Statistical understanding and visualisation of AI embeddings

Supervisor: Nick Whiteley (nick.whiteley@bristol.ac.uk)

Embedding is a key functionality in modern AI technologies which involves mapping tokenised text, images, audio and other real-world data to vectors in Euclidean space, often using some form of trained neural network. The embedding vectors typically have several hundred or thousand dimensions and proximity in this high-dimensional space encodes semantic similarity between the input objects. This is used, for example, by systems such as Chat-GPT to put together informative sentences, by Stable Diffusion to generate realistic images, and by Spotify to make relevant music recommendations.

Although they have proved to be extremely effective in practice, understanding of embeddings is still incomplete in many ways. There are many open questions about what statistical properties of the underlying data and the method used to obtain the embedding influence its ability to convey useful information; how sensitive it is to anomalies and variability in data, and how it should be updated when new training data arrive. The first objective of this project is to investigate such questions from a rigorous statistical point of view. Given the seemingly powerful capacity of embeddings to capture detailed information in data, it is also an open question as to whether they can be used to help humans discover novel and unexpected patterns in other types of data, such as genetic disease states and molecular interactions in pharmacology. The second objective of this project is to investigate embedding visualisation methods to enable such discovery. This is a challenging task because the useful information in embeddings is encoded in a much higher dimensional space than the three-dimensional world we live in, so that principled and carefully designed dimension reduction techniques are needed.

This project will expose you to cutting-edge, rigorous, statistical, mathematical and computational analysis of machine learning methods, especially neural networks other AI technologies. The balance of theoretical and practical aspects of the research can be tuned to suit your strengths, but a solid background in mathematics and statistics is required along with experience and enthusiasm for coding, ideally in Python.

Keywords: autoencoders, dimension reduction, embeddings, high-dimensional statistics, neural networks, self-supervised learning.