Building Blocks for a Complex-Valued Transformer Architecture

Florian Eilers and Xiaoyi Jiang

florian.eilers@uni-muenster.de, xjiang@uni-muenster.de

Introduction

- Most neural network architectures are build for real-valued signals
- However: In many applications complex-valued signals occur
- Complex-valued building blocks have been investigated for CNNs and RNNs, but not yet for the transformer architecture [1], commonly used in signal processing

We **contribute**:

- Derivation of a complex-valued attention mechanism
- Introduction of a complex-valued layer normalization
- Arrangement of a full complex-valued transformer architecture with the prior building blocks

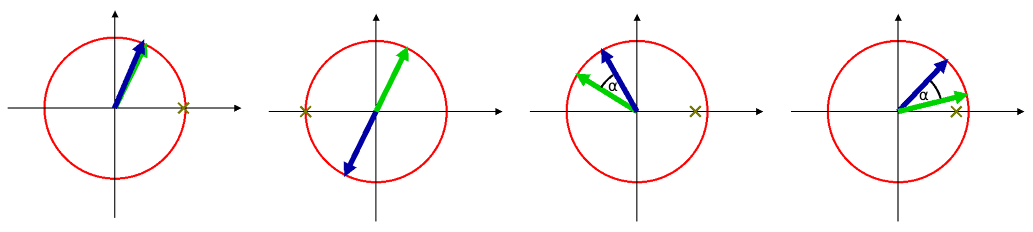
$\mathbb{C}$ Attention

Defining the softmax of a vector X of length n we can formulate the scaled dot-product attention:

$$softmax(X) = \sigma(X) = \frac{\exp(X)}{\sum_{i=1}^{n}\exp(X_i)},\quad Att(Q,K,V) = \sigma\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Properies of the attention mechanism

$Z = \langle Q, K\rangle_{\mathbb{R}^n}$



Let $Q, K \in \mathbb{R}^n$ and Z its Dot-Product. Then core properties of Z are:

- $Z > 0$, iff $Q \angle K < 90°$
- Z is rationally invariant
- Z scales with the length of Q, K
- Z is symmetric

$\mathbb{C}$ Attention:
To preserve aforementioned desired properties (proofs in paper), we define

$$\mathbb{C}Att(A,B) = \sigma\left(\frac{\mathcal{R}\langle Q,K\rangle_{\mathbb{C}^n}}{\sqrt{d_k}}\right)V \tag{2}$$

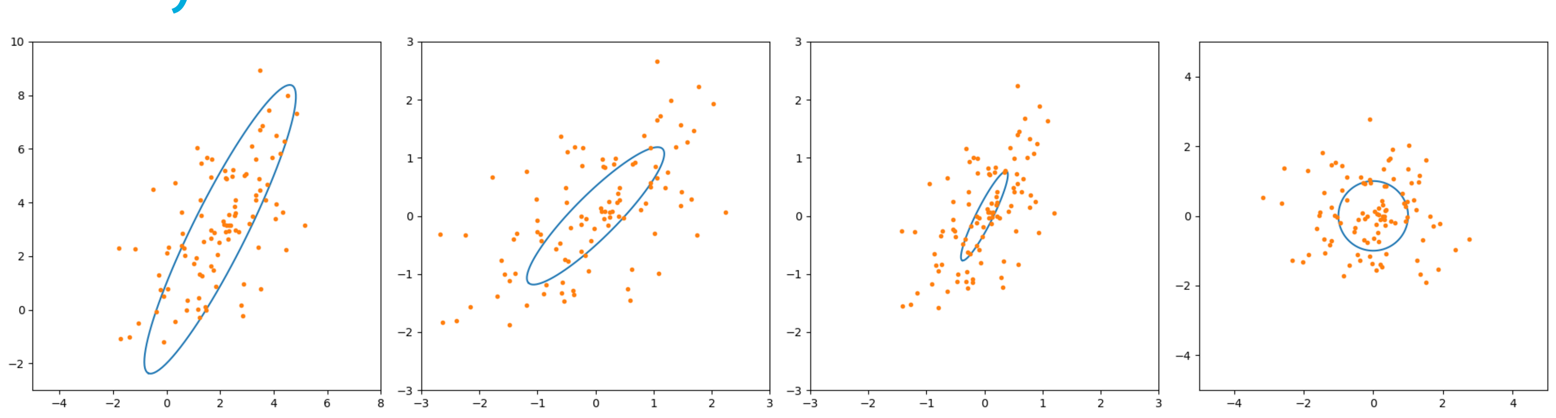
We also test these alternative formulations, even though not satisfying all desired properties:

$$AAtt(A,B) = \sigma\left(\frac{|\langle Q,K\rangle_{\mathbb{C}^n}|}{\sqrt{d_k}}\right)V,\ APAtt(A,B) = \sigma\left(\frac{|\langle Q,K\rangle_{\mathbb{C}^n}|}{\sqrt{d_k}}\right)\operatorname{sgn}(\langle Q,K\rangle)V$$

$$\mathcal{R}\mathcal{I}Att(A,B) = \left(\sigma\left(\frac{\mathcal{R}\langle Q,K\rangle_{\mathbb{C}^n}}{\sqrt{d_k}}\right) + i\,\sigma\left(\frac{\mathcal{I}\langle Q,K\rangle_{\mathbb{C}^n}}{\sqrt{d_k}}\right)\right)V$$

Additionally, we test QK^T instead of $\langle Q, K\rangle_{\mathbb{C}^n}$.

$\mathbb{C}$ Layer normalization

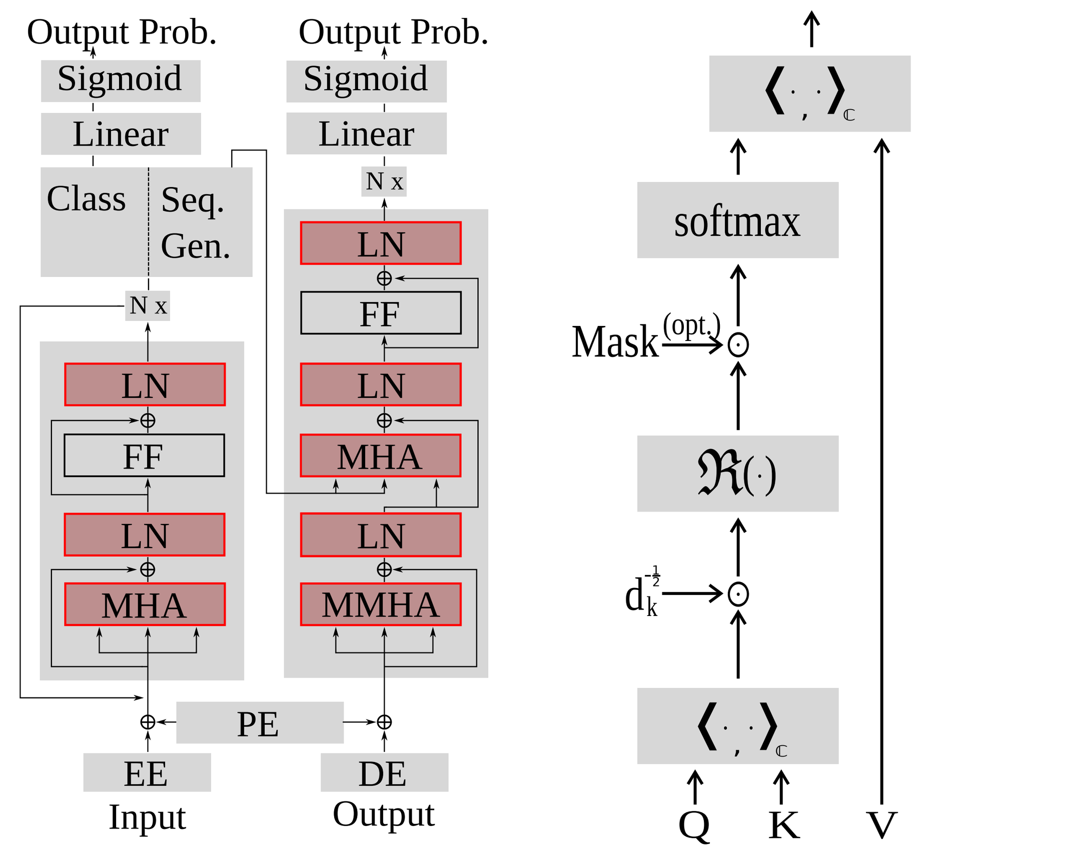


1. Input distribution to be normalized.
2. Separate normalization of $\mathcal{R}$ and $\mathcal{I}$ $\rightarrow$ rotated eliptical output distribution.
3. Normalization with $\mathbb{C}$ variance $\rightarrow$ eliptical output distribution.
4. (Proposed) Normalization with covariance matrix $\rightarrow$ circular output distribution, uncorrelated real and imaginary parts:

$$\begin{pmatrix}\mathcal{R}(\mathbb{C}LN(X))\\ \mathcal{I}(\mathbb{C}LN(X))\end{pmatrix} = \mathrm{Cov}_{\mathbb{C}}^{-\frac{1}{2}}(X)\begin{pmatrix}\mathcal{R}(X-\mathbb{E}(X))\\ \mathcal{I}(X-\mathbb{E}(X))\end{pmatrix} \tag{3}$$

Overview

Legend:
EE = Encoder Embedding
DE = Decoder Embedding
PE = Positional Encoding
(M)MHA = (Masked) Multi-Head Attention
LN = Layer normalization
FF = Feed Forward
N x = repeat N times



Left: The transformer architecture [1], in red: Building blocks derived in our paper
Right: $\mathbb{C}$ attention mechanism

Results

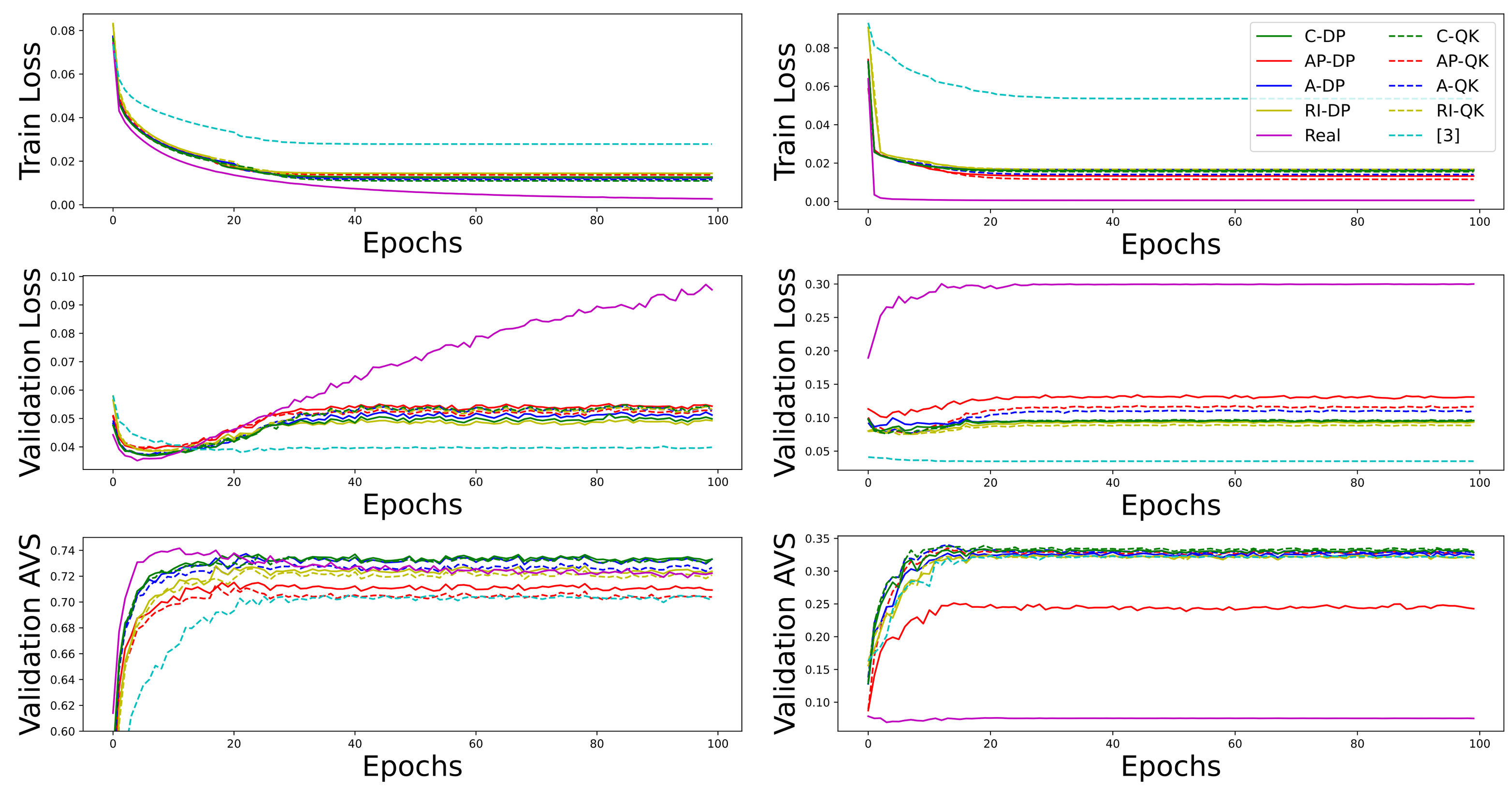
Music dataset [2], 330 pieces divided into 39438 samples, 64 timesteps each.

Classification:

- 128 classes
- Multiclass classification
- Encoder only

Sequence generation:

- Predict last 21 time steps from first 43 timesteps sequentially
- Full transformer architecture



Conclusion

Contributions:

- Derivation of a $\mathbb{C}$ attention mechanism using the $\mathbb{C}$ dot product
- Introduction of a $\mathbb{C}$ layer normalization producing uncorrelated outputs
- Testing the full complex-valued transformer architecture with those building blocks

Results:

- On-par results compared to the real-valued transformer on a real world music dataset
- Improved robustness to overfitting

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NIPS*, 2017.

[2] J. Thickstun, Z. Harchaoui, and S. M. Kakade. Learning features of music from scratch. In *ICLR*, 2017.

[3] M. Yang, M. Q. Ma, D. Li, Y.-H. H. Tsai, and R. Salakhutdinov. Complex transformer: A framework for modeling complex-valued sequence. In *ICASSP*, 2020.