⚡️🚀המאמר היומי של מייק -26.10.24: ⚡️🚀  
Efficient Vision-Language Pre-training by Cluster Masking

היום סוקרים מאמר נחמד בנושא של למידה ניגודית (contrastive learning) אבל הפעם עבור מקרה מולטימודלי. כלומר הפעם מאמנים מודל בסגנון של CLIP הידוע המיועד לבניית ייצוג דאטה ויזואלי (תמונות) והשפה. הרעיון העיקרי בלמידה הניגודית הוא לאמן מודל הממפה קרוב (במרחב האמבדינג) פיסות דאטה דומות ורחוק (באותו המרחב) פיסות דאטה לא דומות.

אבל הפעם מדובר בדאטה מולטימודלי. ב-CLIP המקורי אימנו את המודל לקרב ייצוג של אוגמנטציות שונות של תמונה עם הכותרת שלה ולהרחיק אותן (האמבדינגס של האוגמנטציות השונות של התמונה) מהייצוגים של כותרות שנבחרות באקראי. דבר דומה נעשה דומה לייצוג כותרת של תמונה: מקרבים לאמבדינג של אותה התמונה (עם אוגמנטציות) ולהרחיקו מהייצוגים של השאר.

נציין ש-CLIP המקורי אימן שני מודלי ייצוג שונים(למיטב זכרוני) לתמונות ולשפה אבל יצאו גם שדרוגים שאימנו שני מודלי ייצוג עם הרבה משקלים משותפים (אותה הארכיטקטורה).

הכל טוב ויפה אבל נשאלת השאלה האם ניתן לשפר כאן משהו? מתברר שכן ומהמאמר מציע שכלול קליל ל CLIP.ל-CLIP. כתבתי שאחד הדברים החשוב ב-CL הינה בחירה של הזוגות של פיסות דאטה לא דומות(זוגות שליליות). ככל שיהיה יותר מגוון בזוגות השליליות הייצוג שייבנה יהיה חזק יותר (כי ראה יותר דברים לא דומים ואז יבין יותר טוב איך ״צריך להיראות ייצוג טוב״.

אז המחברים מציעים לקלסטר פאצ'ים בתמונה לקלסטרים וכל פעם לא לבחור את הזוגות החיוביים בצורה אקראית אלא לאפשר בחירה של פאץ' אחד מתוך כל קלסטר. כלומר, לכל באץ' בוחרים רק פאץ' אחד מהקלסטר. כלומר פאצ'ים דומים מדי לא נכנסים לזוגות השליליים ב-CL. הקליסטור יכול להתבצע על הערכים של הפיקסלים בשילוב עם מודל אמבדינג כלשוה.

מאמר פשוט - לקח לי איזה דקה להבין ו 10 דקות לכתוב סקירה. אוהב כאלו…

https://arxiv.org/pdf/2405.08815