Transformer Doctor: Diagnosing and Treating Vision Transformers

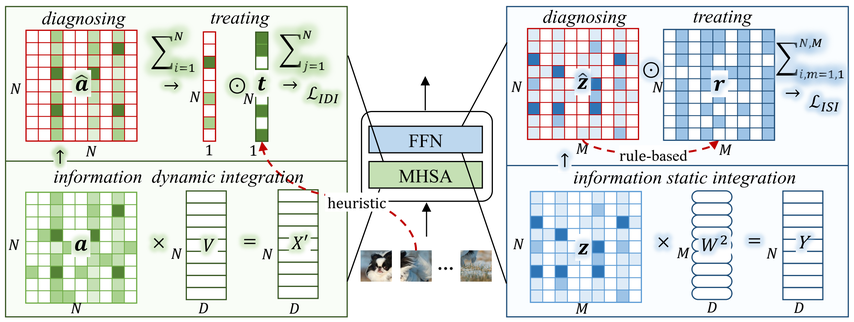
Jiacong Hu, Hao Chen, Kejia Chen, Yang Gao, Jingwen Ye, Xingen Wang, Mingli Song, Zunlei Feng*

ZHEJIANG UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

1. *Introduction*

Inspired by research on information integration mechanisms and conjunctive errors in the biological visual system, this paper conducts an in-depth exploration of the internal error mechanisms of Transformers. We first propose an information integration hypothesis for Transformers in the machine vision domain and provide substantial experimental evidence to support this hypothesis. This includes the dynamic integration of information among tokens and the static integration of information within tokens in Transformers, as well as the presence of conjunctive errors therein. Addressing these errors, we further propose heuristic dynamic integration constraint methods and rule-based static integration constraint methods to rectify errors and ultimately improve model performance. The entire methodology framework is termed as Transformer Doctor, designed for diagnosing and treating internal errors within transformers. Through a plethora of quantitative and qualitative experiments, it has been demonstrated that Transformer Doctor can effectively address internal errors in transformers, thereby enhancing model performance.



2. *Information Integration Hypothesis*

Information Integration Hypothesis: Similar to biological vision, in machine vision, the Transformer continually processes and refines various mixed information in the primary stage, and integrates it in the advanced stage. When erroneous information is integrated, i.e., conjunction errors occur, it leads to incorrect predictions.

Potential Information Integration in MHSA: From the query-key-value mechanism of MHSA below, it can be observed that a certain token X'_i in X' is an integration weighted sum of all tokens in V, where $\mathbf{a}$ can be considered as integration weights within MHSA:

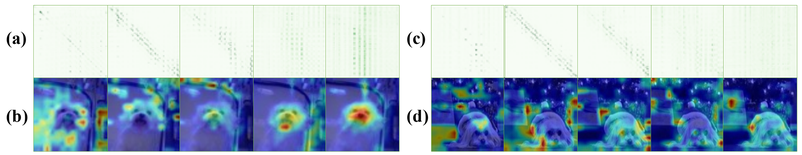
$$X' = softmax\left(\frac{1}{\sqrt{D}}QK^T\right)V \quad \xrightarrow{\text{decompose}} \quad X'_i = \sum_{j=1}^{N} \mathbf{a_{i,j}} V_j,\ \mathbf{a} = sofmax\left(\frac{1}{\sqrt{D}}QK^T\right)$$

Potential Information Integration in FFN: FFN also employs the query-key-value mechanism similar to MHSA. As can be seen from the following equation, a certain token Y_i in Y is an integration weighted sum of all dimensions in $W^{(2)}$, where $\mathbf{z}$ can be referred to as integration weights within the FFN:

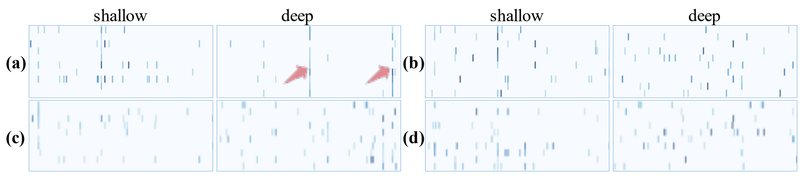
$$Y = gelu\left(X'W^{(1)}\right)W^{(2)} \quad \xrightarrow{\text{decompose}} \quad Y_i = \sum_{m=1}^{M} \mathbf{z_{i,m}} W_m^{(2)},\ \mathbf{z} = gelu\left(X'W^{(1)}\right)$$

3. *Diagnosing*

Inter-token Information Dynamic Integration: The figure below presents a visual analysis of the integration weight **a**.In the initial stages of the Transformer, MHSA primarily mixes and processes adjacent patch information among tokens. However, in the advanced stages of the Transformer, MHSA dynamically and selectively integrates specific patch information among tokens. When integrating incorrect information among tokens, termed conjunction errors, it leads to model mispredictions.



Intra-token Information Static Integration: The figure below illustrates a visual analysis of the integration weight **z**. In the initial stages of the Transformer, the FFN primarily mixes and processes various low-level information within tokens. However, in the advanced stages of the Transformer, the FFN statically and selectively integrates specific category information within tokens. When integrating incorrect information within tokens, termed conjunction errors, it leads to model mispredictions.



4. *Treating*

Heuristic Information Dynamic Integration Therapy: The integrated weight $\hat{\mathbf{a}}$ in a multi-head scenario is obtained by computing the gradient of the predicted probability p_k for the true class k. Then, the integration weight $\hat{\mathbf{a}}$ is constrained by optimizing the loss function $\mathcal{L}_{IDI}$ based on the foreground annotation t.

$$\hat{\mathbf{a}} = \frac{1}{H}\sum_{h=1}^{H}\left(max\left(\frac{\partial p_k}{\partial \mathbf{a}^h}, 0\right) \odot \mathbf{a}^h\right) \qquad \mathcal{L}_{IDI} = \sum_{j=1}^{N}\left(\left(\sum_{i=1}^{N}\hat{\mathbf{a}}_{i,j}\right) \odot \left(1 - t_j\right)\right)$$

Rule-based Information Static Integration Therapy: The integration weight $\mathbf{z}$ is updated to $\hat{\mathbf{z}}$ by establishing a connection with the true class k through gradient-based computation. Subsequently, a binary integration rule r is derived using S high-confidence samples. Finally, the loss function $\mathcal{L}_{ISI}$ is computed based on the integration rule r to constrain the integration weight $\hat{\mathbf{z}}$.

$$\hat{\mathbf{z}} = max\left(\frac{\partial p_k}{\partial \mathbf{z}}, 0\right) \odot \mathbf{z} \qquad r = \mathbb{I}(\bar{\mathbf{z}} \geq \tau), \bar{\mathbf{z}} = \frac{1}{\mathrm{S}}\sum_{\mathrm{s}=1}^{\mathrm{S}}\hat{\mathbf{z}}_{(\mathrm{s})}$$

$$\mathcal{L}_{ISI} = \sum_{m=1}^{M}\sum_{n=1}^{N}\left(\hat{\mathbf{z}}_{n,m} \odot \left(1 - r_{n,m}\right)\right)$$

Joint Therapy of Dynamic and Static Integration: Dynamic integration therapy and static integration therapy can be used in combination. In practical applications, it is recommended to sequentially apply static integration constraint therapy followed by dynamic integration constraint therapy to achieve better results.

5. *Experiments*

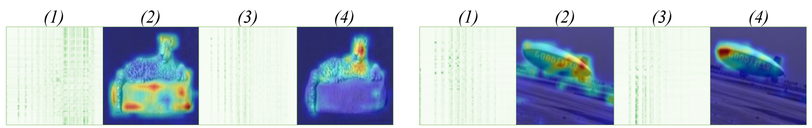
The Performance of Transformer Doctor on SOTA Models: '+Doctor' indicates the performance of model treated with Transformer Doctor. From the table below, it is evident that Transformer Doctor effectively enhances model performance on both small and large-scale datasets.

	CIFAR-10	CIFAR-100	ImageNet-10	ImageNet-50	ImageNet-1K
ViT-Tiny	82.17	56.02	78.8	59.62	64.77
+Doctor	83.00 (+0.87)	58.08 (+2.06)	80.80 (+2.00)	61.02 (+1.40)	68.86 (+4.09)
DeiT-Tiny	82.71	56.97	80.20	61.47	66.83
+Doctor	83.96 (+1.25)	59.49 (+2.52)	81.20 (+1.00)	63.19 (+1.69)	70.75 (+3.92)
CaiT-XXS	82.64	56.10	75.80	58.32	66.28
+Doctor	84.20 (+1.36)	60.00 (+3.90)	77.80 (+2.00)	60.16 (+1.84)	70.25 (+3.97)
TNT-Small	83.31	54.67	81.60	65.45	67.25
+Doctor	84.33 (+1.02)	55.60 (+0.93)	83.00 (+1.40)	67.64 (+2.08)	69.97 (+2.72)
PVT-Tiny	83.27	51.94	82.00	71.81	67.73
+Doctor	84.82 (+1.55)	55.10 (+3.16)	84.20 (+2.20)	74.53 (+2.72)	70.94 (+3.21)
Eva-Tiny	87.56	64.15	83.80	71.23	72.51
+Doctor	88.28 (+0.72)	64.99 (+0.84)	85.80 (+2.00)	72.95 (+1.72)	75.45 (+2.94)
BeiT-Tiny	74.69	49.58	79.80	71.59	70.46
+Doctor	76.20 (+1.51)	51.03 (+1.45)	82.20 (+2.40)	73.57 (+1.98)	71.98 (+3.12)

The Performance of Transformer Doctor with Various Computational Forms: 'mean' denotes directly averaging integration weights across all heads, 'min' and 'max' respectively represent taking the minimum and maximum integration weights within each head. Each row corresponds to ViT-Tiny and PVT-Tiny architectures, with ImageNet-10 dataset used for evaluation.

Base	***+IDI***				***+ISI***	
	$\min(\mathbf{a}^h)$	$\max(\mathbf{a}^h)$	$\mathrm{mean}(\mathbf{a}^h)$	$\hat{\mathbf{a}}$	$\mathbf{z}$	$\hat{\mathbf{z}}$
78.80	79.20 (+0.40)	78.60 (-0.20)	79.20 (+0.40)	**80.20 (+1.40)**	79.00 (+0.20)	**80.40 (+1.60)**
82.00	81.20 (-0.80)	81.20 (-0.80)	80.00 (-2.00)	**83.60 (+1.36)**	82.40 (+0.40)	**83.40 (+1.40)**

Comparison of the Intra-token Integration Weights **a**: (1) and (3) depict integration weights before and after treatment, respectively. (2) and (4) show the corresponding heat map effects of integration weights overlaid onto the original image before and after treatment.



Comparison of the Inter-token Integration Weights **z**: (1) and (2) represent the intra-token integration rules for correct predictions. (3) and (4) depict the intra-token integration weights before and after treatment, respectively.