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

לילה טוב חברים, היום אנחנו שוב בפינתנו deepnightlearners עם סקירה של מאמר בתחום הלמידה העמוקה. היום בחרתי לסקירה את המאמר שנקרא:

Taming Transformers for High-Resolution Image Synthesis

פינת הסוקר:

המלצת קריאה ממייק: חובה ללא ספק!

בהירות כתיבה: גבוהה

ידע מוקדם: הבנה טובה בגאנים, טרנספורמרים ו VQ-VAE די הכרחית להבנת המאמר

יישומים פרקטיים אפשריים: יצירת תמונות באיכות מרהיבה (לא פחות!!)

פרטי מאמר:

לינק למאמר: זמין להורדה.

לינק לקוד: זמין כאן (בנוסף יש עוד 3 מימושים "לא רשמיים")

פורסם בתאריך: 21.06.21, בארקיב (v3)

הוצג בכנס: CVPR 2021

תחומי מאמר:

- מודלים גנרטיביים ליצירת דאטה בתחום הויזואלי

כלים מתמטיים, מושגים וסימונים:

- VQ-VAE
- גאנים
- טרנספורמרים

תמצית מאמר:

Figure 1. Our approach enables transformers to synthesize high-resolution images like this one, which contains 1280x460 pixels.

מודלים גנרטיביים ליצירת פיסות דאטה חדשות בתחום הויזואלי הגיעו לתוצאות מרשימות ב-3 השנים האחרונות. מודלים גנרטיביים כמו StyleGAN3 ו-VQ-VAE2 מצליחים לגנרט תמונות באיכות מאוד גבוהה במגוון רזולוציות. יתרה מזו התמונות הנוצרות באמצעות מודלים אלו נראות ממש פוטוריאליסטיות וכבר לא ניתן להבחין בין תמונה מגונרטת ל"טבעית".

רוב המודלים הגנרטיביים בעלי ביצועי SOTA בדומיין הויזואלי הינם גאנים בעלי ארכיטקטורה מבוססת על שכבות קונבולוציה (למרות שבשנה האחרונה VAE-ים, מודלי דיפוזיה ו- גישות אחרות התחילו להחזיר להם מלחמה). ביצועים עדיפים של רשתות קונבולוציה בדומיין הזה נובעים מה-"inductive bias" האינהרנטי שמאפיין רשתות מסוג זה. Inductive bias של רשתות קונבולוציה מנצל תלויות לוקאליות חזקות הקיימות בתמונות הטבעיות. לעומת זאת לטרנספורמרים אין bias כזה שמקשה עליהם ללמוד את האופיינים של התפלגות הדאטה בדומיין התמונות הטבעיות. עקב כך רוב הרשתות מבוססות טרנספורמרים בדרך כלל:

- או מצוידות ב-backbone הבנוי משכבות קונבולוציה להפקת פיצ'רים "לוקאליים"

- או מוסיפים את ה-inductive bias לטרנספורמרים, כלומר נותנים יותר משקל לקשרים בין פאצ'ים קרובים בתמונה.

המאמר הנסקר שילב את שתי הגישות הנ"ל ועשה את הדבר הבא:

1. אימון של VAE שבו הן המקדד והן המפענח הינם רשתות קונבולוציה. למעשה המחברים השתמשו ב-VAE מקוונטט בו המרחב הלטנטי (המכיל פלטים של המקדד) הוא למעשה אוסף דיסקרטי של וקטורים הנקרא codebook; ארחיב על זה בהמשך).
2. שימוש בטרנספורמר (ובפרט במפענח שלו) עם תיבול קל של "inductive bias" המתאים לדומיין התמונות, בשביל ללמוד את ההתפלגות מעל המרחב הלטנטי הדיסקרטי.
3. גנרוט של תמונה מתחיל מיצירה של וקטור לטנטי באמצעות הטרנספורמר המאומן. לאחר מכן מזינים את הוקטור הנוצר לרשת המפענחת של הטרנספורמר ליצירת תמונה.

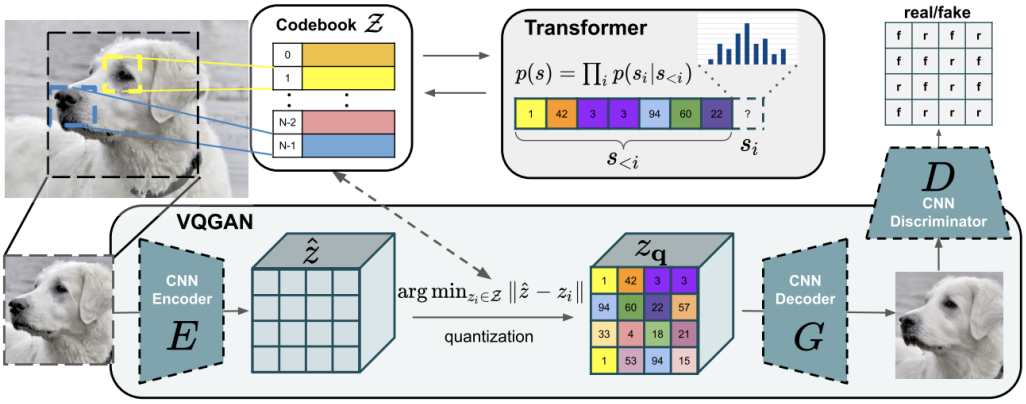

Figure 2. Our approach uses a convolutional *VQGAN* to learn a codebook of context-rich visual parts, whose composition is subsequently modeled with an autoregressive transformer architecture. A discrete codebook provides the interface between these architectures and a patch-based discriminator enables strong compression while retaining high perceptual quality. This method introduces the efficiency of convolutional approaches to transformer based high resolution image synthesis.

עד כאן הכל טוב ויפה אבל בשם של המאמר מופיע גם מילה GAN והוא לא הוזכר עד עכשיו. למעשה VAE מהשלב הראשון מאומן בצורה לא סטנדרטית. לפונקציית הלוס הרגילה של VQ-VAE (שגם עברה מתיחת פנים) מוסיפים את פונקציית הלוס הסטנדרטית של הגאן. כלומר **בנוסף** למקודד ולמפענח של VQ-VAE מאמנים דיסקרימינטור D. מטרתה של D היא לזהות אם התמונה שנוצרה באמצעות המפענח של VQ-VAE או שזו תמונה אמיתית.

למעשה VQ-GAN היא "חתונה משולשת" ומוצלחת במיוחד של VQ-VAE, גאן והטרנספורמר.

הסבר של רעיונות בסיסיים:

אחרי שהבנו מה מה הן אבני הבנייה העקריים של VQ-GAN בואו נצלול לפרטים נבין איך החיה המורכבת הזאת עובדת. בשביל להסביר את אופן האימון של VQ-GAN אנו קודם כל נרענן בזכרון מה זה VQ-VAE שעל בסיסו בנוי שלב האימון הראשון של VQ-GAN.

מה זה VQ-VAE?:

VQ-VAE הינו סוג של Variational AutoEncoder (VAE) בעל מרחב לטנטי סופי (אך מאוד גדול). נזכיר ש- VAE רגיל הוצע ב- 2014 על ידי Kingma ו- Welling. למעשה VAE מהווה הכללה של AutoEncoder סטנדרטי שהוא שיטה להורדת מימד לא לינארית. החידוש של VAE יחסית לאוטו-אנקודר הוא תוספת של הדרישה על התפלגות הייצוגים הלטנטיים של דאטה. כלומר בנוסף לכך ייצוג לטנטי של דאטה צריך לשמר את התכונותיו החשובות, וקטורי הייצוג עצמם צריכים להיות מפולגים לפי התפלגות נתונה (לרוב גאוסית סטנדרטית). תוספת זו מאפשרת לגנרט דאטה חדש באמצעות המפענח של VAE כאשר הקלט אליו הוא וקטורי ייצוג הנדגמים מהתפלגות זו. פונקצית לוס של VAE מורכבת מסכום של לוס השחזור הריבועי (המודד עד כמה טוב הצלחנו לשחזר את הקלט) ואיבר רגולריזציה הכופה התפלגות נתונה על הפלט של האנקודר (מרחק KL).

VQ-VAE הינו מודיפיקציה של VAE שבה המרחב הלטנטי (הנקרא codebook) הוא למעשה דיסקרטי ומכיל מספר סופי של וקטורי הייצוג. כדי לגנרט פיסת דאטה חדשה בוחרים וקטור מהמרחב הדיסקרטי הזה (שמכיל כמות עצומה של וקטורים ולכן בכל זאת מאפשר גנרוט של דאטה מאוד מגוון) ומעבירים אותו דרך המפענח המאומן.

כאשר משתמשים ב- VQ-VAE עבור יצירת תמונות בדרך כלל מחלקים תמונה ל- M פאצ'ים כאשר כל פאץ' מקדד באחד הוקטורים מה-codebook. במקרה הזה תמונה היא בעצם מערך באורך M של וקטורים מ-codebook (עם חשיבות לסדר כמובן!). למשל עבור M=8 (במציאות יש הרבה יותר פאצ'ים) ייצוג של תמונה יכול להיראות כך: [22, 46, 2, 11, 98, 17, 9, 78] כאשר כל איבר במערך זה הוא מספר סידורי של וקטור ייצוג מה-codebook. האימון של VQ-VAE הוא קצת טריקי כי בנוסף לפרמטרים של המקדד והמפענח צריך לאמן גם את וקטורי ה-codebook. וקטורים אלו נבחרים באמצעות **פעולה לא גזירה** מהפלט **z** של המקדד (בוחרים את הוקטור הכי קרוב **z** במונחי מרחק L2) שמקשה על ה-backprop. למי שמתעניין איך מתמודדים עם הסוגיה הזו ממליץ להעיף מבט ב- בבלוג מעולה של ברקלי.

המבנה של פונקציית לוס של VQ-VAE מורכב מלוס השחזור הריבועי של VAE הסטנדרטי והמרחק הריבועי בין פלט של המקדד והוקטור הקרוב מה-codebook (למעשה זה טיפה יותר מורכב עקב הפעולה הלא גזירה שתוארה קודם).

פונקציית לוס של VQ-GAN:

המאמר הנסקר בחר להחליף את לוס השחזור הריבועי בסכום של:

- הלוס הסטנדרטי של גאנים (שכמובן מצריך אימון רשת דיסקרימינטור).
- הלוס הפרספטואלי (perceptual loss).

המאמר טוען כי פונקציית הלוס המוצעת באה לתת "טיפול שורש" בסוגיית הכואבת של VAE: התמונות המטושטשות (יחסית לגאנים למשל) שהוא מייצר. הסיבה לכך טמונה במבנה של איבר השחזור של פונקציית הלוס הריבועית של VAE, שממזערת את השגיאה **הממוצעת** הגורמת לתמונה המשוחזרת להיות קרובה לתמונה המקורית, **"אך רק בממוצע"**. תופעה זו, של קושי של רשתות עמוקות להתמודד עם תדרים גבוהים בקלט ובפלט, ידועה ונכתבו עליה לא מעט עבודות לאחרונה ([1], [2], [3], [4]).

הלוס הפרספטואלי:

כעת נסביר מהו הלוס הפרספטואלי L_{per}. המטרה של L_{per} היא למדוד דמיון בין הפיצ'רים של התמונה המשוחזרת לבין הפיצ'רים של התמונה המקורית. אבל אלו פיצ'רים ניקח בשביל השוואה הזו? הרי המטרה היא למדוד את "רמת פוטוריאליסטיות" של התמונה המשוחזרת אז הפיצ'רים צריכים לשקף את "המאפיינים החשובים" של התמונה המקורית. כדי להפיק פיצ'רים כאלו בדרך כלל לוקחים רשת מאומנת כמו VGG או ResNet50, ומחשבים מרחק (בד"כ L2) בין פלטים של השכבות שלהן עבור התמונה המקורית למשוחזרת.

אציין שהלוס של גאן מחושב כממוצע על פני כל הפאצ'ים של תמונה בדומה ל-PatchGAN. זה כופה על התמונה המגונרטת לא רק להיות כמה שיותר "דומה לאמיתית כמקשה אחת" אלא דורשת שדמיון זה יתקיים בכל פאץ'.

אני מאמין שסכום של שני לוסים אלו מאפשרים למפענח של VQ-VAE ליצור תמונות מרהיבות.

"למידת" מרחב לטנטי של VQ-GAN:

אוקיי, הצלחנו להפיק ייצוג חזק מתמונה המהווה מערך של וקטורים מה-codebook (כל וקטור מיוצג ע"י המספר הסידורי שלו) כאשר כל וקטור מהווה ייצוג של פאץ' של התמונה. בשלב השני של אימון VQ-GAN המטרה היא לאמן מודל ליצירה של ייצוגים אלו (סדרות של ייצוגי פאצ'ים). כך נוכל להשתמש במודל זה לגנרוט של ייצוג לטנטי של תמונה שמוזן לאחר מכן למפענח ליצירת תמונה.

איך עושים זאת? מאחר וניתן ליצור תמונה באופן אוטורגרסיבי (פאץ' אחרי פאץ') מאמנים מפענח של הטרנספורמר (המאמר השתמש בארכיטקטורה דומה לזו של GPT2) בשביל לחזות ייצוג לטנטי (מספרו הסידורי ב-codebook) של פאץ' הבא בהינתן כל הפאצ'ים שכבר גונרטו. במשימה פאצ'ים של תמונה הם "משחקים תפקיד של טוקנים" של משימות של השפה הטבעית.

מעשית **לאחר סיום אימון של VAE בשלב הראשון**, לוקחים את כל הייצוגים הלטנטיים של התמונות מהדאטסט ומאמנים דקודר של הטרנספורמר לחזות ייצוג של פאץ' בהינתן הפאצ'ים הקודמים

הערה: לפני תחילת שלב האימון השני, "מקפיאים" את כל הפרמטרים של המקדד, המפענח ואת ה-codebook.

"תיבול" של inductive bias:

רגע, אבל מה עם התבלין מסוג inductive bias שהבטחתי קודם לכן? המחברים מצאו כי שימוש בפאצ'ים גדולים מ-16x16 פוגע בביצועים של המודל. מצד שני עקב משאבי החישוב המוגבלים שעמדו לרשותם, הם לא הצליחו לאמן טרנספורמר עבור יותר מ-256 פאצ'ים. איך יוצאים מהמצב הזה ומגנרטים תמונות גדולות יותר מ-256x256? פשוט משתמשים רק בפאצ'ים הקרובים לפאץ' הנחזה - והנה קיבלתם ה-inductive bias המובטח :)

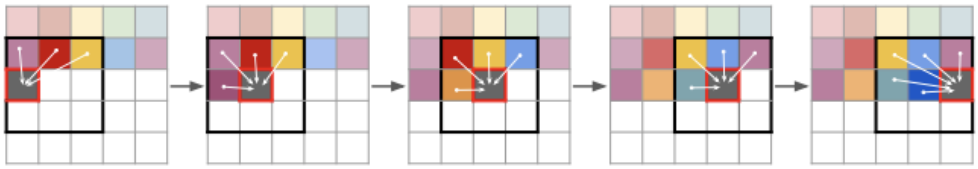

Figure 3. Sliding attention window.

סיכום קצר של שלבי אימון VQ-GAN:

- מאמנים VQ-VAE כאשר פונקצית הלוס היא שילוב הלוס הסטנדרטי של גאן והלוס הפרספטואלי (perceptual loss).
- מקפיאים את כל הפרמטרים של כל הרשתות שאומנו בשלב הראשון
- לוקחים את כל הייצוגים הלטנטיים של התמונות מהדאטהסט
- מאמנים מפענח של הטרנספורמר לחיזוי הוקטורים הלטנטיים מהשלב הקודם
- **מגנרטים תמונות:** יוצרים ייצוג לטנטי של תמונה באמצעות מפענח מאומן של הטרנספורמר ומעבירים ייצוג זה דרך המפענח של VAE שאומן בשלב הראשון **ונותר ללא שינוי מאז**

הישגי מאמר:

המחברים הראו שיפור מבחינת Frechet Inception Distance (FID) ו- Inception Score (IS) מול מודלים גנרטיביים חזקים עבור כמה דומיינים ורזולוציות.

Model	acceptance rate	FID	IS
mixed k, $p = 1.0$	1.0	17.04	70.6 ± 1.8
$k = 973$, $p = 1.0$	1.0	29.20	47.3 ± 1.3
$k = 250$, $p = 1.0$	1.0	15.98	78.6 ± 1.1
$k = 973$, $p = 0.88$	1.0	15.78	74.3 ± 1.8
$k = 600$, $p = 1.0$	0.05	5.20	280.3 ± 5.5
mixed k, $p = 1.0$	0.5	10.26	125.5 ± 2.4
mixed k, $p = 1.0$	0.25	7.35	188.6 ± 3.3
mixed k, $p = 1.0$	0.05	5.88	304.8 ± 3.6
mixed k, $p = 1.0$	0.005	6.59	402.7 ± 2.9
DCTransformer [48]	1.0	36.5	n/a
VQVAE-2 [61]	1.0	$\sim$31	$\sim$45
VQVAE-2	n/a	$\sim$10	$\sim$330
BigGAN [4]	1.0	7.53	168.6 ± 2.5
BigGAN-deep	1.0	6.84	203.6 ± 2.6
IDDPM [49]	1.0	12.3	n/a
ADM-G, no guid. [15]	1.0	10.94	100.98
ADM-G, 1.0 guid.	1.0	4.59	186.7
ADM-G, 10.0 guid.	1.0	9.11	283.92
val. data	1.0	1.62	234.0 ± 3.9

Table 4. FID score comparison for class-conditional synthes
on 256×256 ImageNet, evaluated between 50k samples and th
training split. Classifier-based rejection sampling as in VQVAE-
uses a ResNet-101 [22] classifier. BigGAN(-deep) evaluated vi
https://tfhub.dev/deepmind truncated at 1.0. "Mixed
k refers to samples generated with different top-k values, here k
$\{100, 200, 250, 300, 350, 400, 500, 600, 800, 973\}$.

נ.ב.

ממש אהבתי את המאמר כי הוא משלב גישות מאוד מעניינות בלמידה עמוקה: VQ-VAE, GAN וטרנספורמרים וגם מנצל את ה-inductive bias הקיים בדומיין הויזואלי. מומלץ בחום רב!

מילות תודה: ברצוני להודות לעדו בן-יאיר על עזרתו בהכנת סקירה זו.

#deepnightlearners

הפוסט נכתב על ידי מיכאל (מייק) ארליכסון, PhD, Michael Erlihson.

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