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:

Robust Optimal Transport with Applications in Generative Modeling and Domain Adaptation

---

פינת הסוקר:

המלצת קריאה ממייק: מומלץ למביני עניין בטכניקות מורכבות ל -domain adaptation.

בהירות כתיבה: בינונית

רמת היכרות עם כלים מתמטיים וטכניקות של ML/DL הנדרשים להבנת מאמר: הבנה עמוקה בתכונות של מרחקים שונים בין מידות הסתברות והבנה טובה בבעיות אופטימיזציה עם אילוצים. הבנה בטרנספורט אופטימלי רצויה גם כן.

יישומים פרקטיים אפשריים: ניתן להשתמש בגישה זו לאימון של גאנים כאשר סט האימון חשוד להכיל דוגמאות זרות וגם כן למשימות UDA.

פרטי מאמר:

לינק למאמר: זמין להורדה.

לינק לקוד: זמין כאן

פורסם בתאריך: 12.10.20, בארקיב.

הוצג בכנס: NeurIPS 2020

תחום מאמר:

- מרחק בין דאטהסטים עם אווטליירים (outliers)
- מודלים גנרטיביים (GANs)
- אדפטצית דומיינים בלתי מונחית (unsupervised domain adaptation - UDA)

כלים מתמטיים, מושגים 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 (Lip-1)
- דוגמאות לא טיפוסיות או אווטליירים (OL)

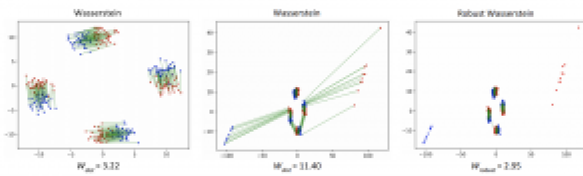
תמצית מאמר:

המאמר הנסקר מציע שיטה לחישוב מרחק בין דאטהסטים, הרובסטי לדוגמאות לא טיפוסיות (,OL outliers). למעשה המרחק המוצע מוגדר עבור כל שתי מידות הסתברות והמרחק בין דאטהסטים הוא המקרה הפרטי שלו. מרחק זה נקרא טרנספורט אופטימלי רובסטי (ROT - Robust Optimal Transport), הוא מבוסס על מרחק OT הסטנדרטי ומנסה להתגבר על רגישותו לדוגמאות OL. המאמר דן ברובו במקרה הפרטי של OT שזה מרחק וסרשטיין (WD- Wasserstein Distance) כך שאתמקד רק במרחק וסרשטיין הרובסטי (RWD) בהמשך הסקירה. רגישות של מרחק OT לדוגמאות OL ניתן לנסח באופן הבא: בהינתן שני דאטהסטים עם WD די נמוך, החלפתו של חלק מאוד קטן של דוגמאות באחד דאטהסטים בדוגמאות OL עלולה להוביל לעלייה בלתי פרופורציונלית ב-WD ביניהם. לטענת המאמר מרבית הדאטהסטים הגדולים מכילים דוגמאות OL, ושימוש במרחק ביניהם שרגיש לדוגמאות אלו, עלול להוביל לתוצאות ירודות במשימות שונות. למשל אימון של GAN עם מטריקת מרחק כזו (כמו וסרשטיין גאן - WGAN) עלול להוביל לכך ש-WGAN 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 הזה ומציע לשערך את המרחק בין התפלגויות P_1 ו- P_2 עי" המרחק בין שתי התפלגויות קרובות אליהן, Q_1 ו-Q_2 בהתאמה, ע"י הוספה של שני איברי רגולריזציה המכילים את סכום המרחקים Div(Q_1,P_1) ו-Div(P_2,Q_2). המרחק Div(P,Q) בין ההתפלגויות P ו-Q מוגדר כמרחק f-divergence עבור פונקציה f נתונה. 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 קנטורוביץ'-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 יישומו לבעיות פרקטיות כמו למשל אימון של GAN.

בשביל להתגבר על קושי זה ולשמר את הרובסטיות של המרחק לגבי דוגמאות OL, המאמר מציע לשנות את ניסוח בעיית אופטימיזציה של UOT באופן הבא: במקום לאפטם על כל ההתפלגויות ה "בערך שוות" ל P_1 ול-P_2, הם ״מגבילים״ (מלמעלה) את המרחקים האלו ע"י קבועים ρ_1 ו-ρ_2. זה כמובן הופך את מרחק ROT המוצע במאמר להיות תלוי באופן ישיר ב-ρ_1 ו- ρ_2 אבל הבעיה הדואלית נהיית יותר פשוטה ותלויה רק בפונקציה אחת (שהיא פונקצית ליפשיץ מסדר 1). מצד שני זה מוסיף אילוץ לבעיית אופטימיזציה הדואלית אך המאמר מוכיח שעדיין ניתן לפתור אותה בדרך יחסית נוחה.



(a) Clean dataset (b) Dataset with outliers (c) Dataset with outliers

Figure 1: Visualizing couplings of Wasserstein computation between two distributions shown in red and blue. In (a), we show the couplings when no outliers are present. In (b), we show the couplings when 5% outliers are added to the data. The Wasserstein distance increases significantly indicating high sensitivity to outliers. In (c), we show the couplings produced by the Robust Wasserstein measure. Our formulation effectively ignores the outliers yielding a Wasserstein estimate that closely approximates the true Wasserstein distance.

תקציר מאמר:

נזכר קודם כל מה זה מרחק OT והמקרה הפרטי שלו מרחק וסרשטיין WD.

טרנספורט אופטימלי OT:

OT הינו מרחק בין שתי מידות הסתברות P_2 ו-P_1, המוגדרות על אותו מרחב X, עבור פונקציית מחיר אי שלילית $c(y_1,y_2)$ נתונה. OT מודד עד כמה מידות הסתברות "קרובות" (כמו מרחק KL או JS). מקרה פרטי של OT שבו פונקציית מחיר הינה מרחק L_p, נקרא מרחק וסרשטיין WD מסדר p. כאשר $p=1$ מרחק זה נקרא מרחק earth mover.

אז מה זה בעצם WD?

בנוסחה עבור WD בין P_1 ו-P_2 מופיע מינימום מעל כל מידות הסתברות על מרחב המכפלה של X עם עצמו, כאשר הפונקציות השוליות שלה הן מידות ההסתברות P_1 ו-P_2 ותחת סימן האינטגרל יש את המרחק בין הנקודות. לפשטות בואו ניקח $p=1$. בנוסף נניח שמרחב X הוא חד מימדי (R). למה זה בעצם נקרא מרחק earth mover? למעשה מרחק זה מגדיר כמה "מסה" אנו צריכים להעביר בשביל להפוך את המידה הסתברות P_2 ל-P_1 כאשר המחיר העברת הנקודה x מהתומך P_2 לנקודה y מהתומך של P_1 הינה $|x-y|$.

למה פעולת מינימום מופיעה בנוסחה עבור WD, אתם שואלים? אפשר "להפוך את P_1 ל-P_2 במספר דרכים ואנחנו רוצים את הדרך הכי קצרה (מבחינת ״המסה המועברת״).

ולמה מופיעה בנוסחה מידת הסתברות M על מרחב המכפלה של X עם עצמו? פונקציה ב-(x,y) זו מגדירה איזה "חלק" מהמסה ההסתברותית בנקודה x אנו מעבירים לנקודה y. נניח שלנקודה x הסתברות 0.5, אנו מעבירים שליש ממנה לנקודה y_1 ושני שליש הנותרים לנקודה y_2. במקרה הזה

$T(x\_1, y\_1) = 0.5 * 1/3 \sim= 0.17$ ו- $T(x\_1, y\_1) = 0.5 * 2/3 \sim= 0.33$. התנאי שהפונקציות השוליות של M צריכות להיות שוות P_2 ו-P_1 נחוץ, כי אנו רוצים להעביר את כל המסה מכל הנקודה של P_1 לנקודות של P_2 בלי לאבד (או להרוויח) מסה נוספת. להבדיל כמעט כל מרחק בין מידות ההסתברות WD לוקח בחשבון של התכונות של הקבוצות שעליהן מידות אלו מוגדרות בצורה מפורשת עי" התחשבות במרחק בין הנקודות שלהם. ולבסוף הצורה הדואליות של WD היא בעצם בעיית אופטימיזציה המנסה למקסם הפרש התוחלות של פונקציית d תחת P_1 ו- P_2 מעל מרחב של כל פונקציות ליפשיץ d מסדר 1.

עכשיו בואו נסביר איך ניתן להגדיר את WD על דאטהסטים:

איך מגדירים WD בין דאטהסטים:

עבור שני דאטהסטים בגודל סופי ניתן להגדיר את מידות ההסתברות עליהם כסכום של פונקציות דלתא על הנקודות של דאטהסט, כאשר ההסתברות של כל נקודה הינה שווה. המרחק בין כל הנקודות בדאטהסטים ניתן ע״י מטריצה ואז בעיית אופטימיזציה הופכת לבעיית תכנות לינארי (המידה על מרחב המכפלה שעליה מבצעים את האופטימיזציה ניתנת לתיאור עי״ מטריצה גם כן).

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

איך מגדירים RWD בין דאטהסטים?

קודם כל נציין כי כל אחת פונקציות התפלגות (מידות הסתברות) Q_1 ו-Q_2 קרובות ל-P_1 ו-P_2 בהתאמה (ראה הסבר בפרק ״רעיון בסיסי) ניתן להגדיר בתור משקול של הסתברויות של דוגמאות (כמובן שגם ב-P_1 וגם ב-P_2 לכל דוגמא של אותה הסתברות) בשני הדאטהסטים כאשר סכום המשקלים בדטאסט הוא 1 (אחרת Q_i לא תהווה מידת הסתברות). ניתן לראות כי בעיית אופטימיזציה שאנו פותרים 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-rho_1 ו- rho_2). המאמר מראה כי תנאים אלו ניתן לתרגם למרחקי chi-2 בין ההתפלגויות הממושקלות Q_1 ו-Q_2 למקוריות P_1 ו-P_2 על הדאטהסטים. בעיה זו למעשה הינה [תכנות קוני מסדר שני](https://en.wikipedia.org/wiki/Second-order_cone_programming) ויש דרכים יעילות לפתור אותה. עבור דאטהסטים גדולים לפתרון זה עלולה להיות עלות חישובית מאוד גבוהה. כדי להקטין את הסיבוכיות החישובית של הפתרון, מחברי המאמר עשו רפרמטריזציה של המשקלים ע" רשתות נוירונים כאשר הקלט לרשתות אלו הוא דוגמאות מהדאטהסטים.

הישגי מאמר:

המאמר השתמש ב-RWD כדי לבנות GAN עם הממזער RWD בין התפלגות פלט הגנרטור לבין דאטהסט האימון. המחברים הראו כי עבור דאטהסטים המכילים דוגמאות OL (או אלו שהם יצרו באמצעות ״לכלוך״ דאטהסטים ״נקיים״ באחוז מסוים של תמונות מדאטהסטים אחרים) התמונות שגונרטו עם RWDGAN נראות יותר ״נקיות״ מבחינה ויזואלית אפילו עבור אחוזי OL יחסית גבוהים. מעניין כי כאשר מאמנים את RWDGAN על דאטהסטים נקיים (עם rho_1 ו-rho_2 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 גאן עם 3 ארכיטקטורות שונות.

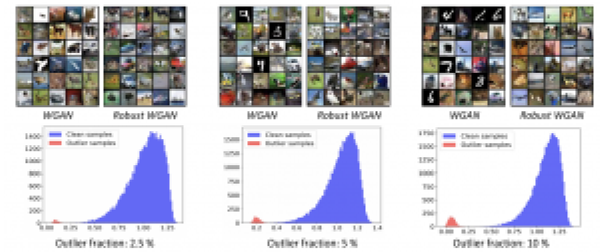
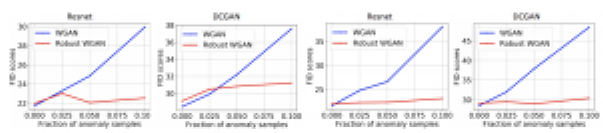


Figure 3: Visualizing samples and weight histograms. In the top panel, we show samples generated by WGAN and robust WGAN trained on CIFAR-10 dataset corrupted with MNIST samples as outliers. WGAN fits both CIFAR and MNIST samples, while robust WGAN ignores the outliers. In the bottom panel, we visualize the weights (output of the W(.) function) for in-distribution and outlier samples. Outlier samples are assinged low weights while in-distribution samples get large weights.

תופעה מעניינת של RWDGAN: משקול אופטימלי של דוגמא נתונה למעשה משקף את "רמת הקושי" של הגנרטור לגנרט אותה (כלומר עד כמה דיסקרימינטור הצליח "לפצח אותה"). אתם שואלים למה בעצם? אם משקל של דוגמא נמוך, זה אומר שהגנרטור "החליט להנמיך בחשיבותה ולהקטין את השפעתה ללוס" מהסיבה שהוא חושב שהדוגמא הזו הינה OL. ד"א המאמר מראה שבדאטהסטים "מלוכלכים" עם דוגמאות מדאטהסטים אחרים המשקלים של הדוגמאות "הזרות" יצאו נמוכות משמעותית מהרגילות.



(a) CIFAR-10 + MNIST (b) CIFAR-10 + Uniform noise

Figure 2: FID scores of GAN models trained on CIFAR-10 corrupted with outlier noise. In (a), samples from MNIST dataset are used as the outliers, while in (b), uniform noise is used. FID scores of WGAN increase with the increase in outlier fraction, while robust WGAN maintains FID scores.

בנוסף המאמר הראה כי שימוש ב-RWD עבור משימות UDA משפר באופן ניכר את ביצועי דיוק עבור 3 ארכיטקטורות רשת שונות (עבור דאטהסט VISDA17).

נ.ב.

המאמר עם רעיון די מעניין, מכיל גם הוכחות ריגורוזיות המסבירות למה הגישה המוצעת עובדת. מה שמטריד אותי טיפה עם RWD זו הבחירה של פרמטר rho. המאמר מוכיח כי עם אחוז דוגמאות OL ידוע אז קיים ביטוי לערך rho אופטימלי. ברוב המקרים זה לא המצב ובחירה של rho עלולה להיות לא טריויאלית.

#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.