המאמר היומי של מייק - 17.03.25  
JanusFlow: Harmonizing Autoregression and Rectified Flow for Unified Multimodal Understanding and Generation

מזמן לא סקרתי מאמר על מודלים גנרטיביים מולטימודליים. מודלים אלו מאומנים לא רק לגנרט דאטה מכמה סוגים (במקרה של JanusFlow של שפה טבעית ותמונות) אלא גם לבצע משימות הכרוכות בהבנה של הקשרים ביו המודליות האלו. למשל מודל מולטימודלי בתחום שפה ותמונות צריך להיות לענות על שאלות על תמונה. המודל מורכב ממודל עיקרי (הנקרא LLM) וכמה אנקודרים ודקורדרים המיועדים לייצוג דאטה ממודליות שונות והפיכתו של ייצוגו לפיסת דאטה (דקודרים). כל המודלים במאמר מבוססים על הטרנספורמרים באופן מאוד לא מפתיע.

המאמר מציע שיטה לאמן מולטימודלי (הנקרא LLM במאמר) כזה כאשר הפרט המעניין לגביו הוא שימוש באנקודרים שונים לשפה ולתמונות (ברוב המודלים המולטימודליים משתמשים באותו מודל backbone). בגדול במהלך האימון המודל לומד לחזות את הטוקנים של תשובה על פרומפט נתון כאשר פרומפט ותשובה יכולים להיות גם טוקן ויזואלי (ייצוג של פאץ' של תמונה) וגם טוקן רגיל(= סדרת אותיות). בנוסף הפרומפט יכול להיות שילוב של טוקנים ויזואליים וטוקנים של השפה במשימת visual question answering. בנוסף (לא מופיע במאמר הזה בצורה מפורשת אך נעשה במודלים מולטימודליים אחרים) המודל מאומן גם על דאטה טקסטואלי בלבד(כמו ב-pretraining של מודל שפה רגיל)

כמה פרטים על המודלים השונים (פרט ל-LLM) המופיעים במאמר. עבור דאטה שפתי הטוקנים עוברים אנדוקר מאומן (נקרא und enc) - אחרי הטוקנים עוברים שכבה לינארית מאומנת. עבור דאטה ויזואלי יש אנקודר סטנדרטי לא מאומן המבוסס על VAE ואחרי יש עוד אנקודר מאומן. מכיוון שהמודל הגנרטיבי לתמונות הינו מודל דיפוזיה שימוש ב-VAE (חלק בלתי נפרד של מודלי דיפוזיה גנרטיביים) לא צריך להפתיע. בנוסף כאמור יש שני דקודר מאומנים שאליהם מוזנים הייצוגים הנבנים על ידי LLM.

המאמר מציע שיטה תלת שלבית לאימון המודלים כאשר כל שלב ״מפשירים״ יותר ויותר מודלים (כולל LLM) כאשר בשלב האחרון מאמנים את כולם פרט ל-VAE.

מודלי דיפוזיה במאמר מבוססי על (rectified flows (RF המנסה למפות את הדאטה מהתפלגות פשוטה (גאוסית) להתפלגות הדאטה בצורה ישרה כלומר המסלול בין x\_0 הגאוסי ל x\_1 של הדאטה הוא ישר. כלומר כל נקודה x\_t במסלול הזה היא צירוף קמור של x\_0 ו-x\_1. בגדול מודל הדיפוזיה מאומן לשערך את המהירות הקבועה v(השווה x\_0 - x\_1 עבור כל נקודה x\_t במסלול. הדגימה מבוצעת על ידי פתרון משוואה דיפרנציאלית המתארת התקדמות של x\_0 ל x\_1 עם מהירות v (שיטת אוילר). מודל דיפוזיה המאומן במאמר הוא לטנטי.

פרט מעניין על המאמר: אחד האיברים בפונקציית לוס של מודל דיפוזיה קונסת אותו על אי התאמה של ייצוג הפנימי המורעש (המחושב על שכבות הביניים של המודל) לייצוג התמונה הנקייה המחושב על אנקודר חזק (understanding encoder). וכמבון יש classifier guidance באימון של מודל דיפוזיה (קלאסי)

מאמר כתוב יפה ודי ברור - מומלץ!

https://arxiv.org/abs/2411.07975