⚡️🚀המאמר היומי של מייק 05.08.24: ⚡️🚀

Improving Text Embeddings for Smaller Language Models Using Contrastive Fine-tuning

חוזרים לסקור מאמרים קלילים על מודלי שפה והיום בפוקוס מודלי שפה קטנים. המאמר שנסקור קצרות היום מציע שיטה לשיפור ייצוג של טקסט המופק על ידי מודל שפה קטן. ידוע שמודל שפה קטן (במאמר שיפרו את הייצוגים של הדקודרים) לא תמיד מצטיין ביצירה של ייצוג (אמבדינג) עוצמתי של טקסט - פשוט בגלל הגודל ו-expressiveness נמוכה יחסית.

אז המאמר מציע להשתמש בשיטת למידה ניגודית (contrastive learning) כדי לשפר את הביצועים. בגדול למידה ניגודית מאמנת מודל (לייצוג דאטה) במטרה לקרב פיסות דאטה (למשל תמונות או טקסט) שהן קרובות (סמנטית או בעלות אותה משמעות) ובאותו הזמן להרחיק את הייצוגים של פיסות דאטה לא דומות. השיטה הוצגה ב- 2018 על ידי Oord האגדי.

המאמר מציע להשתמש בלמידה ניגודית כדי לעשות פיין טיון לייצוגי הדאטה המופקים על ידי מודל שפה בפרט הפלט של השכבה האחרונה עבור טוקן EoS המסמן את סוף המשפט. עדכון משקלי המודל נעשה כמובן עם LoRA על דאטהסט המכיל משפטים בעלי משמעות קרובה וגם זוגות משפטים רחוקים סמנטית. המחברים טוענים שזה משפר את איכות הייצוג המופק על ידי המודל למספר משימות downstream (בפרט סיווג).

מאמר קלילי ונעים לקריאה….

https://arxiv.org/abs/2408.00690