לילה טוב חברים, היום אנחנו שוב בפינתנו DeepNightLearners עם סקירה של מאמר בתחום הלמידה העמוקה היום בחרתי לסקירה את המאמר שנקרא:
Robust Optimal Transport with Applications in Generative Modeling and Domain Adaptation
שיצא לפני שלושה שבועות.

תחום מאמר: מרחק בין דאטה סטים עם אווטליירים, מודלים גנרטיביים, אדפטצית דומיינים בלתי מונחית (unsupervised domain adaptation - UDA).

מאמר הוצג בכנס:NeurIPS 2020

כלים מתמטיים וסימונים: טרנספורט אופטימלי (OT), טרנספורט אופטימלי רובסטי (ROT) ,טרנספורט אופטימלי בלתי מאוזן (UOT), מרחק וסרשטיין (WD), מרחק f ומרחק chi-2 בין מידות הסתברות(f-divergence), בעיות אופטימיזציה מינימקס (minimax problems), פונקציות ליפשיץ עם מקדם 1 (1-Lip), דוגמאות לא טיפוסיות או אווטליירים (OL).

תמצית מאמר: המאמר מציע שיטה לחישוב מרחק בין דאטה סטים, הרובסטי לדוגמאות לא טיפוסיות (OL). בעצם מרחק זה מוגדר במאמר עבור כל שתי מידות הסתברות והמרחק בין דאטה סטים זה המקרה הפרטי שלו. מרחק זה נקרא ROT, הוא מבוסס על OT ומנסה להתגבר על רגישותו לדוגמאות OL. המאמר דן ברובו במקרה הפרטי של OT שזה מרחק וסרשטיין (WD) ואתמקד רק במרחק וסרשטיין הרובסטי (RWD) בהמשך הסקירה. הרגישות לדוגמאות OL ניתן לנסח באופן הבא: בהינתן שני דאטה סטים עם WD די נמוך, החלפתו של חלק מאוד קטן של דוגמאות באחד דאטה סטים בדוגמאות OL עלולה להוביל לעלייה בלתי פרופורציונלית ב- WD. לטענת המאמר מרבית הדאטה סטים הגדולים מכילים דוגמאות OL, ושימוש במרחק ביניהם שרגיש להם, עלול להביא לתוצאות ירודות במשימות שונות. בנוסף אימון GAN עם מטריקת מרחק כזו (כמו GAN של וסרשטיין) עלולה להוביל לכך ש GAN יגנרט "ערבובים" בין הדוגמאות הרגילות לבין דוגמאות OL.

רעיון בסיסי: אחת הדרכים להתמודד עם סוגייה זו היא משקול דוגמאות OL במטרה למזער את השפעתן על המרחק. טרנספורט אופטימלי בלתי מאוזן (UOT) משתמש ברעיון ומציע לשערך את המרחק בין התפלגויות 1_P ו- 2_P עי" המרחק בין שתי התפלגויות קרובות אליהם 1_Q ו 2_Q בהתאמה, עי" הוספת שני איברי רגולריזציה המכילים את סכום המרחקים 1_Div(P) ו- 2_Div(P_2, Q ). המרחק Div במקרה הזה מוגדר כמרחק(סטייה) f בין ההתפלגויות עבור פונקציה f מסוימת. הבעייתיות בגישה הזה נובעת בהיבט המימושי שלה. בדרך כלל לא פותרים את בעיית הטרנספורט האופטימלי בצורה ישירה אלא פותרים את הבעיה הדואלית שלה (הידועה כצורה של קנטרוביץ'-רובינשטיין). להבדיל מבעיית טרנספורט אופטימלי הסטנדרטית, הצורה הדואליות של UOT מכילה שתי פונקציות שאותן צריך לאפטם בו זמנית (כאשר הן תלויות אחת בשנייה בדרך די מורכבת) שמקשה מאוד על שימושו לבעיות פרקטיות כמו למשל אימון של GAN.

בשביל להתגבר על הקושי הזה ולשמר את הרובסטיות של המרחק לגבי דוגמאות OL, המאמר מציע לשנות את ניסוח בעיית אופטימיזציה של UOT באופן הבא: במקום לאפטם על כל ההתלפגויות ה "בערך שוות" ל 1_P טל 2_P הם מגבילים (מלמעלה) את המרחקים האלו עי" קבועים 1_rho ו- 2_rho. זה הופך את מרחק ROT תלוי באופן ישיר ב 1_rho ו- 2_rho אבל הופך את הבעיה הדואלית ליותר פשוטה ותלויה רק בפונקציה אחת (שהיא פונקצית ליפשיץ מסדר 1). מצד שני שמוסיף תנאי (מגבלה) לבעית אופטימיזציה הדואלית אך המאמר מוכיח שעדיין ניתן לפתור אותה בדרך יחסית נוחה.

תקציר מאמר: להמשך הדיון , ניזכר קודם כל מה זה OT והמקרה הפרטי של WD.

<u>טרנספורט אופטימלי:</u>OT הינו מרחק בין שתי מידות הסתברות P_2 ו P_1 המוגדרות על אותו מרחב X עבור פונקציית מחיר חיובית c נתונה. OT מודד עד כמה מידות הסתברות "קרובות" (כמו KL או JS). במקרה הפרטי שבו פונקציית מחיר הינה מרחק (בין שתי נקודות x ו y) בחזקה p כלשהי, OT נקרא WD מסדר p. כאשר p=1 המרחק הזה נקרא מרחק "מזיז הקרקע" (earth mover).

אז מה זה בעצם WD? בנוסחה מופיע מינימום על כל מידות הסתברות על מרחב המכפלה של X עם עצמו כאשר הפונקציות השוליות שלה הן מידות ההסתברות שעבורן אנו מחשבים את המרחק. ותחת סימן האינטגרל יש את המרחק בין הנקודות. לפשטות בואו ניקח p=1 והמרחק האוקלידי כמטריקת המרחק. בנוסף נניח שמרחב X הינן חד מימדי (R). למה זה בעצם נקרא מרחק מזיז הקרקע? בעצם המרחק הזה מגדיר כמה "מסה" אנו צריכים להעביר בשביל להפוך את המידה P_2 ל P_1 כאשר המחיר העברת הנקודה מהתומך P_2 לתומך של P_1 הינה אוקלידית. עכשיו למה המינימום, אתם שואלים? אפשר "להפוך את P_1 ל P_2 במספר דרכים ואנחנו רוצים את הדרך הכי ,קצרה. אז למה בעצם יש את מידת הסתברות על מרחב המכפלה של X? הפונקציה הזו מגדירה איזה "חלק" מהמסה ההסתברותית בנקודה x אני מעביר לנקודה y. כלומר אם יש לך x הסתברות 0.5 אני יכול להעביר שליש ממנה לנקודה y_1 ושני שליש הנותרים לנקודה 2y. התנאי שהפונקציות השוליות של של המידה הזו צריכים להיות P_2 ו P_1 נדרש כי אנו רוצים להעביר את כל המסה מכל הנקודות של P1_ לכל הנקודות של P_2 בלי לאבד (או להרוויח) מסה נוספת. להבדיל כמעט כל מרחק בין מידות ההסתברות מרחק וסרשטיין לוקח בחשבון של התכונות של הסטים שעליהם מידות אלו מוגדרות בצורה מפורשת עי" התחשבות במרחק בין הנקודות שלהם. ולבסוף הצורה הדואליות של WD היא בעצם בעית אופטימיזציה על כל פונקציות ליפשיץ מסדר 1 d על הפרש התוחלות של d תחת P_1 וי P_2.

עכשיו בואו נסביר איך ניתן להגדיר את WD על דיאטה סטים:

<u>WD על דאטה סטים:</u>

עבור שני דאטה סטים בגודל סופי ניתן להגדיר את מידות ההסתברות עליהם כסכום של פונקציות דלתא על הנקודות של הדאטה סט כאשר הסתברות של כל נקודה הינה שווה. המרחק בין כל הנקודות בדאטה סטים מוגדר כמטריצה ואז בעיית אופטימיזציה הופכת לבעיית תכנות לינארי (המידה על מרחב המכפלה שעליה מבצעים את האופטימיזציה ניתנת לתיאור עי״ מטריצה).

הדבר האחרון שנותר לנו זה להבין איך WD הרובסטי (RWD ) מוגדר על דטא סטים:

<u>בעיית אופטימיזציה עבור RWD:</u>

קודם כל נציין הפונקציות התפלגות קרובות ל P_1 ו P_2 זה בעצם משקול של הסתברויות בשני הדאטה סטים כאשר סכום המשקלים חייב להיות 1. אז בעיית אופטימיזציה שאנו פותרים כאן כוללת שני סטים של משקלים המסתכמים ל 1 ועל כל פונקציות lip-1. להבדיל מ- WD מתווספת כאן המגבלה על המרחקים בין ההתפלגויות של הסטים הממושקלים למקוריים (צריכים להיות קטנים מ 1_rho ו- 2_rho עבור שני הדאטה סטים). התנאים האלו זה פשוט מרחקי 2-chi בין שתי ההתפלגויות הממושקלות למקוריות על הדאטה סטים. הבעיה הזו למעשה הינה תכנות קוני מסדר שני ויש דרכים יעילות לפתור אותה. הבעיה שעבור דאטה סטים גדולים לפתרון זה עלולה להיות עלות חישובית מאוד גבוהה. אז מה שמחברי המאמר עשו הם עשו רפרמטריזציה של המשקלים ע" רשתות נוירונים כאשר הכניסה לרשת הינם דוגמאות מהדאטה סטים.

הישגי מאמר: המאמר השתמש ב RWD כדי לבנות GAN כאשר פונקצית המטרה של זה מינימיזציה של RWD עבור התפלגות הגנרטור והדאטה סט. הם הוכיחו שעבור דאטה סטים עם דוגמאות OL (או ״לכלכו דאטה סטים עם אחוז מסוים של תמונות מדאטה סטים אחרים) התמונות שגונרטו עם RWDGAN נראות יותר ״נקיות״ מבחינה ויזואלית אפילו עבור אחוז OL יחסית גבוהים. מעניין שכאשר מאמנים את RWDGAN על דאטה סטים נקיים (עם 1_rho ו 2_rho מסוימים - יש פה שאלה של איך לכייל אותם) אז IS ו- FID של התמונות המגונרטות איתו כמעט ולא השתנה יחסית לאימון עם WD רגיל. ההשוואות נעשו כאן עבור וסרשטיין GAN עם 3 ארכיטקטורות שונות.

דבר מעניין עם RWDGAN הוא שהמשקול האופטימלי של דוגמא נתונה בעצם משקפים את "רמת הקושי" של הגנרטור לגנרט אותה (כלומר עד כמה דיסקרימינטור הצליח "לפצח אותה"). אתם שואלים למה בעצם? אם המשקל של דוגמא הינו נמוך זה אומר שהגנרטור "החליט להנמיך בחשיבותה ולהקטין את השפעתה ללוס" מהסיבה שהוא חושב שהדוגמא הזו הינה OL. המאמר מראה בסטים "המלוכלכים" עם דוגמאות מדאטה סטים אחרים המשקלים את הדוגמאות "הזרות" יצאו נמוכות משמעותית מהרגילות.

בנוסף הם הראה ששימוש ב RWD עבור משימת UDA משפר באופן ניכר את ביצועי דיוק עבור 3ארכיטקטורות רשת שונות (עבור דאטה סט 17VISDA).

לינק למאמר:https://arxiv.org/pdf/2010.05862.pdf

לינק לקוד:https://github.com/yogeshbalaji/robustOT

נ.ב. המאמר עם רעיון די מעניין, מכיל גם הוכחות ריגורוזיות המסבירות למה הגישה שלהם עובדת. מה שמטריד אותי טיפה עם RWD זו הבחירה של פרמטר rho. הם מוכיחים במאמר שעם אחוז דוגמאות OL ידוע אז יש ביטוי לערך rho אופטימלי. ברוב המקרים זה לא המצב ובחירה של rho עלולה להיות לא טריויאלית.