המאמר היומי של מייק - 22.02.25  
When Does Perceptual Alignment Benefit Vision Representations?

ממשיכים בהפסקה קלה ממודלי שפה וסוקרים קצרות מאמר בתחום הראיה הממוחשבת. המחברים מציעים שיטה לשיפור ייצוגים (אמבדינגד) של דאטה ויזואלי קרי תמונות באמצעות טיוב (fine-tune) של האמבדר או באקבון (המודל המפיק ייצוגים אלו). טיוב זה נעשה על הדאטהסט של תמונות המסומנות כדומות (ולא דומות) על ידי בני אדם. טיוב זה הוכח (אמפירית) כתורם לאיכות הייצוגים המופק למגוון משימות בתחום הבנת דאטה ויזואלי כגון ספירת עצמים בתמונה, סגמנטציה, שערוך עומק, אחזור עצמים (instance retrieval).

הטיוב נעשה עם שיטת למידה ניגודות (contrastive learning) שכתבתי עליה לא מעט וסקרתי עשרות מאמרים בנושא המרתק הזה. בגדול מאוד המטרה העיקרית של למידה ניגודית היא לקרב ייצוגים של פיסות דאטה דומים (הקרבה נמדדת בד״כ על ידי מרחק קוסיין אבל יש עוד אופציות) ולהרחיק ייצוגים של פיסות דאטה לא דומות. המחברים השתמשו בדאטהסט NIGHTS המכיל שלישיות של תמונות בעלות בערך אותו תוכן סמנטי; עם זאת, התמונת בדאטהסט שונות בתנוחה, צורה, צבע ומספר של חפצים.

המחברים משתמשים בגישת triple loss ללמידה הניגודית והמודל מאומן על שלישיות של תמונות כאשר כל שלישיה מורכבת מתמונת עוגן (reference) ושתי תמונות נוספות שסומנו על ידי מתייגים אנושיים בנוגע לדמיון לתמונת העוגן. כאן המתייגים התבקשו לסמן איזו משתי התמונות דומה יותר/פחות לתמונת העוגן. המטרה של האימון היא למקסם את ההפרש בין מרחק של ייצוג של תמונה(=טוקן CLS) דומה יותר לתמונת העוגן לזה של התמונה פחות דומה כאשר המרחק נמדד באמצעות מרחק קוסיין.

המחברים גם מציעים לבצע למידה ניגודית על השרשור של ייצוג התמונה עם הייצוג הממוצע של כל הפאצ'ים בתמונה. לטענתם זה משפר את התוצאות במידה מסוימת.ֿ

https://arxiv.org/abs/2410.10817