Eliminating High-Norm Artifacts in Vision Transformers via Gated Attention

Team Members: Vince Wu, Parth Sheth

Emails: vincewu8@stanford.edu, parth2@stanford.edu

1 Motivation

Vision transformers (ViTs) have become a prominent computer vision architecture, yet they display an architectural artifact. ViTs trained on large datasets and/or for many epochs repurpose low-information background tokens (i.e. sky or grass) into "attention sinks" for global information, overwriting local context [1]. Softmax attention forces a nonzero output summing to 1, so when the model has no relevant token to attend to, it must dump probability mass somewhere. These artifact tokens can distort attention maps, contribute to quantization collapse, and degrade performance on downstream tasks like segmentation and object discovery.

The current solution is to append extra register tokens to the image embedding, accommodating the artifacts by providing a designated "trash can" for the attention mechanism. This incurs a quadratic computational penalty $O((N+K)^2)$ as the number of registers K scales, and it introduces a hyperparameter (K) that must be tuned for every task.

Recent work has shown that adding a non-linear gate following scaled dot-product attention (SDPA) addresses "attention sinks" in LLMs by allowing softmax attention to be scaled rather than forcing the output to sum to 1 [2]. We hypothesize that SDPA output gating implemented in ViTs will intrinsically address high norm artifacts without the need for extra register tokens. Moreover, we seek to explore whether SDPA gating will a) stabilize training by preventing gradient explosions associated with attention sinks b) reduce computational cost compared to using register tokens

This is a theoretical result addressing ViT architecture.

2 Methods

We plan to apply and improve upon ViTs with the novel application of SDPA gating to ViTs. Specifically, we will modify the self-attention layer to include a learnable sigmoid gate g applied to the attention output O and hidden states after pre-normalization X:

$$O_{final} = O \odot \sigma(XW_{gate} + b_{gate})$$

We will use the DeiT III framework for ViT training, following this reproducibility study [3]. We plan to use the ImageNet-100 ($\tilde{1}$30k images) dataset for training and validation [4].

3 Intended Experiments

We plan to run three experiments on distinct ViT architectures. During each, we will train for 800 epochs and measure accuracy, training loss, attention map/L2 norm distribution, and FLOPS.

- Experiment 1: Vanilla ViT-Small. Train up to 800 epochs, measure distribution of l2-norms every 10 epochs to find the training epochs ϵ at which high-norm artifacts emerge.
- Experiment 2: ViT-Small with register tokens. Train ViT-small with K register tokens $K \in \{2, 4, 8, 16, 32\}$. Train ViT-small with $K = 4$ register tokens across $LR \in \{1e^{-3}, 3e^{-3}, 5e^{-3}\}$).
- Experiment 3: Gated ViT. Train ViT-small with SDPA gating across $LR \in \{1e^{-3}, 3e^{-3}, 5e^{-3}\}$). Also train ViT-small with gating after Value layer (also shown to cause improvements with LLMs). Additionally record average gating score to measure gating sparsity.
- Experiment 4: Benchmark the Gated ViT using ImageNet-A and ImageNet-R to evaluate whether mitigating attention sinks improves ViT's performance on adversarial data.

4 Team Contributions

- **Vince Wu**: Ideation, experimental design and execution, poster/paper creation
- **Parth Sheth**: Design/execution, poster/paper creation

References

[1] Darcet et al., Vision Transformers Need Registers, ICLR 2024

[2] Gated Attention for Large Language Models: Non-linearity, Sparsity, and Attention-Sink-Free

[3] Xiao et al., Efficient Streaming Language Models with Attention Sinks, ICLR 2024

[4] ImageNet-100 Dataset, Kaggle