Self-attention as an attractor network: pseudo-memories without backpropagation

A path from transformers to statistical mechanics

- A transformer is a **series of q feed-forward transformations blocks**, each consisting in a self-attention layer followed by a multilayer perceptron, applied to a sequence of N embedding vectors $\vec{x}_i \in \mathbb{R}^d$.
- Instead of different blocks, a simplification in the direction of energy models is **recycling**. It consists in the **iteration of a single block for q repetitions** (in the embedding space).
- A further simplification is to iterate q **times the bare self-attention**, without any multilayer perceptron.

A bare self-attention vectorial spins system

The bare self-attention update of vectorial spins $\vec{x}_i \in \mathbb{R}^d$, $i = \{1, ..., N\}$ is

$$\vec{x}_i^{t+1} = \gamma \sum_{j(\neq i)} \alpha_{i \leftarrow j} \, \mathbf{J} \, \vec{x}_j^t + \vec{x}_i^t$$

where are defined attention weights

$$\alpha_{i \leftarrow j} = \mathrm{softmax}_j[\lambda \sum_{\alpha,\beta} x_i^\alpha \, J_{ij}^{\alpha\beta} \, x_j^\beta]$$

The update can be written as the minimization of a negative Pseudo-Likelihood (NPL):

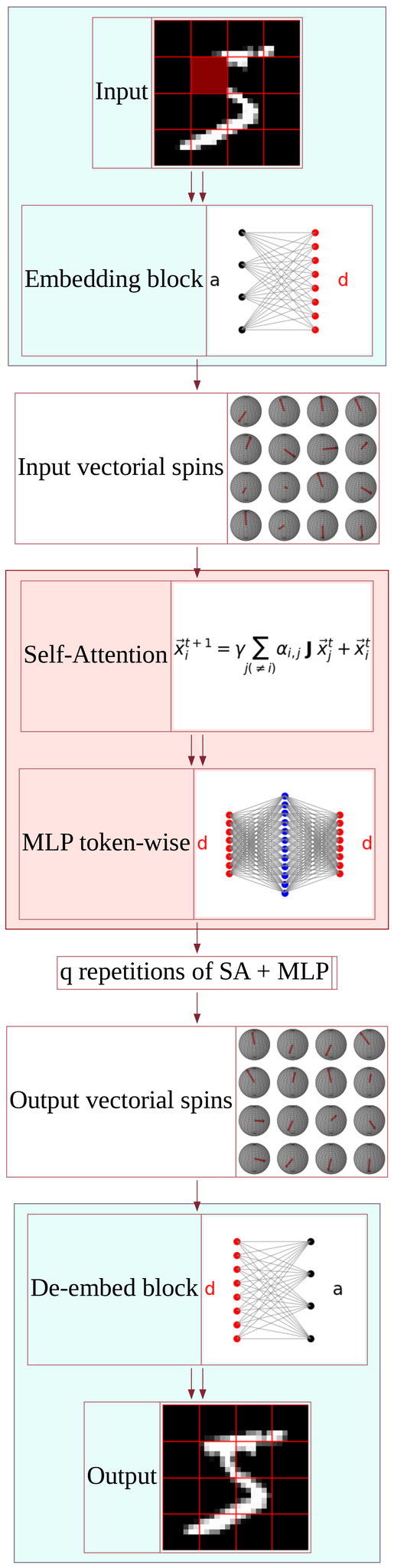
$$\mathrm{NPL}(x) = -\frac{1}{\lambda} \sum_i \log\Big[\sum_{j(\neq i)} \exp\big(\lambda \vec{x}_i \cdot \mathbf{J} \vec{x}_j\big) \Big] = \sum_i e_i(x) \qquad (1)$$

- Difference with energy: each spin $i = 1, ..., N$ is modified to minimize its individual cost e_i, not the global NPL.
- It is not a true pseudo-likelihood since it is not normalized.
- We **train bare self-attention directly minimizing the NPL**: given a dataset of sequences x_μ and i, training consists in finding $\mathbf{J}$ that minimizes $\sum_\mu e_i(x_\mu)$.

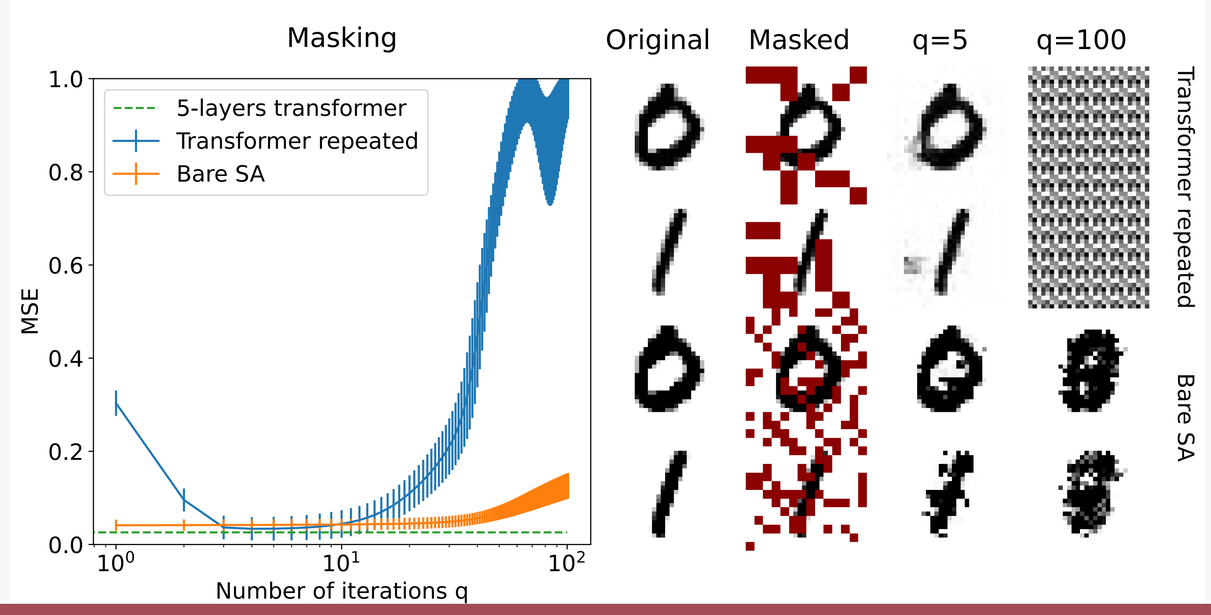
Take-home message

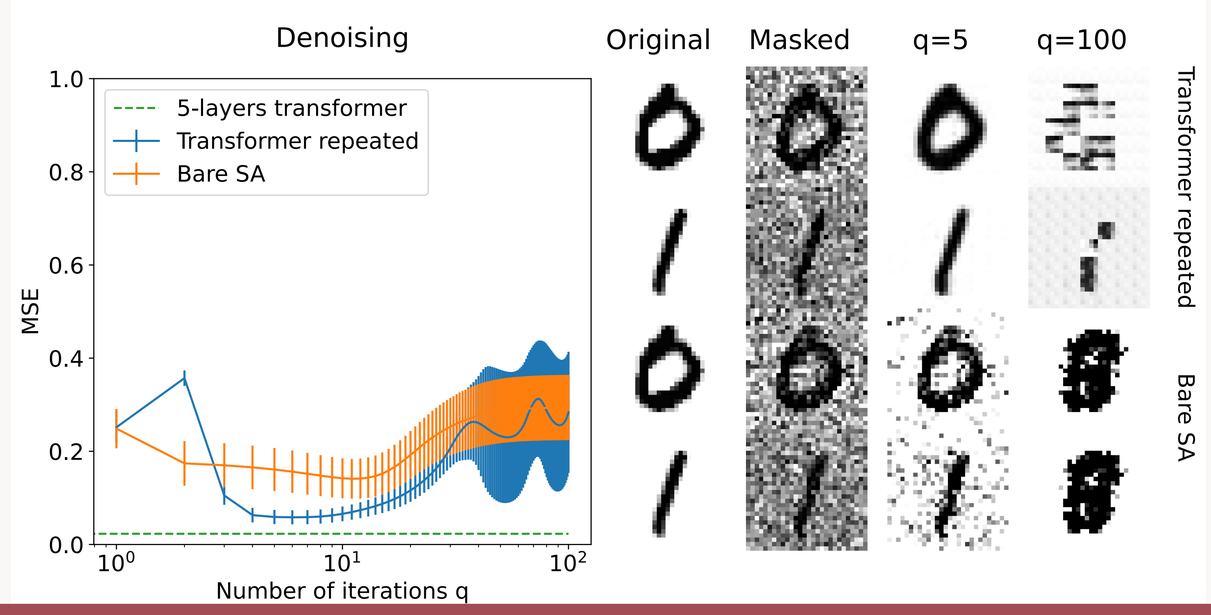
- Self-attention can be written as the optimization of a negative pseudo-likelihood (NPL).
- Optimizing the NPL given a dataset results in an attractive behaviour when bare self attention is iterated.
- A complete transformer, made of a single self-attention + multilayer perceptron block iterated q times (recycling), shows a qualitatively similar attractive behaviour.

The transformer model



Iterative transformers show an attractive behaviour





Francesco D'Amico[1], Matteo Negri[1,2]

francesco.damico@uniroma1.it

[1]Università di Roma Sapienza, [2] CNR-NANOTEC

Funded by the European Union NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

FAIR Future Artificial Intelligence Research

Sapienza Università di Roma