המאמר היומי של מייק - 27.02.25  
Unity by Diversity: Improved Representation Learning for Multimodal VAEs

היום אני חוזר למאמר על (Variational Autoencoder (VAE אחרי תקופה ארוכה מאוד, יותר משנה, אני מניח.

נזכיר ש-VAE הוא סוג של מודל גנרטיבי שלומד לדחוס דאטה למרחב לטנטי נמוך-ממד ומובנה (עם התפלגות מושרית), ולאחר מכן לשחזר אותם. האתגר המרכזי הוא להבטיח שמרחב זה ״יישאר חלק״ כך שדגימה ממנו תייצר דאטה ריאליסטי. כדי להשיג זאת, VAE מאזן בין שני יעדים: שחזור מדויק של הדאטה המקורי תוך שמירה על כך שהמרחב הלטנטי יהיה קרוב(בהתפלגות) להתפלגות פשוטה ומוגדרת היטב, בדרך כלל גאוסיאנית. זה מבטיח שנקודות סמוכות במרחב הלטנטי יתאימו לפיסות דומות, מה שמאפשר יצירה של דוגמאות חדשות קוהרנטיות.

מולטימודל VAEs מרחיבים את הרעיון הזה כדי להתמודד עם סוגים שונים של דאטה, כגון תמונות, טקסט ואודיו, במסגרת מאוחדת. הקושי עם MVAE נובע מכך שמודלים למודים(modes) שונים חולקים מידע מסוים אך גם מכילים פרטים ייחודיים לכל מודל. ישנם מודלי MVAE שמנסים לכפות על כל המודלים לחלוק ייצוג משותף אחד, מה שעלול להוביל לאובדן מידע ייחודי לכל מודל. אחרים שומרים עליהם מופרדים מדי, מה שמונע אינטראקציות משמעותיות בין המודלים. MVAE מושלם חייב למצוא את האיזון: לתפוס גם את המבנה המשותף של המודים השונים תוך שמירה על מה שמייחד כל מודל.

התובנה המרכזית כאן היא דחיית ההנחה שכל המודים יחיו באותו המרחב הלטנטי. MVAEs קודמים פעלו תחת ההנחה כי יש להשתמש בייצוג לטנטי יחיד (מפולג גאוסי סטנדרטי בד״כ) עבור כל המודלים (לכל ה-מודאליות). גישה זו לרוב מובילה למרחב לטנטי משותף שהוא או "מחובר״ מדי - מה שמאלץ מודלים שאינם תואמים להתערבב באופן לא טבעי- או ״חופשי״ מדי, כך שהוא אינו מצליח ללמוד קשרים חשובים בין המודלים השונים. כדי להתגבר על כך המחברים מציעים ללמוד את ההתפלגות הלטנטית מתוך הדאטה עצמו, ובכך יוצר מרחב לטנטי שתחשב בניואנסים הייחודיים של כל מודל תוך שמירה על יכולת העברת מידע בין מודלים שונים.

במקום להשתמש בהתפלגות יעד קבועה (נגיד גאוסית כמו ברוב המקרים), MVVM VAE בונה התפלגות לטנטית בסגנון Mixture of Experts, המתחשב בכל המודלים (למודליות השונים) בזמן האימון. כל מודל תורם להערכה של ההתפלגות הזו, כך שהיא פועלת כהגבלה רכה ולא כהגבלה קשיחה כמו ב-VAE הרגילים. זהו שינוי מהותי לעומת שיטות המבוססות על מיצוע/כפל של התפלגויות - כאן מדובר בהתפלגות דינמית, שתלויה בדאטה ומתעדכנת כל הזמן כי המודל לומד את המודים. למעשה, היא מתפקדת כמו ״שלד הסתגלותי״, המעצב את המרחב הלטנטי כך שיתמוך במבנה משותף, אך מבלי לכפות אותו.

בואו נצלול לפרטים מתמטיים על איך כל זה נעשה בפועל. ללוס השחזור (עד כמה טוב המודל מצליח לשחזר את הדאטה) המחברים מוסיפים איבר רגולריזציה שהיא סכום של JS divergences בין ההתפלגויות הפוסטריוריות המשוערכות (q\_ϕ(z|X של כל מודליות לבין ההתפלגות הלטנטית הממוצעת הנלמדת (h(z|X הממוצעת על כל המודליות. כאן X מסמן את הדאטה ו-z הינו הייצוג הלטנטי.

מה זה אומר בפועל?

כל מודל (מודאליות) מתומרץ להישאר קרוב להתפלגות הממוצעת הנלמדת, שמבוססת על כל המודלים יחד.

כל מודל שומר על המבנה הייחודי בלתי תלוי שלו, מה שמונע קריסת מרחב הלטנטים לתת מרחב קטן מדי (הטענה במאמר.

ההתפלגות הממוצעת (h(z ∣ X פועלת כרגולריזציה דינמית ונלמדת, כך שהייצוגים נותרים משמעותיים ושימושיים.

בסופו של דבר, MMVM VAE אינו מכתיב מבנה לטנטי קשיח(כמו MVAE רגיל) - במקום זאת, הוא מאפשר למבנה להתפתח באופן טבעי מתוך דאטה עצמם. וזה בדיוק מה שהופך אותו לכל כך חזק.

המודל המוצע גם מציג ביצועים טובים במשימות השלמת דאטה חסר. MVAEs מתקשים כאשר מבקשים מהם לשחזר מודל חסר מתוך קלט חלקי. למה? כי הייצוגים המשותפים שלהם נוטים להיות נוקשים מדי - או שהם נשענים באופן מוגזם על מידע משותף (מה שמוביל לדורות גנריים ומטושטשים) או שהם שומרים על ייצוגים נפרדים שאינם מקיימים אינטראקציה משמעותית. לעומת זאת, MMVM VAE מבטיח שהמרחבים הלטנטיים של כל מודל יישארו אינפורמטיביים, גם כאשר מודלים מסוימים חסרים. התוצאה? שחזורים עקביים יותר, רלוונטיים מבחינה הקשרית, ובעלי דיוק גבוה יותר ביחס למבנה הצפוי של הנתונים החסרים.

יש עוד טענה מעניינת במאמר שאני מודה שלא הבנתי עד הסוף: הקשר בין MMVM VAE לבין למידה ניגודית. השימוש ב-JSD כרגולריזציה מקרב את ההתפלגויות הלטנטיות בין המודליות השונות, מבלי לגרום להן לקרוס לתת מרחב קטן (טענה לא מובנת לי). זה דומה לאופן שבו למידה ניגודית פועלת במודלי שפה-ויזן: היא מבטיחה שקלטים דומים יפיקו ייצוגים קרובים, אך תוך שימור ההבדלים ביניהם. עם זאת, בניגוד ללמידה קונטרסטיבית, אשר לרוב דורשת יצירה מפורשת של זוגות דוגמאות חיוביות ושליליות, MMVM VAE מטמיע את תהליך היישור הזה בתוך המודל הגנרטיבי עצמו. המשמעות היא שהייצוגים הלמידתיים מתכנסים באופן טבעי במהלך האימון, ללא צורך בטריקים של דגימה קונטרסטיבית או מטרות נוספות.

כמובן, ישנם גם טרייד-אופים. ההסתמכות של המודל על התפלגויות ראשוניות תלויות-דאטה הופכת גנרוט הבלתי מותנה כמו ב-VAE לבלתי ישים. רוצים לדגום דאטה מולטי-מודאליים חדשים מאפס? חבל, השיטה הזו לא נועדה לכך :). אבל זה לא באג, אלא פיצ'ר. MMVM VAE אינו מנסה להיות מודל גנרטיבי טהור כמו GANs או מודלי דיפוזיה. המטרה שלו היא למידת ייצוגים מובנית וברורה, שבה התלויות בין המודלים נשמרות מבלי לכפות מגבלות מלאכותיות. במרחב הזה, מדובר בהתקדמות מחקרית משמעותי

https://arxiv.org/abs/2403.05300