⚡️🚀המאמר היומי של מייק 30.07.24: ⚡️🚀

Denoising Vision Transformers

מזמן לא סקרנו מאמר בראייה הממוחשבת והיום נתרענן עם סקירה של מאמר די מעניין מהדומיין הזה. המאמר מציע שכלול ל-Vision Transformer או ViT בקצרה. משפחת ViT כוללת מודלים מבוססי טרנספורמרים המיועדים לעיבוד דאטה ויזואלי ולהפקת ייצוג חזק של תמונה. מה אני מתכוון כאשר אני אומר ייצוג חזק של תמונה? למעשה זה ייצוג (לֹטנטי) של תמונה, בעל מימד משמעותית נמוך יותר מהתמונה עצמה בד״כ, שניתן לנצלו לאימון מודלים למגוון משימות downstream (כגון סגמנטציה, זיהוי אובייקטים, סיווג וכדומה).

המאמר טוען שניתן לשפר אתת את הייצוגים המופקים על ידי ViT באמצעות ניקוי רעשים הנוצרים בגלל השימוש ב-positional encoding או קידוד תלוי מיקום. מטרתו של קידוד תלוי מיקום היא להעביר למודל מידע על מיקום של הפאצ'ים של התמונה. אזכיר כדי להזין תמונה ל-ViT אנו מפרקים אותה לפאצ'ים, משטחים אותם ומזינים אותם למודל. לוקטור המייצג כל פאץ' אנו מוסיפים (אשכרה מחברים) וקטור המקודד את מיקומו היחסי בתמונה של הפאץ'.

המאמר טוען שהוקטורים המקודדים מיקום מרעישים את ייצוגי הפאצ'ים ומקשים על שימושם למשימות downstream. לטענת המחברים רעש המתווסף לייצוגי הפאצ'ים מכיל מידע על המיקום של הפאצ' בלבד ולא מכיל שום מידע על התוכן של הפאץ'. לעומתו שני החלקים האחרים בייצוג הפאץ' מכילים מידע על התוכן הסמנטי של הפאץ' והשני מכיל מידע המערבב את ייצוג התוכן וייצוג המיקום. המחברים טוענים שניקוי הייצוג מהרעש המידע על המיקום בלבד תורם לעוצמתו של הייצוג.

כדי לאתר את הארטיפקט המיקומי הזה בייצוג הפאץ' המאמר מציע לאמן מודל המזהה את שלושת החלקים של הייצוג שהזכרנו בפסקה הקודמת. זה נעשה עלי די אוגמנטציה של תמונה (הזזה, קרופ וכדומה) דרך ניצול התכונות האינהרנטיות של הרעש המיקומי ושל הייצוג התוכן. כלומר המידע המיקומי בייצוג ״זזה יחד עם הפאץ״ כאשר המידע המייצג את התוכן לא משתנה אם מזיזים את הפאץ' בתמונה. החלק שמערבב את המידע על המיקום והתוכן היא פשוט הפרש בין ייצוג של ViT לבין סכום של שני החלקים האחרים.

בשלב השני מאמנים מודל המזהה את הרעש המיקומי בייצוג הפאץ'. לאחר מכן באינפרנס מחסירים את הרעש הזה מהייצוג של הפאץ' וכדי לקבל ייצוג יותר נקי ועוצמתי.

https://arxiv.org/abs/2401.02957