# Review 198: Improving Text Embeddings with Large Language Models

**Paper: https://arxiv.org/abs/2402.15449v2**

https://arxiv.org/abs/2401.00368

הסקירה הזו תהיה לא סטנדרטית ואתחיל אותה משאלה: למה בחרתי לסקור את המאמר הזה?

לא בגלל שמדובר במודלי שפה - הרי כל יום יוצאים עשרות מאמרים על LLMs. גם לא בגלל שהמאמר הזה מציע שיטה לשיפור ביצועים של פתרונות (RAG (Retrieval Augmented Generation. הסיבה האמיתית היא שימוש בטכניקה שאני מאוד אוהב הנקראת (Contrastive Learning (CI או למידה ניגודית בשפת הקודש.

בזמן האחרון אני לצערי לא רואה יותר מדי עבודות שמשתמשות בפרדיגמה היפה הזו שתיארתי אותה בהרחבה בלא מעט סקירות בשנים האחרונות. בד״כ משתמשים בגישות השונות של הלמידה הניגודית כדי להפיק ייצוגי דאטה עוצמתיים במרחב בעל מימד נמוך או embeddings. ייצוג דאטה עוצמתי - הוא ייצוג שמצליח ״לשמור״ על פיצ'רים האינהרנטיים של פיסת דאטה, כלומר כזה שדוחס את הדאטה בצורה יעילה.

איך השיטה הזו עובדת? העיקרון הוא די פשוט - לכל פיסת דאטה יוצרים פיסת דאטה קרובה (נגיד סמנטית בשפה הטבעית או שתי פאצ'ים של אותה תמונה בראייה הממוחשבת). לאחר מכן יוצרים מספר של זוגות של פיסות דאטה שהן לא קשורות אחת לשניה (נגיד פאצ'ים מתמונות שונות). בגדול מאוד המטרה של CI היא לבנות(= לאמן מודל המפיק אותו) ייצוג הדאטה הממזער מרחק בין פיסות דאטה דומות(זוגות חיוביים) וממקסם אותו בין פיסות הדאטה הלא דומות (זוגות שליליים). לפעמים משתמשים במרחק קוסיין, לפעמים מרחק אוקלידי ויש עוד וריאנטים; יש מגוון שיטות לבנות זוגות חיוביים ושליליים אבל העיקרון נותר על כנוץ

אחד המאמרים הראשונים שהשתמשו בלמידה הניגודית בהקשר של בניית ייצוג דאטה הוא InfoNCE של Oord ושותפיו. לאחר מכן השיטה שימשה מחברים של מאמרים מאוד מפורסמים כמו SimSCE ו-MoCo. עכשיו אתם בטח רוצים לשאול איך CI קשור למאמר שנסקורו היום? כמובן כדי לשפר את האמבדינגס של הטקסט עבור משימות RAG.

ל-RAG לרוב יש שתי בעיות משמעותיות:

לא תמיד טקסט בעל ייצוג קרוב (בד״כ לפי מרחק קוסיין) לייצוג השאלה מכיל תשובה על השאלה (או משהו שניתן לגזור ממנו תשובה). ייתכן מאוד שהטקסט שייך לאותו התחום(דומיין) אבל לא מכיל תשובה על השאלה.

המרחב של ייצוגי הטקסטים (המופקים על ידי LLM-ים עוצמתיי) הוא מרחב לא טריוויאלי מבחינת המרחק בין הייצוגים ולפעמים מרחקים בין זוגות טקסט קשורים ולא מאוד קשורים עלולים להיות קרובים זה לזה. ז״א כאשר אתם בוחרים פיסות טקסט עם ייצוג הקרובים ביותר לייצוג השאלה לא תמיד מקבלים פיסות טקסט רלוונטיות.

אז הפתרון הגיוני ביותר הוא לכייל (fine-tune) מודל שפה על דאטהסט שאלות ותשובות מדומיין (משימה) כך שיקרב את ייצוגים של שאלות ותשובות רלוונטיות וירחיק את הייצוגים של השאלות והתשובות הלא רלוונטיות. הבעיה הגדולה עם הגישה הזו היא שבנייה ידנית של דאטהסט כזה היא יקרה ולוקחת הרבה זמן.

אז למה לא לרתום LLMs למשימה הזו? זה בדיוק מה שעשו המחברים של המאמר המסוקר. הם ביקשו LLM מצוי ליצור זוגות של שאלות עם תשובות נכונות ובנוסף גם זוגות עם תשובות לא נכונות, אבל נראות ״דומה לנכונות״ (hard negatives). היתה שם הנדסת פרומפטים חמודה אבל לא משהו מהפכני במיוחד.

ולבסוף השתמשו בגישת CI סטנדרטית כדי לכייל מודל שפה בדיוק באופן הסברתי לפניכן. לא מצאתי במאמר איך בדיוק מתבצע האימון (הוסיפו שכבות או אימנו כמה מהן) אבל הרעיון ברור.

וזה וזה…