Deep Integrated Explanations

ACM Reference Format:
. 2023. Deep Integrated Explanations. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 GRADIENT ROLLOUT IMPLEMENTATION

The Gradient Rollout (**GR**) technique is a modified version of the Attention Rollout (**AR**) [1] method, which differentiates itself by including a Hadamard product between each attention map and its gradients in the computation, rather than relying solely on the attention map. The GR method can be expressed mathematically as follows:

$$A'_b = I + E_h(A_b \circ G_b), \tag{1}$$

$$GR = A'_1 \cdot A'_2 \cdots A'_B. \tag{2}$$

where A_b is a 3D tensor consisting of the 2D attention maps produced by each attention head in the transformer block b, G_b is the gradients w.r.t. A_b. I is the identity matrix, B is the number of transformer blocks in the model, E_h is the mean reduction operation (taken across the attention heads dimension), and $\circ$ and $\cdot$ are the Hadamard product and matrix multiplication operators, respectively.

2 COMPUTATIONAL COMPLEXITY OF DIX

The computational complexity of DIX is dominated by the order of the iterated integral to be computed. Practically, one will employ the discrete version from Eq. **??** which involves nested sums (each with n summands). Each summand requires the application of $\mathbf{q}^l$, whose computational complexity depends on the specific implementation. For example, in Sec. **??**, $\mathbf{q}^l$ combines both activation and attention maps, respectively, and their gradients, hence the computation involves both forward and backward passes. Therefore, assuming the computational complexity of $\mathbf{q}^l$ is $O(Q)$, the overall computational complexity of Eq. **??** is $O(n^\beta Q)$, with $\beta = \sum_{i=0}^{l} b_i$.

REFERENCES

[1] Samira Abnar and Willem Zuidema. 2020. Quantifying Attention Flow in Transformers. *arXiv preprint arXiv:2005.00928* (2020).

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn