# Review 32: GAN-Control: Explicitly Controllable GANs

**Paper: https://arxiv.org/abs/2101.02477v2**

פינת הסוקר:

המלצת קריאה ממייק: חובה לאוהבי גאנים.

בהירות כתיבה: טובה מאוד.

רמת היכרות עם כלים מתמטיים וטכניקות של ML/DL הנדרשים להבנת מאמר: נדרשת הבנה טובה בארכיטקטורות עכשוויות של הגאנים (StyleGAN2) וידע בסיסי בנושא אימון של הגאנים. בנוסף נדרשת הבנה בסיסית של עקרונות הלמידה הניגודית.

יישומים פרקטיים אפשריים: יצירה של תמונות פוטוריאליסטיות בעלות מכלול מוגדר של פיצ'רים ויזואליים כגון גיל, תנוחת ראש, צבע שיער וכדומה בכמה דומיינים כמו תמונות פנים מצוירות ותמונות פרצופים של חיות.

פרטי מאמר:

פורסם בתאריך: 07.01.21, בארקיב

הוצג בכנס: לא ידוע

תחומי מאמר:

גאנים (GANs).

כלים מתמטיים, טכניקות, מושגים וסימונים:

למידה ניגודית (contrastive learning)

אימון של גאן עם פיצ'רים מופרדים בצורה מפורשת.

מבוא והסבר כללי על תחום המאמר:

ליצירה של תמונות פוטוריאליסטיות בהינתן פרמטרים ויזואלים (כמו גיל, צבע שער, תאורה וכדומה) באיכות גבוהה יש שימושים רבים במגוון תחומים כגון עיצוב גרפי, משחקי וידאו, קולנוע, תחום הצילומים הרפואיים ועוד. בשנים האחרונות נרשמו כמה בפריצות דרך בתחום הזה במיוחד בפיתוח מודלים יצירת צילומי פנים (face images) בעלות מכלול מוגדר של פיצ'רים ויזואליים. מודלים אלו בדרך כלל משתמשים בשיטות מידול 3D וסובלים לרוב מעלויות יצירה גבוהות ובעלות שונות נמוכה (יוצרות תמונות דומות אחת לשנייה). מצד שני מודלי גאנים עכשווים כמו StyleGAN2 מפגינים יכולת מרשימה ביצירת תמונות פוטוריאליסטיות באיכות מאוד גבוהה ובעלות יצירה סבירה אך מתקשות ליצור תמונות בעלות פיצ'רים ויזואליים נתונים. יש עבודות המשלבות את שתי הגישות האלו ומצליחות ליצור תמונות פנים פוטוריאליסטיות באיכות מרשימה בעלות תכונות נשלטות כמו תנוחה, תאורה וסוג הבעת פנים. אך מכיוון שמודלים אלו מתבססים על מידול 3D הם לא מאפשרים ליצור, למשל, תמונה של אדם בגיל מסוים כי מודלי 3D אינם מאפשרים זאת. בנוסף קשה להרחיב שיטות אלו לדומיינים קרובים כמו תמונות מצוירות או צילומים של פרצופי חיות אם אין ברשותנו מודלים של 3D המתאימים לתחומים אלו.

המאמר הנסקר מציע גישה, הנקראת GAN-Control, הנותנת מענה לחולשה זה ומציע שיטה המאפשרת ליצור תמונות בעלות תכונות ויזואליות מוגדרות בצורה מפורשת. השיטה מצליחה ליצור תמונות פוטוריאליסטיות ב 3 דומיינים שונים: צילומי פנים, תמונות מצוירות וצילומי פרצופים של חיות. הגישה שלהם מאפשרת ליצור תמונות מגוונות לגיל, תאורה, תנוחה וצבע שער נתונים בצורה מפורשת לתחומים הנ״ל.

השיטה המוצעת מורכבת משני שלבים עיקריים:

אימון של גאן עם פיצ'רים מופרדים בצורה מפורשת(explicitly disentangled features): בגדול (ההסבר המפורט יינתן בפרק הבא) מחלקים את המרחב הלטנטי הראשון Z (ראה הערה על המרחבים הלטנטיים למטה) לתת-מרחבים Z\_i כאשר כל תת מרחב אחראי על פיצ'ר ויזואלי ספציפי של התמונה המגונרטת (תת-המרחב האחרון אינו אחראי על פיצ'ר ספציפי ואחראי על שאר הפיצ'רים הלא נשלטים של תמונה)

אימון של רשת הממפה פיצ'רים ויזואליים נתונים (בצורה מפורשת כמו למשל גיל או זווית צילום של תמונת פנים) לתת-מרחב הלטנטי "שלה". כתוצאה מכך תת מרחב Z\_age , למשל, האחראי על הגילאים יחולק ל"איזורים" כך שכל ״איזור״ אחראי על גיל מסוים. שלב זה יאפשר לגנרט תמונה בעלת פיצ'רים ויזואליים נתונים:

הערה לגבי המרחבים הלטנטיים של GAN-Control:

המאמר משתמש בארכיטקטורה של StyleGAN2 שיש לו שני מרחבים לטנטיים Z ו- W כאשר הראשון הינו מרחב הלטנטי "הגולמי" הסטנדרטי של הגאנים והשני W הינו מרחב הסגנון הלטנטי עם פיצ'רים מופרדים האחראים על אספקטים ויזואליים שונים של התמונה המגונרטת). המיפוי מ- Z ל- W ממודל עי״ רשת MLP בעלת 8 שכבות(עם משקלים נלמדים כמובן). המאמר מציע למדל כל מיפוי מתת-מרחבים Z\_i ל W\_i עי״ רשת MLP שלה.

הסבר של רעיונות בסיסיים:

נדון כעת איך מאמנים את GAN-Control לגנרט תמונות מפיצ'רים מופרדים.

שלב אימון 1: המאמר מציע להשתמש בלוס הניגודי לכל פיצ'ר. בשביל כך כל באטץ' נבנה מזוגות של וקטורים לטנטיים z\_b המכילים תת-וקטור משותף (z\_bk המתאים לתת-מרחב הלטנטי Z\_k) כאשר שאר התת-וקטורים הינם שונים. נציין שעבור זוג דוגמאות כאלו, הגנרטור של GAN-Control אמור ליצור שתי תמונות דומות בפיצ'ר ה- k בלבד ושונות בשאר הפיצ'רים. בשביל לבנות הלוס הניגודי המאמר מגדיר מרחק בין זוג תמונות I\_1 ו- I\_2 מבחינת הפיצ'ר k, המסומן (D\_k(I\_1, I\_2 המודד דמיון בין התמונות בפיצ'ר k. למשל עבור הפיצ'ר שהוא זהות של אדם בתמונה D\_k מודד ״עד כמה I\_1 ו- I\_2 מתארות אותו אדם).

אז איך בעצם מחשבים את הלוס הניגודי כאן?

עבור זוג תמונות עם פיצ'ר k זהה, אנו מנסים להגדיל את הדמיון בין התמונות הנוצרות מבחינת פיצ'ר k. כלומר המטרה הינה להקטין את המרחק ביניהן, הניתן עי״ D\_k מוגדר כמקסימום בין הפרש של (D\_k(I\_1, I\_2 וקבוע c\_k , לבין אפס. כלומר אנו ״שואפים״ שהמרחק המקסימלי בין I\_1 ו- I\_2 מבחינת פיצ'ר k יהיה c\_k לכל היותר.

עבור זוג תמונות עם פיצ'ר שונה j, המטרה למקסם את המרחק בין התמונות מבחינת פיצ'ר זה. הלוס במקרה הזה מוגדר עי״ המקסימום בין אפס לבין ההפרש בין קבוע \*c\_j ו- (D\_j(I\_1, I\_2. בדומה לסעיף הקודם המטרה כאן לגרום למרחק D\_j להיות לכל הפחות \*c\_j.

פונקציית לוס מוגדרת כסכום של הלוסים על כל הפיצ'רים

בסוף, כמקובל בגאן, מוסיפים ללוס הניגודי את הלוס האדברסריאלי של StyleGAN2.

חישוב מרחקי D\_k: בשביל למדוד מרחק בין תמונות מבחינת פיצ'ר k המאמר משתמש בייצוג במימד נמוך M\_k של תמונה הנבנה עי״ רשת המאומנת למטרת זיהוי פיצ'ר k. למשל עבור דמיון בין תמונות מבחינת גיל, המאמר משתמש בייצוג מימד נמוך שנבנה עי״ הרשת לזיהוי גיל, למדידת דמיון בין תמונות מבחינת זהות של אדם המצולם, לוקחים את הייצוג מהרשת לזיהוי פנים (face detection). המאמר משתמש במרחקי L1, L2 ומרחק cosine למדידת מרחק בין וקטורי ייצוג של תמונות מבחינת פיצ'רים ויזואליים שונים.

שלב אימון 2: המאמר מציע לבנות מיפוי(לאמן רשת) נפרד לכל פיצ'ר k ויזואלי כמו גיל (20, 30, 40 ,…) או זווית צילום (0ֿ, 5, 15,…) למרחב ״סגנון״ לטנטי שלו W\_k. אבל איך מאמנים את הרשתות אלו? המאמר מציע את דרך אלגנטית לעשות זאת:

מרגילים מספר וקטורי z.

ממפים אותם למרחב ״סגנון״ W (זוכרים שכבר אימנו את המיפוי הזה בשלב הראשון).

מגנרטים תמונות מוקטורי הסגנון w ומעבירים את התמונות דרך הרשתות לזיהוי כל פיצ'ר y\_k.

בונים דאטהסט מזוגות (w\_k, y\_k) לכל פיצ'ר k.

מאמנים רשתות מקודדות E\_k הממפות y\_k ל- w\_k (גם לכל פיצ'ר בנפרד).

אז מה קורה בזמן האינפרנס? זה מאוד פשוט – מזינים פיצ'רים ויזואליים y\_k לרשתות מקודדות E\_k ובונים וקטורי הסגנון w\_k. בסוף משתמשים בגנרטור המאומן בשביל ליצור תמונה.

הישגי מאמר:

המחברים השוו את התמונות הנבנות עם GAN-Control עם אלו שנוצרו עם השיטות DFG ו- CONFIG (שיוצרות תמונות פנים עם פרמטרים ויזואליים נשלטים ומבוססות על גישות מידול 3D). המאמר הצליח להוכיח את עליונותה של GAN-Control על שיטות אלו ב- 3 דומיינים שונים בשני היבטים הבאים:

מרחק inception של פרשה (FID) משופר(נמוך יותר) המצביע בדרך כלל על תמונות יותר פוטוריאליסטיות.

דיוק משופר מבחינת התאמה של תמונה לפרמטרים הוויזואליים שאיתם היא נבנתה. למשל עבור תמונה הנוצרת עם זווית צילום של 30 מעלות, GAN-Control הצליח ליצור תמונות עם זוויות צילום קרובה יותר ל 30 מעלות בממוצע ועם שונות נמוכה יותר (זווית צילום של תמונה נמדדת עי״ רשת ייעודית מאומנת למשימה זו). דיוק מבחינת פרמטרים אחרים נמדד בצורה דומה.

דאטהסטים: FFHQ, MetFaces, AFHQ

נ.ב. המאמר מציע גישה מאוד יעילה ואינטואיטיבית ליצירת תמונות בעלות פיצ'רים ויזואליים נשלטים באיכות גבוהה יותר מהשיטות המתחרות. הגישה מסוגלת ליצור תמונות פוטוראליסטית בעלות תכונות ויזואליות נתונות ב 3 דומיינים שונים: צילומי פנים, צילומי פרצופי חיות ותמונות מצוירות. הקוד לא שותף כרגע (אני מניח שהמחברים טרם הספיקו למלא בקשת פטנט – זה חשוב לחברה כמו אמזון). אני די בטוח הקוד יפורסם ממש קרוב. אני גם מחכה לראות את השימושים של גישה זו בדומיינים נוספים.