ViT-Calibrator: Decision Stream Calibration for Vision Transformer

Lin Chen[1], Zhijie Jia[1], Lechao Cheng[2], Yang Gao[1], Jie Lei[3], Yijun Bei[1]*, Zunlei Feng[1]

[1]Zhejiang University
[2]Zhejiang Lab
[3]Zhejiang University of Technology
{lin_chen, jiazhijie, roygao, beiyj, zunleifeng}@zju.edu.cn, 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 class token. 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.

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 (Feng et al. 2021), security monitoring (Litjens et al. 2017), and autonomous driving (Bojarski et al. 2016). 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, Transformers model's self-attention mechanism captures global semantic information (Vaswani et al. 2017), allowing Transformers to be more proficient in handling long sequential data, which opens up new possibilities for visual feature learning (Dosovitskiy et al. 2020; Liu et al. 2021; Touvron et al. 2021a; Yuan et al. 2021). 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 (Kulesza et al. 2015). The latest work (Feng et al. 2022) on optimizing deep models exploit gradient constraint strategies to diagnose and repair convolutional neural network models. In addition, some works (Bastani, Kim, and Bastani 2017; Lei et al. 2018) apply interpretable decision tree methods to approximate deep learning models for analysis and detection. Furthermore, other works, such as, visualization techniques (Krause, Perer, and Ng 2016) have also explored to explain and diagnose (Krause et al. 2017; Zhang et al. 2018b) 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 attention-based Transformers.

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 empirical experiments, one of which is derived from the biological neural feedback principle (Demos 2005), 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, irrelevant dimensions can harm classification performance. Consequently, we devise a two-stage information propagation calibration 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 Calibration 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

*Corresponding author.

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 feeds back advantageous deep features to the shallow network based on similarity. To provide feedback on deep semantics, fusion tuning is performed for both attention and token features in the feedback mechanism.

In the second stage, 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 serves 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 therefore the first decision stream calibration technique for the vision transformer, termed as ViT-Calibrator. Two new perspectives are provided for optimizing the Transformer model: token feedback and dimension constraints. Extensive experiments show that the proposed ViT-Calibrator can effectively calibrate the erroneous 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.

Related Work

Transformer-based Classification

Inspired by the tremendous success of Transformer in natural language processing (Devlin et al. 2018; Radford et al. 2018), it has also been widely used in computer vision. To enhance the model's receptive field and global dependency, ViT (Dosovitskiy et al. 2020) 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. ConViT (d'Ascoli et al. 2021) combined CNN and Transformer to improve computational efficiency and classification performance through an adaptive feature importance weighting mechanism. Dai et al. (2021) designed a lightweight visual Transformer network CoAtNet based on multi-resolution input and grouped convolution for effective feature extraction and 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 (Liu et al. 2021). 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. Yuan et al. (2021) 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, Wang et al. (2021) used non-overlapping patch partitioning to reduce feature size in model PVT. Touvron et al. (2021b) proposed a cross-scale attention mechanism, CaiT, that simultaneously considers global and local information to improve performance in image classification tasks. Additionally, some other research attempts to design various self-supervised learning schemes (Bao et al. 2021; Caron et al. 2021) for ViT in a generative and discriminative manner.

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, Gilad-Bachrach, and Zhu (2016) 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 (Brooks et al. 2015; Krause, Perer, and Ng 2016; Kulesza et al. 2010) 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, Kim, and Bastani (2017) proposed a method of using symbolic regression to explain deep learning models. Using interpretable random forests (Lei et al. 2018) 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 (Zhang et al. 2018b) 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. (2022) proposed a gradient-constrained convolutional neural network model optimization method that uses gradient constraints to optimize convolutional neural networks. Unlike the above methods, we focus on automatically processing Transformer models based on diagnostic results.

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 (Fong, Patrick, and Vedaldi 2019; Fong and Vedaldi 2017), backpropagation-based (Zhang et al. 2018a), saliency map-based (Dabkowski and Gal 2017; Zeiler and Fergus 2014; Zhou et al. 2018, 2016), and Shapley value-based

methods (Chen et al. 2018; Lundberg and Lee 2017). Among these, LRP (Bach et al. 2015) 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 and Zuidema (2020) 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, Gur, and Wolf (2021) proposed a 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 methods can be only used to analyze some failure cases or understand the decision-making mechanism. These methods can't be used to diagnose and treat the deep model automatically.

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 will be eliminated gradually in the forward propagation.*

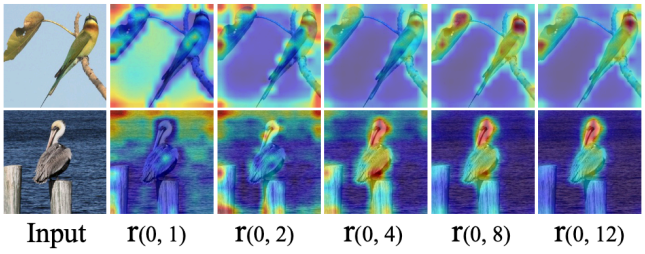

Figure 1: Visualization results of the information propagation relationship, where $r(l, l')$ represents the attention propagation results from layer l to layer l'.

In this section, we analyze 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.

To verify the information flow mechanism of the attention, we calculate the attention rollout (Abnar and Zuidema 2020) from the input layer to the different output layers. Assuming we want to compute the information flow from the i-th layer to the j-th layer in Transformers, we need to recursively multiply all attention weight matrices of the layers, and the calculation is given as follows:

$$\widetilde{A}_{ij} = \begin{cases} A_i \widetilde{A}_{i-1}, & i>j \\ A_i, & i=j \end{cases},$$

where A represents the original attention matrix, and the multiplication operation denotes matrix multiplication.

Figure 1 shows the information flow relationship from the input to each layer's output in DeiT (Touvron et al. 2021a) on a correctly classified image from the ImageNet-1K dataset. We can observe that it gradually shifts focus from the background to the foreground 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. 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 class token, and irrelevant dimensions interfere with the model's prediction results.*

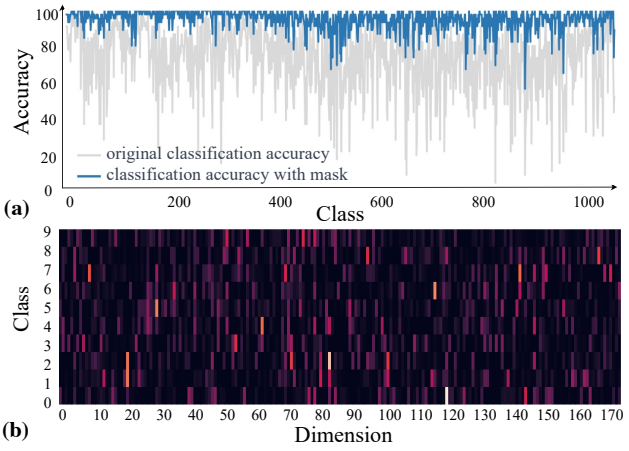

Figure 2: (a) The original classification accuracy and the accuracy after removing dimensions unrelated to the category of DeiT on ImageNet-1K. (b) Statistical correlations of different dimensions with the last layer's class token for ten classes on ImageNet-1K.

In this paragraph, we adopt the relevance propagation technique LRP (Bach et al. 2015) to attribute the vision Transformer classification results. We assume that on the j-th neuron of layer l in the network, the LRP rule is employed to decompose the relevance coefficients of the previous layer $R_j^{(l+1)}$, in order to obtain the relevance score $R_i^{(l)}$ for the neuron i within the current layer l. This decomposition is based on the localized preactivation z_{ij} and the aggregation z_j at the layer's current output. In essence, this rule can be expressed as:

$$R_i^{(l)} = \sum_j \frac{z_{ij}}{z_j} R_j^{(l+1)},$$

where localized preactivation z describes the values propagated through the model during forward prediction, for instance, $z_{ij} = x_i \cdot w_{ij}$, and z denotes the aggregated preactivation within a neural network layer with learned weight parameters w_{ij}. In the propagation mechanism, the total sum of relevance coefficients across layers must be equal. Therefore, the following conditions are satisfied:

$$\sum_i R_i^{(l)} = f(x),$$

where $f(x)$ represents the final prediction result. Therefore, utilizing the same attribution technique, the relevance coefficient $(R_n^l)_d$ for the d-th dimension of the n-th token's output in layer l satisfies the following conditions:

$$\sum_n \sum_d (R_n^l)_d = f(x).$$

Therefore, $(R_{cls}^l)_d$ can represent the correlation between the target category and a specific dimension of the class token in a particular layer.

Figure 2(b) shows the statistical correlation between all dimensions of the last layer's class token of DeiT (Touvron et al. 2021a) and the categories on 10 classes of the ImageNet-1K dataset. For each category, we computed the sum $\sum_1^{100} R_{cls}^l$ of the correlation coefficients of the class token 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 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 with 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.

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 calibration, 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.

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 interact to derive the feedback offset for 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 is nonlinearly projected onto the semantic space using a Multi-Layer Perception (MLP) module. The L-th layer vector is used to obtain the i-th bias b_i for the projected feature x_i^L:

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

In order to integrate more effective features, we assign distinct attention weights to tokens, utilizing inter-layer similarity to construct weight matrices that correspond to individual 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 class 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

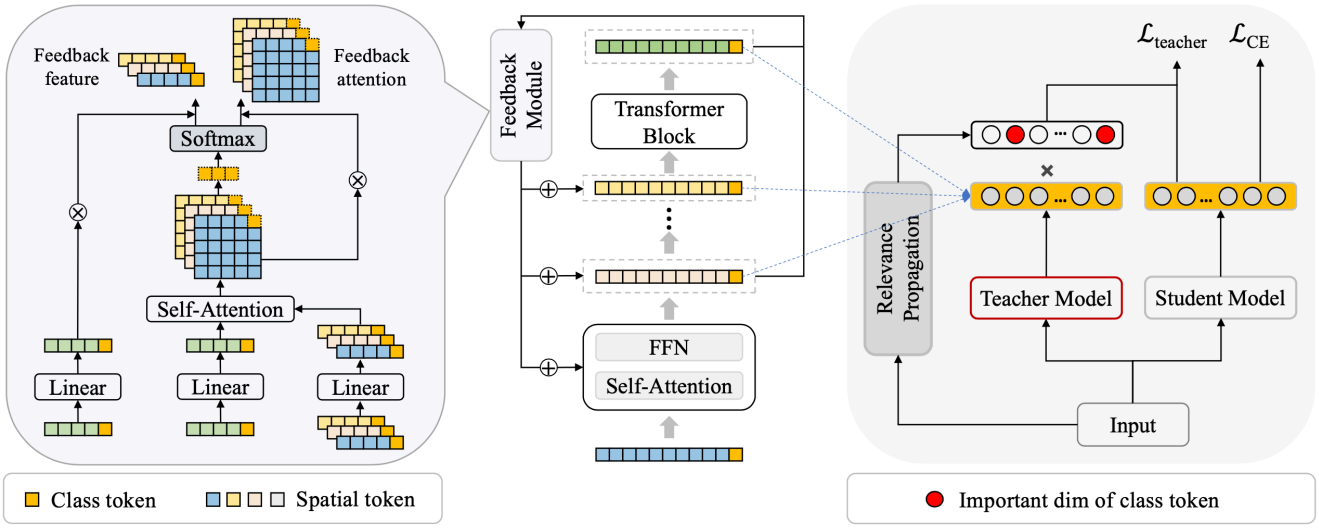

Figure 3: The framework of Vision Transformer Calibrator is composed of token calibration and dimension calibration, which can calibrate the incorrect feature stream from token-level and dimension-level of the class token, respectively.

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.

Dimension-level Decision Stream Calibration

Inspired by *discovery 2*, we accumulated the correlation $(R_{cls}^l)_d$ between the dimension of each layer's class token 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 layer's class token. 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. $\mathcal{L}_{MSE}$ denotes the mean square error, $Z_{student}^l$ 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.

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 (Dosovitskiy et al. 2020), DeiT (Touvron et al. 2021a), Flexivit (Beyer et al. 2022), Eva (Fang et al. 2022).

Dataset	Method	ViT-T	ViT-S	DeiT-T	DeiT-S	Flexivit-S	Flexivit-B	Eva-L
CIFAR-100	Ori.	87.44	90.24	84.32	87.41	89.66	91.51	95.88
	+Token	88.01(**+0.57**)	90.50(**+0.26**)	85.10(**+0.78**)	87.54(**+0.13**)	89.74(**+0.08**)	91.61(**+0.10**)	95.93(**+0.05**)
	+Dim	88.12(**+0.68**)	90.55(**+0.31**)	84.81(**+0.49**)	87.67(**+0.26**)	89.79(**+0.13**)	91.72(**+0.21**)	95.99(**+0.11**)
	+All	88.20(**+0.76**)	90.59(**+0.35**)	85.24(**+0.92**)	87.62(**+0.21**)	89.85(**+0.19**)	91.67(**+0.16**)	95.95(**+0.07**)
ImageNet-1K	Ori.	75.45	81.40	72.17	79.86	82.53	84.67	87.94
	+Token	77.03(**+1.58**)	82.33(**+0.93**)	73.84(**+1.67**)	81.01(**+1.15**)	83.16(**+0.63**)	85.43(**+0.76**)	88.18(**+0.24**)
	+Dim	75.87(**+0.42**)	81.76(**+0.36**)	72.54(**+0.37**)	80.07(**+0.21**)	82.86(**+0.33**)	84.81(**+0.14**)	87.99(**+0.05**)
	+All	77.19(**+1.74**)	82.43(**+1.03**)	74.19(**+2.02**)	81.14(**+1.28**)	83.26(**+0.73**)	85.11(**+0.44**)	88.13(**+0.19**)
ImageNet-Real	Ori.	82.07	84.54	82.00	85.68	87.82	88.65	90.53
	+Token	83.08(**+1.01**)	85.24(**+0.70**)	83.45(**+1.45**)	86.81(**+1.13**)	88.64(**+0.82**)	89.13(**+0.48**)	90.69(**+0.16**)
	+Dim	82.44(**+0.37**)	84.95(**+0.41**)	82.44(**+0.44**)	85.97(**+0.29**)	88.13(**+0.31**)	89.00(**+0.35**)	90.65(**+0.12**)
	+All	83.22(**+1.15**)	85.47(**+0.93**)	83.62(**+1.62**)	87.03(**+1.35**)	88.70(**+0.88**)	89.15(**+0.50**)	90.66(**+0.13**)

Table 1: The base and improved accuracy of 7 classifiers on 3 mainstream classification datasets. Here, 'Ori.' represents the use of standard model architecture. All scores are in %.

Layer Number n	1	2	3	4	5
Top-1 Accuracy (%)	73.4	73.5	**73.8**	73.7	73.6
Top-5 Accuracy (%)	91.3	91.6	**91.9**	91.6	91.5

Table 2: Ablation study on the number n of feedback output layers when the starting feedback output layer is set to 3.

Layer Index	1	3	5	7	9
Top-1 Accuracy (%)	72.1	**73.1**	73.0	72.8	72.4
Top-5 Accuracy (%)	91.0	**91.3**	91.2	91.2	91.1

Table 3: Ablation study on the model output layer index when the feedback output layer is set to one layer.

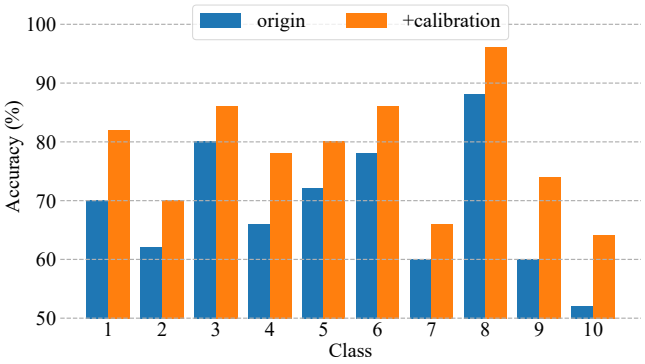

Figure 4: The accuracy and increased accuracy of each category with two stage calibration on the ImageNet-1K dataset.

Dataset. The datasets contain CIFAR-100 (Krizhevsky, Hinton et al. 2009), ImageNet-1K (Deng et al. 2009) and ImageNet-Real (Hendrycks et al. 2021).

Experiment setting. For 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. When training the feedback module, we keep the backbone network fixed. For parameter settings in the dimension calibration stage, we use the mean correlation of J images as the threshold selection criterion for the selection threshold v corresponding to specific category dimensions. In the category-related dimension filtering step, according to different models, we select the dimension within the range of 40% to 60% as relevant dimensions by default. In the baseline setting, for ImageNet, we use the pre-training weights publicly available in the timm library (Wightman 2020). Additionally, we fix the random seed to ensure the stability of the experiment.

Effectiveness of Vision Transformer Calibrator

This section presents the classification performance of 7 mainstream models on 3 datasets. We provide the combined results of two-stage training in Table 1. 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 on 10 categories. Among them, the 10 categories are selected by proportional indexing from the 1000 categories on ImageNet-1K, and the optimized Transformer classification performance is further improved. These validate the effectiveness of the proposed model optimization approach.

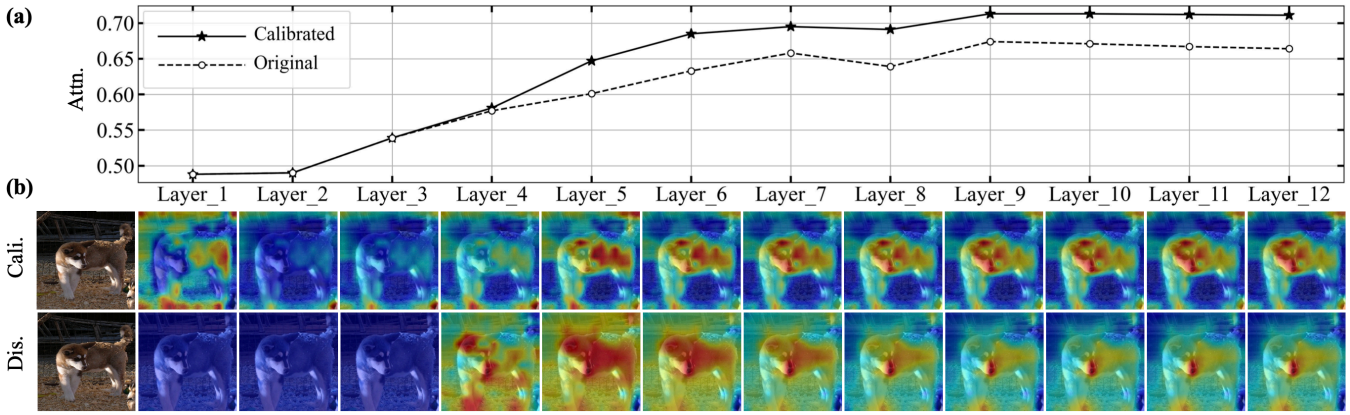

Figure 5: Visualization of DeiT model's attention to image foreground before and after calibration on an ImageNet-1K image across different layers. (a) The proportion of attention to the image foreground at different layers as the input image is processed. (b) "Cali." indicates the attention distribution across different layers after calibration, while "Dis." represents the difference in attention distribution before and after calibration.

Visual Results

We compute the propagation relationships from the input layer to each output layer using attention rollout and generate visual heatmaps. Deeper colors in the heatmap indicate a greater contribution of the corresponding region to the final classification result. Figure 5(a) illustrates proportion of attention to the image foreground at different layers before and after calibration, revealing a significant increase in attention towards the image foreground after calibration. In Figure 5(b), we visualize the difference between after calibration and before. Due to setting our feedback module's starting layer as the fourth layer, the model's attention region remains unchanged before and after calibration in the first three layers. After the fourth layer, more information is directed towards the image foreground after calibration.

Ablation Study

In this section, we conduct ablation studies on ImageNet-1K using DeiT-T. In the token calibration stage, we conduct ablation experiments to explore the feedback target layer and the total number of dynamic feedback layers. For the feedback target layer search, Table 3 indicates that setting the feedback target layer to the starting layer leads to lower model accuracy. This could be due to the starting layer primarily capturing basic visual information, and introducing high-level semantics could interfere with classification. As the feedback target layer deepens, the model's performance improves, potentially because middle layers extract relevant semantic information. High-level semantic feedback at this point can better handle image features. Regarding the total dynamic feedback layers n, we set the third layer as the initial output. Data in Table 2 suggests that model performance first improves and then stabilizes as layer count increases. Excessive feedback doesn't significantly enhance performance but affects training speed. Thus, in order to balance the training speed and the overall performance of the model, the feedback output layers should be reasonably set.

Discussion and Future Work

In summary, our experiments show that ViT-Calibrator has limited performance improvements for large-parameter models. As the number of heads increases, the benefit of token calibration diminishes. This could be because the excessive attention heads disrupt the transmission of token feedback. Moreover, the feedback module design introduces some extra overhead during training and inference, which is a challenge that requires further attention in the future.

In the dimension calibration stage, we observed that by generating important dimensions for specific categories and removing unimportant dimensions based on selected training set dimensions, we achieved a significant 10%-20% improvement in classification performance. However, in practical applications, the calibrated model often performs lower. Future work will focus on learning more precise category-aware dimensions to greatly enhance model performance.

The ViT-Calibrator we introduced has shown significant effectiveness in classification networks, providing novel insights for other tasks like segmentation and object detection, leading to improved performance across diverse domains.

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.

Acknowledgments

This work is funded by National Natural Science Foundation of China (U20B2066, 62106235, 62202015), Zhejiang Province High-Level Talents Special Support Program "Leading Talent of Technological Innovation of Ten-Thousands Talents Program" (No. 2022R52046), and Ningbo Natural Science Foundation (2022J182).

References

Abnar, S.; and Zuidema, W. 2020. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7): e0130140.

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Bastani, O.; Kim, C.; and Bastani, H. 2017. Interpretability via model extraction. *arXiv preprint arXiv:1706.09773*.

Beyer, L.; Izmailov, P.; Kolesnikov, A.; Caron, M.; Kornblith, S.; Zhai, X.; Minderer, M.; Tschannen, M.; Alabdulmohsin, I.; and Pavetic, F. 2022. FlexiViT: One Model for All Patch Sizes. *arXiv preprint arXiv:2212.08013*.

Bojarski, M.; Del Testa, D.; Dworakowski, D.; Firner, B.; Flepp, B.; Goyal, P.; Jackel, L. D.; Monfort, M.; Muller, U.; Zhang, J.; et al. 2016. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*.

Brooks, M.; Amershi, S.; Lee, B.; Drucker, S. M.; Kapoor, A.; and Simard, P. 2015. FeatureInsight: Visual support for error-driven feature ideation in text classification. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 105–112. IEEE.

Cadamuro, G.; Gilad-Bachrach, R.; and Zhu, X. 2016. Debugging machine learning models. In *ICML Workshop on Reliable Machine Learning in the Wild*, volume 103.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chefer, H.; Gur, S.; and Wolf, L. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 782–791.

Chen, J.; Song, L.; Wainwright, M. J.; and Jordan, M. I. 2018. L-shapley and c-shapley: Efficient model interpretation for structured data. *arXiv preprint arXiv:1808.02610*.

Dabkowski, P.; and Gal, Y. 2017. Real time image saliency for black box classifiers. *Advances in neural information processing systems*, 30.

Dai, Z.; Liu, H.; Le, Q. V.; and Tan, M. 2021. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34: 3965–3977.

Demos, J. N. 2005. *Getting started with neurofeedback*. WW Norton & Company.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

d'Ascoli, S.; Touvron, H.; Leavitt, M. L.; Morcos, A. S.; Biroli, G.; and Sagun, L. 2021. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, 2286–2296. PMLR.

Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2022. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*.

Feng, Z.; Hu, J.; Wu, S.; Yu, X.; Song, J.; and Song, M. 2022. Model doctor: A simple gradient aggregation strategy for diagnosing and treating cnn classifiers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 616–624.

Feng, Z.; Wang, Z.; Wang, X.; Zhang, X.; Cheng, L.; Lei, J.; Wang, Y.; and Song, M. 2021. Edge-competing pathological liver vessel segmentation with limited labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1325–1333.

Fong, R.; Patrick, M.; and Vedaldi, A. 2019. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2950–2958.

Fong, R. C.; and Vedaldi, A. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, 3429–3437.

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; et al. 2021. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8340–8349.

Krause, J.; Dasgupta, A.; Swartz, J.; Aphinyanaphongs, Y.; and Bertini, E. 2017. A workflow for visual diagnostics of binary classifiers using instance-level explanations. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 162–172. IEEE.

Krause, J.; Perer, A.; and Ng, K. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, 5686–5697.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Kulesza, T.; Burnett, M.; Wong, W.-K.; and Stumpf, S. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, 126–137.

Kulesza, T.; Stumpf, S.; Burnett, M.; Wong, W.-K.; Riche, Y.; Moore, T.; Oberst, I.; Shinsel, A.; and McIntosh, K. 2010. Explanatory debugging: Supporting end-user debugging of machine-learned programs. In *2010 IEEE Symposium on Visual Languages and Human-Centric Computing*, 41–48. IEEE.

Lei, J.; Wang, Z.; Feng, Z.; Song, M.; and Bu, J. 2018. Understanding the prediction process of deep networks by forests. In *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, 1–7. IEEE.

Litjens, G.; Kooi, T.; Bejnordi, B. E.; Setio, A. A. A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J. A.; Van Ginneken, B.; and Sánchez, C. I. 2017. A survey on deep learning in medical image analysis. *Medical image analysis*, 42: 60–88.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021a. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Touvron, H.; Cord, M.; Sablayrolles, A.; Synnaeve, G.; and Jégou, H. 2021b. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 32–42.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, 568–578.

Wightman, R. 2020. PyTorch Image Models (Timm). https://github.com/rwightman/pytorch-image-models.

Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.-H.; Tay, F. E.; Feng, J.; and Yan, S. 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, 558–567.

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, 818–833. Springer.

Zhang, J.; Bargal, S. A.; Lin, Z.; Brandt, J.; Shen, X.; and Sclaroff, S. 2018a. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10): 1084–1102.

Zhang, J.; Wang, Y.; Molino, P.; Li, L.; and Ebert, D. S. 2018b. Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models. *IEEE transactions on visualization and computer graphics*, 25(1): 364–373.

Zhou, B.; Bau, D.; Oliva, A.; and Torralba, A. 2018. Interpreting deep visual representations via network dissection. *IEEE transactions on pattern analysis and machine intelligence*, 41(9): 2131–2145.

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*, 2921–2929.