ID de Contribution: **45** Type: **Non spécifié**

Dynamic metastability in the self-attention model

vendredi 11 octobre 2024 11:10 (45 minutes)

The pure self-attention model is a simplification of the celebrated Transformer architecture, which neglects multi-layer perceptron layers and includes only a single inverse temperature parameter. Despite its apparent simplicity, the model exhibits a remarkably similar qualitative behavior across layers to that observed empirically in a pre-trained Transformer. Viewing layers as a time variable, the self-attention model can be interpreted as an interacting particle system on the unit sphere. We show that when the temperature is sufficiently high, all particles collapse into a single cluster exponentially fast. On the other hand, when the temperature falls below a certain threshold, we show that although the particles eventually collapse into a single cluster, the required time is at least exponentially long. This is a manifestation of dynamic metastability: particles remain trapped in a "slow manifold" consisting of several clusters for exponentially long periods of time. Our proofs make use of the fact that the self-attention model can be written as the gradient flow of a specific interaction energy functional previously found in combinatorics.

Orateur: M. GESHKOVSKI, Borjan