Enhancing Accuracy through Architectural Modifications in Vision Transformers

Abstract for the MLDM Workshop of the AI*IA Conference

Mauro Mezzini[1], Alessio Ferrato[2], Carla Limongelli[2], and Giuseppe Sansonetti[2]
1- Department of Education, Roma Tre University,
Via del Castro Pretorio, 20, 00185, Rome, Italy
2- Department of Engineering, Roma Tre University,
Via Vito Volterra 62, 00146, Rome, Italy
mauro.mezzini@uniroma3.it,alessio.ferrato@uniroma3.it,
carla.limongelli@uniroma3.it,giuseppe.sansonetti@uniroma3.it

Abstract

The Transformer architecture represents one of the most significant advancements in Artificial Intelligence in the last decade. Its evolution started from Language Translation [1] and ended as the main tool in Natural Language Processing [2], eventually giving rise to the Large Language Models [3]. One of the most critical components of the Transformer architecture is the *multi-head attention mechanism*. In the self-attention mechanism, the output sequence $y = (y_0, \ldots, y_{N-1})$ (with N denoting the number of tokens of the input sequence) of a self-attention module is obtained as a linear combination of the input sequence $x = (x_0, \ldots, x_{N-1})$, that is, $y_j = \sum_{i=0}^{N-1} \alpha(x_j, x_i)x_i$. The most popular choice for the coefficients α is the one in which $\sum_{i=0}^{N-1} \alpha(x_j, x_i) = 1$ and $\alpha(x_j, x_i) \geq 0$ for all $i, j = 0, 1, \ldots N-1$. To ensure this, a *score function* $a(x_j, x_i)$ is provided and the softmax operation on a ensures that the coefficient α forms a convex combination, that is, $\alpha(x_j, x_i) = softmax(a(x_j, x_i))$. In the multi-head self-attention module, there are h scores matrices where h is the number of heads, and the scores of different sequences are batched together to speed up the computation.

We propose a methodology consisting of applying a convolutional filter to the batch of scores, followed by a rectified linear unit operation. We tested this modification using a Vision Transformer architecture on the CIFAR-10 dataset[1], obtaining encouraging results.

References

1. D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).
2. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 6000–6010.
3. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), ACL, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[1] https://www.cs.toronto.edu/~kriz/cifar.html