Topology-based Data Reduction for Green Deep Learning

Javier Perera-Lago

Abstract

The training of Deep Learning models requires to collect massive amounts of real-world labelled data. The use of these large amounts of data may bring some problems, such as prolonged training periods, significant energy consumption and substantial carbon emission. To reduce this high consumption of resources, many researchers are looking for ways to reduce the size of these datasets, by simplifying the feature space or by removing redundant or noisy items. In this talk we will introduce the Dominant Datasets algorithm, which preserves topological concepts such as persistent homology, to compress the training datasets while preserving their intrinsic properties, in such a way that the model training requires less computations but achieves similar results.