לילה טוב חברים, היום אנחנו שוב בפינתנו DeepNightLearners עם סקירה של מאמר בתחום הלמידה העמוקה
היום בחרתי לסקירה את המאמר שנקרא:
Unsupervised Discovery of Interpretable Directions in the GAN Latent Space
שיצא לפני כ 4 חודשים.

תחום מאמר: GANs, חקר של המרחב הלטנטי של GANs

מאמר הוצג בכנס: **IMCL**

תמצית מאמר: המאמר מציע שיטה למצוא וקטורים (כיוונים) "ברי פירוש" (interpretable") במרחב הלטנטי ב GAN מאומן. כלומר אם נחבר וקטור בר פירוש לכל וקטור במרחב הלטנטי השינוי בין התמונות המגונרטות עי" וקטורים אלה יהיה רק <u>במאפיין אחד בלבד</u> של התמונה כמו צבע גוון עור, צורת גבות, רקע וכדומה. השיטה לא תלויה בארכיטקטורה של GAN ולא דרושת שום supervision(!!)

רעיון בסיסי: הנחת הבסיס של המאמר שכיוונים ברי פירוש שונים גורמים לטרנספורמציות בעלות שוני רב של התמונה שניתן להבחין בין טרנבפטרמציה אחד לאחרת. אז התהליך הלמידה הם מנסים לאתר (ללמוד) כיוונים במרחב הלטנטי הגורמים לטרנספורמציות בעלות שונות גבוהה.

תקציר מאמר: כמו שצוין במאמר כל המחקרים המתמקדים בחיפוש כיוונים ברי פירוש במרחב הלטנטי כולל supervision. מה שנהוג לעשות בדרך כלל זה לעשות כל מיני טרנספורמציות ברות פירוש לתמונות (סיבוב, הקטנה, הוספת משקפיים וכדומה) ולראות איזה כיוונים במרחב הלטנטי גורמים לשינויים האלו. ניתן לראות שבעיית חיפוש כיוונים ברי פירוש במרחב הלטנטי ניתן לגשת בדרך קונסטרוקטיבית קרי לבנות GAN עם גנרטור בעל פיצ'רים מופרדים (disentangled) כמו ב StyleGAN ודומיו. בדרך כללי זה מוביל לארכיטקטורה מאוד מורכבת וקשה לאימון של הגנרטור.

אז מה שהמאמר מציע הוא בעצם לקחת גנרטור מאומן ולנסות למצוא K (שווה בדרך כלל זה מימד של הקלט של הגנרטור) וקטורים ברי פירוש במרחב הלטנטי. זה נעשה בדרך מאוד פשוטה האינטואיטיבית. הם <u>מאמנים</u> רשת כאשר וקטורים ברי פירוש הם חלק מהשקלים שלה. למעשה הם לא עושים שום טרנספורמציה מפורשת לתמונה כמו שנהוג המאמרים הקודמים אלא רק "משחקים" עם המרחב הלטנטי. בגדול הם מגרילים וקטור במרחב הלטנטי, מגרילים את התוספת לוקטור הזה (המחושבת עי" שימוש במטרימה המורכבת מוקטורי ברי פירוש) , מגנרטים תמונות עם גנרטור מאומן עם הוקטור הראשון ועם הסכום שלהם, מזינים את התמונות שיצאו לרשת שמנסה לשערך את וקטור השני (התוספת). ביותר פירוט תהליך האימון נראה כך:

<u>ארכיטקטורת ואימון הרשת למציאת וקטורי ברי פירוש:</u>

1. מגרילים את מספר הכיוון k (בין 1 ל K) ויוצרים מזה וקטור one-hot
2. מגרילים גודל (אורך) a של וקטור הזה (מהתפלגות יוניפורמית סימטרית סביב 0)
3. מכפילים את הוקטור הזה במטריצה A עם <u>משקלים מאומנים</u> המכילה וקטורי בר פירוש שאנחנו מנסים לשערך
4. מגרילים וקטור z במרחב הלטנטי התפלגות גאוסית סטנדרטית
5. מגנרטים שתי תמונות, הראשונה מ z והשנייה מהסכום של z ושל הוקטור המתקבל בסוף שלב 3,עי" הזנתם לגנרוטור מאומן
6. מעברים את שתי התמונות האלו דרך רשת עם משקלים מאומנים R כאשר מטרתה לשערך את מספר הכיוון k ואת הגודל שלו a (ז"א הפלט של הרשת זה וקטור הסתברויות K-מימדי ומספר ממשי)

חשוב להבין שהמשקלים של R והמשקלים של מטריצת הכיוונים A מאומנים ביחד. עקב כך תוך כדי תהליך האימון העמודות של מטריצה A מנסות "לשפט" את בעיית הסיווג עבור שהרשת R מנסה לפתור, עי" התכנסות לכיוונים קלים יותר להבחנה.

פונקציית לוס: פונקציית לוס מורכבת משני מחוברים: הראשון הינו קרוס-אנטרופי סנדרטי על השערוך של k והשני זה הלוס הריבועי על השערוך של a. האיבר השני הוא מין רגולריזציה בשביל אורך של וקטור הכיוון ישפיע בצורה רציפה על התמונה המגונרטת כלומר למנוע מיפוי של כל הכיוונים לקבוצה קטנה של תמונות.

הערה לגבי מטריצת הכיוונים: הם ניסו לבחור מטריצת כיוונים משתי צורות - בעלת עמודות עם אורך 1 ומטריצה אורתונורמלית (עמודות אורתוגונליות בעלות נורמה 1). עבור דאטה סטים שונים אופציות שונות הציגו ביצועים יותר טובים אך לא מצאתי התייחסות לכך במאמר

ביצועים: בסופו של דבר חלק הכיוונים (לא מצאתי מה האחוז) שהתקבלו כתוצאה מהתהליך זה נמצאו גורמים לשינויים במאפיין אחד של התמונה, הניתנים להבחנה עי" עין אנושי כגון צבע שער, סיבוב של תמונה, גוון עור, אודם וכדומה). דבר מעניין שהם הצליחו לגלות שאחד הכיוונים שהם מצאו הינו אחראי על הרקע של התמונה שאיפשר להם לטעון שהם מצאו דרך לעשות אוגמנטציה טובה לדאטה סטים למשימות אוגמנטציה. אציין המאפיינים הניתנים להבחנה ויזואלית של שמתאימים לכיוונים שנמצאו משתנים (!!) בין מודלי GAN שונים ובין דאטה סטים שונים. אוסיף בנוסף להבחנה האנושית, מטריקה נוספת לשערוך ביצועים שהם השתמשו בה הינה דיוק שחזור (RCA) הכיוון k ברשת שלהם. כמובן ש ל RCA גבוה לא מעיד על כך שמצאנו כיוון בר פירוש חזק כי קיימים שילובים של כמה מאפיינים בתמונה קלים להבחנה אבל בשילוב עם ההבחנה האנושית כנראה עובד לרע (מצוין במאמר)

דאטה סטים: MNIST, AnimeFaces, Imagenet and CelebA-HQ

סוגי **GAN** שנבדקו:Spectral Norm GAN, ProGAN, BigGAN

לינק למאמר: מאמר

לינק לקוד: קוד

נ.ב. מאמר עם רעיון נחמד, אך קצת לא מבושל. מעבר לאינטואיציה הבסיסית לא מצאתי הסברים למה הרעין שלהם עבד. גם הייתי רוצה לראות דיון מעמיק בתלות בין הכיוונים (המאפיינים של תמונה) שנמצאו לבין הדאטה סט שעליו אומן GAN, ארכיטקטורה ותכונותיו האחרות. בקיצור נראה שהכיוון הזה רק בתחילת דרכו ומקווה לראות את ההמשך.