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:

TransGAN: Two Transformers Can Make One Strong GAN

פינת הסוקר:

המלצת קריאה ממייק: חובה בהחלט (בכל זאת גאן ראשון מבוסס על טרנספורמרים).

בהירות כתיבה: גבוהה.

רמת היכרות עם כלים מתמטיים וטכניקות של ML/DL הנדרשים להבנת מאמר: הבנה טובה בטרנספורמרים וידע בסיסי בגאנים.

יישומים פרקטיים אפשריים: TransGAN יודע לייצר תמונות כמו כל גאן אך בינתיים התוצאות אינן נראות בקנה מידה של SOTA בתחום כמו StyleGAN2.

פרטי מאמר:

לינק למאמר: זמין להורדה.

לינק לקוד: זמין כאן.

פורסם בתאריך: 16.02.21, בארקיב.

הוצג בכנס: טרם ידוע

תחומי מאמר:

- טרנספורמרים (Transformers)
- גאנים (GANs)

כלים מתמטיים, מושגים וסימונים:

- טרנספורמר לתמונות (visual transformers).
- שיטות אוגמנטציה גזירות (differentiable augmentations).
- הוספה של משימה self-supervised (סופר-רזולוציה) לתהליך אימון.
- אתחול לוקאלי של משקולות self-attention.
- Frechet Inception Distance (FID).
- Inception Score.

תמצית מאמר:

כפי שאתם בטח יודעים, ב-3 השנים האחרונות הטרנספורמרים השתלטו על עולם ה-NLP. בעקבות המאמר המפורסם "Attention is All You Need", רובם המוחלט של מאמרי ה-NLP משתמשים בארכיטקטורה של הטרנספורמר בצורה זו או אחרת. בשנה האחרונה הטרנספורמרים החלו את פלישתם גם לתחום הראיה הממוחשבת (לדוגמא An image is worth 16×16 words, המאמר שסקרתי לאחרונה Pretrained Image Transformer). הטרנספורמרים הצליחו להפיק ייצוגים (representations) חזקים לתמונות המשמשים לאחר מכן למגוון משימות דיסקרימינטיביות.

המאמר הנסקר מנסה להמשיך לקדם את מהפכת הטרנספורמרים לדומיין הויזואלי ומציע מודל גנרטיבי ראשון שהארכיטקטורה שלו מורכבת מהטרנספורמרים בלבד – ללא שימוש בקונבולוציות. בניית מודל גנרטיבי טוב בדומיין התמונות ללא קונבולוציות זה אכן דבר די מהפכני. הרי הקונבולוציות מהוות כלי אולטימטיבי להפקת פיצ'רים מהתמונות, המנצלות את התלות הלוקאלית החזקה שקיימת באופן אינהרנטי בתמונות. המאמר מצליח להסתדר בלעדיהן וזו אכן בשורה גדולה, אולם יש כאן קאטץ' קטן. המחברים מצהירים באופן מפורש שהארכיטקטורה שלהם "נטולת קונבולוציות" (CNN-free), ואכן אתם לא תמצאו שם שכבות קונבנציונליות. אבל, וזה אבל די גדול, לקראת סוף הסקירה אסביר איך הם בכל זאת הצליחו להכניס "חיה מאוד דומה ל- CNN" בדלת האחורית של המודל שלהם.

הסבר של רעיונות בסיסיים:

המאמר מציע מודל של גאן (GAN) ליצירה של תמונות שהגנרטור והדיסקרימינטור שלו מבוססים על הטרנספורמרים.

קצת רקע על גאנים: כפי שאתם זוכרים, גאן מורכב מרשת הגנרטור G, שמטרתה ליצור תמונות ורשת הדיסקרימינטור D, שמטרתה להבחין בין תמונות אמיתיות לבין אלו שנוצרו ע"י הגנרטור G (מבצע משימת סיווג בינארית). G מנסה לבלבל את הדיסקרימינטור ולגרום לו לסווג את התמונות

שהוא יוצר כאמיתיות. במילים אחרות, הגנרטור מנסה לשפר את איכות הדוגמאות שהוא יוצר על סמך הציון שהוא מקבל מהדיסקרימינטור D.

כאמור, המאמר הנסקר מציע להיפטר מקונבולוציות שהתרגלנו לראות הן בגנרטור G והן בדיסקרימנטור D (קונבולוציות משוחלפות – transposed convolutions). במקום זאת, המאמר מציע לבנות את G ואת D מהטרנספורמרים וגם להוסיף רובד נוסף לתהליך האימון של הגאן שלהם, שקיבל באופן לא מפתיע את השם TransGAN.

קודם כל, בואו נבין איך ניתן לגנרט תמונה באמצעות הטרנספורמר.

מבנה הגנרטור:

הקלט לגנרטור הינו וקטור רעש גאוסי z כמו שמקובל גם בגאנים הסטנדרטיים. לאחר מכן התמונה נבנית באופן הבא:

- מעבירים את z דרך MLP בשביל לבנות את התמונה ברזולוציה נמוכה (8×8) כאשר כל פיקסל מיוצג ע"י כמות גדולה של ערוצים, המסומנת כ- C.
- לוקחים את הווקטורים המתאימים לכל ערוץ ומכניסים אותם למקודד (encoder) של טרנספורמר (כל וקטור כאן מייצג פיצ'רים של פיקסלים בתמונה שתיווצר בהמשך) ביחד עם הקידוד המיקומי הנלמד (learnable positional encoding). בסך הכל זה מאוד דומה לאיך שאנחנו עובדים עם הטרנספורמר במשימות NLP, כאשר שם אנו מזינים לטרנספורמר וקטורים המייצגים מילים (או תתי-מילים).
- מבצעים upsampling 2×2 באמצעות שיטת pixelshuffle. כתוצאה מכך מתקבלים וקטורים המייצגים פאטצ'ים של תמונה בגודל 16×16, עם מחצית הערוצים המקוריים C/2.
- חוזרים על שני השלבים האחרונים ומקבלים כתוצאה מכך וקטורים של פאטצ'ים עבור תמונה בגודל של 32×32, עם C/4 ערוצים.

Figure 1. The pipeline of the pure transform-based generator and discriminator of TransGAN. Here $H = W = 8$ and $H_T = W_T = 32$. We show 9 patches for discriminator as an example while in practice we use 8×8 patches across all datasets.

- מפעילים הטלה לינארית על הערוצים הנותרים בשביל לבנות תמונה בגודל 32x32x3.

מבנה הדיסקרימינטור:

מכיוון שהדיסקרימינטור צריך בסך הכל להבחין בין תמונה סינתטית (המיוצרת על ידי גנרטור) לתמונה אמיתית, מספיק לקחת פאטצ'ים של תמונה ולהכניס אותם למקודד של הטרנספורמר (עם משקלים משלו כמובן). וקטור ייצוג של פאטץ' מחושבים באמצעות טרנספורמציה לינארית של הפיקסלים של פאטץ'. לאחר מכן לווקטורי ייצוג אלו מתווסף קידוד מיקומי נלמד, והם מוכנסים למספר מקודדים של טרנספורמר אחד אחרי השני. נציין שבדומה ל-An image is worth 16×16 words מוסיפים לוקטורי ייצוג טוקן [cls], שמשמש בסופו של דבר לסיווג של תמונה.

איך מאמנים TransGAN?

כאן המחברים עשו משהו מעניין. כנראה שבהתחלה הם ניסו לאמן את TransGAN כמו שמאמנים גאנים רגילים אבל התוצאות היו מאכזבות (ניחוש שלי). בניסיון להבין את מקור הביצועים החלשים הם החליפו את הגנרטור והדיסקרימינטור של TransGAN (לסירוגין) באלו המבוססים על רשתות קונבולוציה (מ-AutoGAN, WGAN-GP ו-StyleGAN v2), וגילו שמקור החולשה נמצא דווקא בדיסקרימנטור שלא מצליח "לנווט" את הגנרטור שייצור תמונות איכותיות. עקב כך המחברים הוסיפו כמה אלמנטים לתהליך האימון של TransGAN שבפועל שיפרו את ביצועיו בצורה ניכרת.

תוספות לתהליך האימון:

1. שימוש בטכניקות אוגמנטציה כבדות: זאת, על מנת ליצור כמות מאוד גבוהה של דוגמאות. הסיבה לכך כנראה טמונה בעובדה שלאחר הסרה של שכבות הקונבולוציה, human-designed bias, המנצל את התכונות האינהרנטיות של דומיין התמונות, TransGAN לא מצליח ללמוד את התכונות האלו בצורה טובה מספיק.
2. אימון משותף של TransGAN עם משימה self-supervised: בנוסף לאימון הרגיל של גאן, המחברים הציע לאמן אותו למשימה של סופר-רזולוציה. כלומר מורידים את הרזולוציה של תמונה (downsampling) מסט האימון ומנסים לשחזר את התמונה המקורית תוך כדי הוספה של לוס השחזור (MSE) ללוס הרגיל של גאן.

Table 1. Inception Score (IS) and FID results on CIFAR-10. The first row shows the AutoGAN results (Gong et al., 2019); the second and thirds row show the mixed transformer-CNN results; and the last row shows the pure-transformer GAN results.

GENERATOR	DISCRIMINATOR	IS↑	FID↓
AUTOGAN	AUTOGAN	8.55± 0.12	**12.42**
TRANSFORMER	AUTOGAN	**8.59± 0.10**	13.23
AUTOGAN	TRANSFORMER	6.17± 0.12	49.83
TRANSFORMER	TRANSFORMER	6.95 ± 0.13	41.41

Table 2. The effectiveness of Data Augmentation (DA) on both CNN-based GANs and TransGAN. We used the full CIFAR-10 training set and DiffAug (Zhao et al., 2020b).

METHODS	DA	IS ↑	FID ↓
WGAN-GP	×	**6.49** ± 0.09	39.68
(GULRAJANI ET AL., 2017)	√	6.29 ± 0.10	**37.14**
AUTOGAN	×	8.55 ± 0.12	**12.42**
(GONG ET AL., 2019)	√	**8.60 ± 0.10**	12.72
STYLEGAN V2	×	9.18	11.07
(ZHAO ET AL., 2020B)	√	**9.40**	**9.89**
TRANSGAN	×	6.95 ± 0.13	41.41
	√	**8.15 ± 0.14**	**19.85**

3. אתחול לוקאלי של משקלי מנגנון self-attention: אתם זוכרים שאמרתי לכם שלמרות שלא תמצאו שכבות קונבולוציה ב-TransGAN, הן כן הוכנסו פנימה בדלת האחורית? תיכף אסביר זאת. שתי התוספות הראשונות לאימון (סעיפים 1 ו-2) הצליחו לשפר את הביצועים של TransGAN אך הוא עדיין נשאר מאחור שיטות SOTA מבחינת FID ו-IS. עקב כך המחברים הציעו לאתחל משקלים של מנגנון self-attention באופן הבא:

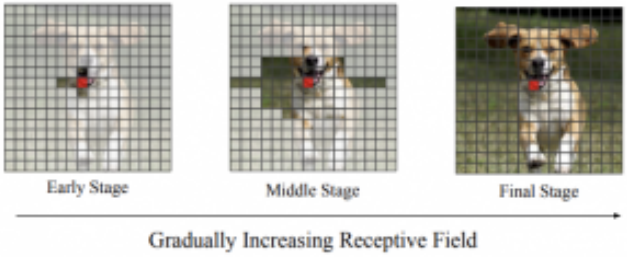

Figure 3. Locality-aware initialization for self-attention. The red block indicates a query location, the transparent blocks are its allowable key locations to interact with, and the gray blocks indicate the masked region. TransGAN gradually increases the allowable region during the training process.

באיטרציות הראשונות מאמנים רק את הקשרים הלוקאליים: כלומר מפעילים מסכה של אפסים על מטריצת משקלים של query כך שוקטור ייצוג של טוקן (פאטץ') "יראה רק את השכנים הקרובים שלו". זה קצת מזכיר את מה שעושים בדקודר של הטרנספורמר הקלאסי בשביל למנוע ממנו להתחשב בטוקנים העתידיים בטקסט בפענוח. כאן לעומת זאת מונעים מטוקן (= פאטץ') להתחשב

בפאטצים רחוקים ממנו. ככל שמתקדמים עם איטרציות האימון מחלישים את המסכות ונותנים ייצוגי פאטצ'ים להתחשב בפאטצ'ים רחוקים יותר. לקראת סוף האימון, מבטלים את המסכות לגמרי ומאמנים את משקלי ה-self-attention בצורה רגילה.

תוספת זו למעשה מאפשרת ל-TransGAN להגיע לתוצאות של שיטות SOTA, המוזכרות לעיל.

פינת האינטואיציה לאתחול לוקאלי של self-attention:

איך שיטת אימון זו קשורה לקונבולוציות אתם שואלים? התשובה פשוטה: כאשר מונעים מהטוקנים (פאטצ'ים) להתחשב בטוקנים רחוקים, אנו למעשה מעניקים ל-TransGAN את מה שנקרא human designed bias: אנו "רומזים" לו שקשרים לוקאליים מאוד חשובים בתמונות. למעשה, אותו bias מוביל אותנו להשתמש ברשתות מבוססות שכבות קונבולוציה כמעט לכל המשימות של הראייה הממוחשבת. כלומר את הקונבולוציות אנחנו לא רואים כאן, אך human-designed bias נותר בעינו.

הישגי מאמר:

TransGAN מצליח להגיע לביצועים דומים של שיטות SOTA חזקות כמו AutoGAN, WGAN-GP ו- StyleGAN v2 מבחינת FID ו- IS.

Table 5. Unconditional image generation results on CIFAR-10.

METHODS	IS	FID
WGAN-GP (GULRAJANI ET AL., 2017)	6.49 ± 0.09	39.68
LRGAN (YANG ET AL., 2017)	7.17 ± 0.17	-
DFM (WARDE-FARLEY & BENGIO, 2016)	7.72 ± 0.13	-
SPLITTING GAN (GRINBLAT ET AL., 2017)	7.90 ± 0.09	-
IMPROVING MMD-GAN (WANG ET AL., 2018A)	8.29	16.21
MGAN (HOANG ET AL., 2018)	8.33 ± 0.10	26.7
SN-GAN (MIYATO ET AL., 2018)	8.22 ± 0.05	21.7
PROGRESSIVE-GAN (KARRAS ET AL., 2017)	8.80 ± 0.05	15.52
AUTOGAN (GONG ET AL., 2019)	8.55 ± 0.10	12.42
STYLEGAN V2 (ZHAO ET AL., 2020B)	**9.18**	**11.07**
TRANSGAN-XL	8.63 ± 0.16	11.89

נ.ב.

מאמר מאוד מעניין המציע גאן ראשון מבוסס כולו טרנספורמרים המגיע לביצועי SOTA. תוצאה זו הושגה בזכות שימוש בכמה טריקים מעניינים במהלך האימון.

#deepnightlearners

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

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