

סקירה זו היא חלק מפינה קבועה בה אני סוקר מאמרים חשובים בתחום ה-DL/ML, וכותב גרסה פשוטה וברורה יותר שלהם בעברית. במידה ותרצו לקרוא את המאמרים הנוספים שסיכמתי, אתם מוזמנים לבדוק את העמוד שמרכזו אוטם תחת השם deepnightlearners.

---

לילה טוב חברים, היום אנחנושוב בפינטנו deepnightlearners עם סקירה של מאמר בתחום הלמידה העמוקה. היום בחרתי לסקירה את המאמר שנkirא:

### Diffusion Models Beat GANs on Image Synthesis

---

#### פינת הסוקר:

**המלצת קריאה ממיק:** חובה למי שרצה למדוד מודלים גנרטיביים פרט לגאנים-VAE.

**בஹרות כתיבה:** ביןונית.

**רמת היכרות עם כלים מתמטיים וטכניקות של DL/ML הנדרשים להבנת מאמר:** הבנה טובה של עקרונות VAE, הבנה של שיטות דגימה מתקדמות כמו דינמיקה של לנגן.

**ישומים פרקטיים אפשריים:** יצירת תמונות יותר "aicottiyot" מהగישות המתחזרות, קרי גאנים-VAE.

---

#### פרטי מאמר:

**לינק למאמר:** [זמן להורדה](#).

**לינק לקוד:** [זמן כאן](#)

**פורסם בתאריך:** 21.06.01, בארכיב.

**הציג בכנס:** טרם ידוע.

---

## תחומי מאמר:

- מודלים דיפוזיוניים כלומרים - DDPM - Diffusion Denoising Probabilistic Models -  
לగנרטית של DATA ויזואלי.

## ידע מוקדם:

- הבנה טובה בטכניקות מבוססות variational inference לניתוח פונקציות נראות
- מירבית (כמו ב-VAE).
- רקע טוב בהסתברות לא-ז'יק (latent space knowledge).

---

## מבוא:

מודלים גנרטיביים מבוססי רשותות ניירונים לייצור DATA ויזואלי רשמו התקדמות מרשימה בשנים האחרונות. מודלים כמו [VQ-VAE2](#) ו- [StyleGAN2](#) מסוגלים לגורט תמונות מגוונות באיכות מרשימה בדומיננסים שונים. כרגע רוב המודלים הגנרטיביים עם תוצאות SOTA הם מסוגGAN ו-VAE (עם יתרון ניכר לגאנים). מלבד גאנים ו-VAEs קיימים סוגים נוספים של מודלים גנרטיביים מבוססים על גישות אחרות כמו מודלים דיפוזיאוניים ומודלים מבוססי זרימה (flow). עד כה מודלים אלו לא הצליחו (לפחות מבחינת המדרדים המקובלים כמו FID ו-Inception Score) להציג ביצועים בררי השווה עם תוצאות SOTA. נציין כי לפחות מבחינה ויזואלית איכות התמונות הנוצרות באמצעות מודלים דיפוזיוניים ומושגים זרימה לא נופלת מזו של אלו הנוצרות באמצעות גאנים ו-VAE-ים המתקדמים ביותר (דעה אישית).

המאמר הנסקר הוא הראשון (למייטב ידיעתי) שבו הצליח מודל דיפוזיאני להגיעה לביצועים טובים יותר ממודלים גנרטיביים, אשר נתונים כיום את התוצאות הטובות ביותר. זו בשורה ממשמעותית עד כדי כך שמחברי המאמר צינו אותה ישירות בכותרת (באנגלית):

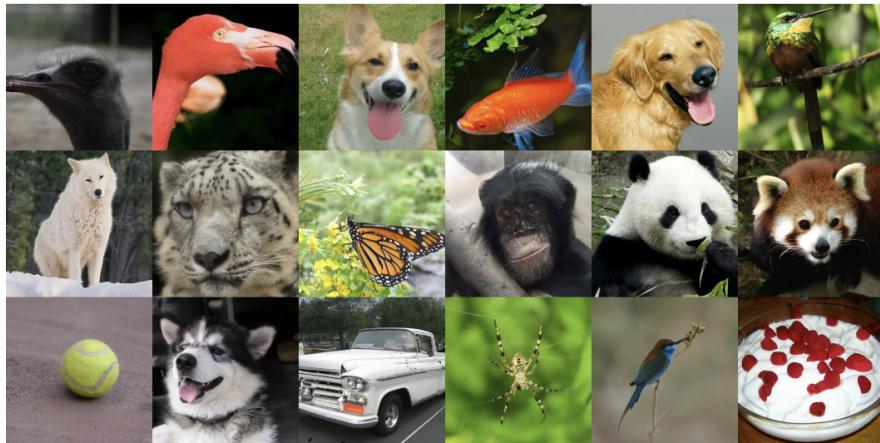


Figure 1: Selected samples from our best ImageNet 512×512 model (FID 3.85)

### תמצית מאמר:

המאמר הנסקר מtabסס על שני מאמרים קודמים ומציע שורת שיפורים שהצליחו "להרים את הביצועים של DDPM" לרמה שלGANים ומעבר לכך:

- מאמר [רקע 1](#) למעשה הציע את מה שנקרא **Probabilistic Model** או בקיצור **DDPM**. מעניין כי מודלים דיפוזיאוניים לגנרטוט נתונים הומצאו עוד ב-2015 ב- [מאמר רקע 0](#).
- מאמר [רקע 2](#) הציע רפרמטריזציה של פונקציית הלוס, שינוי של תהליך האימון (יפורט בהמשך) וכמה טרייקים נחמדים נוספים שבפועל שיפורו את יכולות התמונהות המוגנרטומות באמצעות המודל.
- המאמר הנסקר מציע דרך לנצל דата מתואג לאימון מודל דיפוזיאוני לצד כמה שיפורים ארכיטקטוריה של רשתות ניירונים המעורבות בתהליך הganרטוט.



Figure 6: Samples from BigGAN-deep with truncation 1.0 (FID 6.95, left) vs samples from our diffusion model with guidance (FID 4.59, middle) and samples from the training set (right).

כאמור, המאמר הנסקר מציג שורת שיפורים ל[מאמר רקע 2](#) שבעצמו מהוות גרסה משודרגת של [מאמר רקע 1](#). עקב כך אתחיל מסקירה מפורטת ועמינית של מודל דיפוזיאוני שהוצע ב[מאמר רקע 1](#), לאחר מכן אסקור את השדרוגים של [מאמר רקע 2](#) של המאמר הנסקר.

## תקציר מאמר רקע 1:

**מודל דיפוזיאוני DDPM לגינרט דאטה:** הרעיון של DDPM הוא די פשוט. לוקחים תמונה, מוסיפים אליה רעש גאוסי במשר כמה איטרציות (מאות או אלפיים) עד שהתמונה הופכת להיות לרעש גאוסי איזוטרופי  $(\mathbf{I}, 0)$  - זה נקרא תהליך קדמי (forward process). המטרה של מודל דיפוזיאוני הוא למדლ (ללמד) את התהליך הפוך (reverse process) - כלומר לגנרט תמונה מרעש גאוסי איזוטרופי צעד אחריו צעד.

**מטרת אימון DDPM:** המטרה היא למדיל את התפלגות  $Pr(x_t|x_{t-1})$  כאשר  $x_t$  היא התמונה המתקבלת באיטרציה  $t$  של התהליך הקדמי המתואר לעיל. באופן פורמלי, אם נסמן את התפלגות התמונות מהדעתהstry ב-  $(x_0|q) \sim x_0$ , אז התהליך הקדמי יתואר באופן הבא:

$$q(x_1, \dots, x_T | x_0) := \prod_{t=1}^T q(x_t | x_{t-1}) \quad (1)$$

$$q(x_t | x_{t-1}) := \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t \mathbf{I}) \quad (2)$$

כאשר  $\beta_t \in (0, 1)$  הם סדרה של קבועים דטרמיניסטיים ו-  $T$  מסמן את מספר האיטרציות של התהליך.

**חיזוי:** כמו שכנראה כבר ניחתט זמן החיזוי הוא עקב האכילס של מודלים דיפוזיוניים. כדי לבנות תמונה מרעש אנו צריכים לשחרר את כל הצעדים של התהלייר הפוך. המאמר הנסקר מדבר על בערך 4000 איטרציות המצריכים הריצה של 4K רשות אחת אחרי השניה שזה מבון מאוד בעיתוי.

**פרטים על הרעש המוסף:** תוחלת (פרק פיקסל) של רעש גausi המוסף בכל איטרציה תלויות בערך של הפיקסל. רעש המוסף עבור פיקסל  $\{x_t\}$  באיטרציה  $t$  מוגדר באמצעות התפלגות נורמלית  $(\beta_t, \sqrt{\alpha_t} x_{t-1}, \alpha_t = 1 - \beta_t)$  הינו ערך הפיקסל  $\{x_t\}$  בתמונה מורעתשת מאיטרציה  $t-1$ .

**נקודת חשובה: מידול של התהלייר הפוך** עשוי להיראות פשוט לאור העובדה שהטהלייר הקדמי (המתואר באמצעות התפלגות  $(x_t | x_{t-1})_q$ ) מתפלג גausi. אולם השערור של  $(x_t | x_{t-1})_q$  אינו מושימה פשוטה והתפלגות זאת אינה גausi. הסיבה לכך היא שלהבדיל מהטהלייר הקדמי שהוא הוספה של רעש גausi בעל תוחלת ושונות ידועות לתמונה, התהלייר הפוך הוא למעשה ניקוי של תמונה מורעתשת מחלוקת של הרעש שיש בה (מכאן באה המילה denoising במשמעותה הנדרשת "הבנייה" של התפלגותות של תמונות בשם של המודל). כדי לבצע denoising כזה נדרשות "הבנייה" של המתפלגותות של תמונות המתקבלות בשלבים השונים של הטהלייר דיפוזיוני.

עקב המורכבות הטמונה במידול של  $(x_t | x_{t-1})_q$ , משערכים אותה באמצעות התפלגות גausi פרמטרית  $(x_t | x_{t-1})_q$  הממודלת ע"י, מי היה מנחש, רשות ניירונית. פורמלית:

$$p_\theta(x_{t-1} | x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (3)$$

כאן  $I_t \gamma = (t, x_{t-1}) \Sigma$  (כלומר הרשות חוזה רק את סקלר  $\gamma$ ).

**נקודת חשובה:** למה ניתן לקרב  $(x_t | x_{t-1})_q$  באמצעות  $p_\theta(x_{t-1} | x_t)$  גausi בדיק טוב?

הרי כבר אמרנו ש- ♦ "טומנת בה ידע על התפלגות התמונות" של הדאטסט עליו מאומן DDPM. מתרברר כי קירוב זה עובד טוב כאשר הרעש המוסף בכל שלב של הטהלייר קדמי הוא בעל תוחלות ושונות נמוכות מספיק (אחד מהמאמרי רקע מצין כי קיימת הוכחה של גausיות תחת תנאים מסוימים על התפלגות של  $(x_t | x_{t-1})_q$  אך לא ראייתי אותה).

**DDPM מול מודלים גנרטיביים אחרים:** ברמת העיקרון DDPM דומה למודלים גנרטיביים אחרים כמו GAN, VAE או מודלי זרימה שגם יוצרים תמונה מרעש. אבל כאן הדמיון בין גישות אלו נגמר כי הדריכים בהן הן מדילות מיפוי מרעש לתמונה הן מאוד שונות (למרות ShVAE ו-DDPM משתמשים ב-ELBO לבנייה של פונקציית המטרה שלהם).

**איך מאמנים מודל דיפוזיוני?** מטרת האימון של מודל דיפוזיוני היא מינימום לוג של נראות מירבית (likelihood) של הדאטסהט ביחס לוקטור פרמטרים  $\theta$ . כמובן לוג של נראות מירבית של דאטסהט נתון הוא סכום של  $\log(p_\theta(x))$  עבור כל התמונות  $x$  מהדאטסהט. בדומה ל-VAE (אך עם קצת סיבוך עקב איטרציות מרובות המעורבות בתהילר), משתמשים בחסם תחתון (ELBO) כדי לקבל את פונקציית מטרה  $L_{\text{vlb}}$  של בעיית אופטימיזציה עבור מודל דיפוזיוני:

$$\begin{aligned} L_{\text{vlb}} &:= L_0 + L_1 + \dots + L_{T-1} + L_T \\ L_0 &:= -\log p_\theta(x_0|x_1) \\ L_{t-1} &:= D_{KL}(q(x_{t-1}|x_t, x_0) \parallel p_\theta(x_{t-1}|x_t)) \\ L_T &:= D_{KL}(q(x_T|x_0) \parallel p(x_T)) \end{aligned}$$

כאן  $(x_T)p$  הוא רעש גaussiano איזוטרופי.

**הסביר על האיברים של  $L_{\text{vlb}}$ :**

- $L_1$  - מודד עד כמה "סביר" לקבל את התמונה המקורית  $x_0$  מתוך  $x_1$  שהתקבלה בשלב לפני האחרון של התהילר ההפכי.
- $L_t < T$  - מודד דמיון בין התפלגות המשערכת  $(x_{t-1}|x_t)p$  לבין "התפלגות האמיתית"  $(x_{t-1}|x_t, x_0)q$  הנדגמת לתמונה  $x_0$  מהדאטסהט.
- $L_T$  - מודד עד כמה  $x_T$ , המתקיים בשלב האחרון של התהילר הקדמי, "קרובה" (במשמעותו) לרעש Gaussiano איזוטרופי.

**טהילר אימון של DDPM בגודל:** פונקציית הלוא שلنנו היא סכום של  $T$  מחוברים אי שליליים. כדי למצוור אותה, דוגמים  $T \leq t \leq 0$  ומבצעים איטרציה של gradient descent על האיבר  $h_t$  של הסכום. כאמור אנו מאמנים רשת  $N_\theta$  כדי לחזות את התפלגות  $(x_{t-1}|x_t)p_\theta$  לכל  $t < T$ . בכל איטרציה מאפטמים את הפרמטרים של  $N_\theta$  כדי למצוור את הלוא  $L_t$  עבור  $t$  הנdegם ( $t$  מזון לתוך הרשת).

**פלט של הרשת:** הדרך הטבעית היא לאמן את הרשת לחזות את  $(x_t, t)$  מ- $I_t$  ו-  $\gamma_t := (x_{t-1}, t)$  תחולת ומטריצת הקוריאנס של  $(x_t|x_{t-1})p$ . אך ניתן גם לאמן  $N_\theta$  לחזות עוד פרמטרים המעורבים בתהילר (כמו התפלגות  $(x_0)p_\theta$  של התמונה המקורית  $x_0$ ) מהם (יחד עם  $x_t$ ) ניתן לגזר את  $(x_t, t)$  מ-  $I_t$  ו-  $\gamma_t$ .

**הערה:**  $\gamma_t$  לא נחזה באמצעות רשת ניורונים במאמר [פרק 1](#) אלא משתמשים בקירוב שלו - הסיבובות לכך יפורטו בהמשך.

מאמר רקע 1 בחר לאמן  $\theta$  כדי להזות פרמטר אחר שנייתן לגזר ממנה את  $(t, x_t)$  מ-  $x_0$  תוך שימוש בתוכנות של התהיליך הקדמי. כעת נרחב איר ניתן לעשות זאת. ניתן לבצע את רפרמטריזציה הבאה להתפלגות:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I}) \quad (8)$$

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon \quad (9)$$

כאשר  $\bar{\alpha}_t = \prod_{j=0}^t \alpha_j$  והוא רעש גauss סטנדרטי  $(0, 1)$ . אינטואיטיבית די ברור כי  $x_0$  מתפלג גaussית כי  $x_t$  נבנה מ-  $x_0$  באמצעות הוספת רעשים גaussים בעלי תוחלות שונות ושוריות ידועות. בנוסף מתקיים:

$$\tilde{\beta}_t := \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t \quad (10)$$

$$\tilde{\mu}_t(x_t, x_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}x_t \quad (11)$$

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}(x_t, x_0), \tilde{\beta}_t \mathbf{I}) \quad (12)$$

במאמר רקע 1 מאמנים  $\theta$  להזות את רעש  $\epsilon$  המוסף בפועל בשלב  $t$  (המחברים טוענים שזה מושפר את איכות התמונהות המיוצרות) שמננו ניתן לגזר  $(t, x_t)$  מ-  $x_0$  באופן הבא:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left( x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) \quad (13)$$

למעשה פונקציית LOSS שהרשות  $N$  מاؤמתת למצער היא:

$$L_{\text{simple}} = E_{t, x_0, \epsilon} [||\epsilon - \epsilon_\theta(x_t, t)||^2] \quad (14)$$

**הערה:** כמו שכבר ציינו המאמר לא משער  $\gamma$  אלא משתמש רק בקירוב  $\beta_t$  (שינוי של  $x_t|x_{t-1}$ ). למעשה ניתן לראות כי  $\beta_t < \bar{\beta}$  (ערכים דטרמיניסטיים) אך בפועל שימוש בכל אחד חסמים אלו הוביל לתוצאות מאוד דומות. צריך לציין שימוש ב-  $L_{\text{simple}}$  שקול למשך של המחברים בפונקציית המטריה המקורית  $L_{\text{vlb}}$  (זה נובע מהצורה של מרחוק KL בין התפלוגיות גaussיות).

## ארכיטקטורת רשת:

mbosst על זו של [Wide ResNet](#) + [PixelCNN](#) שהוא שילוב [U-Net](#). כדי לקודד את מספר איטרציה  $t$  משתמשים בקידוד מיקומי (positional encoding) מהמאמר המקורי על הטרנספורמרים (Attention is All You Need). המחברים גם משתמשים במנגנון self-attention בין שכבות קונבולוציה ברוחות שונות.

בכך סיימו לתאר את DDPM כמו שהוזג במאמר [רף 1](#). עת נעבור לשינויים שהוצעו למודל זה במאמר [רף 2](#) ובמאמר הנסקה.

## תקציר שיפורים/שינויים של DDPM:

למעשה יש ארבעה סוגים של שיפורים שבזכותם DDPM הצליח להcout את הגאנים:

### שינויים בפרמטרים של התהיליך הקדמי:

- [מאמר רף 2](#) (פרק 3.2): קבועי  $T \leq t$ ,  $\beta$  וקביעים באופן שונה. המחברים שמו לב כי השלבים האחרונים של התהיליך הקדמי יוצרים תМОנות רועשות מדי ולא תורמים לאיכות התמונה המוגנרטת. עקב כך הוצע קבוע קבועים אלו כדי "להאט הפיכת של תמונה לרעש".

### שינויים בפונקציית loss ובתהליך אימון של $N_\theta$ :

- [מאמר רף 2](#) (פרק 3.1): כאמור בגרסה המקורי של DDPM המחברים החליטו לא לשער שונות  $\gamma_t$  של  $x_{t-1|t}$  והסתפקו בשערור של תוחלתו (באופן עקייף דרך  $\bar{\beta}_t$ ). ההסבר שלהם לגבי למה זה עובד מספיק טוב היה טמון בעובדה כי  $\gamma_t < \bar{\beta}_t$  אך  $\bar{\beta}_t$  ו-  $\bar{\beta}_t$  הם מאוד קרובים עבור רוב ערכי  $t$ . [מאמר רף 2](#) נקט בגישה אחרת והציג רפרמטריזציה קמורה של  $\gamma_t = \exp(v \log(\bar{\beta}_t) - 1)$ , כאשר  $v \in (0, 1)$  ואימנו רשות לשערור של  $v$ . נציין כי פונקציית הלוס הקודמת  $L_{simple}$  לא מכילה את  $\gamma_t$  אך המחברים השתמשו צירוף לינארי של  $L_{vlb}$  ו-  $L_{simple}$  בתור פונקציית loss חדשה.
- [מאמר רף 2](#) (פרק 3.3) מחליף דגימה יונייפורמית ב-  $t$ - sampling的重要性 sampling. ההסתברות של בחירת ערך  $t$  פרופורציאונלית לערך  $L_t$  המומוצע. לטענת המחברים זה מקטין את התנודתיות של הגראדיאנטים שלהם.

- [המאמר הנסקה](#) משתמש בDATA מתואג לאימון של DDPM. הרעיון הוא לנצל תМОנות מתואגות ל"ניוט של מודל דיפוזיוני לכיוון" שבו תМОנות שהוא מייצר בתהיליך הופכי, יסוויגו עם הקטגוריה נכונה בWOODOT גובהה באמצעות מסוג מאומן מראש. ככלומר לכל ערך של  $t$  מאמנים רשות מסווגת  $N_{\varphi,t}$  שהפלט שלה עבור תמונה  $x_t$  (ማיטרציה  $t$ ) הוא  $p_\varphi(y|x_t)$

עבור קטגוריה  $y$ . במהלך האימון לתמונה בעלת קטגוריה  $y$ , "מתקנים" את ההתפלגות  $x_t|x_{t-1}$  באופן כך שההתמונות תקבלנה ערך גבואה של  $p_\varphi(y|x_{t-1})$ . במקרה לשער  $p_\theta(x_{t-1}|x_t, y) = Z p_\theta(x_{t-1}|x_t) p_\varphi(y|x_t)$  ( $Z$ -Dogmix מ- $\theta$ ) (זאת שאותם יכולים לנחש שעורך זה לא גמור קל ומערב מתמטיקה לא טריוויאלית (זה מבוסס על [score-based generative models](#) לדינמיקה של לנגבין). יותר פרטים נמצאים בפרק 4 של המאמר הנסקר.



Figure 3: Samples from an unconditional diffusion model with classifier guidance to condition on the class "Pembroke Welsh corgi". Using classifier scale 1.0 (left; FID: 33.0) does not produce convincing samples in this class, whereas classifier scale 10.0 (right; FID: 12.0) produces much more class-consistent images.

### שיפורים בארכיטקטורה של $N_\theta$ : (המעוניינים ביניהם)

- מגנון attention בעל רזולוציות רבות (multi-resolution).
- שימוש בבלוקים residual של BigGAN ל-up/downsampling.
- Adaptive group normalization (AdaGN)

**דיזון תהליכי החיזוי** שינוי בהגדרת תהליכי הופכי שמאפשר חיזוי מדויק של  $x_{t-1+m}$  עבור  $0 < m$ . שינוי זה מאפשר לדוגם את  $x_t$  כל  $m$  צעדים ול- $m$  גדול מקטין את זמן החיזוי באופן משמעותי. המתמטיקה העומדת מאחורי ההגדירה החדשה זו די לא טריוויאלית ובנוסף התהליכי הקדמי מאבד את המركוביות שלו כי  $x_t$  תלוי באופן מפורש גם ב-  $x_{t-1}$  וגם ב-  $x_0$ .

### הישגי המאמר:

כאמור המודל הדיפוזיאוני המוצע הצליח להוכיח את הగנים המוביילים מבחינת FID. זמן החיזוי עדין נותר די גבוה יחסית לגאנ אבל יש שיפור ניכר יחסית למודלים דיפוזיאוניים קודמים.

### ג.ב.

מאמר ממש מגניב המציריך הבנה עמוקה של 3 מאמרם שקדמו לו בנושא של מודלים דיפוזיאוניים (עוד שניים בנושא סמכים). המתמטיקה לא טריוויאלית אבל היה שווה את המאמץ.

#deepnightlearners

הפוסט נכתב על ידי מיכאל (מייק) ארליךソン, PhD, Michael Erlhson.

מיכאל עובד בחברת הסיבר Salt Security בתפקיד Principal Data Scientist. מיכאל חוקר ופועל בתחום הלמידה העמוקה, ולצד זאת מרצה ומנגיש את החומרים המדעיים לקהל הרחב.