Differential Gated Self Attention

Motivation. Transformers excel across tasks but remain vulnerable to corrupted inputs because standard self-attention treats all query–key interactions uniformly, allowing sensor noise or spurious tokens to be propagated and amplified. Prior differential-attention approaches provide noise cancellation but lack input-dependent, head-wise control and therefore cannot adapt inhibition at the granularity of tokens. This gap motivates an attention mechanism that delivers context-aware, per-head suppression to realize lateral-inhibition-style contrast enhancement within self-attention, while remaining drop-in compatible and lightweight.

Method. We propose *Multi-head Differential Gated Self-Attention (M-DGSA)*. Building on the Differential Transformer's [1] differential attention - which contrasts two parallel softmax maps to suppress common-mode noise - we replace the fixed global subtraction with a learned, *per-token, per-head* gate. For each head, inputs are projected into *excitatory* (Q^+, K^+) and *inhibitory* (Q^-, K^-) branches that yield $A^+ = \mathrm{softmax}(Q^+ K^{+\top}/\sqrt{d})$ and $A^- = \mathrm{softmax}(Q^- K^{-\top}/\sqrt{d})$, where d is half the head dimension. A gate $g = \sigma(XW_g + b_g)$ then performs subtractive fusion,

$$A = g \odot A^+ - (1 - g) \odot A^-,$$

and the fused map attends shared values V. Each head's output is normalized with head-wise GroupNorm and scaled by $(1 - \lambda)$ with $\lambda = 0.8$. The design tiles across h heads and is concatenated as in standard multi-head attention, preserving $O(hN^2)$ complexity with only lightweight gating overhead. Ablations indicate a single-layer gate is most effective. Taken together, these choices implement *input-conditioned lateral inhibition* inside self-attention. ***Setups.*** We instantiate M-DGSA in a Transformer [2] encoder for text (DGT) and a Vision Transformer (ViT) [3] for images (DGViT). Models are trained from scratch (PyTorch) on common benchmarks without external pretraining. Vision: CIFAR-10/100, Fashion-MNIST, SVHN. Language: Rotten Tomatoes, IMDB, AG News, 20 Newsgroups. Architectures match vanilla baselines in width/depth. For text, we report both diverse FFN setups and a controlled setting with identical FFN (SwiGLU) to isolate attention effects while for vision, we keep the original ViT FFN with GeLU, which performed best in our trials versus SwiGLU.

Results. Across five seeds, M-DGSA improves accuracy and robustness over vanilla Transformer/ViT and the Differential Transformer baseline. Examples: on CIFAR-10, DGViT yields $\approx +2\%$ absolute over a matched ViT; on 20 Newsgroups, DGT achieves up to 63.5% test accuracy, a $\approx$ 12–17% gain over matched baselines. On Rotten Tomatoes/IMDB/AG News, DGT provides consistent +0.5–1.5% improvements, while DGViT also outperforms a matched ViT on CIFAR-100, Fashion-MNIST, and SVHN across all seeds. Attention-rollout visualizations show sharper focus on salient structures and suppression of background clutter in both images and text.

Takeaways. Input-conditioned, head-wise inhibitory gating is a simple, biologically inspired addition that (i) improves noise resilience, (ii) sharpens attention maps, and (iii) generalizes across language and vision classification within a self-contained attention module.

[1] Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer. *arXiv preprint arXiv:2410.05258*, 2024.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.