Your Paper

You

June 15, 2025

1 Introduction

The article All Bark and No Bite: Rogue Dimensions in Transformer Language Models Obscure Representational Quality discusses so-called rogue dimensions that interfere with standard measures of representational similarity, such as cosine similarity and Euclidean distance. Words can be represented as vectors in a continuous space, which allows language to be analyzed geometrically. Static embeddings (Word2Vec, GloVe) reflect semantic similarity well, but in recent years they have been replaced by contextual embeddings from transformers (BERT, GPT-2, etc.), which have achieved high results in NLP. However, there is still little understanding of how these models represent the meaning of words. Research shows: The cosine similarity in BERT and other models is sensitive to the position of words. The later layers of transformers reflect semantics less well — their embeddings hardly correlate with human evaluations. Contextual embeddings are anisotropic: all words become similar to each other in the last layers. Self-similarity decreases: the same word in different contexts is almost unrecognizable by the model. The authors note that the blame for all this lies with the 1-5 dominant dimensions. These measurements are concentrated far from the origin and have a disproportionately high dispersion. Anomalous measurements can cause the cosine similarity and Euclidean distance to depend on less than 1% of the embedded space. Although rogue dimensions strongly influence metrics such as cosine similarity and Euclidean distance, they do not affect the behavior of the models themselves when performing real tasks.The authors of the article considered several ways to combat such measurements: standardization, removal of principal components (all-but-the-top), and rank correlation (Spearman). Standardization showed the best correspondence with human judgments about word similarity.