🚀המאמר היומי של מייק -06.10.24: ⚡️🚀  
CONTRASTIVE LOCALIZED LANGUAGE-IMAGE PRE-TRAINING

ממשיכים הפסקה בסקירות על מודלי שפה ועוברים לסקירות על מודלים מולטימודליים (שפה ותמונות). טוב, הפסקה למחצה. אתם בטח זוכרים את המודל שנקרא CLIP שעשה הרבה רעש לפני כמה שנים.

CLIP הוא אחד המודלים מולטימודליים הראשוניים שהצליח לייצר אמבדינגס חזקים ומיושרים (aligned) של טקסט ושל תמונות. מיושרים הכוונה של הייצוגים של תמונה וטקסט שמתאר את תוכנה קרובים אחד לשני בזמן שהייצוגים של תמונה וטקסט לא מתאימים רחוקים אחד מהשני (במקרה הזה ביחס למרחק קוסיין ביניהם).

המודל הזה אומן על דאטהסט ענק של תמונות והכותרות שלהם (או טאגים) מהאינטרנט כאשר אימנו אותו תוך שימוש בטכניקה למידה ניגודית (contrastive learning או CL). בגדול מאוד טכניקות CL מאומנות להפיק ייצוג סמנטי מדאטה (מסוגים שונים) כאשר המטרה היא לקרב את הייצוגים (אמבדינגס) של פיסות דאטה קרובות (או חיוביות) ולהרחיק ייצוגים של פיסות דאטה לא דומות (שליליות). במקרה של CLIP פיסות דאטה חיוביות הם הייצוגים של תמונה והכותרת שלה ואילו הזוגות השליליים בנויים מכותבות ותמונות שנבחרו באקראי.

המאמר שנסקור אחד כאמור משכלל את CLIP על ידי הקניה של יכולות לוקליזציה לייצוג. הכוונה כאן שהמחברים מאמנים ייצוגים של תמונה ושל טקסט באופן כזה שבהינתן ייצוג התמונה I וייצוג התיאור של פאץ' ב I המכיל אובייקט מסוים יהיה ניתן להפיק ב״קלות״ את מיקום האובייקט בתמונה.

במילים פשוטות נניח שיש לנו אריה עומד ושואג בתמונה הנמצא ב-bounding box (המוגדר על ידי רביעיה של קואורדינטות שלו בתמונה) המסומן ב- B. המחברים מאמנים רשת אנקודר לתמונות f\_I רשת אנקודר לטקסט f\_T כך שייצוג התמונה R\_I ייצוג ״אריה עומד ושואג״ R\_T, המופקים על ידי שני האנקודר האלו (בהתאמה) כך שרשת רדודה יחסית (נקראת prompter במאמר), המקבלת אותם, תוכל לחזות את מיקום האריה B בתמונה. דרך אגב המיקום כאן לא חייב להיות מתואר על ידי bounding box אלא יכול להיות מוגדר (בערך) על ידי כמה ניקודת, תיאור כללי (נגיד חיה, בלי להזכיר שזה אריה) ובעוד צורות.

האימון נעשה כמו בלמידה הניגודית כמו ב-CLIP המקורי. אבל בנוסף ללוס הרגיל שלו יש כאן עוד לוס ניגודי המקרב את ייצוגים של כותרת הפאץ' בתמונה לייצוג המופק על Prompter מייצוג התמונה ומהמתאר של הפאץ' (נגיד BB) ומרחיק את הייצוגים האלו לפאצ'ים שונים. כמובן שה-Prompter גם מאומן תוך כדי,

המאמר משתמש במודלים מאומנים למטרת זיהוי אובייקטים בתמונה (OWLv2) ובמודלים מאומנים אחרים (VeCap) למתן כותרות לפאצ'ים האלו.

מאמר די חמוד וקליל…

https://arxiv.org/pdf/2410.02746