ID de Contribution: **2** Type: **Non spécifié**

Synthetic Data –Friend or Foe in the Age of Scaling?

jeudi 23 mai 2024 10:30 (1 heure)

As AI and LLM model size grows, neural **scaling laws** have become a crucial tool to predict the improvements of large models when increasing capacity and the size of original (human or natural) training data. Yet, the widespread use of popular models means that the ecosystem of online data and text will co-evolve to progressively contain increased amounts of synthesized data.
In this talk we ask: **How will the scaling laws change in the inevitable regime where synthetic data makes its way into the training corpus?** Will future models, still improve, or be doomed to degenerate up to total **(model) collapse**? We develop a theoretical framework of model collapse through the lens of scaling laws. We discover a wide range of decay phenomena, analyzing loss of scaling, shifted scaling with number of generations, the ‘’un-learning” of skills, and grokking when mixing human and synthesized data. Our theory is validated by large-scale experiments with a transformer on an arithmetic task and text generation using the LLM Llama2.

Orateur: KEMPE, Julia (NYU Center for Data Science and Courant Institute of Mathematical Sciences)