Neural networks with no priviledged basis

Aurélien Garivier and Jean-Christophe Mourrat

This internship aims to invesigate, both theoretically and experimentally, the potential of original activation functions respecting the symmetries of the neural networks. One of the simplest possible "neural network architectures" is the function

$$f : x \mapsto V\sigma(Ux),$$

where $x \in \mathbb{R}^d$, $U \in \mathbb{R}^{\ell\times d}$, $V \in \mathbb{R}^{k\times\ell}$, and the mapping $\sigma : \mathbb{R} \to \mathbb{R}$ is applied coordinate by coordinate on Vx, so that for every $i \in \{1\dots,\ell\}$, we have

$$(\sigma(Ux))_i = \sigma\left(\sum_{j=1}^{d} U_{ij}x_j\right). \tag{1}$$

Suppose for a moment that the mapping σ is just the identity, and that $k = d$. The singular value decomposition of the matrix VU gives us two families of orthonormal vectors in $\mathbb{R}^d$, say $(u_1,\dots,u_d)$ and $(v_1,\dots,v_d)$, as well as a family of numbers $\lambda_1 \geqslant \cdots \geqslant \lambda_d \geqslant 0$, such that

$$f(x) = VUx = \sum_{i=1}^{d} \lambda_i(x\cdot u_i)v_i.$$

In situations in which f is learnt from data, one can expect that many of the directions $(u_1,\dots,u_d)$ and $(v_1,\dots,v_d)$ will be highly interpretable, each one encoding one particular "feature". On the other hand, the output of each "neuron" indexed by i in (1) does not have any special meaning in this case.

This simple example shows that in general, it does not really make sense to hope that each individual "neuron" will naturally be associated with a particular feature. When σ is a non-linear function such as the ReLU, applying the non-linearity coordinate by coordinate breaks the invariance by change of basis, and the interpretation of singular value directions may become compromised. On the other hand, attention layers are defined without reference to a priviledged basis, and the singular vectors of the weight matrices are indeed highly interpretable in this case [1].

The goal of the proposal is to look for ways to replace "perceptron" layers of the form $x \mapsto V\sigma(Ux)$ by alternative non-linear functions that would not require to choose a priviledged basis. One candidate is to use mappings of the form

$$x \mapsto (W_1x) \otimes (W_2x),$$

since the tensor product $\otimes$ is "basis-agnostic". The goal would be to determine, through theorical arguments and numerical experiments, how these tensor-product non-linearities compare with classical perceptron layers, in terms of their expressivity, trainability, and interpretability.

Another potential benefit of using tensor-product non-linearities is that they are polynomial functions of the data. As such, these non-linearities can be computed efficiently even when operating on data that, for privacy purposes, is only available through homomorphic encryption.

References

[1] Beren Millidge and Sid Black. The singular value decompositions of transformer weight matrices are highly interpretable. In *AI Alignment Forum*, 2022, https://www.alignmentforum.org/posts/mkbGjzxD8d8XqKHzA/the-singular-value-decompositions-of-transformer-weight.