CHRISTOS THRAMPOULIDIS, University of British Columbia

Implicit Geometry of Next-token Prediction: From Language Sparsity Patterns to Model Representations

How do language models map linguistic patterns to their representations? Specifically, can the geometry of context and word embeddings be characterized by the structure of the training data? We demonstrate that, in a large-model regime trained sufficiently long, context and word embeddings emerge from matrix factorization of a logit matrix that decomposes into sparse and low-rank components. As training progresses, the low-rank component becomes dominant and can be computed solely from the sparsity pattern of the training data, determined by unique word-context pairings across the dataset.