Parameter estimation of long memory stochastic processes with deep neural networks

Csanády Bálint, Boros Dániel, Ivkovic Iván, Nagy Lóránt, Kovács Dávid, Tóth-Lakits Dalma, Márkus László, Lukács András

Eötvös Loránd Tudományegyetem, Matematikai Intézet,
Mesterséges Intelligencia Kutatócsoport

We present a purely deep neural network-based approach for estimating long memory parameters of time series models that incorporate the phenomenon of long-range dependence. Parameters, such as the Hurst exponent, are critical in characterizing the long-range dependence, roughness, and self-similarity of stochastic processes. The accurate and fast estimation of these parameters holds significant importance across various scientific disciplines, including finance, physics, and engineering. We harnessed efficient process generators to provide high-quality synthetic training data, enabling the training of scale-invariant 1D Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) models. Our neural models outperform conventional statistical methods, even those augmented with neural network extensions. The precision, speed, consistency, and robustness of our estimators are demonstrated through experiments involving fractional Brownian motion (fBm), the Autoregressive Fractionally Integrated Moving Average (ARFIMA) process, and the fractional Ornstein-Uhlenbeck process (fOU).

LlamBERT: Large-scale low-cost data annotation in NLP

Csanády Bálint, Muzsai Lajos, Vedres Péter, Nádasdy Zoltán, Lukács András

Eötvös Loránd Tudományegyetem, Matematikai Intézet,
Mesterséges Intelligencia Kutatócsoport

Large Language Models (LLMs), such as GPT-4 and Llama 2, show remarkable proficiency in a wide range of natural language processing (NLP) tasks. Despite their effectiveness, the high costs associated with their use pose a challenge. We present LlamBERT, a hybrid approach that leverages LLMs to annotate a small subset of large, unlabeled databases and uses the results for fine-tuning transformer encoders like BERT and RoBERTa. This strategy is evaluated on two diverse datasets: the IMDb review dataset and the UMLS Meta-Thesaurus. Our results indicate that the LlamBERT approach slightly compromises on accuracy while offering much greater cost-effectiveness.