ZigMa: A DiT-style Zigzag Mamba Diffusion Model

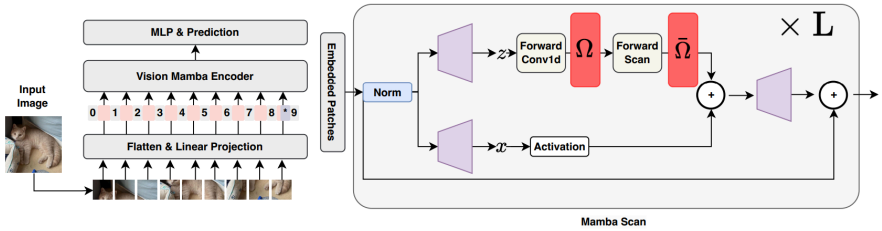


Figure 2: ZigMa. Our backbone is structured in L layers, mirroring the style of DiT [65]. We use the single-scan Mamba block as the primary reasoning module across different patches. To ensure the network is positionally aware, we've designed an arrange-rearrange scheme based on the single-scan Mamba. Different layers follow pairs of unique rearrange operation Ω and reverse rearrange $\bar{\Omega}$, optimizing the position-awareness of the method.

המאמר הזה משך את תשומת ליבי מכמה סיבות:

1. יש מודלי דיפוזיה - האהבה הקודמת שלי שבקרוב מאוד אחדש את הקשר איתם
2. יש כאן SSMs (State-Space Models) בדמות Mamba - האהבה הנוכחית שלי שתיכף אני מסיים להכין עליה מצגת די רצינית ובתקווה ישמע אותי מציג אותה בפורומים השונים
3. המאמר פורסם בראשון לאפריל ובהתחלה קצת חשדתי 🙂

בנוסף יש במאמר גם קצת מהטרנספורמרים (cross-attention) שעוד מוסיף לשלמותו. אוקיי, אז מה יש לנו במאמר הזה מעבר לכמה מילים "באזזיות". המאמר מציע ארכיטקטורה מעניינת המיועדת לגינרוט תמונות וגם וידאו. כאמור הארכיטקטורה היא שייכת למשפחה של מודלי דיפוזיה גנרטיביים אבל מכילה חלקים המורכבים מ-SSMs (ממבה) בנוסף ל-cross-attention הלב של הטרנספורמרים. ויש כאן חידוש מעניין לגבי הסדר שבו מכניסים פאצ'ים של תמונות (או פריימים של וידאו) במהלך אימון המודל.

נתחיל מהסבר קצר על מודלי דיפוזיה גנרטיביים. בהינתן דאטהסט (של תמונות או/ו סרטונים) אנו מאמנים את רשת באופן הבא:

1. מוסיפים כמויות קטנות של רעש גאוסי לפיסת דאטה עד שהיא היא הופכת לרעש טהור
2. מאמנים רשת נוירונים (עם Mamba ו-cross-attention במקרה שלנו) כדי למדל של התהליך ההפוך: כלומר מפיסת דאטה מורעשת מאיטרציה n לחזות אותה באיטרציה n-1.

כאשר יש בידינו מודל כזה אנו למעשה מסוגלים לגנרט תמונה מרעש גאוסי טהור בצורה הדרגתית, איטרציה אחרי איטרציה. עם השנים צצו שיטות רבות ומגוונות מאוד לאיך להוסיף רעש ומה בדיוק כדאי לחזות עם הרשת שלנו.

בשנה וחצי האחרונות היו כמה חידושים מעניינים במודלי דיפוזיה ומכיוון שהמאמר משתמש בהם אני חייב לספר לכם בגדול במה מדובר (כאמור הולך לדבר על זה בהרחבה בסקירות הבאות).

לאחרונה יצאו כמה מאמרים מעניינים (למשל https://arxiv.org/abs/2210.02747 ו-https://arxiv.org/abs/2303.08797 אבל יש עוד עשרות אחרים) המכליל מודלי דיפוזיה לתהליך רציף של מיפוי התפלגות פשוטה (כגון גאוסית סטנדרטית) להתפלגות של הדאטה (ההתפלגות המורכבת). תהליך זה נקרא

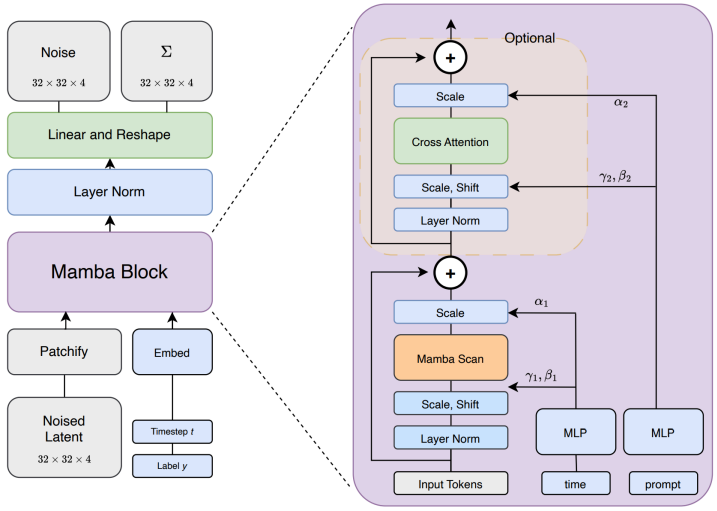
זרימה רציפה (continuous flow) הדיסקרטיזציה שלו (במישור הזמן כלומר האיטרציות) היא מודל דיפוזיה גנרטיבי עבור מיפוים מסוימים. יש לנו חופש לבחור את המיפוי (זרימה) בין התפלגות דאטה להתפלגות הפשוטה ויש לא מעט מחקרים על איך לבחור אותו בצורה אופטימלית (למקסם את איכות הדאטה המגונרטת, לייצב את התהליך, ליצור מיפוי כמה שיותר פשוט או ישר וכדומה).

אז איך כל המתמטיקה הזו קשורה לגנרוט דאטה? אז יש כאן עוד קצת מתמטיקה שנצטרך לצלול בה. בגדול הזרימה הרציפה בין להתפלגות הפשוטה להתפלגויות הדאטה (לפעמים נקראת reverse-time) ניתן לתאר על ידי משוואה דיפרנציאלית סטוכסטיות (SDE) שמכילה:

1. הדאטה המורעש עצמו x_t
2. מהירות או השתנות(נגזרת בזמן) של הזרימה בזמן v_t(x) (תחשבו על זה כמו על תנועה במרחב בין 2 עננים של נקודות שניתן להגדיר אותה על ידי המהירות הכיוונית ונקודה ההתחלה או הסוף).
3. פונקציית score (s_t(x) שהיא בעצם לוגריתם של p_t(x) - פונקציית התפלגות של הדאטה המורעש
4. יש גם תהליך reverse-time Wiener המהווה את החלק אקראי (סטוכסטי) ב-SDE הזה

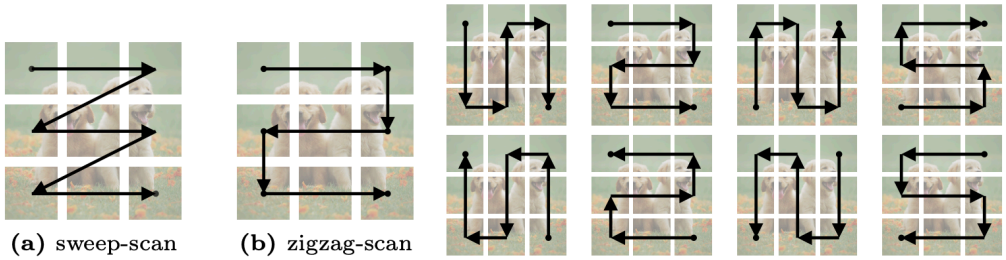
אז מה אפשר לעשות עם ה-SDE הזה, למה צריך אותו? מתברר כי עבור פרמטרים שהזרימה בין התפלגות הדאטה להתפלגות הפשוטה ניתן לנסח בעיות אופטימיזציה המאפשרות שערוך של s_t(x) ו- v_t(x) בהינתן דאטה לאימון. אחרי שנשערך אותן ניתן לפתור את ה-SDE שדיברנו עליה נומרית (נגיד בשיטת אוילר-מרואימה) כלומר מנקודת התחלה הנדגמת מההתפלגות הפשוטה (גאוסית) נוכל לגנרט דאטה צעד אחרי צעד. וזה בדיוק מה שעושים במאמר.

אוקיי, שרדנו את המתמטיקה - עכשיו מה הקשר ל-SSMs כאן? בשביל כך צריך להיזכר בארכיטקטורה של Diffusion Transformer או DiT, זה שעליו מבוסס המנוע של SoRA של OpenAI. למעשה DiT מורכב מהבלוקים של טרנספורמרים שמטרתם היא למדל את הפרמטרים (s_t(x) ו- v_t(x)) (כמובן לאחר דיסקרטיזציה במישור הזמן, כלומר איטרציות). המאמר המסוקר מחליף את בלוקי הטרנספורמר ב-Mamba (בנוסף הם גם לוקחים cross-attention שזה הלב של הטרנספורמר אך לפי הציור שלהם החלק הזה הוא אופציאונלי).



אבל כאן יש לנו בעיה. מכיוון שממבה היא ארכיטקטורה מיועדת לסדרות בעלת מימד הזמן חד מימדי (למשל טוקנים של טקסט) כאן יש לנו תמונות ובהן קיימים קשרים דו מימדיים בין הפאצ'ים (טוקנים ויזואליים) בתמונה

וקשרים תלת מימדיים בוידאו (בנוסף בין הפריימים). המאמר מתאים את המבנה של SSM עבור הקלט בעל קשרים רב מימדיים על שילוב של SSM-s שכל אחד מקבל את הקלט בסדר שונה (תראו בתמונה). כלומר שכבות של ממבה מוערמות (stacked) אחת מעל השנייה כל הקלט נכנס לכל אחד מהם בסדר שונה (למיטב הבנתי כל הממבות עובדות עם אותם מטריצות הפרמטרים A, B, C). זה מאפשר ל- ZigMa להתחשב בקשרים האלו. המאמר מרחיב את הגישה הזו לגנרוט וידאו (עבור קשרים תלת-מימדיים).



(a) sweep-scan **(b)** zigzag-scan **(c)** zigzag-scan with 8 schemes

אציין שבדומה ל-DiT המודל המוצע פועל במרחב הלטנטי כלומר הקלט למודל דיפוזיה הוא ייצוג לטנטי של הדאטה אחרי האנקודר. DiT משתמש באנקודר ובדקורר של VAE (אחד השכלולים שלו) אך במאמר הזה לא הצלחתי להבין האם המחברים לקחו VAE. במקום אחד במאמר רומזים שהאנקודר גם מכיל SSM אבל לא מצאתי לזה אזכורים נוספים.

התוצאות נראות לא רע, לשנות 2020 ככה אבל מכיוון שזה אחד המאמרים הראשונים המשלבים SSMs ומודלי דיפוזיה נסלח להם על כך.

יצאה סקירה ארוכה אבל מובנת פחות או יותר בתקווה...