המאמר היומי של מייק: 21.06.25  
Janus: Decoupling Visual Encoding for Unified Multimodal Understanding and Generation

המאמר הזה לא חדש אבל פתאום גיליתי שהתחלתי לסקור אותו בקובץ דוקס נידח ונתקלתי בו בצורה די אקראית. תוך כדי חיפוש בערוץ הטלגרם שלי גיליתי שעשיתי סקר(בסוף ינואר) ורוב המנוים (יותר מ 85%) רצו שאסקור אותו. מקיים את ההבטחה הפעם בדליי של 5 חודשים.

המאמר אימן מודל מולטימודלי לשפה ולתמונות. להבדיל מרוב העבודות בתחום המאמר מציע להפריד בין הבנה של טקסט וקלט ויזואלי לבין גמרוט של טקסט ותמונות. כלומר המחברים מאמנים 3 מודלים שונים(אמנם עם רכיבים משותפים) להבנה וגנרוט של טקסט, הבנה וגנרוט טקסט עבור תרחישים מולטימודליים והשלישי עבור גנרוט של תמונות. הבנה כאן הכוונה קידוד של קלט למחרב ייצוג וקטורי משלו ואדפטר הממפה אותו למרחב הלטנטי של מודל שפה L שהוא backbone של ל Janus. יש עוד 2 מודלים קטנים (heads) הממפה את הפלט של מודל שפה L לפני הפיכתו לטוקנים (של השפה ושל התמונות).

מאמנים את Janus על מגוון משימות כמו הבנת התמונה וכל האובייקטים בה, דיאלוג בהתבסס על מה שמופיע בתמונה, יצירת תמונה מקלט טקסטואלי וקלט ויזואלי כמו עריכת תמונות על בסיס קלט טקסטואלי ועוד. אציין שהאחרי האנקודרים והאדאפטרים הייצוגים שלהם מוזנים למודל שפה גדול (שהוא גם מאומן מהשלב השני של האימון של Janus).

יש 3 שלבים עיקריים באימון Janus. המטרה העיקרית של שלב הראשון היא ליצור ״חיבור״ מושגי בין רכיבים(מודלים) ויזואליים לשפתיים בתוך מרחב האמבדינג, כך שמודל שפה יוכל להבין את היישויות המוצגות בתמונות ולפתח יכולת ראשונית ליצירת תמונות. בשלב זה אנו משאירים את מקודדי התמונה ואת ה-LLM, ומאמנים את הרדאפטרים (עבור הטקסט ועבור התמונות) וגם מודל הראש (head) עבור המודל לגנרוט טקסט.

בשלב זה אנו מבצעים אימון מאוחד על גבי קורפוס מולטימודלי, כדי לאפשר ל-Janus ללמוד גם הבנה וגם גנרוט מולטימודלית. באופן פרקטי מאמנים את כל הרכיבים של Janus חוץ מאשר שני אנקודרים: השפתי והטקסטואלי. בשלב האחרון אנו עושים פיין טיון למודל המאומן בעזרת דאטה מבוססי הנחיות, במטרה לחזק את היכולת לעקוב אחר הנחיות ולנהל דיאלוגים כדי להבטיח שמודל Janus יהיה מיומן גם בהבנה וגם בגנרוט מולטימודליים, הם לא מאמנים מודלים נפרדים לכל משימה. במקום זאת, המחברים משתמשים בשילוב של דאטהסטים של דיאלוגים טקסטואליים בלבד, דאטהסטים של משימות הבנה מולטימודלית וכאלו של גנרוט של תמונות מטקסט, כדי להבטיח גמישות במגוון תרחישים.

https://arxiv.org/abs/2410.13848