# A Simple Method for Attention Visualization

Xiaokai Zhang[⋆]

School of Computer Engineering and Science, Shanghai University
**Project:** https://github.com/BitSecret/SimpleAttnViz

**Abstract.** This paper formally defines the concepts of intra-token information mixing and inter-token information mixing, while proposing a visualization analysis method based on attention matrix forward propagation. The proposed method achieves interpretable analysis of Transformer model decision-making processes by quantitatively measuring the contribution of different input tokens to model outputs. To validate the method's effectiveness, we constructed an image classification model based on the ViT and conducted experiments on the CIFAR-10 dataset. The proposed method generated attention heatmaps that visualize the model's focus regions during the classification process.

**Keywords:** Transformer · Attention visualization · Interpretability.

## 1 Introduction

The Transformer [8] architecture revolutionized deep learning by replacing traditional RNN [2]/CNN [6] paradigms with self-attention mechanisms, thereby establishing the foundation for modern large-scale models. Subsequent developments, including BERT [3] and GPT [1] built upon Transformer, propelled natural language processing into the pre-training era. The Vision Transformer (ViT) [4] further demonstrated Transformer's capability to surpass CNN performance, emerging as a new paradigm in computer vision. Post-2021 witnessed exponential growth in large model research, with Transformer becoming the core architecture and spawning numerous efficient variants. Current research continues to advance Transformer-based architectures across multiple frontiers, including multimodal learning (e.g., CLIP [7]) and reasoning (e.g., CoT [9]), maintaining its position as one of the most active research domains in AI.

Deep neural networks represented by Transformer models are often regarded as "black boxes" due to their complex nonlinear structures and massive parameters, making the decision-making process between inputs and outputs inherently uninterpretable. This characteristic typically exhibits an inverse relationship between model performance and interpretability. Such lack of interpretability raises significant concerns in high-stakes domains like healthcare and finance. However, the attention mechanism possesses intrinsic interpretability advantages by explicitly modeling correlation weights between input elements, thereby providing a transparent window into model decisions.

[⋆] Email: XiaokaiZhang@shu.edu.cn

This paper first reviews the computational process of standard Transformer architecture, then formally defines the concepts of intra-token information mixing and inter-token information mixing. Building upon this foundation, we propose an attention matrix forward propagation method to compute global attention relationships between outputs and inputs. Finally, we implement a ViT-based image classifier and conduct experiments on the CIFAR-10 dataset [5]. The proposed method generates attention heatmaps that visualize the model's focus regions during classification.

## 2 Transformer Encoder Layer

This section revisits the computational process of the Transformer Encoder Layer, while adopting a modified notation and formulation scheme distinct from the original paper.

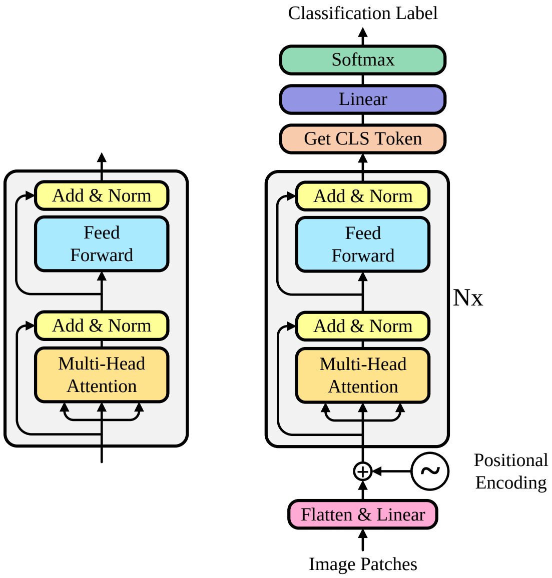

**Fig. 1.** Transformer Encoder Layer (Left) and ViT-CIFAR (right)

As illustrated in Figure 1, the standard Transformer Encoder Layer processes an input $X \in R^{n \times d_{model}}$ representing embedded representations of one $CLS$ token and $n-1$ image tokens. The input $X$ first undergoes three linear transformations $W^Q \in R^{d_{model} \times d_k}$, $W^K \in R^{d_{model} \times d_k}$ and $W^Q \in R^{d_{model} \times d_v}$ to derive the query $Q \in R^{n \times d_k}$, key $K \in R^{n \times d_k}$, and value $V \in R^{n \times d_v}$ matrices, as formalized in Equations 1-3.

$$Q = XW^Q \tag{1}$$

$$K = XW^K \tag{2}$$

$$V = XW^V \tag{3}$$

The self-attention scores $A \in R^{n \times n}$ are computed using Formula 4, and the self-attention layer output output $Y \in R^{n \times d_v}$ is obtained via Formula 5.

$$A = \text{softmax}(\frac{QK^T}{\sqrt{d_k}}) \tag{4}$$

$$V' = AV \tag{5}$$

Additionally, Multi-Head Attention (MHA) is employed to project $X$ into different feature subspaces. Here, we adopt a formulation different from the original Transformer paper, as shown in Equation 7, where $V_i'$ is derived from Formula 5, $h$ is the number of heads, weight $W_i^O \in R^{d_v \times d_{model}}$ and output $X' \in R^{n \times d_{model}}$.

$$X' = \sum_{i=1}^{h} V_i' W_i^O \tag{6}$$

The above describes the computational process of the Multi-Head Attention Sublayer. The output of this sublayer is then passed through a residual connection followed by Layer Normalization, as formalized in Equations 7.

$$Y = \text{LayerNorm}(X + X') \tag{7}$$

Layer Normalization (LayerNorm) operates exclusively on the feature dimension within a single token, and is completely independent of other tokens, as formalized in Equations 8, where $\mu_i$ and $\sigma_i^2$ represent the mean and variance of each feature dimension in $x_i$ respectively, $\gamma$ and $\beta$ is a learnable scaling factor, and $\epsilon$ is a small constant to prevent division by zero.

$$y_i = \gamma \cdot \frac{x_i - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} + \beta, \quad \forall x_i = X'[i] + X[i], i = 1, 2, \ldots, n \tag{8}$$

The input is then passed through a feedforward neural network with ReLU activation, as shown in Formula 9, where $W_1 \in R^{d_{model} \times d_{ff}}$, $W_2 \in R^{d_{ff} \times d_{model}}$, $b_1 \in R^{d_{ff}}$ and $b_2 \in R^{d_{model}}$ are the learnable weights and bias of the linear transformation. A residual connection and layer normalization are subsequently applied again.

$$Y' = \max(0, YW_1 + b_1)W_2 + b_2 \tag{9}$$

$$Z = \text{LayerNorm}(Y + Y') \tag{10}$$

## 3 ViT-CIFAR

We propose an image classifier based on the Transformer Encoder Layer and ViT architecture, with its detailed structure illustrated in the right panel of

Figure 1. After passing through the ViT-CIFAR network's hybrid image feature processing, the label-relevant features are aggregated into the CLS token. We extract this *CLS* token while disregarding the other tokens.

$$X_{cls} = X[0] \tag{11}$$

The *CLS* token then undergo a linear transformation $W^L \in R^{d_{model} \times l}$ followed by softmax normalization to predict the image label across $l$ classes:

$$\hat{X}_{label} = \text{softmax}(X_{cls} W^L) \tag{12}$$

The loss function is shown in Formula 13, where $\hat{X}_{label}$ is the predicted image label, $X_{label}$ is the ground truth, $\sigma$ is the sigmoid activation function.

$$\mathcal{L} = -\frac{1}{l} \sum_{i=1}^{l} X_{label,i} \cdot \log(\sigma(\hat{X}_{label,i})) + (1 - X_{label,i}) \cdot \log(1 - \sigma(\hat{X}_{label,i})) \tag{13}$$

## 4 Attention Matrix Forward Propagation

We first formally define the concepts of *intra-token* information mixing and *inter-token* information mixing. Given a matrix $X \in R^{n \times d_{model}}$ composed of embeddings from n tokens, where each row vector $x_i$ represents the $d_{model}$-dimensional embedding of token $i$. Left-multiplying matrix $X$ by matrix $W \in R^{k \times n}$ is termed inter-token information mixing, where $k$ denotes different modes of information mixing. Right-multiplying matrix $X$ by matrix $W \in R^{n \times d}$ is termed intra-token information mixing, where $d$ represents the dimensionality of the new feature space after linear transformation.

Since we focus on the global attention distribution across tokens, we retain the left-multiplication matrix while ignoring the right-multiplication matrix. Additionally, operations on individual tokens, such as LayerNorm and activation functions, do not alter the inter-token attention distribution and can thus be disregarded. Substituting Equations 3, 5, and 6 into 15 yields:

$$Y = \text{LayerNorm}(X + \sum_{i=1}^{h} A_i X W_i^V W_i^O) \tag{14}$$

The symbol $\overset{\triangle}{\approx}$ represents the computed result after disregarding both intra-token information mixing and operations on individual tokens, allowing us to derive Equation (10) as follows:

$$Y \overset{\triangle}{\approx} (I + \sum_{j=1}^{h} A_j) X \tag{15}$$

Substituting Formula 9 into 10, we obtain Formula 16. This indicates that the feedforward layer does not alter the attention weights between tokens.

$$Z = \text{LayerNorm}(Y + \max(0, Y W_1 + b_1) W_2 + b_2) \overset{\triangle}{\approx} Y \tag{16}$$

Based on the above discussion, we derive the attention forward propagation formula for ViT-CIFAR:

$$X_{label} \overset{\triangle}{\approx} \prod_{i=N}^{1} (I + \sum_{j=1}^{h} A_{ij})X \tag{17}$$

## 5 Experiment Results and Discussion

The model has an embedding dimension of 256, 8 attention heads, a dropout rate of 0.1, and 6 stacked layers. The CIFAR-10 dataset contains 50,000 training images and 10,000 test images, for which we adopt the original dataset split. For training, we employ the Adam optimizer with a batch size of 100, a learning rate of 0.0001, and train for 100 epochs. We conducted the training on a single NVIDIA GeForce RTX 4070 GPU.

Under this configuration, the training process took approximately 10 minutes and achieved a classification accuracy of 99.22% on the test set. Subsequently, we evaluated the proposed attention visualization method using the trained model and several test set samples.

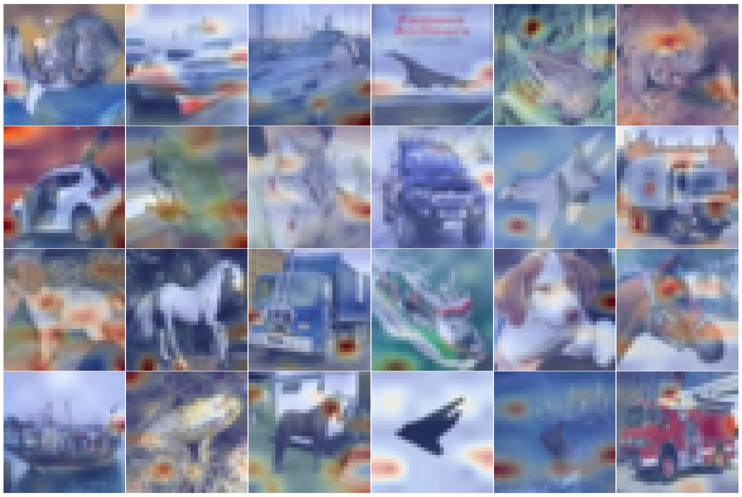

**Fig. 2.** Attention Visualization.

The experimental results indicate that the proposed method does not appear to demonstrate interpretability as intended. This limitation could stem from either inherent flaws in the methodology itself or potential issues in the implementation. Subsequent efforts will focus on refining the theoretical framework and improving the programming implementation.

## 6 Conclusion

This paper proposes an attention visualization method and conducts experiments on the CIFAR-10 dataset. However, the experimental results demonstrate that this is not an effective attention visualization approach. I will further refine both the theoretical analysis and engineering experiments to improve this method comprehensively.

## References

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
2. Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1724–1734. Association for Computational Linguistics, Doha, Qatar (Oct 2014)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Communications of the ACM **60**(6), 84–90 (2017)
7. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
9. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)