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:

Explaining in Style: Training a GAN to explain a classifier in StyleSpace

---

פינת הסוקר:

המלצת קריאה ממייק: כמעט חובה (לא חייבים אך ממש מומלץ).

בהירות כתיבה: גבוהה.

רמת היכרות עם כלים מתמטיים וטכניקות של ML/DL הנדרשים להבנת מאמר: נדרשת היכרות די מעמיקה עם עקרונות StyleGAN2 והבנה בסיסית במושגי Model Explainability.

יישומים פרקטיים אפשריים: המאמר מאפשר לאתר פיצ'רים ויזואליים, הגורמים לשינוי המשמעותי ביותר בהתפלגות התוצאה של רשת הסיווג עבור תמונה זו.

פרטי מאמר:

לינק למאמר: זמין להורדה.

לינק לקוד: לא הצלחתי לאתר.

פורסם בתאריך: 27.04.21, בארקיב.

הוצג בכנס: טרם ידוע.

תחומי מאמר:

- Model Explainability
- GANs

כלים מתמטיים, מושגים וסימונים:

- StyleGAN2
- Path length regularization Loss
- LPIPS
- KL divergence
- Style reconstruction loss

מבוא:

בשנים האחרונות רשתות נוירונים השתלטו על עולם הראייה הממוחשבת. רובן המוחלט של תוצאות SOTA במגוון רחב של משימות ראייה ממוחשבת הושגו באמצעות שימוש ברשתות נוירונים. אולם פתרונות אלו, ובפרט רשתות נוירונים המשמשות למשימות סיווג שונות - עדיין מהווים סוג של "קופסא שחורה", במובן שההחלטות של הרשת לא תמיד "מובנות" לבני אדם. למשל, קשה לנו להסביר אילו תכונות ויזואליות של תמונה הובילו לסיווג כזה או אחר על ידי הרשת. כלומר אנו לא תמיד יודעים מה גרם לרשת המסווגת לזהות חתול בתמונה בהסתברות גבוהה: צורה של אוזניים, שפם חתולי או צורה של אף.

קיימות שיטות המנסות "להסביר" את הסיווג המופק באמצעות הרשת (נקרא להן שיטות הסבר) עבור תמונה מסוימת על ידי זיהוי "איזורים" בתמונה, אשר המשפיעים באופן משמעותי על הפלט בשכבה האחרונה של הרשת (שכבת הסיווג). כלומר, שינויים ב"איזורים" אלו משנים באופן משמעותי את הסיווג, הניתן על ידי הרשת (קרי, ההסתברות הנחזית עבור אחת מקטגוריות הסיווג). איזורים אלו נקראים מפות חום (heatmaps).

לגישות כאלו יש שתי מגבלות עיקריות:

- שיטות אלו מזהות איזורים (אובייקטים) לוקליים של תמונה המשפיעים על החלטות הרשת באופן ניכר. אולם יכולתן של שיטות אלו לזהות "תכונות" (אטריבוטים - attributes) יותר גלובליות של תמונה כמו גדלים של אובייקטים שונים או צבעים, המשפיעות בצורה משמעותית על הסיווג, הינן מוגבלות.
- שיטות אלו מצליחות לזהות את האיזורים "החשובים לסיווג" של תמונה אך לא מספקים אינדיקציה איזה שינוי באיזורים הללו נדרש בשביל להביא לשינוי כזה או אחר של פלט הרשת.

שיטות הסבר ממשפחת CE - counterfactual explanations מתגברות על קשיים אלו באמצעות זיהוי תכונות (להבדיל מאיזורים, אטריבוטים) של תמונה, המשפיעות באופן ניכר על פלט הרשת. זיהוי זה נעשה באמצעות ניתוח של פלט הרשת עבור תמונות, השונות מתמונה נתונה בכמה תכונות בודדות בלבד. לבסוף נבחר מספר קטן של תכונות המשפיעות באופן מקסימלי על הסיווג הניתן ע"י הרשת.

באופן טבעי שיטות CE בתחום הויזואלי עושות שימוש נרחב ב-GAN-ים, הידועים ביכולתם ליצור תמונות מווקטור "פיצ'רים חבויים" **z** בעל מימד נמוך הרבה יותר מהתמונה. כדי לזהות אטריבוטים של תמונה אשר "חשובים" לזיהוי של קטגוריה מסוימת, ניתן "לשחק" עם רכיביו של וקטור **z** כדי לראות אלו מהם משפיעים על פלט הרשת עבור קטגוריה זו באופן המשמעותי ביותר. חשוב לציין כי כאשר "הפיצ'רים" (הרכיבים) של וקטור **z** הם מעורבבים (כלומר כל רכיב של וקטור "אחראי" על קומבינציה מסוימת של אטריבוטים ויזואליים של תמונה), קשה "לבודד" אטריבוט ויזואלי המשפיע ביותר על פלט הרשת.

תמצית מאמר:

כידוע StyleGAN היה אחד הגאנים הראשונים שהצליח "להפריד" (disentangle) את הפיצ'רים של וקטור קלט **z** כך שכל תת-קבוצה של רכיביו הינה "אחראית" על פיצ'ר ויזואלי מסוים של תמונה (כגון צבע שיער ועיניים, אורך שיער, גוון של עור). יותר ספציפית, בשלב הראשון וקטור קלט של StyleGAN מוזן לרשת, המפיקה ממנו את הפיצ'רים הויזואליים המופרדים (הפלט של רשת זו נקרא וקטור סגנון - style vector). עקב כך StyleGAN הופך לכלי עזר טבעי לבנייה של "מסביר החלטות של הרשת", המבוסס עם אטריבוטים ויזואליים.

המאמר הנסקר מציע שיטה, הנקראת StyleEx, שבליבה נמצא StyleGAN2, לזיהוי פיצ'רים ויזואליים של תמונה, המשפיעים ביותר על החלטה המופקת על ידי רשת מסווגת נתונה. נציין כי גאן, המאומן על דאטהסט נתון של תמונות, אינו בהכרח "יתפוס" פיצ'רים ויזואליים רלוונטיים למסווג נתון. למשל גאן, המאומן על דאטהסט תמונות של מכוניות עשוי שלא לגלות פיצ'רים משמעותיים למסווג של דגם של מכונית. כדי להתגבר על קושי זה, המאמר משלב את המסווג המאומן באימון של StyleGAN2. כך StyleGAN2 המאומן לומד "להפיק" את הפיצ'רים הויזואליים הרלוונטיים למסווג נתון. לאחר מכן, עבור כל קטגוריית סיווג y בוחרים את כל התמונות p_y המסווגות כ-y. בשלב האחרון מאתרים קבוצה של כמה פיצ'רים (style coordinates) שהשינוי בהם מקטין את ההסתברות הממוצעת של y מעל p_y.

הסבר של רעיונות בסיסיים:

כמו שתואר בפרק הקודם, האימון של StyleEx מורכב משני שלבים. כעת נתאר כל שלב בצורה מפורטת יותר:

שלב 1 : אימון משותף של StyleGAN2 ביחד עם רשת מסווגת נתונה C

נזכיר כי כדי לאמן גאן, אנו מאמנים יחד את שתי רשתות:

- רשת הגנרטור G, שמטרתה ליצור תמונה מוקטור (זה יכול להיות וקטור גאוסיאני או וקטור קבוע במקרה של StyleGAN שבו האלמנטים האקראיים "מוזרקים" ישירות לשכבות של G).
- רשת הדיסקרימינטור D, המאומנת כדי להבחין בין דוגמא מסט האימון לדוגמא מלאכותית שגונרטה על ידי G.

נזכיר שאנו מעוניינים "לעצב" את מרחב הפיצ'רים (הנקרא מרחב הסגנון ל-StyleGAN) כך שיכלול אטריבוטים רלוונטיים לרשת מסווגת נתונה. דרך אגב, ה"עיצוב" של מרחב הסגנון המקורי מתבצע באמצעות של טרנספורמציה אפינית (הנלמדת) של מרחב הסגנון המקורי של StyleGAN2. להשגת מטרה זו, המאמר מציע את השינויים הבאים ל-StyleGAN2:

- הוספת רשת מקודדת (Encoder) המיועדת לבנייה של וקטור סגנון מתמונה. הרשת המקודדת E מאומנת יחד עם G תוך כדי מזעור של לוס השחזור (תמונה מוזנת ל-E ולאחר מכן G משחזרת אותה והלוס מודד עד כמה טוב הצלחנו לשחזר את התמונה). כלומר E ו-G יחד מהוות auto-encoder.
- הגנרטור G מקבל את כקלט גם את הסיווג עבור התמונה שהוא מייצר. תוספת זו מאפשרת להכניס פיצ'רים, רלוונטיים לסיווג, למרחב הסגנון החדש.
- "לוס הסיווג" הוסף לפונקציית לוס של StyleGAN2. לוס זה מודד מרחק בין הסיווג של התמונה המקורית (הקלט ל-E) לבין הסיווג עבור תמונה, המגונרטת באמצעות G מהקלט של E. מרחק זה בין הסיווגים נמדד על ידי KL-divergence.

מבנה של פונקצית לוס של עבור שלב 1:

פונקצית לוס מורכבת מ-4 איברים:

1. הלוס האדברסרי (הלוגיסטי) הרגיל של גאן מהמאמר המקורי של Goodfellow et al.
2. הלוס שמטרתו לגרום לכך שכל שינוי של וקטור הסגנון יביא לשינוי פרופורציונלי בתמונה הנוצרת. כלומר שינוי קטן בוקטור הסגנון צריך לגרום לשינוי קטן בתמונה הנוצרת ממנו וככל שוקטור הסגנון משתנה יותר, השינוי בתמונה הנוצרת ממנו יהיה גדול יותר. כהערת אגב, לוס זה הוצע לראשונה ב-StyleGAN2.
3. לוס השחזור שהוסבר לעיל מורכב מ-3 המחוברים הבאים:

a. איבר המודד מרחק L1 בין התמונה המקורית לתמונה המשוחזרת (המתקבלת באמצעות העברתה של התמונה המקורית דרך המקודד E ולאחר מכן דרך הגנרטור G).
b. איבר LPIPS המודד מרחק perceptual בין התמונה המקורית למשוחזרת. איבר זה מודד מרחק בין ייצוגי התמונות הללו המתקבלים באמצעות רשתות מאומנות כמו VGG או SqueezeNet.
c. איבר המודד מרחק L1 בין וקטורי הסגנון של התמונה המקורית למשוחזרת. איבר זה הוא גרסא של לוס שחזור הסגנון שהוצע ב-StarGANv2.

4. מרחק KL בין הפלטים של הרשת המסווגת עבור התמונה המקורית והתמונה המשוחזרת.

שלב 2: איתור של אטריבוטים ויזואליים "המשפיעים" על הסיווג לקטגוריות

המטרה של שלב זה היא לאתר את הכיוונים במרחב הסגנון הגורמים לשינויים משמעותיים בפלט של רשת הסיווג. חיפוש זה מתבצע באופן הבא:

- לכל קטגוריית סיווג בוחרים את כל התמונות המסווגות עם קטגוריה זו על ידי הרשת המסווגת.
- מבצעים חיפוש של מספר (שנקבע מראש) קואורדינטות של וקטור הסגנון הגורמים לירידה הבולטת ביותר של ההסתברות הממוצעת של קטגוריה זו (המחושבת על ידי הרשת המסווגת).
- לכל קואורדינטה של וקטור הסגנון מוצאים את הכיוון של וקטור הסגנון (+1 או -1) הגורם לירידת הסתברות של קטגוריה זו.

נציין שניתן ליישם שיטה זו גם למציאת של הכיוונים "המשפיעים" ביותר של תמונה נתונה.

הישגי המאמר:

חייב להודות שהתוצאות לא פחות ממרשימות, לפחות מבחינה ויזואלית. StyleEx הצליח לזהות פיצ'רים ויזואליים לא מעורבבים קלים להבנה עבור מגוון רשתות מסווגות בתחומים מגוונים. למשל, הגישה המוצעת הצליחה לזהות פיצ'רים המשפיעים ביותר על זיהוי גיל, מין וגם על זיהוי מחלות של עיניים ואפילו מחלות של עלים.

נ.ב.

מאמר מגניב עם תוצאות מרשימות שבבסיסו רעיון הגיוני וקל להבנה בתחום ה-explainability של רשתות נוירונים.

#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.