המאמר היומי של מייק - 25.12.24:  
Vision language models are blind

מאמר נחמד הטוען שמודלי שפה ויזואליים הם די עיוורים כלומר אין להם סיכוי לעבור בדיקה אצל אופטומטריסט מורשה. הנה כמה עובדות על המבחנים הכושלים שלהם:

מודלי שפה ויזואליים או VLMs לא יכולים לקבוע באופן אמין האם שני קווים (או שני מעגלים) נחתכים, במיוחד כשהם קרובים זה לזה. הדיוק בזיהוי 0, 1 או 2 נקודות חיתוך בין שתי פונקציות לינאריות למקוטעין בעלות 2 מקטעים נע בין 47% ל-85%. באותה משימת שני המעגלים, המודלים מתפקדים טוב יותר (דיוק של 73-93%) אך עדיין רחוק מה-100% המצופה.

מודלי שפה ויזואליים יכולים לזהות בצורה מושלמת מעגל ומילה בנפרד אך כאשר המעגל המילה נמצאת בתוך המעגל המודלים נוטים להתקשות בזיהוי איזו אות מוקפת במעגל.

מודלי ראייה-שפה יכולים לספור צורות במדויק, למשל, מעגלים , ריבועיים כאשר הם נפרדים ורחוקים זה מזה. עם זאת, כל המודלים מתקשים לספור מעגלים חותכים (כמו הלוגו האולימפי), ובאופן כללי, צורות בסיסיות שהן חופפות או מקוננות.

בסידור ריבועים בצורה של רשת, אנו מגלים ש-VLMs נכשלים באופן מפתיע בספירת מספר השורות או העמודות ברשת, בין אם היא ריקה או מכילה טקסט. זה מפתיע בהתחשב בכך שהמודלים מתפקדים כל כך טוב (דיוק ≥ 90%) על הדאטהסט ב-DocVQA הכולל שאלות רבות עם טבלאות(אוברפיט כנראה).

כאשר המודל מתבקש לעקוב אחר מסלולים צבעוניים במפת רכבת תחתית של עד 8 מסלולים וסך הכל 4 תחנות, VLMs לעתים קרובות נכשלים בזיהוי היכן מסלול מסתיים, כלומר, ומפגינים דיוק של 23% עד 50% .

המודל GPT-4o עולה בביצועיו על Gemini-1.5 Pro ו-Claude-3 Sonnet ב-7 בנצ'מרקים מורכבים עבור VLMs אך מתפקד באופן משמעותי פחות טוב במשימות הנבחנות במאמר, שבהן Gemini-1.5 Pro ו-Sonnet-3.5 הם הטובים ביותר. כלומר, המאמר מגלה מגבלות מפתיעות של מודלי ראייה-שפה שלא נמדדו בבנצ'מרקים רגילים.

בקיצור אולי VLMs האלו צריכים משקפיים…

https://arxiv.org/abs/2407.06581