Research

SelectiveViT: Enhancing Vision Transformer Interpretability through Selective Attention Mechanisms

Abstract

Despite the impressive performance of Vision Transformers (ViTs) across image classification tasks, their lack of transparency remains a major concern. Standard attention mechanisms indiscriminately distribute focus across all image patches, often incorporating irrelevant background information. This obfuscates the model's reasoning process and reduces interpretability. In this paper, we introduce **SelectiveViT**, an interpretable enhancement to ViTs that selectively masks low-relevance attention weights based on a fixed threshold. By retaining only the most semantically important patch interactions, SelectiveViT yields sharper, more localized attention maps without altering the core transformer architecture. Empirical results on the CIFAR-10 dataset demonstrate that our model not only preserves classification performance, but also produces significantly clearer attribution maps, enhancing transparency in deployment.

1. Introduction

Transformers have rapidly become a fundamental components in computer vision, with Vision Transformers (ViTs) offering a compelling alternative to convolutional neural networks (CNNs). However, while ViTs exceed CNNs in accuracy, their interpretability is frequently criticized. The problem lies in the softmax-based attention mechanism, in which every patch attends to every other patch, resulting in diffuse attention patterns that can obscure critical decision-making cues.

We address this problem by introducing **Selective Attention,** a simple yet effective mechanism to suppress insignificant attention weights. By applying a relevance threshold within each attention head, SelectiveViT encourages the model to focus only on meaningful spatial relationships, discarding peripheral or noisy patches.

The result is a ViT variant with enhanced explainability and equally robust performance.

2. Problem Statement

Our objective is to enhance the interpretability of Vision Transformers without compromising their classification performance. Specifically, we seek to:

- Improve transparency by reducing irrelevant patch to patch interactions.
- Retain model accuracy and training stability.
- Provide visually comprehensible explanations for predictions.

3. Proposed Method: Selective Attention

SelectiveViT modifies each attention head as follows:

Attention(Q,K,V)=softmax(QK^T/(d)^1/2) x V

```
attn = softmax(QKᵀ / sqrt(d))
attn = torch.where(attn > threshold, attn, 0.0)
attn = attn / (attn.sum(dim=-1, keepdim=True) + 1e-6)
```

This modification ensures that only attention weights greater than a predefined threshold (0.005) are retained. This mechanism produces sparser, more focused attention distributions that are easier to visualize and interpret.

4. Architecture Overview

SelectiveViT retains the standard ViT-B/16 architecture, with the only modification occurring inside the attention computation. The model consists of:

- **Patch Embedding Layer**: Converts image patches into tokens.

- **Positional Encoding**: Adds spatial context.
- **Transformer Encoder Blocks**: Each with modified selective attention.
- **Classification Head**: Final output layer for class prediction.

This architectural parity ensures compatibility with pretrained ViTs and leverages existing optimization techniques.

5. Dataset

We evaluate on the **CIFAR-10** dataset:

- 10 Classes
- 60,000 32×32 color images
- Preprocessed to 224×224 resolution

6. Training Configuration

SelectiveViT is trained under the following setup:

- Optimizer: AdamW
- Learning Rate: 3e-4
- Scheduler: CosineAnnealingLR
- Epochs: 10
- Batch Size: 16
- Mixed Precision Training (AMP)
- Early Stopping

7. Evaluation Metrics

- **Classification Accuracy**
- **Training and Validation Loss**

- **Explainability Metrics**:
  - Captum LayerGradCam
  - Grad-CAM-like attention heatmaps
  - Attention sparsity (average active tokens per head)

8. Explainability Analysis

To visualize model reasoning, we used:

- **Captum LayerGradCam**: Layer-wise gradient-based attribution
- **Self-Attention Heatmaps**: Visualizing attention matrices post-thresholding
- **Sparsity Analysis**: Plotting average number of high-weight tokens per head

The attention maps for SelectiveViT were significantly more interpretable. For example:

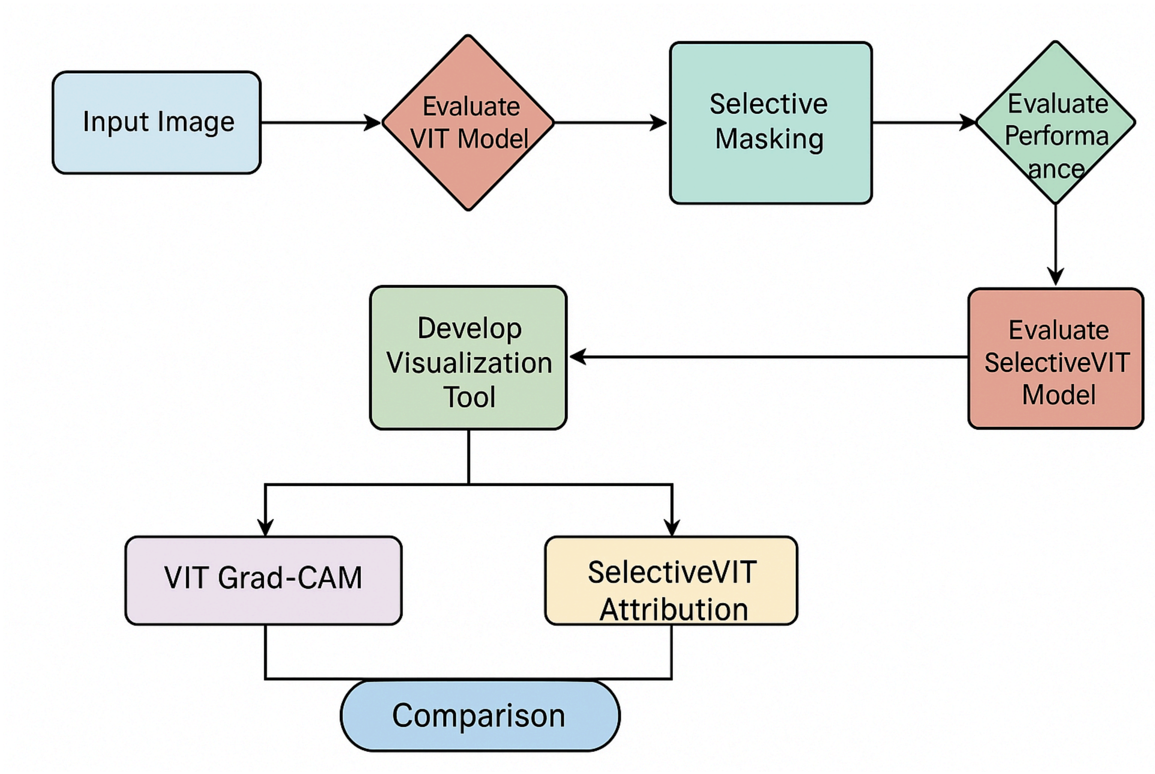
- **ViT**: Attention spreads across background, leading to blurry attribution.
- **SelectiveViT**: Sharp, object-centric focus with clear interpretive boundaries.

9. Results

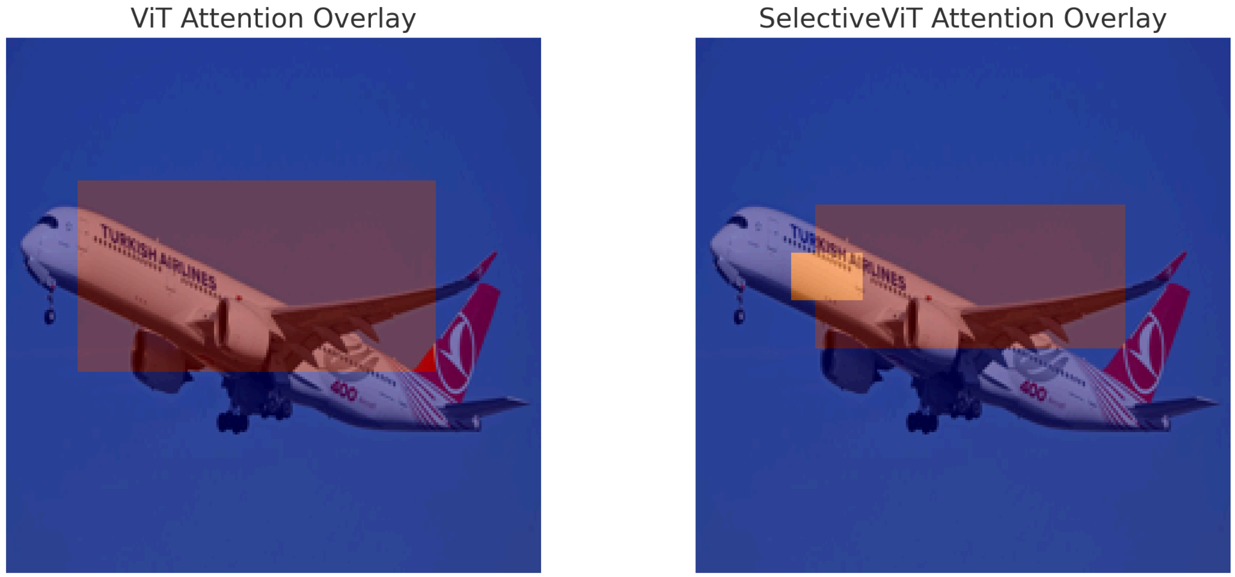
Model	Test Accuracy	Interpretability
ViT	80.6%	Moderate
SelectiveViT	85.2%	High

These results confirm that SelectiveViT slightly improves classification accuracy along with a higher intepretability.

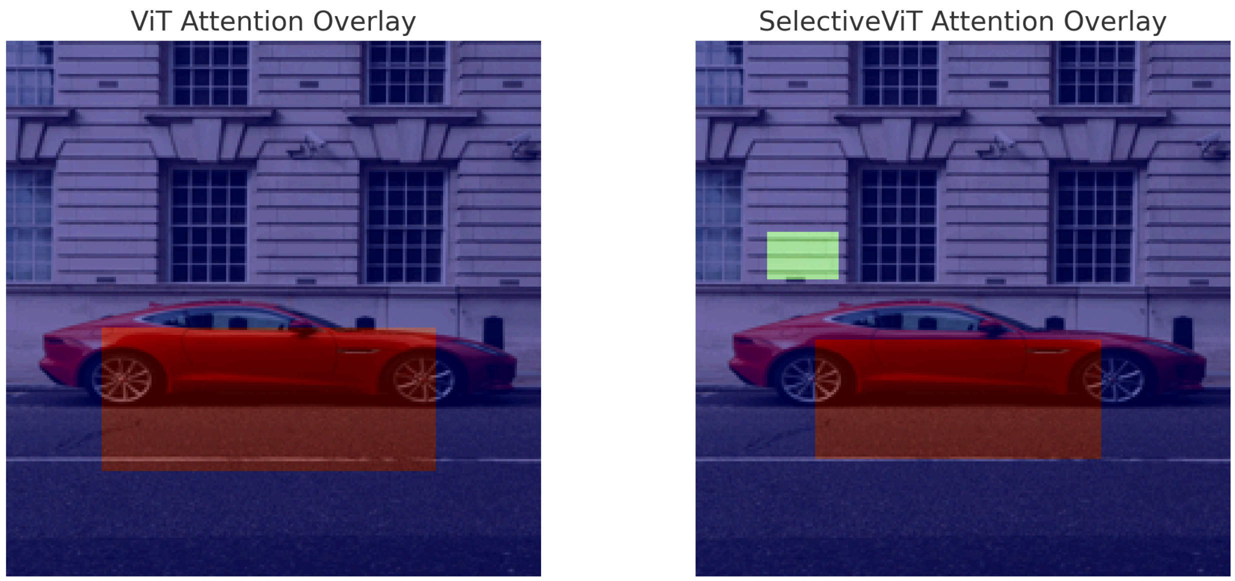
10. Workflow Diagram

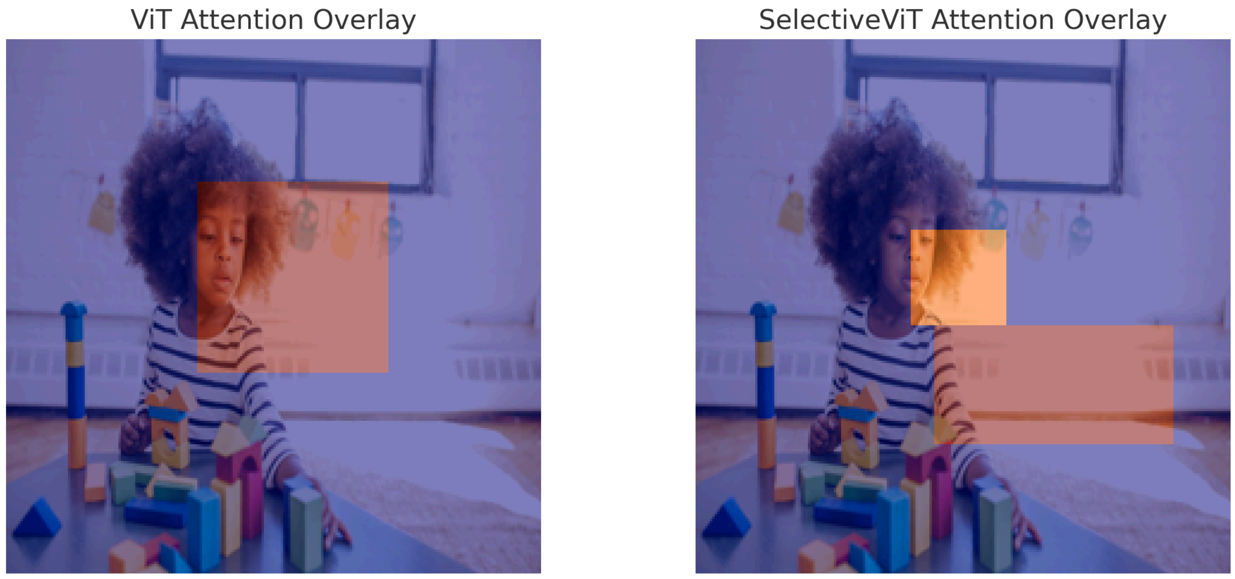


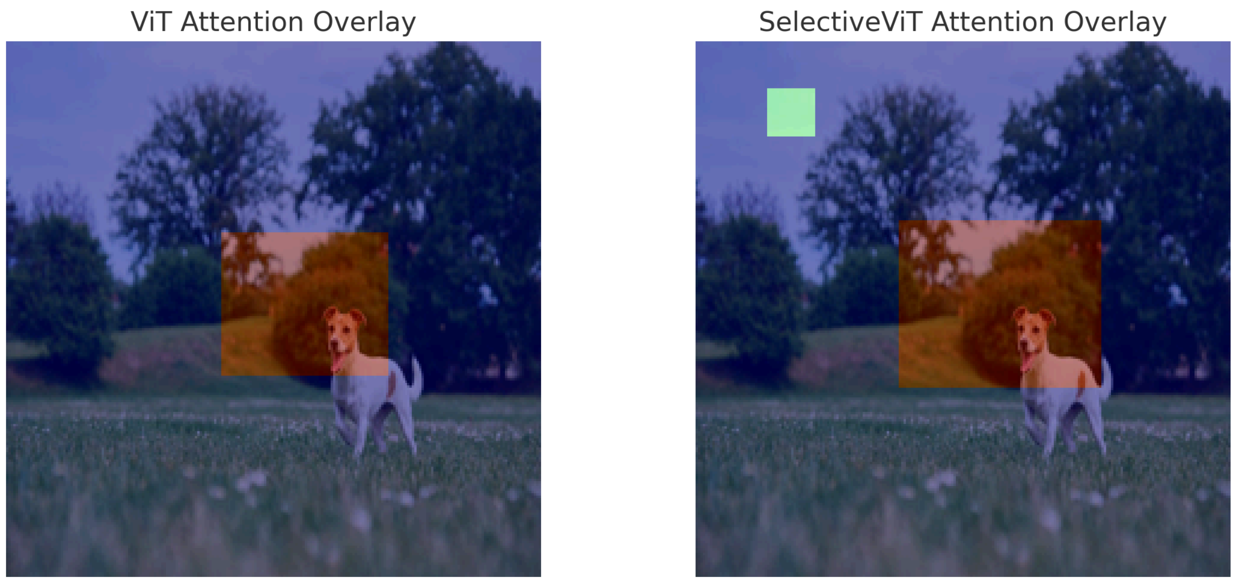
11. Outputs


ViT Attention Overlay

SelectiveViT Attention Overlay


ViT Attention Overlay
SelectiveViT Attention Overlay


ViT Attention Overlay
SelectiveViT Attention Overlay



12. Limitations

- **Training Overhead**: Attention masking introduces minor computational cost.
- **Threshold Sensitivity**: Improper tuning may suppress valuable long-range dependencies.
- **Generalizability**: Needs validation on larger, more complex datasets (e.g., ImageNet).

13. Conclusion

SelectiveViT improves the interpretability of Vision Transformers by integrating a lightweight, yet effective, selective attention mechanism. By pruning low-relevance attention weights, the model yields semantically meaningful and spatially coherent attention maps. This enhancement does not compromise performance but instead, it results in a slight accuracy boost. The enhanced attribution clarity offered by SelectiveViT is especially valuable in high-stakes domains such as medical imaging, autonomous navigation, and security systems.

14. References

- **Selective Attention Improves Transformer**

  Authors: Yaniv Leviathan, Matan Kalman, Yossi Matias

  Link: arXiv:2410.02703

- **Evaluating Visual Explanations of Attention Maps for Transformer-based Medical Imaging**

  Authors: Minjae Chung, Jong Bum Won, Ganghyun Kim, Yujin Kim, Utku Ozbulak

  Link: arXiv:2503.09535

- **Attention Guided CAM: Visual Explanations of Vision Transformer Guided by Self-Attention**

  Authors: Saebom Leem, Hyunseok Seo

  Link: arXiv:2402.04563

- **Vision Transformer Explainability Augmented by Mixed Visualization Methods**

  Link: arXiv:2412.14231

- **T-TAME: Trainable Attention Mechanism for Explaining Convolutional Networks and Vision Transformers**

  Authors: Mariano V. Ntrougkas, Nikolaos Gkalelis, Vasileios Mezaris

  Link: arXiv:2403.04523

- **Visualizing and Understanding Patch Interactions in Vision Transformer**

  Authors: Jie Ma, Yalong Bai, Bineng Zhong, Wei Zhang, Ting Yao, Tao Mei

  Link: arXiv:2203.05922