FEED FORWARD THAT PRODUCES SHAKESPEARE...

Timur Ishuov, Gergely Acsai, Richárd Farkas
Department of Computer Algorithms and Artificial Intelligence
University of Szeged

ABSTRACT

When we use a Feed Forward in General Purpose Transformer neural network, it is the last dimension's vector that participates in the vector and weight matrix multiplication (parallel to Time domain). Hence, the Feed Forward neural network calculates different interconnections between Embedding nodes, but the Time domain is left untouched. But if you transpose input between Time domain and Embedding domain, now the Time domain nodes participate in interconnection computation. We only needed to shut down the future nodes...

Keywords Self-Attention · General Purpose Transformer · Natural Language Processing

Methods and Materials

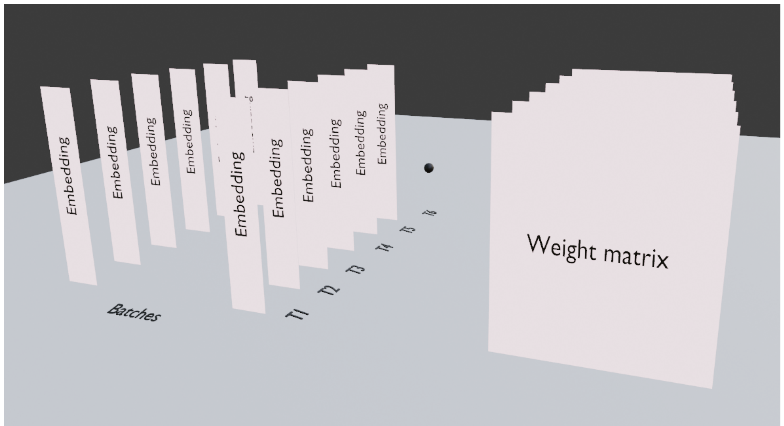
"All Shakespeare" dataset, data type - float16-mix, token size – 3 symbols, vocabulary size – 8507 tokens, token embedding – 800, number of heads – 20 (head size – 40), layers – 20, time intervals (context block) - 384, number of parameters 41.5M, optimizer - Adam, learning rate – 3e-5, dropout – 0.1, batch size – 64, pytorch library, videocard - NVidia GPU 3060 RTX

1 Introduction

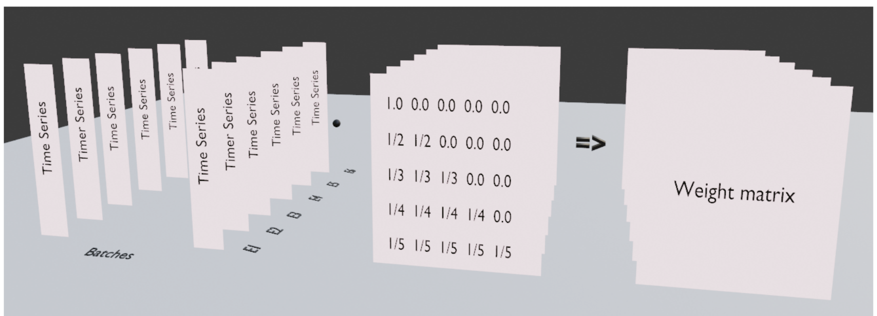
This work is based on the recent achievements of the simplification of Self-Attention[1] mechanism[2][3][4]. Though we started our research independently, the closer alternatives are work by [2] applied for the speech recognition task, unpublished work of Abdulrahman, H [5] where the author passed the united token and position embedding values through 2 weight matrices independently and used masked second output as an attention for the first output and pNLP-Mixer[6] , the project based on MLP-Mixer[7] which studied the replacement of Convolution[8] and Self-Attention with Multilayer Perceptron[9] with Channel and Spatial resolutions mixing. They use different tools, e.g. minimum hash fingertip in non-trainable projection layer to enhance MLP-Mixer, and the work can be considered as more advanced step in comparison to ours achieving 99.4% performance of mBERT on MTOP with less parameters involved. Our work is a simple showcase that it is possible to use Feed Forward layer instead of Self-Attention in Language Model[10] to generate a reasonable text as an output, we only need to suppress nodes that are responsible for future tokens in computational communication. This results in only one triangular linear layer for each head. We were able to put 20 heads and 20 layers inside single NVidia 3060 RTX with 12GB memory.

2 Feed Forward network that produces Shakespeare

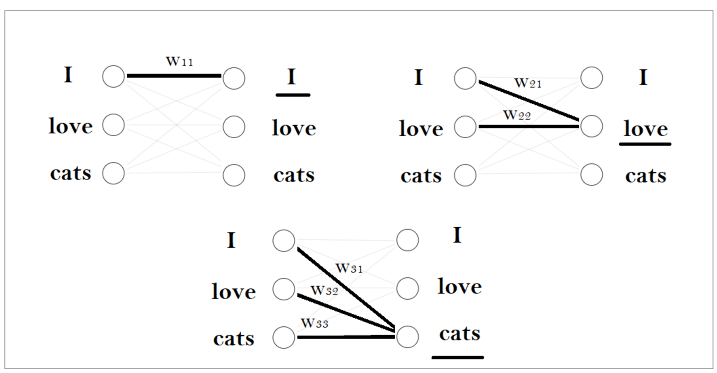
When you have "3D" Tensor consisting of: Batch, Time, Embedding (which in general case can be thought as a description of each time series), Linear Layer just uses the last "1D" vector "Embedding" and multiplies it by "2D" weight matrix producing "1D" vector output, but does it for all "Time" series in parallel. Hence, TIME does not participate in the linear layer. So, there is no interconnection between tokens or words in the sentence. But what Linear Layer does, it learns interconnection (through weights) between each embedding value or node in the embedding vector to better suit the target:



Embedding domain. Conventional computation



Time domain with Lower Triangular mask applied to Weight Matrix



Resultant operation in Time domain

Figure 1: Processing of Embedding Vectors by an interconnection with the output layer [1]

Though we encounter limitations because we do not consider TIME domain. Let us transpose the (Batch, Time, Embedding) into (Batch, Embedding, Time). In this way Time series will participate in computation. What is still missing is the idea that the current Word (Embedding) cannot talk to the future Words (Embeddings). We can apply lower triangular matrix with distributed values as the mask to the weights as in Fig.1.

Averaging weights in this manner helps the neural network to adjust faster in the beginning, but weights are still learned, not to mention, that with a lower triangular matrix we do not consider future words. Instead of 1.0, one can also use the

[1] for illustration purposes only, vectors should be transposed in the real calculation

decreasing discount factor with an initial value of $\gamma = 0.9999$, to make the latest words more important, but 1.0 works perfectly fine and faster.

This resembles the Attention mechanism with straightforward simplifications.

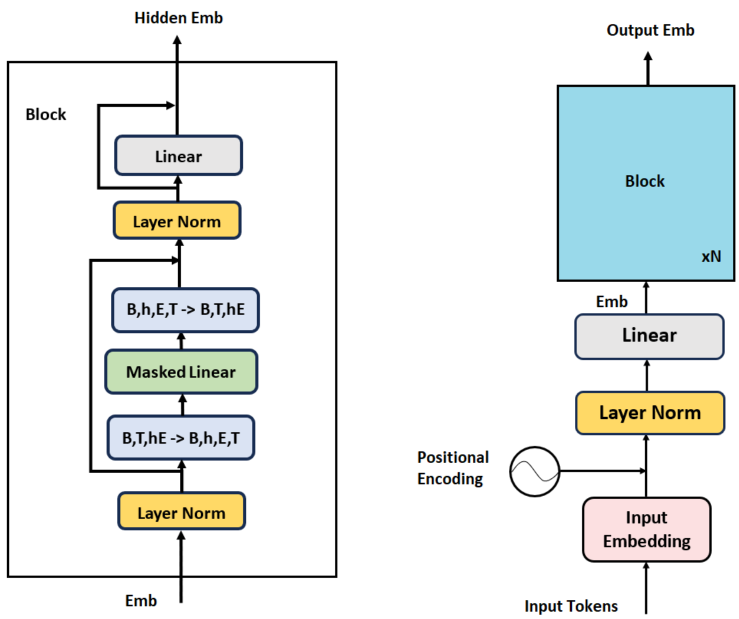


Figure 2: Feed Forward Transformer Architecture

3 Conclusion

We tested our algorithm on "All Shakespeare" Dataset. The model had only 34.2M parameters, with context size of 512 tokens of size-1. It started producing meaningful words at 1000 iteration, after 2000 - some phrases, and after 10000 iterations it started producing small sentences without particular meaning.

```
AUFIDIUS:
Would they often fair lady G. Give me and Lord Northumberland,
Is Bring me go.

ROMEO:
Come, no; for them quite lodge, the feast, with
aqua-vitae, I pray you adorn; and she hath still.

FLORIZEL:
I will were seen in blood when night,
And for
dest gentleman:
I know not much despare this merry fool.
By the earless.

ARCHBISHOP OF GAUNT:
Believe me, already: shed by riches himself of sorrows and envy's blood and pi
that makes one with
```

Figure 3: Word completion

We believe it can be further improved with a Large Language model and/or with Reinforcement Learning. There is one unsolved drawback however: with single value network we need to re-think cross-attention mechanism.

Acknowledgments

This work was done by means of computational resources of the Department of Computer Algorithms and Artificial Intelligence, University of Szeged. The authors are the Stipendium Hungaricum and Hungary State Scholarship holders. We are grateful for the opportunity that was provided.

References

[1] 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.

[2] Shucong Zhang, Erfan Loweimi, Peter Bell, and Steve Renals. When can self-attention be replaced by feed forward layers? *arXiv preprint arXiv:2005.13895*, 2020.

[3] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. *Advances in Neural Information Processing Systems*, 35:4203–4217, 2022.

[4] Haoneng Luo, Shiliang Zhang, Ming Lei, and Lei Xie. Simplified self-attention for transformer-based end-to-end speech recognition. *arXiv preprint arXiv:2005.10463*, 2020.

[5] Hunar Abdulrahman. Reweight gpt, 2023.

[6] Francesco Fusco, Damian Pascual, and Peter Staar. pnlp-mixer: an efficient all-mlp architecture for language. *arXiv preprint arXiv:2202.04350*, 2022.

[7] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.

[8] Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel. Handwritten digit recognition with a back-propagation network. In *Neural Information Processing Systems*, 1989.

[9] F. Rosenblatt. *The Perceptron, a Perceiving and Recognizing Automaton Project Para*. Report: Cornell Aeronautical Laboratory. Cornell Aeronautical Laboratory, 1957.

[10] Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*, 2020.