המאמר היומי של מייק: 13.06.25  
Inference-Time Scaling for Diffusion Models beyond Scaling Denoising Steps

המחברים של המאמר הזה עושים משהו די לא שגרתי בדומיין של תמונות. הם שואלים שאלה פשוטה לכאורה: נניח שכבר אימנו מודל דיפוזיה מצוין, האם אפשר להפיק ממנו יותר בזמן הריצה? האם ניתן בשלב ההיסק לשפר את איכות התמונה הנוצרת מבלי פשוט להוסיף עוד ועוד צעדי denoising? התשובה של מחברי המאמר, מסתבר, היא כן. אבל הדרך לשם עוברת דרך מהלך מקורי לגמרי: חיפוש אחר רעש טוב יותר.

מי שעבד עם מודלי דיפוזיה יודע שבסופו של דבר, התהליך כולו מתגלגל קדימה מתוך וקטור רעש התחלתי. הוקטור הזה נבחר בדרך כלל בצורה אקראית, ונראה שכל מה שצריך ממנו זה להיות "רעש לבן". אבל מה אם לא כל רעש נברא שווה? מה אם אפשר לבחור רעש "חכם יותר" – כזה שידרדר את המודל לתמונה איכותית יותר, אפילו בלי לשנות את הארכיטקטורה, את מספר השלבים או את משקליו של המודל?

המאמר מציע בדיוק את זה: במקום להמשיך להאריך את מסלול הדיפוזיה (כלומר להגדיל את מספר השלבים), נוכל להשקיע את אותו תקציב חישובי בחיפוש סלקטיבי אחר רעש ראשוני שמניב תוצאה טובה יותר. מדובר בשינוי קונצפטואלי די משמעותי: אנחנו מפסיקים לחשוב על denoising כציר השיפור היחיד, ומתחילים לראות את הסטוכסטיות עצמה כלומר את הרעש, כמשהו שאפשר לא רק לדגום ממנו אלא לכוון אותו.

כדי שזה יעבוד, צריך שני דברים: קודם כל, דרך למדוד איכות של תוצאה סופית. לזה קוראים במאמר הזה verifier וזו יכולה להיות פונקציה כמו CLIPScore, aesthetic predictor, FID או כל מדד איכות אחר שתואם את מטרת הדגימה. המרכיב השני הוא אלגוריתם חיפוש כלומר דרך לבחור או לייצר רעשים חדשים, להשוות ביניהם לפי הפלטים שהם יוצרים, ולמצוא רעש שנותן תוצאה טובה יותר.

המבנה הזה, של verifier לצד אלגוריתם(שיטה) חיפוש, הוא הליבה של החדשנות כאן. מדובר במסגרת גנרית מספיק כך שהיא לא תלויה בארכיטקטורה של מודל הדיפוזיה, ולא דורשת fine-tuning. כל מה שצריך זה פונקציית דירוג, ויכולת להריץ כמה דגימות. מכאן פשוט מתחילים לחפש.

החיפוש יכול להיות פשוט כמו לקחת 64 רעשים ולבחור את הכי טוב. אבל הוא גם יכול להיות מתוחכם יותר, למשל, לנסות וריאציות על רעש נתון בכיוונים רנדומליים (שיטה שנקראת Zero-Order Optimization), או אפילו להוסיף רעש רק בחלק מהשלבים, ולאתחל מחדש את תהליך הדיפוזיה ממקום אחר במסלול (מהלך שהם קוראים לו Search-Over-Paths). במילים אחרות, מדובר כאן לא רק על שיפור איכות, אלא על גישה חדשה להבנת המסלולים שמודלי דיפוזיה צועדים בהם והאופן שבו הרעש משפיע עליהם.

אבל אולי המרכיב העמוק ביותר של המאמר הזה הוא מה שהוא לא מנסה לעשות. הוא לא טוען שצריך לשנות את המודל. הוא לא טוען שצריך לשפר את הרשת או לאמן אותה טוב יותר. כל החדשנות שלו טמונה בהכרה שעצם ההחלטה באיזה רעש להתחיל היא פרמטר אקטיבי בזמן ריצה. וזה חשוב, כי עד כה ההתמקדות בשלב ההיסק של מודלי דיפוזיה הייתה מוגבלת: ניסו רק לקצר אותו, לשפר את מסלול השחזור (הסרת רעש) אבל לא לגעת ברעש ההתחלה. המאמר הזה מפרק את ההנחה הזו.

במובן מסוים, מדובר כאן על הכנסת אלגוריתמיקה לשלב שאמור להיות פסיבי השלב שבו המודל כבר קיים ואנחנו רק "מריצים אותו". אבל ברגע שאנחנו מקבלים את הרעיון שאפשר לעשות אופטימיזציה בזמן הריצה על פרמטרים כמו הרעש, אנחנו פותחים דלת לא רק לשיפור תוצאות, אלא להבנה עמוקה יותר של המנגנונים הפנימיים של דיפוזיה.

ולכן לדעתי, התרומה המרכזית של המאמר הזה איננה בגרף כזה או אחר של FID. היא בשינוי החשיבה שהוא מגלם: ממודלים סטוכסטיים שמתפקדים כקופסה שחורה, למודלים שבהם הסטוכסטיות עצמה נעשית ניתנת לשליטה, לאופטימיזציה, ולעיצוב מחדש בזמן אמת. האם זו תהיה פרקטיקה רווחת? אולי רק במקרים יוקרתיים מאוד של גנרוט, שבהם כל שיפור קטן שווה הרבה. אבל כקונספט, זה צעד נוסף בהפיכת היסק מסטטית לאינטליגנטית, וזה מהלך מרתק בפני עצמו.

https://arxiv.org/abs/2501.09732