Convolution Revolution: Wiener Filters for Better Attention

Andrei Danila

June 14, 2024

Abstract

The objective of this paper is to integrate Wiener filters, as described by Pelacani-Cruz et al, 2023 in "Convolve and conquer: Data Comparison with Wiener Filters" ([CSB+23]) in the Attention mechanism of the Transformer. The goal of this modification is to explore the potential benefits of convolutional similarity metrics in the comparison of the Query and Key vectors of tokens, working under the assumption that there is a spatial/contextual correlation between embeddings.

1 Literature Review

Recently, the field of natural language processing (NLP) has witnessed groundbreaking advancements, particularly since the Transformer model introduced in [VSP+23]. The Transformer's innovative self-attention mechanism has redefined the landscape of NLP by enabling efficient and scalable sequence processing, surpassing the capabilities of recurrent and convolutional neural networks. Concurrently, novel approaches in data comparison techniques, such as those proposed by Cruz et al. in 2023 ([CSB+23]), have emphasized the importance of maintaining global spatial correlations in data analysis, challenging the adequacy of traditional metrics like mean squared error (MSE). This paper seeks to integrate these advancements by exploring the potential benefits of convolutional similarity metrics, specifically Wiener filters, within the attention mechanism of the Transformer model. This integration aims to enhance the model's ability to capture spatial and contextual correlations between embeddings, ultimately improving the awareness of data comparisons in NLP tasks.

The paper "Attention is All You Need" ([VSP+23]) (hereafter sometimes referred to as "the Attention paper") by Vaswani et al., 2017, introduced the Transformer model. The core innovation of this model is the self-attention mechanism, which allows the model to weigh the relevance of different words in a sequence relative to each other, regardless of their distance. This mechanism operates through the calculation of scaled dot-product attention, where Query, Key, and Value vectors are derived from the input embeddings. These vectors are projected into multiple attention heads, enabling the model to capture various aspects of contextual relationships simultaneously. By applying these attention weights to the Value vectors, the model effectively captures and emphasizes the most contextually relevant parts of the sequence. The result is a model that can process sequences in parallel, significantly improving computational efficiency and allowing for greater scalability in handling longer sequences compared to traditional models. This attention mechanism lies at the heart of the Transformer's ability to understand and generate natural language with high accuracy, paving the way for subsequent advancements and applications in language modeling and beyond.

Recent advancements in data comparison techniques have highlighted the limitations of traditional methods like mean squared error (MSE), particularly in capturing global spatial structures and textures in data. In "Convolve and Conquer: Data Comparison with Wiener Filters" ([CSB+23]) (hereafter sometimes referred to as "the Wiener paper") by Cruz et al. (2023), the authors propose a novel approach using Wiener filters to measure data similarity. The convolutional nature of Wiener filters enables a more holistic comparison of data samples by maintaining global correlations, thereby addressing the inadequacies of MSE which assumes data points are independent and of equal importance. Wiener filters, traditionally used in image deconvolution, are repurposed here to globally match sample pairs through convolution. This method demonstrates significant improvements in applications such as data compression, medical imaging imputation, and generative modeling, showing higher resolution and better perceptual quality in reconstructed images. The Wiener-based similarity measure also enhances robustness against data translations. The research underscores the potential of Wiener filters to serve as a versatile tool for data comparison across various machine learning tasks, promoting a shift from local to global data analysis perspectives.

In terms of modified attention mechanisms, the following are the most cited. [CGRS19] introduced the strided attention mechanism, which is a factorized attention pattern where one head attends to the previous l locations, and the other head attends to every l-th location, optimizing for a stride close to $\sqrt{n}$. This mechanism was introduced as part of the Sparse Transformer architecture and is particularly effective for data with naturally aligned structures like images or music. However, it may struggle with non-periodic data such as text, where spatial coordinates do not necessarily match future relevance. In efficiency improvement of the attention mechanism, the most known paper is [BPC20], which introduced sparse attention to improve computation time for long-form document retrieval, and demonstrated state-of-the-art results on character-level language modeling, outperforming RoBERTa on long document tasks, and validated the Longformer-Encoder-Decoder (LED) variant on the arXiv summarization dataset.

The integration of Wiener filters into the attention mechanism presents a promising and innovative avenue for enhancing data comparison techniques in natural language processing. By leveraging the convolutional similarity metrics proposed by Cruz et al. ([CSB+23]), it is possible to improve the spatial and contextual correlation between embeddings, potentially leading to more robust and accurate models. This fusion of methodologies not only builds on the strengths of the Transformer's self-attention mechanism but also addresses the limitations of traditional data comparison metrics. As the field of NLP continues to evolve, such innovative integrations will be crucial in driving further advancements and achieving higher levels of performance and reliability in language modeling and other related tasks.

2 Tentative Methodology

The following methodology, as the heading implies, is tentative. Therefore, it will be approached in a cyclical, evolving manner.

2.1 Developing the new attention equation

At the moment of writing this plan, the modified attention mechanism is as follows: each entry in the attention scores matrix (which was originally populated by the Dot product of the pairs of Qs and Ks) will be the Wiener Similarity Metric (WSM) between each pair of K and Q. Whether this attention scores matrix will be scaled or not remains to be seen. Afterwards, the scores are softmin'ed (because Wiener Similarity Metric is 0 when the two elements are identical) and multiplied by the values matrix.

Therefore, the modified attention formula (without scaling) is:

$$\text{softmin}\left(\frac{WSM(Q,K)}{\sqrt{d_k}}\right)V$$

where

$$WSM(Q,K)$$

is the Wiener Similarity Metric between all possible combinations of Queries and Keys for one attention head. This returns a matrix of dimensions (ntokens, ntokens), retaining the dimensions of the original attention mechanism.

2.2 Writing code to reflect modifications

To implement the new attention mechanism, I will use the Pytorch Python library along with the AWLoss repository which accompanies the Wiener paper. Specifically, I will take the transformer code from the Annotated transformer paper ([Rus22])

Since the new attention mechanism requires all possible combinations of Queries and Keys, I anticipate needing to write custom CUDA/GPU/Triton code to accommodate that.

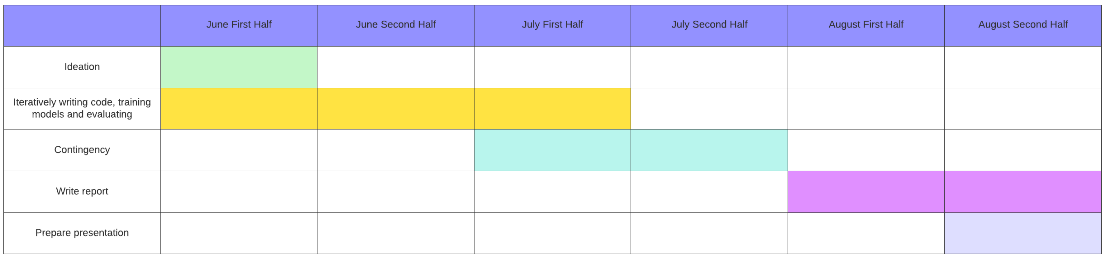
2.3 Training model

The model will be trained on GPUs, either from the college's HPC resource or elsewhere. The original transformer took 12 hours on 8xP100 GPUs to train, so we can assume that the Wiener transformer will take at least as long (because overhead is unknown at this point).

2.4 Evaluating Results

Due to constraints on model size, the evaluation of the proposed changes will be done using the English to French WMT2014 dataset and then scored using the BLEU metric. To ensure good comparisons, the trained model will reflect the number of parameters specified in the original "Attention is All You Need" paper (approx. 65M). The scores reported in the "Attention is All You Need" paper will serve as the control result, while the scores from the modified Wiener attention transformers (all the configurations) will be the variant results.

3 Workplan and Methodology



	June First Half	June Second Half	July First Half	July Second Half	August First Half	August Second Half
Ideation	🟩					
Iteratively writing code, training models and evaluating	🟨	🟨	🟨			
Contingency			🟦	🟦		
Write report					🟪	🟪
Prepare presentation						▨

Figure 1: Project Gantt Chart

3.1 Workplan

The workplan is as follows:

1. In the initial stages of the project, I will spend time thinking about the architecture of the Wiener transformer (as described in point 2.1), checking ideas with my supervisor.

2. Towards the end of this initial stage, I will also start writing code to reflect this new architecture. Once I have achieved inference, I will start training.

3. After training is possible (no bugs, it runs relatively efficiently), I will focus on hyperparameter optimization, looking at different settings I can modify, such as whitening factor, penalty function or scaling.

4. Towards the end of the training period (around mid July) I and my supervisor will make an assessment of whether or not it's worth continuing with this approach (the Wiener transformer). If not, I will switch to the contingency described below.

3.2 Contingency

While the original plan described above is relatively risky and relies on certain assumptions (most importantly, that embeddings are contextually or spatially linked), I propose a second approach to be tried in case the first approach fails. This second approach is to use the Wiener Loss (as described in [CSB+23]) as the loss function of a classic transformer in order to compare the output and target of the translation task (i.e. *ypred* and *ytrue*), instead of Cross Entropy Loss. This approach is more intuitive, and early research points to it being a valid approach. However, to push for novelty, this approach will be attempted only if the first approach has failed.

The contingency plan will be considered if there is no working/useful Wiener attention model by July 1st, with a full switch to this methodology by July 15th. I anticipate that this change will be much easier to implement than the Wiener attention.

References

[BPC20] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020.

[CGRS19] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers, 2019.

[CSB$^+$23] Deborah Pelacani Cruz, George Strong, Oscar Bates, Carlos Cueto, Jiashun Yao, and Lluis Guasch. Convolve and conquer: Data comparison with wiener filter, 2023.

[Rus22] Sasha Rush. The annotated transformer, 2022. Accessed: 2024-06-11.

[VSP$^+$23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.