המאמר היומי של מייק - 05.04.25  
GIVT: Generative Infinite-Vocabulary Transformers

היום חוזרים כמה שנים שנים אחורה בו מילים VAE, VQ-VAE, VQ-GAN היו מושכים אותה תשומת לב כמו שמקבלים היום מודלי דיפוזיה גנרטיביים (אמנם פחות מאג'נטים אבל בכל זאת). המאמר שנסקור היום מציע שכלול מעניין ל-VQ-VAE שמשך את עיניי כי כאמור מאמרים בנושא זה הפכו להיות ״ציפור נדירה״ בנוף שלנו (של AI).

קודם כל אתן הקדמה קצר לגבי VQ-VAE. נתחיל את ההסבר מ-VAE שזה ראשי תיבות של Variational AutoEncoder שהומצא אי שם ב-2014 על ידי Kingma האגדי. בגדול VAE מורכב משתי רשתות, אנקודר ודקודר כאשר הראשון מפיק את הייצוג הלטנטי (או אמבדינג) של פיסת דאטה כאשר הדקודר הופך את הייצוג הלטנטי לתמונה. הפלט של האנקודר הוא הפרמטרים של ההתפלגות הגאוסית (וקטור תוחלות ומטריצת קווריאנס אלכסונית) ממנה דוגמים את הוקטור הלטנטי המוזן לדקודר לשחזור תמונת הקלט לאנקודר.

פונקציית הלוס של VAE נבנית על בסיס ELBO (שזה Evidence Lower Bound) ומכילה 2 איברים. הראשון לוס השחזור של-VAE קלאסי הוא הנורמה של הפרש התמונה המשוחזרת ביחס לתמונה המקורית (בגרסאות מתקדמות יותר התווספו לזה לוס perceptual ולוס בסגנון GAN) והאיבר השני הוא KL divergence בין ההתפלגות של הייצוג הלטנטי המופק מהדאטה (המיוצג על ידי וקטור תוחלות ומטריצת קווריאנס אלכסונית) לבין התפלגות נורמלית סטנדרטית. באינפרנס אנו דוגמים וקטור לטנטי מהתפלגות נורמלית סטנדרטית ומזינים אותו לדקודר.

שכלול מעניין שהפך להיות מאוד פופולרי של VAE הוא VQ-VAE. במקום להגדיר מרחב לטנטי בתור התפלגות גאוסית - אלא מגדיר אותו בצורה דיסקרטית. כל פאץ' בתמונה מתואר(במרחב הלטנטי) על ידי וקטור מה-codebook בגודל סופי שמאומן יחד עם האנקודר והדקודר. כלומר יש מספר סופי של הייצוגים הלטנטיים עבור כל פאץ' (נחזור על זה עוד מעט). כאשר אימון האנקודר, הדקודר והוקטורים מה-codebook מסתיים אנו מאמנים מודל נוסף לחיזוי ייצוג לטנטי של פאצ'ים, על כל הייצוגים הלטנטיים של הפאצ'ים של הדאטהסט. מודל זה (נגיד טרנספורמר) מאומן לחזות באופן אוטורגרסיבי את הוקטור מה-codebook (כלומר מספרו) של הפאץ' הבא בהינתן הפאצ'ים הקודמים שכבר גונרטו. לאחר מכן הווקטורים הלטנטיים של הפאצ'ים מוזנים לדקודר לגנרוט דאטה (תמונה).

כאמור יש מספר מוגבל וסופי של הייצוגים הלטנטיים עבור כל פאץ' ושזה די מגביל את העושר הסמנטי של התמונות ש-VQ-VAE ושיטות דומות מסוגלות לגנרט. וזה בדיוק המקום שהמאמר שנסקור עכשיו מחדש - הוא מציע שיטה לעבור לייצוג רציף (ללא codebook) של הוקטורים הלטנטיים. אבל איך אפשר לעשות זאת? נזכיר ש-VQ-VAE אנו כל פעם חוזים התפלגות קטגוריאלית מעל ה-codebook כלומר השכבה האחרונה במודל אוטורגרסיבי היא סופטמקס בגודל של ה-codebook.

האם ניתן לאמן מודל שיוצר ייצוגים לטנטיים רציפים בצורה אוטורגרסיבית? התשובה היא כן - בשלב הראשון המאמר מאמן את ה-VAE הסטנדרטי שהסברתי עליו בתחילת הסקירה. הדבר הזה נעשה ברמה של פאץ' כלומר הייצוג הלטנטי של תמונה מורכב מהייצוגים של הפאצ'ים שלה. בשלב השני המחברים מאמנים מודל טרנספורמר סיבתי שחוזה את הייצוג של הטוקן הבא באמצעות חיזוי פרמטרים של התפלגות gaussian mixture שממנו נדגם הוקטור הלטנטי עצמו. כלומר כל פעם הטרנספורמר הסיבתי (לוקח בחשבון רק את הוקטורים שנוצרו כבר) חוזה את וקטורי התוחלות, פרמטרים של מטריצות קווריאנס אלכסוניות של כל משתנה ב-mix ומשקלי הערבוב). אחרי שהוקטורים הלטנטיים נחזו ונדגמו הם מוזנים לדקודר לגנרוט תמונה.

נציין כי GIVT להבדיל מ-VQ-VAE ניתן לאמן במלאו יחד עם האנקודר והדקודר שלטענת המחברים יכול להיות בעייתי. המחברים מציעים (במקום הטרנספורמר הסיבתי) לאמן מודל (הנקרא adapter) של Normalized Flow לגנרוט של ייצוג לטנטי כולו של הדאטה לאחר שהאנקודר והדקודר כבר אומנו וככה להפריד את שני השלבים.

בנוסף המאמר מציע לאמן טרנספורמר לא סיבתי לחיזוי ייצוגים לטנטיים של פאצ'ים (מאומן דומה לmasked languaged modeling או MLM). הגישה הזו שלא ידעתי עליה לפני הוצגה במאמר MaskGit.

https://arxiv.org/pdf/2312.02116