⚡️🚀המאמר היומי של מייק 11.06.24:⚡️🚀

The Geometry of Categorical and Hierarchical Concepts in Large Language Models

המאמר חוקר כיצד קונספטים ומושגים מקודדים במרחבי הייצוג (embeddings)של מודלים של שפה גדולה. הכותבים חוקרים 2 שאלות מרכזיות: הייצוג של קונספטים קטגוריים והקידוד של יחסים היררכיים בין קונספטים.

הם מרחיבים את ההסתכלות הלינארית הרגילה על הקונספטים כדי להראות שהקונספטים קטגוריים מיוצגים כסימפלקסים, קונספטים היררכיים הם אורתוגונליים, וקונספטים מורכבים מיוצגים כפוליטופים שנבנים מסכומים ישירים של סימפלקסים.

המחקר בוחן 957 קונספטים היררכיים עם נתונים מ- WordNet באמצעות מודל ג'מה. הכותבים מראים שקונספטים סמנטיים high-level יכולים להיות מנוטרים ומנוהלים על ידי מדידה ועריכה ישירה של הייצוגים הווקטוריים הפנימיים של ה-LLMs. התוצאות התיאורטיות מגלות מבנה פשוט שבו קונספטים קטגוריים מיוצגים גיאומטרית כסימפלקסים ומושגים היררכיים מקודדים כאורתוגונליות.

https://arxiv.org/pdf/2406.01506