# Entanglement-based attention for transformers

Poojith U Rao[1,2], Florian Speelman[1,2], Rahaf Aljundi[3], and Sachin Kinge[3]

[1] Qusoft, Amsterdam
[2] University of Amsterdam, Amsterdam
[3] Toyota Motor Europe

Machine learning has revolutionized numerous domains by enabling machines to learn complex patterns and relationships from vast amounts of data. This capability stems from the utilization of neural networks, which are inspired by the structure of the human brain. Transformers have emerged as a particularly successful architecture, excelling at tasks such as Natural Language Processing (NLP) [1], computer vision tasks [2, 3], audio processing [4] and numerous other domains. They have become the backbone of many state-of-the-art NLP models like BERT [5], GPT [6], T5 [7], etc. Transformers depend on the attention mechanism to capture dependencies and relationships within the data and focus on "important" input segments that produce an output.

In the seemingly unrelated world of quantum mechanics, physicists use quantum mechanical wave functions to model complex relations between particles in a multi-particle quantum system. The repeated interactions between particles often create quantum correlations or entanglement between them. The wave function has become an indispensable tool for predicting the properties of quantum mechanical systems.

Similarities between a quantum-mechanical wave function modeling relationships between quantum particles and a deep neural network modeling the relationship between segments of a high-dimensional input are studied in [8, 9]. In particular, [8] studies the structural equivalence between a function modeled by a Convolutional Arithmetic Circuit (ConvAC) and a many-body quantum wave function using the underlying Tensor Network (TN) structure. They make an important observation that the ability of a ConvAC to represent correlations between input regions is strongly related to quantum entanglement. Similarly, the expressiveness of a CNN, or equivalently of a many-body wave function, is related to their ability to model the intricate correlation between the inputs [8]. Hence, it is understandable that deep learning models such as CNN and recurrent neural networks (RNN) can efficiently represent highly entangled quantum systems [9]. An Attention-based Quantum Transformer (AQT) was demonstrated to not only outperform other neural-network-based models for Quantum State Tomography (QST) but also accurately reconstruct the density matrix associated with a noisy quantum state experimentally realized on an IBMQ quantum computer [10]. The success of the AQT is speculated to come from its ability to model quantum entanglement across the entire quantum system, much as the attention model for NLP captures the correlations among words in a sentence [10]. Similarly, a quantum-aware transformer (QAT) proposed to capture the complex relationship between measured frequencies highlights the similarity between highly structured sentences in NLP and intrinsically structured measurements in QST.

These works surmise a relation between inputs modeled by a deep learning model and the quantum entanglement between particles in a quantum wave function. This leads to a natural question: *Can the attention mechanism in the transformer, which captures the relation between the input segments, take advantage of quantum entanglement?* To answer this, we attempt to incorporate quantum entanglement into the attention mechanism of a transformer.

## 1 Entanglement-based attention

This work only considers a transformer encoder composed of attention and fully connected layers. The attention mechanism operates similarly to an information retrieval system, where the output is *value vectors $V$* whose associated *key vector $K$* is similar to the *query vector $q$*. The similarity function commonly used is the dot product similarity. This is expressed as follows:

$$\text{Attention(q, K, V)} = \sum_i \text{Similarity}(q, K_i) \times V_i \tag{1}$$

We incorporate quantum entanglement into the attention mechanism by replacing the similarity function used to compute the attention coefficients with a measure of entanglement. The methodology followed is as follows:

1. **Quantum embedding:** The query and key vectors are encoded as quantum states using a Quantum Feature Map (QFM). We use a QFM adapted from the *ArbitraryStatePreparation* circuit in PennyLane [11].
2. **Entangle quantum states:** A Parameterized Quantum Circuit (PQC) is applied to entangle the quantum embeddings of query and key vectors. We use a modified version of the "circuit 14" as suggested in [12] due to its favorable expressibility and entangling capability, along with reasonable circuit costs considering the number of parameters and circuit depth.
3. **Measure entanglement:** Attention coefficient $\alpha_{i,j}$ is computed as a measure of entanglement between the query $|Q\rangle$ and key $|K\rangle$ states. The function to compute a measure of entanglement using the PQC is as follows:

$$\alpha_{i,j} = \text{Entanglement Measure}(\text{PQC}(|Q_i\rangle, |K_j\rangle)) \tag{2}$$

$$\text{Attention}(Q_i, K, V) = \sum_j \text{Softmax}(\frac{\alpha_{i,j}}{\sqrt{d_k}})V_{i,j} \tag{3}$$

Here, the dimension of the key vector $d_k$ is used to scale the attention coefficients. We consider the following functions between query and key embeddings.

(a) **Bipartite entanglement from FST:** The entanglement entropy between sub-systems gives a measure of entanglement between query and key state.

Table 1: Text datasets

| Model | MC Train | MC Test | RP Train | RP Test |
|---|---|---|---|---|
| Classical | 100 | 100 | 86.48 | 74.19 |
| **Bipartite entanglement** | 100 | **100** | 81.08 | **70.96** |
| Multipartite entanglement | 80.0 | 73.33 | 75.96 | 70.96 |
| Swap test | 82.85 | 73.33 | 79.72 | 67.74 |
| QSANN (altered) | 58.57 | 56.66 | 67.57 | 54.84 |
| QSANN (original) | 100.00 | 100.00 | 95.35 | 67.74 |

Table 2: MNIST dataset

| Model | MNIST Train | MNIST Test |
|---|---|---|
| Classical | 54.27 | 51.23 |
| **Bipartite entanglement** | 43.96 | **48.48** |
| Swap test | 42.19 | 46.00 |
| QSANN (altered) | 10 | 13 |
| QKSAN | 13 | 8 |

Table 3: A comparison of accuracy on test datasets revealed that the bi-partitie entanglement-based approach outperforms existing quantum attention models.

(b) **Swap test:** The controlled SWAP test is a prominent method to determine the similarity of two pure states $|\phi\rangle_{query}$ and $|\phi\rangle_{key}$. We use the swap test as a base method to compare the efficacy of entanglement measures.

(c) **Multipartite entanglement using swap test:** The swap test is modified to compute a measure of entanglement, concurrence $C_n$ [13].

## 2 Experiments and results

The proposed methodology was implemented using the PennyLane [11] and PyTorch [14] libraries. The hyperparameters of the hybrid quantum-classical architecture were selected to ensure that the performance depends significantly on the attention layer. It was compared with a classical attention layer, Quantum Self-Attention Neural Networks [15], and Quantum Kernel Self-Attention Networks [16]. The approaches were tested on text classification datasets MC and RP [15] and the MNIST dataset. We used an altered version of QSANN to restrict the power of classical components. Table 1 and Table 2 compare the performance of all the methods.

The architecture we use critically tests the attention layer, and we observe that the entanglement-based approach performs better than the SWAP test and Gaussian projected quantum self-attention. It outperforms existing quantum attention-based methods but still lags behind classical attention. *The results indicate that entanglement measures can better capture relationships between the words in a sentence.* We observed that the time to compute entanglement measures is comparably higher. We plan to improve the efficiency of computing entanglement measures using classical shadow techniques. We hope this work motivates other researchers to explore how entanglement can be better incorporated into the attention mechanism.

## References

1. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
2. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
3. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, pp. 213–229, Springer, 2020.
4. L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5884–5888, IEEE, 2018.
5. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
6. A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training," 2018.
7. C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.
8. Y. Levine, D. Yakira, N. Cohen, and A. Shashua, "Deep learning and quantum entanglement: Fundamental connections with implications to network design," *arXiv preprint arXiv:1704.01552*, 2017.
9. Y. Levine, O. Sharir, N. Cohen, and A. Shashua, "Quantum entanglement in deep learning architectures," *Physical review letters*, vol. 122, no. 6, p. 065301, 2019.
10. P. Cha, P. Ginsparg, F. Wu, J. Carrasquilla, P. L. McMahon, and E.-A. Kim, "Attention-based quantum tomography," *Machine Learning: Science and Technology*, vol. 3, no. 1, p. 01LT01, 2021.
11. V. Bergholm, J. Izaac, M. Schuld, C. Gogolin, S. Ahmed, V. Ajith, M. S. Alam, G. Alonso-Linaje, B. AkashNarayanan, A. Asadi, *et al.*, "Pennylane: Automatic differentiation of hybrid quantum-classical computations," *arXiv preprint arXiv:1811.04968*, 2018.
12. S. Sim, P. D. Johnson, and A. Aspuru-Guzik, "Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms," *Advanced Quantum Technologies*, vol. 2, no. 12, p. 1900070, 2019.
13. S. Foulds, V. Kendon, and T. Spiller, "The controlled swap test for determining quantum entanglement," *Quantum Science and Technology*, vol. 6, no. 3, p. 035002, 2021.
14. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.
15. G. Li, X. Zhao, and X. Wang, "Quantum self-attention neural networks for text classification," *arXiv preprint arXiv:2205.05625*, 2022.
16. R.-X. Zhao, J. Shi, and X. Li, "Qksan: A quantum kernel self-attention network," *arXiv preprint arXiv:2308.13422*, 2023.