המאמר היומי של מייק: 28.05.25  
Jasper and Stella: distillation of SOTA embedding models

מאמר די מעניין שמציע שיטה די פשוטה אך עובדת (כנראה) לזיקוק(distillation) ידע מכמה מודלים(מורים) מולטימודליים גדולים למודל אחד קטן (סטודנט). כמובן שהרציונל כאן טמון בכך שהמודל הקטן יצליח ללמוד(בתקווה) את העושר הייצוגי משני מודלים גדולים מצד אחד ויהיה קטן מצד שני שזה גם מבורך כי מקל על שימושו עם ראגים. הרי עבור מודלים בעלי מימד הייצוג קטן יותר צריך פחות פעולות אריתמטיות לחישוב דמיון בין ייצוג דאטה נתון לבין הייצוגים השמורים בראג.

זיקוק ידע מתבצע ב 3 שלבים עיקריים. בשלב הראשון המחברים מנסים לקרב את הייצוגים המופקים על ידי שרשור (ונרמול) של כמה מודלי מורה (הגדולים וחזקים) למודל קטן אחד. בשלב הזה המימד של וקטור הייצוג המופק על ידי המודל הקטן שווה שסכום של אלו המופקים על ידי המודלים הגדולים. פונקציית לוס מורכבת מ- 3 חלקים.

הראשון מנסה לקרב את המכפלה הפנימית של ייצוגי המורים המשורשרים וייצוג הסטודנט לאחד (כלומר למצב שהם שווים). החלק השני מנסה לקרב את הקורלציות בין הייצוגים של פיסות דאטה השונות על ידי המודלים - זה נעשה ברמת הבאצ'ים על ידי מזעור מרחק ריבועי בין ״מטריצת קווריאנס לא ממורכזת״ של המודלים הגדולים לבין המודל הקטן. גם ״מטריצת קווריאנס״ הוא המכפלה של מטריצה המכילה ייצוג של הפיסות דאטה בבאץ' בשחלוף שלה. הלוס האחרון הוא סוג של לוס ניגודי (contrastive loss) המנסה להשרות קרבה בין ייצוגים קרובים (לפי מודל המורה) עבור הייצוגים של מודל הקטן ובאותו הזמן להרחיק ייצוגיים של פיסות דאטה לא דומות במרחב האמבדינג שלו (של המודל המזוקק).

בשלב השני מקטינים את מימד האמבדינג של המודל הקטן (הוא היה שווה לסכום המימדים של המודלים הגדולים) תוך שימור של תכונותיו. איך עושים זאת? מוסיפים 3 שכבות למודל הסטודנט מהשלב האחרון ומאמנים רק אותם עם שני הלוסים האחרונים מהשלב הקודם.

בשלב השלישי מאמנים את האנקודר הויזואלי (של תמונות) של מודל הסטודנט מולטימודלי כאשר כל החלקים האחרים מוקפאים. כאן מנסים לקרב את הייצוג המופק על ידי האנקודר הויזואלי עבור תמונה לזה של כותרת התמונה המופק על ידי המודלים המורים. כאן משתמשים ב 3 הלוסים מהשלב הראשון.

זהו זה - מאמר קליל וקל להבנה…

https://arxiv.org/abs/2412.19048