לילה טוב חברים, היום אנחנו שוב בפינתנו DeepNightLearners עם סקירה של מאמר בתחום הלמידה העמוקה היום בחרתי לסקירה את המאמר שנקרא:

Geometric Dataset Distances via Optimal Transport

שיצא לפני כחודש.

תחום מאמר: אדפטציה של דומיינים(domain adaptation), חקר דמיון בין דאטה סטים, transfer learning

תמצית מאמר: המאמר מציע שיטה למדידת "דמיון" (מרחק) בין דאטה סטים מתויגים. השיטה המוצעת הינה אגנוסטית למודל, לא דורשת אימון, לא מחייבת שום דמיון בין הלייבלים ומתבסת על גישה הנקראת טרנספורט אופטימלי. הטענה במאמר שלמרחק זה יש קורלציה גבוה למידת התרומה של domain adaptation ביניהם לביצועים של הדאטה הסט השני (קרי אימון קלאסיפייר על אחד ורק כיול (fine-tuning) של השני).

כלים מתמטיים שהשתמשו בהם במאמר: אופטימל טרנספורט (optimal transport), ומקרה פרטי שלו הנקרא מרחק וסרשטיין (.(WD

תקציר מאמר: אני רוצה להתחיל הסבר על המושגים המתמטיים הנדרשים להבנת המאמר. נתחיל OT שזה המושג החשוב ביותר במאמר.

<u>טרנספורט אופטימלי: טרנספורט(OT)</u>) אופטימלי זה בעצם מרחק בין שתי מידות הסתברות P ו Q המוגדרות על אותו מרחב X לפונקציית מחיר חיובית c נתונה. זה נשמע קצת מפחיד אבל בסך הכל הדבר הזה 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 מרחק וסרשטיין מסדר p. כאשר p=1 המרחק הזה נקרא מרחק "מזיז הקרקע" (earth mover). אז בואו נבין מה זה בעצם מרחק וסרשטיין המתואר ע"י נוסחה (1) במאמר. יש שם משהו קצת מפחיד: יש שם איזה המינימום על כל מידות הסתברות על מרחב הפרודאקט של X עם עצמו כאשר הפונקציות השונות של המידה הזו הם מידות ההסתברות שעבורן אנו מחשבים את המרחק. ותחת סימון אינטגרל יש את המרחק בין הנקודות. בשביל להבין את הנוסחה זו 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 ל Q כאשר המחיר העברת הנקודה מהתומך P לתומך של Q הינה אוקלידית במקרה הזה. עכשיו למה יש שם מינימום, אתם שואלים? כמו שאתם מבינים אפשר "להפוך את P ל Q במספר דרכים ואנחנו רוצים את הדרך הכי זולה. אז למה בעצם יש את מידת הסתברות על מרחב הפרודאקט של X? הפונקציה הזו מגדירה איזה "חלק" מהמסה ההסתבורתית 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 ו Q נדרש כי אנו רוצים להעביר את כל המסה מכל הנקודות של P לכל הנקודות של Q 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.

<u>מציאת מרחק וסרשטיין:</u> למרות האינטואיטיביות הרבה שיש בהגדרה של WD מציאתו אינה טריוויאלית ברוב המקרים. עבור p=1 ניתן להשתמש (כמו שעשו ב Wasserstein GAN) בתצוגה הדואלית (רובינשטיין-קנטורוביץ) המחליפה את המינימום על מידות הסתברות על מרחב הפרודאקט למקסימים של הפרש התוחלות על בין מידות הסתברות אלו מעל מרחב של פונקציות ליפשיץ עם מקדם 1, אולם גם במקרה הזה בעיית אופטימיזציה זו רחוקה מלהיות פשוטה. במקרה של שני דאטה סטים בגודל סופי (המקרה שלנו) ניתן להגדיר את מידות ההסתברות עליהם כסכום של פונקציות דלטה על הנקודות של הדאטה סט). במקרה המרחק בין כל הנקודות בדאטה סטים מוגדר

כמטריצה ואז בעיית אופטימיזציה הופכת לבעיית תכנות לינארי (המידה על מרחב הפודאקט שעליה מבצעים את האופטימיציה ניתנת לתיאור עי״ מטריצה). עדיין לדאטה סטים גדולים הפתרון דורש משאבי חישוב אדירים לא פיזיבילי. ב 2013 הוצע (Sinkhorn) להוסיף לבעיה זו איבר רגולריזציה המודד KL בין המידה על הפרודאקט (שלפיה מאפטמים) לבין המכפלה הקרטזית של P ו- Q. תוספת זו איפשרה לפתור את הבעייה בצורה יותר יעילה.

<u>מרחק בין דאטה סטים דרך מרחק וסרשטיין:</u> אז בואו נחזור לבעיה שלנו ונראה איך מגדירים את מרחק ביניהם דרך כל המושגים הנ״ל. קודם כל בהינתן שני סטים מתויגים מרחב ההגדרה של מידות ההסתברות עליהם זה המכפלה הקרטזית של מרחב הפיצ'רים ומרחב הלייבלים(נסמן אותו ב Z) נציין שמרחבי הלייבלים אינם חייבים להיות זהים אך לפשטות ההצגה נניח שהם כן). המאמר מציע להגדיר את המרחק בין שתי דוגמאות ב Z המסומנות: z1=(x1, 1y) ו- 2z2=(x2, y) כסכום של המרחקים בין 1x ל- 2x (במרחב הפיצירים) ולבין 1y ל- 2y במרחב הלייבלים (בעצם זה שורש p מהסכים של חזקות p של המרחק הראשון ושל המרחק השני). אז המרחק בין הפיצ'רים (הראשון) מחושב בצורה ישירה (אוקלידי או כל מרחק אחר). המרחק בין הלייבלים קצת יותר בעייתי. הדבר הפשוט ביותר הוא לתאר כל לייבל כממוצע של הפיצ'רים של כל הדוגמאות עם הלייבל הזה אך זה לא מספיק מייצג את הלייבל. הדרך היותר טובה היא לחשב אותה כמרחק וסרשטיין בין התפלגויות מותנות של הפיצ'רים בהינתן הלייבלים. עם המרחק בין 1z ו- 2z מוגדר כך ניתן להוכיח שזה מטריקת מרחק תקינה, ומוגדרת על סטים דיסקרטיים כמו שאנחנו צריכים. בסוף המרחק בין הדאטה סטים מוגדר (בדומה ל OT) כמינימום על כל מידות הפרודאקט על Z עם עצמו. את הבעייה הזו ניתן לפתור עם הוספת איבר רגולריזציה KL כמו שהזכרתי קודם. לצערנו אפילו לפתרון הזה יש סיבוכיות של n^5*logn (n - -גודל הדאטה סט) שעושה אותו לא ישים לדאטה סטים גדולים. במקום זאת המחברים מציעים לשערך את ההתפלגות המותנית של פיצ'ר בהינתן לייבל עי״ גאוסיאן שעבורו יש בוטוי מדויק ל WD כאשר סיבוכיות החישוב במקרה הזה יורדת ל 2^n. הם מוכיחים שהמרחק בין דאטה סטים במקרה זה חסום עי״ המרחק המקורי מלמעלה)

הישגי מאמר: עבור שני דאטה סטים הם משווים את הירידה בשגיאת בטסט עבור הדאטה סט השני בין שני תרחישים: אימון רגיל מאפס מול אימון של הראשון וכיול של השני (מאותחל עם המשקלים של הראשון). הם מראים שככל שהמרחק OT המוגדר במאמר קטן יותר הירידה בשגיאה גדולה יותר קרי יש יותר דמיון בין הדאטה סטים.

דאטה סטים: MNIST, FASHION-MNIST, KMNIST, letters EMNIST

לינק למאמר: **paper**

לינק לקוד: לא מצאתי

נ.ב. מאמר עם רעיון מאוד מעניין. מסקרן לראות האם גישה זו תעבוד על דאטה סטים יותר רציניים. בנוסף במרחק המתואר במאמר אין התחשבות לא בפוקציית לוס ולא בסוג המודלים של משתשמים בהם אחר-כך לסיווג. מקווה שנרא הרחבות בקרוב

נ.ב. 2 לינק לספרייה מעולה בפייטורץ' הממשת מרחקים גאומטריים בין דאטה סטים: ספריית פייטורצי (תודה ל