Feed Forward that produces Shakespeare...

Anonymous ACL submission

Abstract

When we use a Feed Forward in the 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...

Methods and Materials

"English Classic Writers" and "All Shakespeare" text corpora, data type - float16-mix, token size – 2 symbols, vocabulary size – 1092 tokens, vocabulary embedding - 1024, token embedding – 768, time intervals (context block) - 430, layers – 17, number of parameters 36.3M, learning rate – 3e-5, batch size – 64, pytorch library, videocard - NVidia GPU 3060 RTX.

1 Introduction

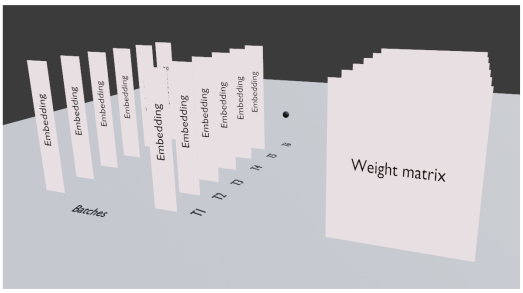
This work is a demonstration of potential and due to computational limitations and narrow data volume no comparison with state-of-the-art language models was made. For the latter, one can follow the recent achievements of the simplification of Self-Attention(Vaswani et al., 2017) mechanism, namely, S4-diagonal(Gupta et al., 2022), SGConv (Li et al., 2022), MEGA(Ma et al., 2022), SPADE (Zuo et al., 2022), FocalNet(Yang et al., 2022), previously, (Zhang et al., 2020), (Luo et al., 2020), more recently, Mamba(Gu and Dao, 2023). Though we started our research independently, the closer alternatives are work by (Zhang et al., 2020) applied for the speech recognition task, unpublished work of Abdulrahman, H (Abdulrahman, 2023) where the author passed the united token and position embedding values through 2 weight matrices independently and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(Fusco et al., 2022) , the project based on MLP-Mixer(Tolstikhin et al., 2021) which studied the replacement of Convolution and Self-Attention with Multilayer Perceptron with Channel and Spatial resolutions mixing(Chen et al., 2021). 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(Schick and Schütze, 2020) 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 weight matrix in place of Self-Attention. Because of computational and theoretical simplicity, our work can be used for tutorials and demonstrations.

2 Feed Forward network that produces Shakespeare

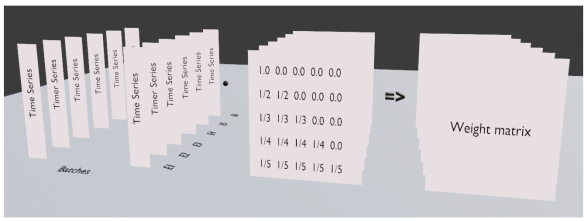
When you have "3D" Tensor consisting of: Batch, Time, Embedding (which in general case can be thought as a description of each time series), Linear Layer just uses the last "1D" vector "Embedding"[1]and multiplies it by "2D" weight matrix producing "1D" vector output, but does it for all "Time" series.

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

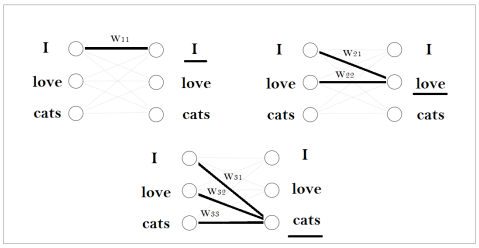
[1]"1D" vector is an abstraction for separate processing of Embedding columns for each Time Series row in the input matrix



Embedding domain. Conventional computation



Time domain with Lower Triangular mask applied to Weight Matrix



Resultant operation in Time domain

Figure 1: Separate processing of Embedding and Time Series Vectors [2]

the embedding vector to better suit the target.

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 (and we do not use bias that can distort interconnections between tokens). Instead of 1.0, one can also use the 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.

[2] for illustration purposes only, vectors and weight matrix should be transposed in the real $x \cdot W^T$ calculation

This resembles the Attention mechanism with straightforward simplifications.

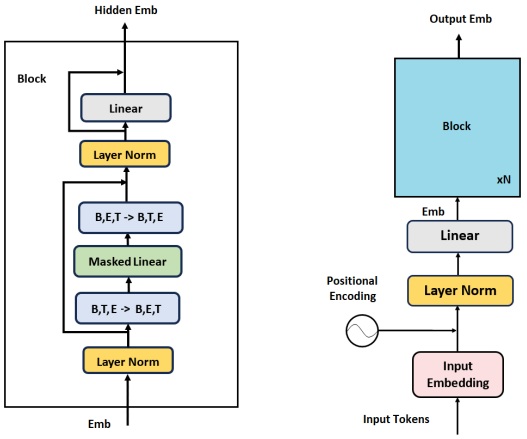


Figure 2: Feed Forward Transformer Architecture

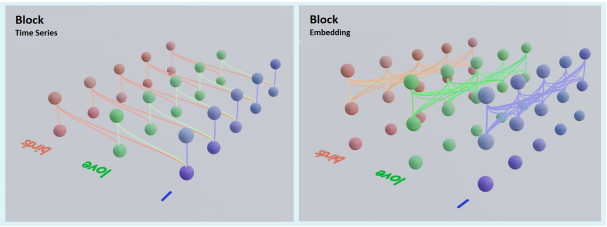


Figure 3: Operations inside Block

Each Time domain vector represents some characteristic of the token and can be assessed as an independent head. Linear layer prior to sequential blocks processes sum of vocabulary embedding and positional encoding producing more uniform vector of token embedding.

3 Advanced architecture

If our triangular linear layer with a periodic activation function like sin is followed by the second triangular linear layer, it will result in a Fourier Series (FS) like transformation: $b_2 + sin(xW_1^T + b_1)W_2^T$, or for Fourier Transform (FT): $b_2 + tanh(xW_1^T + b_1)W_2^T$. We tried this approach but without activation function in between so that the relationship between tokens is not distorted or squashed. Still, the loss dropped down faster. It is worth noting that the second linear layers should also have triangular weight matrix, since each its input node still corresponds to the respective token.

Considering Fourier Series and recent discoveries with Sine activation function(Sitzmann et al., 2020), we applied this methodology to Fully-Connected Linear Layers that follows Time domain

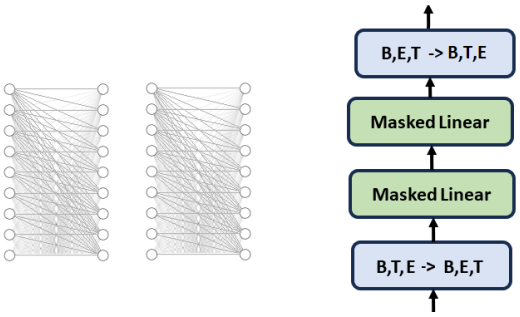


Figure 4: Fourier Transformation

processing. But instead of Sine and Cosine parts in Fourier Series, we used only Sine activation function due to presence of phase shift in the form of bias vector in the first Linear Layer, which to some extent replaces lack of Cosine part.

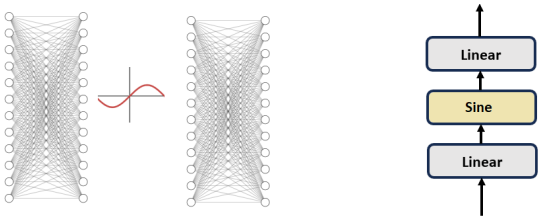


Figure 5: Sine Activation function in between Linear Layers

The model had 36.3M parameters, with vocabulary embedding size of 1024, token embedding size 768, context size of 430 tokens, and 17 sequential layers.

4 Early experiments and preliminary conclusion

We united, cleaned and processed the English Classic writers text corpora into a single dataset (1092 tokens, 1 token = 2 symbols) and pre-trained our model on this dataset. It started producing meaningful words at 1,000 iteration, after 2,000 - some phrases, after 10,000 iterations it started producing small sentences without particular meaning. And after approximately 100,000 iterations it produced sentences with clear understanding of the past context.

Then we fine-tuned our model on smaller "All Shakespeare" text corpus. As we tested different approaches before the architecture finalization which resulted in failures, based on the generated text, we can make a preliminary conclusion that the algorithm works, Fig 6.

```
FLORIZEL: Fortune speed us! Thus we set on, Camillo, to the sea-side.
CAMILLO: The swifter speed the better.
AUTOLYCUS: I understand the business, I hear it: to have an open ear, a quick eye,
 and a nimble hand, is necessary for a cut-purse; a good nose is requisite also, t
o smell out work for the other senses. I see this is the time that the unjust man
doth thrive. What an exchange had this been without boot! What a boot is here with
 this exchange! Sure the gods do this year connive at us, and we may do any thing
extempore. The prince himself is about a piece of iniquity, stealing away from his
 father with his clog at his heels: if I thought it were a piece of honesty to acq
uaint the king withal, I would not do't: I hold it the more knavery to conceal it;
 and therein am I constant to my profession. Aside, aside; here is more matter for
 a hot brain: every lane's end, every shop, church, session, hanging, yields a car
eful man work.
Clown: See, see; what a man you are now! There is no other way but to tell the kin
g she's a changeling and none of your flesh and blood.
Shepherd: Nay, but hear me.
Clown: Nay, but hear me.
Shepherd: Go to, then.
Clown: She being none of your flesh and blood, your flesh and blood has not offend
ed the king; and so your flesh and blood is not to be punished by him. Show those
```

Figure 6: Word completion

Limitations

While the Token Embedding size was 768 and the output of the model was 1092 to suit token numbers in vocabulary, there is respective necessity to increase embedding size (width) to 2048-4096 to better suit the state-of-the-art vocabulary's dimension of 50k-100k tokens. To increase number of time series (length) is also important as well as to increase number of layers (depth). So width, length and depth play a vital role. With only Feed Forward architecture and with a single embedding vector, there is one unsolved drawback: we need to re-think the cross-attention mechanism.

Ethics Statement

This work offers a solution that would help the scientific and industrial community reduce the size and ease of computation of large linguistic models that do not require cross-attention mechanism, which may ultimately have an environmental impact.

Acknowledgements

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

Hunar Abdulrahman. 2023. Reweight gpt.

Weihuang Chen, Fangfang Wang, and Hongbin Sun. 2021. S2tnet: Spatio-temporal transformer networks for trajectory prediction in autonomous driving. In

Asian Conference on Machine Learning, pages 454–469. PMLR.

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

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Ankit Gupta, Albert Gu, and Jonathan Berant. 2022. Diagonal state spaces are as effective as structured state spaces. *Advances in Neural Information Processing Systems*, 35:22982–22994.

Yuhong Li, Tianle Cai, Yi Zhang, Deming Chen, and Debadeepta Dey. 2022. What makes convolutional models great on long sequence modeling? *arXiv preprint arXiv:2210.09298*.

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

Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Luke Zettlemoyer. 2022. Mega: moving average equipped gated attention. *arXiv preprint arXiv:2209.10655*.

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

Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. 2020. Implicit neural representations with periodic activation functions. In *Proc. NeurIPS*.

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

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

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

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

Simiao Zuo, Xiaodong Liu, Jian Jiao, Denis Charles, Eren Manavoglu, Tuo Zhao, and Jianfeng Gao. 2022. Efficient long sequence modeling via state space augmented transformer. *arXiv preprint arXiv:2212.08136*.