המאמר היומי של מייק: 26.04.25  
Multimodal Latent Language Modeling with Next-Token Diffusion

היום שבת והסקירה של היום תהיה קלילה ודי קצרה. הסקירה תתמקד במודלים מולטי-מודליים גנרטיביים המסוגלים "להבין״ וליצור דאטה מכמה מודליות כלומר טקסט, תמונות, אודיו וכדומה. המאמר למעשה בעצם משדך מודלים לטנטיים גנרטיבים עבור דאטה טקסטואלי ועבור דאטה רציף יותר (למרות שגם הוא discretized). המחברים עושים זאת באמצעות אימון של מודלי דיפוזיה גנרטיבים עבור סוגי דאטה שונים במרחב הלטנטי. כלומר המודל מאומן לגנרט ייצוגים לטנטיים עבור דאטה טקסטואלי ועבור דאטה כמו אודיו ותמונות.

להבדיל ממאמרים רבים אחרים המחברים מאמנים לא רק את המודל הגנרטיבי המולטימודלי אלא מאמנים גם מודל אמבדינג להפקה של ייצוגים לטנטיים של דאטה ממודליות שונות. בדרך כלל מודל האמבדינג במודלי דיפוזיה הוא מסוג VAE (שזה Variational Autoencoder) והמחברים מציעים מודיפיקציה קלה ל-VAE. במקום שהאנקודר (הקלט אליו הוא דאטה) של VAE יגנרט את וקטורי התוחלות השונויות של הווקטור הלטנטי הוא מגנרט רק וקטור התוחלות כאשר השוניות מוגרלות ההתפלגות גאוסית עם שונות נתונה (הייפרפרמטר). לדעת המחברים זה מונע קריסה(איפוס) של וקטור השונויות הנוצר על ידי האנקודר שפוגע בגיוון התמונות שהמודל מגנרט.

המחברים מאמנים VAE עבור דאטה לא טקסטואלי בלבד. תמונה או אודיו מחולקת לטוקנים (פאצ'ים לתמונות ולמקטעים בזמן לאודיו) ומוזנים למודל כדאטה סדרתית. שימו לב המודל מסתכל על דאטה בכל מודליות כמו דאטה סדרתי. זה מאוד טריוויאלי לדאטה טקסטואלי ולאודיו כי יש שם סדר אינהרנטי ברור. בתמונות גם יש סדר אבל הוא יכול לבוא בכמה צורות: כלומר ניתן לתאר תמונה כסדרה של פאצ'ים בכמה צורות (למשל משמאל לימין ולמעלה למטה וגם מימין לשמאל ומלטה למעלה).

מודל דיפוזיה לדאטה לא טקסטואלי מאומן לנקות את הרעש מהדאטה (denoising) בהינתן הייצוג הלטנטי שלו(המורעש) ושל ההקשר (כל המודלים במאמר כמובן אוטורגרסיביים). לאחר מכן הוקטור הלטנטי הנקי מוזן לדקודר של VAE לשחזור הדאטה כאשר המטרה של המודל המאומן היא לשחזר את הדאטה כמה שיותר טוב. עבור דאטה טקסטואלי ההרעשה מופעלת על האמבדינג של הטוקנים הטקסטואליים ומודל דיפוזיה מאומן לשחזר אותם. עבור דאטה טקסטואלי מאמנים עוד שכבה לינארית שמטרתה למפות את הוקטור הלטנטי למרחב הטוקנים הטקסטואליים (סופטמקס בגודל של מילון). דרך אגב מודלי דיפוזיה מאומנים יחד עם ה-VAE (אנקודר ודקודר).

כדי להפריד בין דאטה טקסטואלי ולא טקסטואלי המחברים מאמנים טוקנים המפרידים בין דאטה השייך למודליות שונות.

https://arxiv.org/abs/2412.08635