1] proposed a method that combines attention scores across layers. However, it cannot distinguish the positive and negative contributions to decision-making, leading to the accumulation of cross-layer relevance scores. To address this issue, Chefer et al. [9] proposed a new method for information propagation within Transformer model components based on LRP attribution, which comprehensively understands the decision-making and inference processes within the model. Those model interpretability can be only used to analyze some failure cases or understand the decision-making mechanism. Those methods can't be used to diagnose and treat the deep model automatically.

3. Decision Stream Mechanism of ViT

Discovery 1. *The final decision is associated with tokens of foreground targets, while the token features of the foreground target will be transmitted into the next layer as much as possible, and the useless token features of the background area will be eliminated gradually in the forward propagation.*

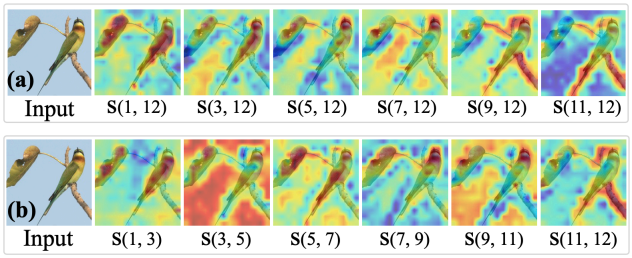

Figure 1: Visualization results of similarity across different layers, where $S(l, l')$ represents the correlation calculation between the output of layer l and the output of layer l'.

In this section, we analyzed the different contributions of deep-level tokens to the final classification results. For an input image I, the output of the $(l-1)$-th layer is denoted as $\{x_1^{l-1}, x_2^{l-1}, ..., x_N^{l-1}, x_{cls}^{l-1}\}$, and the l-th layer of the Transformer encoder module is denoted as $B^{(l)}$, where N denotes the total number of spatial tokens. Therefore, the output of the l-th layer can be calculated as follows:

$$x^l = B^l(x^{l-1}).$$

When l is the final layer, the predicted category $\overline{y}_c$ is determined by $\overline{y}_c = f(x^l)$, where f represents the fully connected layer.

Since deep features contain more semantic information, the output feature x^L of the final layer is directly related to the classification result. Therefore, we calculate the correlation matrix between x^L and the shallow output feature x^l to indicate the correlation between different tokens and the classification result:

$$S(L, l) = \frac{(x_i^L)^T x_i^l}{||x_i^L||\ ||x_i^l||},$$

where i represents the index of tokens between different layers. Based on the similarity of tokens between shallow and deep layers, we found that tokens have different propagation relationships for different layers.

Figure 1(a) shows the cross-layer similarity between the output token of the last layer and the shallow token in Deit [36] on a correctly classified image from the ImageNet dataset. We can observe that there is a combination of foreground and background information during the propagation of image information. As the information is transmitted to deeper layers, there is an even greater emphasis on foreground information, which is most relevant to the image category in terms of semantics. Figure 1(b) shows the similar relationship of tokens between neighboring layers. As can be seen in the figure, there are different transmission mechanisms for image information in different layers. For shallow layers, image information may focus on background information in neighboring layers, which is related to the fact that shallow layers are primarily responsible for processing basic image information. As the number of layers increases, the proportion of foreground information transmitted by the image gradually increases, and the background information is gradually filtered out, resulting in a reduction of information that interferes with the classification decision. For correctly predicted images, foreground tokens have a more significant impact on the classification results, and the foreground token feature will be transmitted into the following layers as much as possible.

Discovery 2. *Each category is only related to a specific dimension in the CLS token, and irrelevant dimensions interfere with the model's prediction results.*

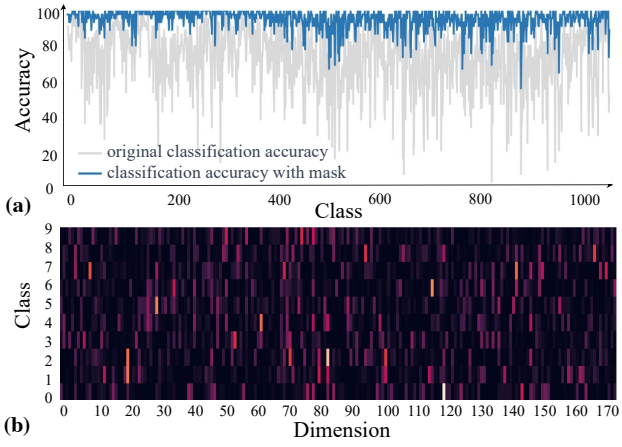

Figure 2: (a) The original classification accuracy and the accuracy after removing dimensions unrelated to the category of Deit on ImageNet. (b) Statistical correlations of different dimensions with the last layer's CLS token for 10 classes on ImageNet.

In this paragraph, we use the relevance propagation technique LRP [3] to attribute the vision Transformer classification results. We assume that the relevance of the actual predicted output of the final layer classifier is R_{final}. In the propagation mechanism, the total sum of relevance coefficients across layers must be equal. Therefore, the following conditions are satisfied:

$$R_{final} = \sum_{e=1}^{V(l+1)} R_e^{(l+1)} = \sum_{e=1}^{V(l)} R_e^{(l)} = \cdots = \sum_{e=1}^{V(1)} R_e^{(1)},$$

where e represents the index of the corresponding vector dimension, and l represents different layers. Under the condition of satisfying the equal sum of relevance, the relevance propagation between the neurons in two consecutive layers is given as follows:

$$\sum_{e'}^{V(l)} R_{e' \leftarrow e}^{(l,l+1)} = R_e^{(l+1)},$$

$$\sum_{e}^{V(l+1)} R_{e' \leftarrow e}^{(l,l+1)} = R_{e'}^{(l)}.$$

We define $R_{e' \leftarrow e}^{(l,l+1)}$ as the portion of relevance that flows from e-th neuron to e'-th neuron. Then, we apply the LRP [3] to calculate the relevance of Transformer.

As described in *discovery 1*, the output of layer l, denoted as $x^l = \{x_1^l, x_2^l, ..., x_N^l, x_{cls}^l\}$, is a collection of d-dimensional feature vectors, where N denotes the total number of spatial tokens. Using the same attribution technique, assuming the relevance coefficient of the target prediction output is R_c, and the relevance coefficient of the d-th dimension of the n-th token output of layer l is $(R_n^l)_d$, satisfying certain conditions:

$$\sum_{n=0}^{N} \sum_{d=0}^{D} (R_n^l)_d = R_{final},$$

where n represents the token index, and N represents the total number of tokens output by a particular layer. Therefore, $(R_{cls}^l)_d$ can represent the correlation between the target category and a specific dimension of the CLS token in a particular layer.

Figure 2(b) shows the statistical correlation between all dimensions of the last layer CLS of Deit [36] and the categories on 10 classes of the ImageNet dataset. For each category, we computed the sum $\sum_1^{100} R_{cls}^l$ of the correlation coefficients of the CLS dimensions for 100 images whose prediction confidence was higher than 0.90. We can observe that each category only correlates with sparse and specific dimensions (bright colors), while most dimensions are not correlated (dark colors). For different high-confidence images in each category, they maintain consistent dimension

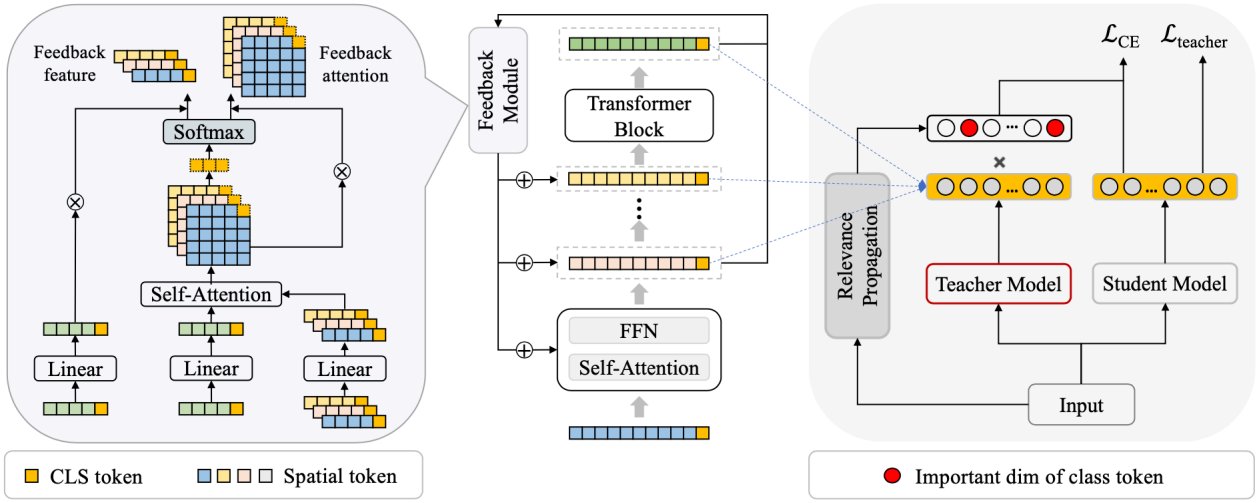

Figure 3: The framework of Vision Transformer Calibrator is composed of token calibration and dimension calibration, which can calibrate the incorrect feature stream from token-level and dimension-level of the class token, respectively.

correlations for that specific category. Figure 2(a) shows the change in accuracy when only relevant dimensions for each class are retained, and other dimensions are set to 0, compared to the original accuracy. We can see that the accuracy significantly increases, indicating that irrelevant dimensions greatly interfere with the classification performance. Meanwhile, the dimensional consistency of the deep network is more regular than that of the shallow network.

4. Vision Transformer Calibrator

Based on the two discoveries mentioned above, we propose a two-stage optimization method for Transformer models to calibrate Transformer classification decision stream. In the first stage of diagnosis and treatment, according to *discovery 1*, the foreground features of the image contribute more to the final classification result. Therefore, we assign greater weights to dominant features by comparing the similarities between tokens. Then we use deep features to guide shallow features and implement the fusion of deep and shallow features through a feedback module. In the second calibration stage, we accumulate the correlation between deep-dimensional features and the predicted category *(discovery 2)*, and based on this, we use a distillation scheme to retain more dominant features while constraining interfering features for the predicted category. The two-stage calibration is described in the following sections.

4.1. Token-level Decision Stream Calibration

Inspired by *discovery 1*, we guide shallow features with deep advantageous features. As shown in Figure 3, we assume that the set of output vectors of the l-th layer of the Transformer block is $x^l = \{x_1^l, x_2^l, ..., x_N^l, x_{cls}^l\}$, denoted as x^l for shallow layers and x^L for deep layers.

For the first feedback mode, we first project the feature x_i^l of i-th token in the l-th layer and the deep feature x_j^L of j-th token in the L-th layer into the same semantic subspace. Because the feature vectors of different layers are actually in different vector spaces, additional vector alignment operations are required, and linear projection is used here for alignment. Finally, the projected vectors are interacted with each other to obtain the feedback offset of the self-attention connection:

$$a_{ij} = (Ux_i^l)^T(Vx_j^L),$$

where U and V are weight matrices used for vector alignment, and a_{ij} represents the semantic similarity of feature vectors between layers after projection into the same space, T denotes matrix transposition.

For the second feedback mode, the deep feature vector is projected nonlinearly by a Multi-Layer Perception (MLP) module onto the semantic space where the L-th layer vector is located to obtain i-th bias b_i for the projected feature x_i^L as follows:

$$b_i = MLP(x_i^L).$$

In addition, to incorporate more advantageous features, we assign different attention weights to tokens by using inter-layer similarity to generate weight matrices corresponding to different tokens. For the deep layer output features, denoted as x^L, and the corresponding shallow layer features, denoted as x^l. Then, the similarity w_i between i-

th token x_i^l in l-th layer and i-th token x_i^L in L-th layer is calculated as follows:

$$w_i = (x_i^l)^T x_i^L.$$

After obtaining the correlation w_i between the deep and shallow layer corresponding tokens, we normalize it and then multiply it with the feedback bias to obtain the final feedback result $b_i = b_i \cdot w_i$.

We combine the above two feedback modes and apply them to the Transformer network. We set a dynamic feedback adjustment coefficient to control the output of the feedback information. Based on the correlation matrix A ($A[i,j] = a_{ij}$) obtained from feedback mode one, we extract the score $A[cls, cls]$ corresponding to the CLS token as the basis for dynamic adjustment, representing the global similarity between the input and output layers. For all $l \in \{0, 1, 2, ..., L\}$, we calculate the selection score s^l for dynamic feedback in the l-th layer:

$$s^l = \frac{\exp A^l[cls, cls]}{\sum_{l=0}^{L} \exp A^{l'}[cls, cls]}.$$

we use s^l as the dynamic feedback layer selection coefficient for the l-th layer. The larger s^l is, the more feedback of high-level semantic information is required for the l-th layer. Then, we apply s^l to the two feedback modes and adjust the corresponding layer feedback:

$$\widetilde{x}_i^{l+1} = \sum_{j=1}^{N} \frac{\exp\left(h_{ij}^l + s^l A^l[i,j]\right)}{\sum_{k=1}^{N} \exp\left(h_{ik}^l + s^l A^l[i,k]\right)} \cdot (x_j^l + s^l b_j^l),$$

$$h_{ij}^l = (K_i^l)^T Q_j^l,$$

where K_i^l represents the key feature of the self-attention layer, and Q_j^l represents the query feature of the self-attention layer. We use $\widetilde{x}_i^{l+1}$ to denote the feature vector after feedback.

4.2. Dimension-level Decision Stream Calibration

Inspired by *discovery 2*, we accumulated the correlation $(R_{cls}^l)_d$ between the dimension of each CLS token layer and the target class. To reduce the interference of misclassified features, we used mean statistics to accumulate this correlation, calculated as follows:

$$\overline{R}_{cls}^l = \frac{1}{J} \sum_{j=1}^{J} R_{cls}^l.$$

For each class, J samples are used to calculate the average correlation distribution $\overline{R}_{cls}^l$, which can explain the relationship between the target class and the dimension of each CLS token layer. We partitioned different dimensions based on this correlation and filtered out the dimensions that significantly contributed to specific classes. For a specific class, we generated a mask for high-contribution dimensions, which was calculated as follows:

$$mask_{cls}^l = \begin{cases} 1, & \overline{R}_{cls}^l \geqslant v \\ 0, & \overline{R}_{cls}^l < v \end{cases},$$

where v is an adjustable threshold. Based on the mask, we can identify the important dimensions for a specific class. Then, we used a self-distillation method as shown in Figure 3, where the teacher model was used to guide the student model in fine-tuning by removing the interfering features using the mask. The total training loss $\mathcal{L}$ composed of the normal cross-entropy loss $\mathcal{L}_{CE}$ and distillation loss $\mathcal{L}_{MSE}$ is calculated as follows:

$$\mathcal{L} = \mathcal{L}_{CE}(f(Z_{student}^L), Y) + \lambda \mathcal{L}_{MSE}(Z_{student}^l, Z_{teacher}^l),$$

where λ denotes the balance parameter, Y denotes the ground truth label, f denotes the following softmax and MLP function, $Z_{student}^L$ denotes the predicted latent representation of student model at the L-th layer, $Z_{student}^l$ and $Z_{teacher}^l$ denote the predicted latent representations of student model and teacher model at the l-th layer ($l \in \{0, 1, 2, ..., L\}$), respectively. The latent representation $Z_{teacher}^l$ of teacher model is obtained by multiplying the original representation of the teacher model at a specific layer with a category-aware mask, that is, $Z_{teacher}^l = Z_{teacher}^{original} \odot mask_{cls}^l$, where $\odot$ denotes the point-wise multiplication.

5. Experiments

In the experiment, the adopted classifiers, datasets, and experiment settings are listed as follows.

Classifier. The selected 4 classifiers cover mainstream classification network architectures, which are listed as follows: ViT [17], Deit [36], Flexivit [4], Eva [19].

Dataset. The datasets we adopted contain CIFAR-100 [26] and ImageNet [15].

Experiment setting.

For the parameter settings in the token calibration stage, we chose the third layer as the starting output layer for feedback and the final layer as the feedback input layer, with a total of three feedback layers. For the parameter settings in the dimension calibration stage, we used the mean correlation of J images as the threshold selection criterion for the selection threshold v corresponding to specific category dimensions. In the baseline setting, for the ImageNet dataset [15], we used the pre-training weights publicly available in the timm library [40]. In the experimental setup, we fixed the random seed to ensure the stability of the

Backbone	Baseline	+Token	+Dim	+All
ViT-T	87.44	88.01(+0.57)	88.12(+0.68)	88.20(+0.76)
ViT-S	90.24	90.50(+0.26)	90.55(+0.31)	90.59(+0.35)
Deit-T	84.32	85.10(+0.78)	84.81(+0.49)	85.24(+0.92)
Deit-S	87.41	87.54(+0.13)	87.67(+0.26)	87.62(+0.21)
Flexivit-S	89.66	89.74(+0.08)	89.79(+0.13)	89.85(+0.09)
Flexivit-B	91.51	91.38(-0.13)	91.72(+0.21)	91.67(+0.16)
Eva-L	95.88	95.54(-0.34)	95.70(-0.18)	95.29(-0.59)

Table 1: The base and improved accuracy of 7 classifiers on CIFAR-10 dataset.

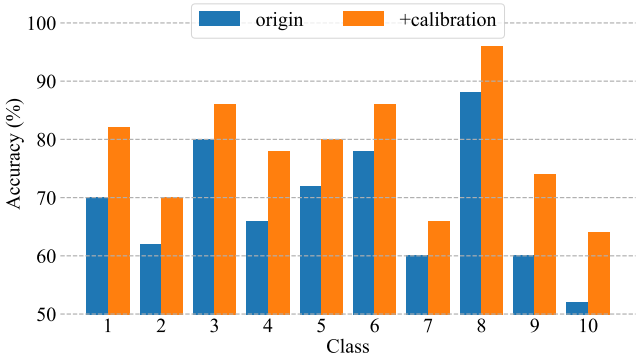

Figure 4: The accuracy and increased accuracy of each category with two stage calibration on the ImageNet dataset.

experiment. More experimental details and source codes are provided in the *supplements*.

5.1. Effectiveness of Vision Transformer Calibrator

This section presents the classification performance of 7 mainstream models on 2 datasets. We provide the combined results of two-stage training in Table 1 and Table 2. It can be seen from the table that Vision Transformer Calibrator improves the accuracy of mainstream visual transformer classifiers by $1\% \sim 2\%$. Figure 4 shows the original and improved classification accuracy of Deit-T [36] on 10 categories in the ImageNet dataset [15]. Among them, the 10 categories are selected by proportional indexing from the 1000 categories in the ImageNet dataset, and the optimized Transformer classification performance is further improved. These validate the effectiveness of the proposed model optimization approach.

5.2. Visual Results

We generated a visualization heatmap by comparing the similarity of corresponding tokens between the deep feedback input layer and the shallow feedback output layer. The higher the similarity, the more features of the corresponding token should be fed back to the shallow layer. Figure 5 shows the visualization results before and after adding token calibration. By calculating the similarity between the output of the last layer and that of a shallow layer, we can examine the image regions the model focuses on. As can be seen from the figure, after calibration, the visually highlighted areas focus more on the image's foreground, demonstrating that the foreground often contributes more to the classification results. Meanwhile, our method assigns greater weight to foreground information during token calibration, further improving model performance.

Backbone	Baseline	+Token	+Dim	+All
ViT-T	75.45	77.03(+1.58)	75.87(+0.42)	77.19(+1.74)
ViT-S	81.40	82.33(+0.93)	81.76(+0.36)	82.43(+1.03)
Deit-T	72.17	73.99(+1.82)	72.44(+0.27)	74.19(+2.02)
Deit-S	79.86	81.01(+1.15)	81.22(+0.21)	81.14(+1.28)
Flexivit-S	82.53	83.16(+0.63)	82.86(+0.33)	83.26(+0.73)
Flexivit-B	84.67	85.43(+0.76)	84.53(-0.17)	85.11(+0.44)
Eva-L	87.94	88.18(+0.24)	84.67(+0.05)	88.13(+0.19)

Table 2: The base and improved accuracy of 7 classifiers on ImageNet dataset.

n	1	2	3	4	5
top-1	73.4	73.5	73.7	73.7	73.6
top-5	91.3	91.6	91.9	91.6	91.5

Table 3: Ablation study on the number n of feedback output layers when the starting feedback output layer is set to 5.

layer	1	3	5	7	9
top-1	72.1	73.0	73.1	72.8	72.4
top-5	91.0	91.3	91.3	91.2	91.1

Table 4: Ablation study on the model output layer index when the feedback output layer is set to one layer.

5.3. Ablation Study

In this section, we conduct two-stage ablation studies on ImageNet using Deit-T [36]. The results of the two-stage calibration experiments can somewhat improve the model performance. The two stages act on the token level and the dimension level, respectively, allowing important tokens and dimensions to be strengthened.

In the token calibration stage, we conduct ablation experiments on the search for the feedback target layer and the total number of dynamic feedback layers. For the feedback target layer search, we set the total number of feedback layers to 1. The data in Table 4 show that the model

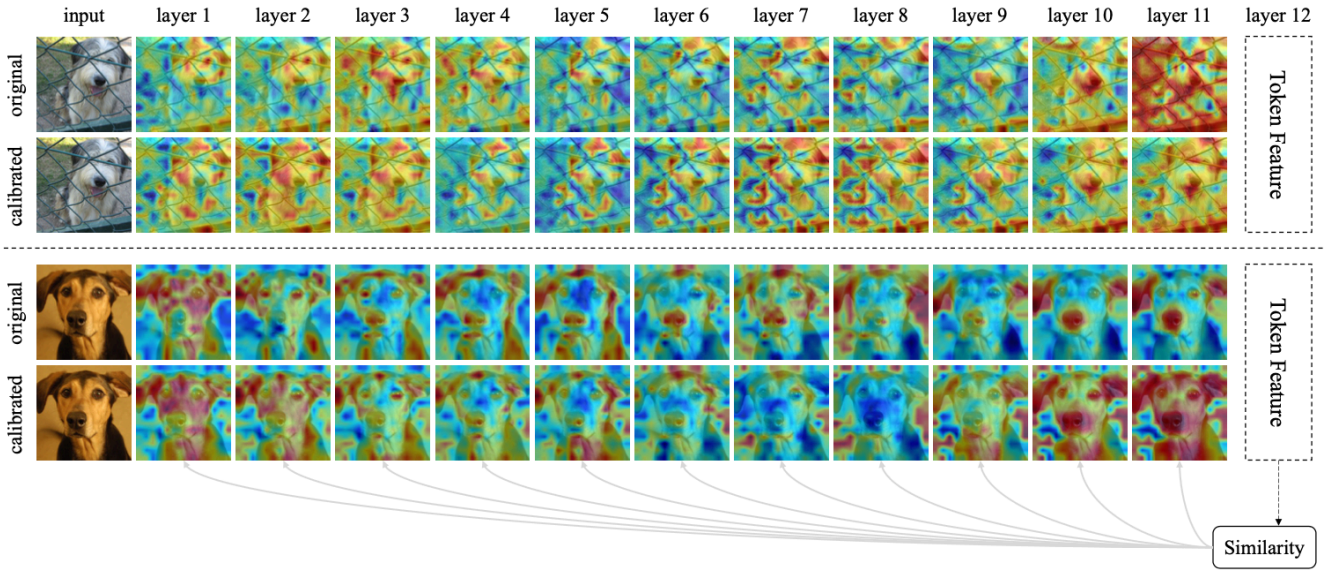

Figure 5: Visualization of the similarity between the last layer features and shallow features of the Deit model before and after calibration for two image examples from ImageNet.

accuracy is lower when the feedback target layer is set to a shallow layer. It may be because the shallow layers of the model extract only very basic visual information, and introducing high-level semantic information may interfere with the model's classification performance. When the feedback target layer is deepened to the middle layer, the model performance improves the most. It may be because the middle layers of the model extract some semantic information directly related to classification. At this time, feedback assistance with high-level semantic information can better handle image features. As for the total number of dynamic feedback layers n, we set the fifth layer as the starting output layer. The data in Table 3 indicates that as the number of layers increases, the model performance first improves and then remains basically unchanged. Too many feedback outputs cannot significantly improve model performance but can instead affect model training speed. Therefore, the model's total number of feedback output layers should be set reasonably.

6. Dicussion and Future Work

Our experiments showed that ViT-Calibrator has limited performance gains for some large-parameter models. For the token calibrate stage, the performance gain of the model decreases as the number of heads increases, which may be because too many heads disrupt the transmission of token feedback information. For the dimension calibration stage, the model's performance is often better on small datasets than on large ones. Because we optimize the model using the sparsity associated with dimensions and categories, distillation can filter out dimensions related to classification results for datasets with too many image categories, causing interference features to be filtered out along with some effective features, resulting in performance degradation.

In addition, we found that when generating important dimensions for specific categories, the classification performance can be significantly improved by 10%-20% if we remove unimportant dimensions for the category based on the important dimensions selected from the training set. However, in practice, the performance of the corrected model is often much lower. It is because there are still redundant features in the model correction process, which can significantly interfere with the model performance.

The Vit-calibrator we proposed has achieved particular effectiveness in classification networks, which also provides a new idea for other tasks. For example, tasks such as segmentation and object detection can be optimized in different tasks based on the inspiration from our work, which will improve the performance of other tasks in the future.

7. Conclusion

In this article, we propose a new paradigm dubbed Decision Stream Calibration that boosts the performance of general Vision Transformers.We shed light on the information propagation mechanism in the learning procedure by exploring the correlation between different tokens and the relevance coefficient of multiple dimensions. Research on Transformers typically starts with attention mechanisms; However, attention mechanisms only reflect one aspect of the Transformer network. We explore and propose two new

research directions based on Transformers: Token-level Decision Stream Calibration through token feedback and Dimension-level Decision Stream Calibration, which increase interpretability while improving model performance. These can serve as references for future research.

References

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

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[3] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. Interpretability via model extraction. *arXiv preprint arXiv:1706.09773*, 2017.

[4] Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. Flexivit: One model for all patch sizes. *arXiv preprint arXiv:2212.08013*, 2022.

[5] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[6] Michael Brooks, Saleema Amershi, Bongshin Lee, Steven M Drucker, Ashish Kapoor, and Patrice Simard. Featureinsight: Visual support for error-driven feature ideation in text classification. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 105–112. IEEE, 2015.

[7] Gabriel Cadamuro, Ran Gilad-Bachrach, and Xiaojin Zhu. Debugging machine learning models. In *ICML Workshop on Reliable Machine Learning in the Wild*, volume 103, 2016.

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[9] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021.

[10] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021.

[11] Jianbo Chen, Le Song, Martin J Wainwright, and Michael I Jordan. L-shapley and c-shapley: Efficient model interpretation for structured data. *arXiv preprint arXiv:1808.02610*, 2018.

[12] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. *Advances in neural information processing systems*, 30, 2017.

[13] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977, 2021.

[14] John N Demos. *Getting started with neurofeedback*. WW Norton & Company, 2005.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[17] 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.

[18] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021.

[19] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022.

[20] Zunlei Feng, Jiacong Hu, Sai Wu, Xiaotian Yu, Jie Song, and Mingli Song. Model doctor: A simple gradient aggregation strategy for diagnosing and treating cnn classifiers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 616–624, 2022.

[21] Zunlei Feng, Zhonghua Wang, Xinchao Wang, Xiuming Zhang, Lechao Cheng, Jie Lei, Yuexuan Wang, and Mingli Song. Edge-competing pathological liver vessel segmentation with limited labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1325–1333, 2021.

[22] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2950–2958, 2019.

[23] 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.

[24] Josua Krause, Aritra Dasgupta, Jordan Swartz, Yindalon Aphinyanaphongs, and Enrico Bertini. A workflow for visual diagnostics of binary classifiers using instance-level explanations. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 162–172. IEEE, 2017.

[25] Josua Krause, Adam Perer, and Kenney Ng. Interacting with predictions: Visual inspection of black-box machine learning

models. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 5686–5697, 2016.

[26] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[27] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 126–137, 2015.

[28] Todd Kulesza, Simone Stumpf, Margaret Burnett, Weng-Keen Wong, Yann Riche, Travis Moore, Ian Oberst, Amber Shinsel, and Kevin McIntosh. Explanatory debugging: Supporting end-user debugging of machine-learned programs. In *2010 IEEE Symposium on Visual Languages and Human-Centric Computing*, pages 41–48. IEEE, 2010.

[29] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[30] Jie Lei, Zhe Wang, Zunlei Feng, Mingli Song, and Jiajun Bu. Understanding the prediction process of deep networks by forests. In *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, pages 1–7. IEEE, 2018.

[31] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[33] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[34] Jose Gustavo S Paiva, William Robson Schwartz, Helio Pedrini, and Rosane Minghim. An approach to supporting incremental visual data classification. *IEEE transactions on visualization and computer graphics*, 21(1):4–17, 2014.

[35] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[36] 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.

[37] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 32–42, 2021.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[39] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021.

[40] Ross Wightman. Pytorch image models (Timm). `https://github.com/rwightman/pytorch-image-models`, 2020.

[41] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 558–567, 2021.

[42] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

[43] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018.

[44] Jiawei Zhang, Yang Wang, Piero Molino, Lezhi Li, and David S Ebert. Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models. *IEEE transactions on visualization and computer graphics*, 25(1):364–373, 2018.

[45] Bolei Zhou, David Bau, Aude Oliva, and Antonio Torralba. Interpreting deep visual representations via network dissection. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2131–2145, 2018.

[46] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[47] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021.

A. Implementation Details

In the experiment, we employ two datasets and seven models, the detailed specifications of which are provided in the Table 5. During training, we uniformly utilize AdamW as the optimizer with a batch size of 128 and adopted a cosine learning rate schedule. The specific training settings for the two datasets are as follows:

CIFAR-100. The CIFAR-100 dataset, a subset of the Tiny Images dataset, comprises 60000 32x32 color images. We fine-tune the pre-trained model for 200 epochs on CIFAR-100 to obtain our baseline. Then we adopt our calibration approach to refine the model. During the token calibration stage, we fine-tune the baseline for an additional 100 epochs. We set the index of the input layer of the feedback module as the last layer, the number of output layers for the feedback module as three. For a model with a total of 12 layers, the index candidate layers for feedback output are $\{3, 4, ..., 9\}$. During the dimension calibration stage, we fine-tune the model for an additional 100 epochs based on the token calibration, and in the category-related dimension filtering step, we select 40% of the dimensions as relevant dimensions by default.

ImageNet-1K. This dataset is a ubiquitous subset of ImageNet, spanning by 1000 object classes and encompassing 1,281,167 images for training, 50,000 for validation and 100,000 for testing. To build up the baseline, we employ a pre-trained model from the *timm* library and perform a two-stage fine-tuning procedure that involves token and dimension calibrations. The training parameters remain consistent with those mentioned above, and we reduce the training duration to 30 epochs per calibration stage to mitigate the computational overhead.

B. More experimental results

B.1. Dimension-level Decision Stream Calibration

For demonstrating the universality of *Discovery 2 that each category is only related to a specific dimension in the CLS token, and irrelevant dimensions interfere with the model's prediction results*, we provide more statistical correlation diagrams from two different perspectives on ImageNet dataset.

Figure 6 illustrates that, for distinct images within a particular category, the dimensions with high correlation remain relatively consistent, while the dimensional correlations vary among different categories. This finding substantiates the efficacy of establishing consistent relevant dimensions for images belonging to a particular category.

Figure 7 illustrates the mean correlation between dimensions and distinct classes in the ImageNet dataset. The result demonstrates that each class has a unique association with specific dimensions, and some dimensions are sparsely related to specific classes. We validate the generality and universality of the *Discovery 2* using two distinct approaches: internal variations among images of the same category and variations among images of different categories.

In our experiment, an important dimension discrimination threshold is required for model dimension calibration. Therefore, we investigate the number of dimensions to be retained for specific categories. We select dimensions associated with specific categories from the training set and evaluate their performance on the corresponding test set. The accuracy is calculated under varying numbers of dimensions, and the classification performance is examined later.

Table 6 reveals how the model's average accuracy for specific categories changes as the percentage of dimensions retained for the category decreases. The trend indicates an initial increase in accuracy followed by a decrease. This implies that as the number of dimensions used to represent a category decreases, irrelevant dimensions that can impede category performance are removed, resulting in performance gains. However, it is necessary to keep the number of dimensions within a reasonable range to avoid losing information that pertains to the original category, leading to poor classification performance. In our experiment, we have determined a threshold of 40% to retain the relevant dimensions while eliminating the irrelevant ones.

Furthermore, we explore the impact of preserving different dimensions on the final classification results. We use the same dimension preservation strategy for each layer's output token in our experiment. Table 7 demonstrates that removing redundant dimensions significantly enhances the performance of the model for deep layers. For shallow layers, the dimension response pattern is not apparent. This is likely due to the fact that deep networks contain more semantic information than shallow networks.

B.2. Token-level Decision Stream Calibration

Based on *Discovery 1 the final decision is associated with tokens of foreground targets, while the token features of the foreground target will be transmitted into the next layer as much as possible, and the useless token features of the background area will be eliminated gradually in the forward propagation*, we employ a fusion of deep and shallow features for Token-level Decision Stream Calibration. Additionally, we assign different weights to the ratio of deep features mapped to shallow features for different tokens based on cross-layer similarity.

Moreover, we discover some peculiar experimental phenomena and provided partial explanations for them. Furthermore, we present more visualization results to further support our findings.

Additional experimental phenomena. During our experiment, we observe that fixed patterns are responsible for certain patches. As shown in Figure 8, we visualize the

Backbone	Number of Layer	Head Numbers	Resolution	Params(Mb)
ViT-T/16 [17]	12	3	224×224	5.72
ViT-S/16 [17]	12	6	224×224	22.05
Deit-T [36]	12	3	224×224	5.72
Deit-S [36]	12	6	224×224	22.05
Flexivit-S [4]	12	6	240×240	22.06
Flexivit-B [4]	12	12	240×240	86.59
Eva-L [19]	24	16	196×196	304.14

Table 5: Detailed parameters of the seven models used for the two-stage calibration method.

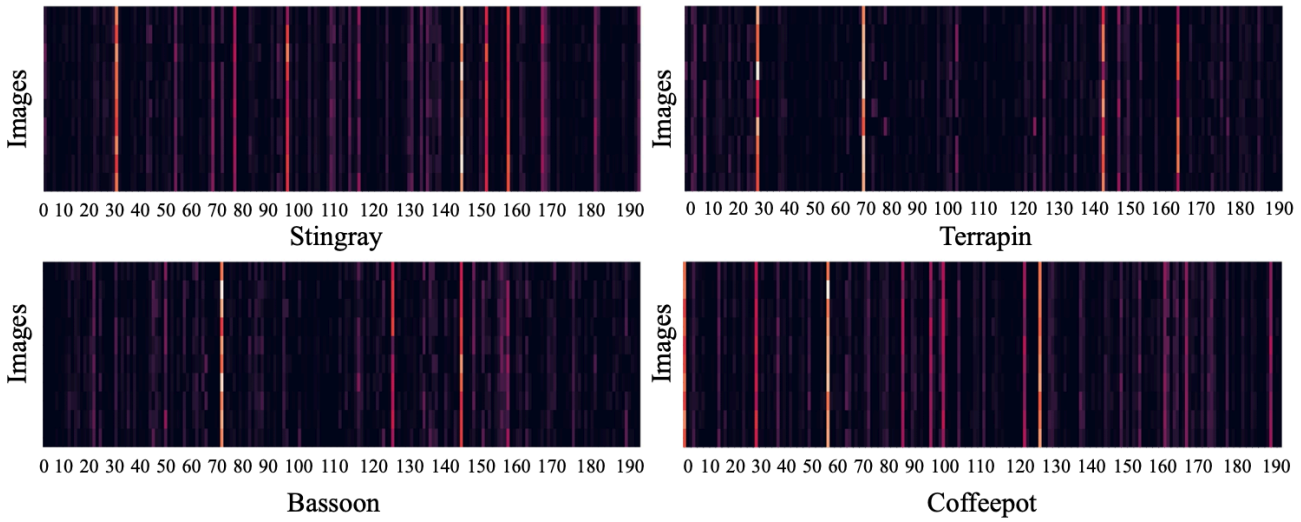

Figure 6: The four images in the figure respectively represent the relationship between randomly selected four categories in ImageNet and specific dimensions of response. In each image, the horizontal axis represents the dimension, the vertical axis represents the total number of images in the current category, and there are 10 images in each category.

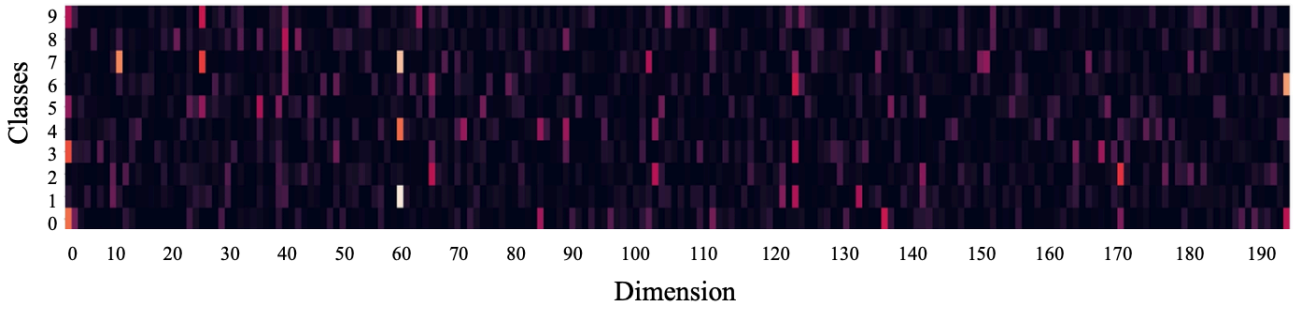

Figure 7: Average statistical correlation graphs for different dimensions on ImageNet's ten categories.

similarity measure between the last layer and the shallow layer without any preprocessing. We notice that starting from the sixth layer, the response patterns of the four corners in the image remain consistent, with their response values typically being the maximum or minimum among all tokens. We postulate that there may be two underlying reasons for this phenomenon. Firstly, deeper layers contain more semantic information, and the four corners contain less category-related semantic information, thereby resulting in dissimilar response patterns from other tokens. Secondly, the positional encoding learned for the four corners may differ significantly from other tokens, thereby leading

Percentage	100%	90%	80%	70%	60%	50%	40%	30%	20%	10%
Accuracy	75.45	81.22	85.19	88.40	91.06	93.14	94.63	95.34	95.61	93.76

Table 6: The impact of retaining dimensions in different proportions on the final classification results.Percentage represents the proportion of dimensions to be retained, and the values of the remaining dimensions are set to 0.

Layer	1	2	3	4	5	6	7	8	9	10	11	12
Accuracy	39.82	34.41	39.06	39.65	46.87	58.64	65.42	81.26	91.02	93.34	95.32	94.63

Table 7: Filtering the impact of preserving dimensions at 40% on the final classification results by selecting different layers with preserved dimensions.

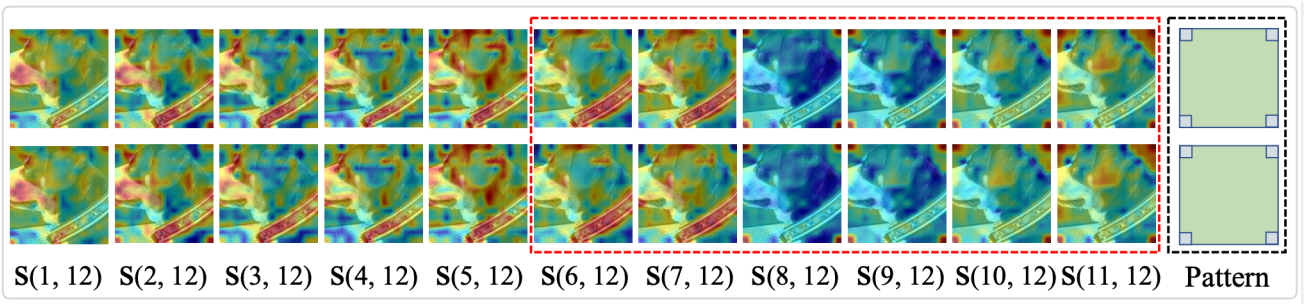

Figure 8: Visualization of last layer features and shallow features similarity, including high response patterns of specific patches.where $S(l, l')$ represents the correlation calculation between the output of layer l and the output of layer l'.

to distinct responses.

More visualization results. We provide more visualization results in the experiment. As shown in Figure 9, the original visualization has many noisy regions, making it difficult to observe the response of each region directly from the original image. Therefore, we first preprocess the interlayer correlations by removing the regions with low responses and retaining more regions with high responses. We can observe that the high response regions in the image are more concentrated in areas that have specific semantic meanings in the image. For example, in the image of the electric vehicle, the high response regions are mainly concentrated on the wheel and seat areas, both of which are closely related to the final classification. At the same time, in the shallow layers of the image, the high response regions are combined with some background areas, while in the deeper layers, they are combined with more semantic information. In addition, the classification of some categories may have a strong correlation with the background, which may be related to the model itself and the data.

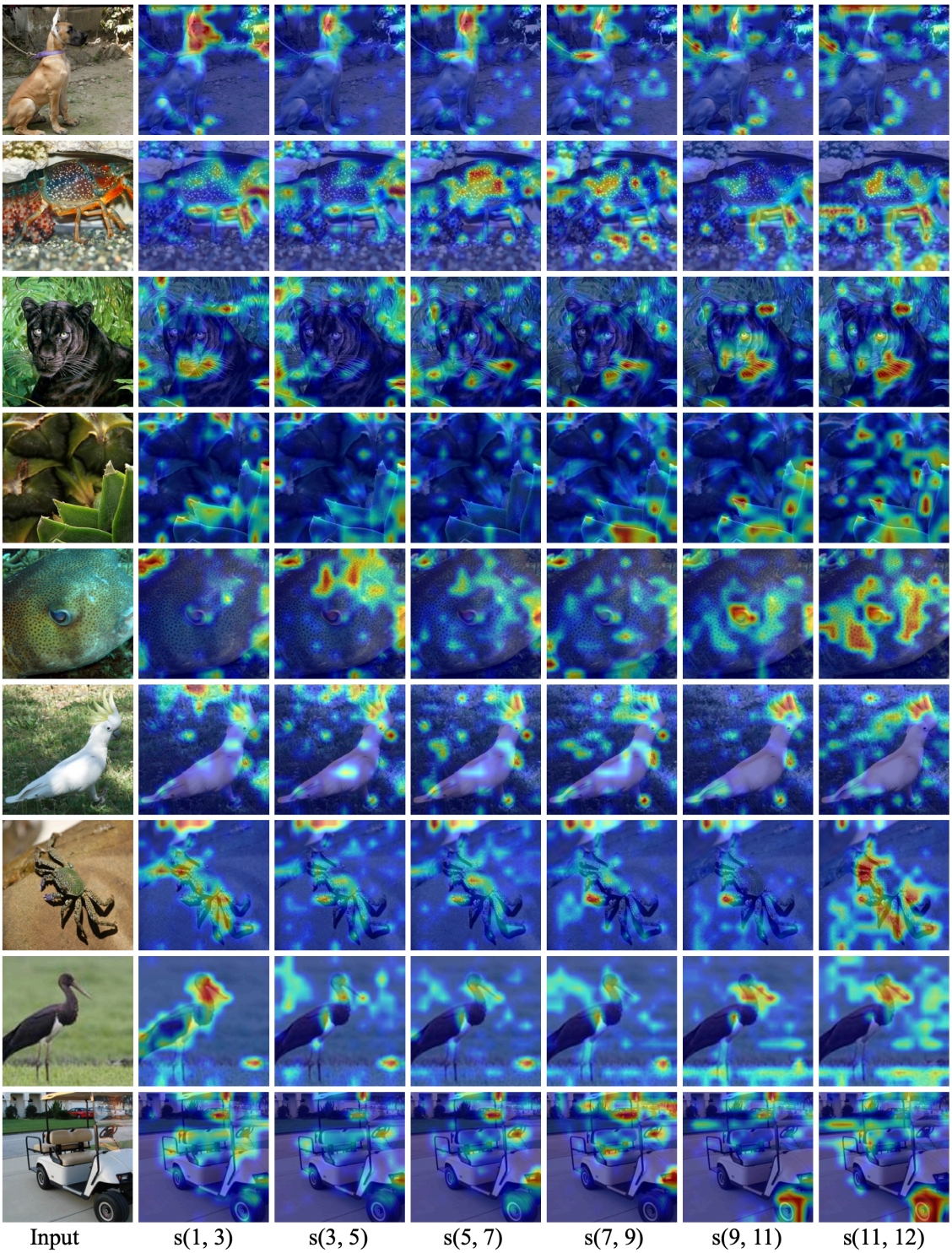

Figure 9: Visualize the similarity between deep and shallow layers while removing partially activated responses in the visualization results.