⚡️🚀המאמר היומי של מייק 12.08.24: ⚡️🚀

Img-Diff: Contrastive Data Synthesis for Multimodal Large Language Models

מודלי דיפוזיה גנרטיביים הגיעו לתוצאות מרשימות לאחרונה והפגינו יכולת לגנרט תמונות באיכות מרהיבה. למרות זאת מודלים אלו מתקשים לפעמים במשימות של עריכת תמונות ולא מצליחים להחליף אובייקטים לא גדולים בתמונה תוך שמירה של כל המאפיינים האחרים של התמונה.

המאמר המסוקר מציע שיטה ליצירת דאטהסט של זוגות תמונות שכל זוג מכיל תמונות זהות פרט לאובייקט אחד בתמונה. כל זוג תמונות מלווה בתיאור של האובייקטים שהוחלפו בשתי התמונות וגם במיקומם בתמונות. בין השאר דאטהסט זה יכול לשמש חוקרים ומהנדסים לאימון מודלים לעריכת תמונות.

איך הם עשו זאת? האמת הפייפליין שלהם די מורכב מכיל הפעלה לא מעט מודלים מולטימודליים, ומודלים לזיהוי ותיאור אובייקטים בתמונה כמו LLAVA, FastSAM, BLIP, CLIP וכדומה. נתאר רק את ה 3 השלבים של התהליך.

בשלב הראשון לקחו כמה עשרות אלפי תמונות מהדאטהסט הידוע MS COCO ויצרו זוגות של תמונות דומות על ידי החלפה של אובייקטים מסוימים באובייקטים אחרים בתמונה עם המודל שנקרא ViCUNA (ההחלפה עצמה בוצעה עם המודל הנקרא InstructPix2Pix).

בשלב השני אנו מפעילים כמה מודלים מולטימודליים כדי לזהות את האיזורים בתמונות שעברו שינוי (בזוגות מהשלב הראשון). קודם כל המחברים את התמונות הלא דומות עם CLIP (כלומר בהתבסס על דמיון של ייצוגי התמונות). לאחר מכן שוב מפלטרים את הדאטהסט על ידי התאמה של תיאורם של האובייקטים והימצאותם בשתי התמונות עם BLIP. בסוף מזהים את מיקום האיזורים בתמונה שבהם הוחלפו האובייקטים (כלומר bounding boxes שלהם).

בשלב האחרון מפיקים תיאור טקסטואלי של כל החלפות של בוצעו בתמונה הראשונה בזוג שהפך אותה לתמונה השנייה בזוג. עושים זאת עם שילוב של LLAVA ו- CLIP.

וככה מקבלים דאטהסט איכותי של זוגות תמונות דומות שמה שהשינוי ביניהם מתואר על ידי התוצאה של השלב האחרונה (כולל מיקום השינוי).

https://arxiv.org/abs/2408.04594