Preliminary Research Idea (8): Particle Transformer

Professor Richard Yi Da Xu
HKBU and TadReamk Limited

Prelude

First, an aside: In 2012, I visited Professor Arnoud Doucet's Statistical Department at Oxford University for half a year. During that time, I systematically learned about Particle Filter+ and its related theories, which provided great inspiration and help. I originally planned to return to Oxford to work with Professor Doucet for another half-year in 2015, but due to the university's request to lead a series of collaborative projects with the industry, the plan was cancelled. This led me down a path different from the initial "pure" academic plan.

Problem Statement and Proposed Solution

Deep Transformer models often face problems of "Attention Collapse" or "Oversmoothing." This means that as the number of layers increases, Token representations become similar and difficult to distinguish. This phenomenon is mathematically highly similar to Weight Degeneracy in Sequential Monte Carlo (SMC) filtering.

By viewing the Transformer's forward propagation process as a series of filtering steps, we propose introducing mature SMC techniques—specifically Resampling and Proper Weighting—to maintain the diversity of Token representations and focus computational resources on the "most important" Tokens (i.e., particles).

Formalizing the Analogy

The analogy between SMC and Transformer concepts can be formalized as follows:

1. **Time step** t in SMC corresponds to **Transformer layer** l.
2. **Particles** in SMC correspond to **Token representation vectors** (or attention heads) in Transformers.
3. **Importance weights** in SMC correspond to **Attention scores** (Softmax output) in Transformers.

Key Components of the Particle Transformer

Resampling Layer

A "Resampling Layer" is introduced after every K Transformer modules. This can be implemented via Gumbel-Softmax.

Unlike standard attention mechanisms that pass the weighted average of all Values, this layer samples indices based on attention weights. It duplicates highly attended Tokens and eliminates (kills) lowly attended Tokens. This acts as a "hard" attention mechanism that can effectively remove noise, prevent irrelevant contextual information from mixing in, and dilute primary signals.

MCMC Moves

In SMC: After resampling, particles are usually "jittered" (MCMC Move) to prevent complete overlap.

In Transformers: This step corresponds to the Feed-Forward Network (FFN).

Theoretical View: The FFN can be theorized as a Transition Kernel, whose purpose is to increase Token diversity after resampling, preventing representation uniformity.