Contribution ID: **16** Type: **Talk**

Dynamical low-rank training of neural networks

Neural networks have achieved tremendous success in a wide variety of applications. However, their memory footprint and computational demand can render them impractical in application settings with limited hardware or energy resources. At the same time, overparameterization seems to be necessary in order to overcome the highly non-convex nature of the training optimization problem. An optimal trade-off is then to be found in order to reduce networks' dimension while mantaining high performance.
Popular approaches in the current literature are based on pruning techniques that look for subnets capable of maintaining approximately the initial performance. Nevertheless, these techniques often are not able to reduce the memory footprint of the training phase.
In this talk we will present DLRT, a training algorithm that looks for "low-rank lottery tickets" by interpreting the training phase as a continuous ODE and by integrating it within the manifold of low-rank matrices.
These subnetworks and their ranks are determined and adapted already during the training phase, allowing the overall time and memory resources required by both training and evaluation phases to be reduced significantly.
The talk is based on [1].

[1] S. Schotthöfer, E. Zangrando, J. Kusch, G. Ceruti, F. Tudisco,
"Low-rank lottery tickets: finding efficient low-rank neural networks via matrix differential equations", NeurIPS, 2022.

Primary authors: ZANGRANDO, Emanuele (Gran Sasso Science Institute); TUDISCO, Francesco (GSSI)

Presenter: ZANGRANDO, Emanuele (Gran Sasso Science Institute)