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:

Sequence-to-Sequence Contrastive Learning for Text Recognition

---

פינת הסוקר:

המלצת קריאה ממייק: כמעט חובה (לא חייבים אך מומלץ בחום לחסידי למידת הייצוג ואוהבי OCR).

בהירות כתיבה: גבוהה.

רמת היכרות עם כלים מתמטיים וטכניקות של ML/DL הנדרשים להבנת מאמר: בינוני (נדרשת הבנה מסוימת במושגי למידת הייצוג).

יישומים פרקטיים אפשריים: שיפור ביצועים למשימות OCR כמו זיהוי לוחות רישוי, זיהוי של תמרורים עבור מערכות רכב אוטונומי, הקטנת גודל סט אימון מתויג הנדרש לרמת ביצועים נתונה.

פרטי מאמר:

לינק למאמר: זמין להורדה.

לינק לקוד: לא הצלחתי לאתר.

פורסם בתאריך: 20.12.20, בארקיב.

הוצג בכנס: NeurIPSi 2020.

תחומי מאמר:

- למידת ייצוגים במימד נמוך למשימות זיהוי טקסט (כתב יד) בתמונה.

- למידה מנוגדת (contrastive learning - CL) למשימות מיפוי סדרה לסדרה (sequence-to-sequence tasks - StST) .

כלים מתמטיים, טכניקות, מושגים וסימונים:

- לוס מנוגד (contrastive loss).
- אוגמנטציה של דאטה ליצירה של דוגמאות "דומות".
- רשתות לעיבוד שדאטה סדרתי (sequential) כמו LSTM.

מבוא והסבר כללי על תחום המאמר:

שיטות לבניית ייצוג במימד נמוך של דאטה עבור דאטהסטים לא מתויגים, המבוססות על למידה מנוגדת, תפסו פופולריות רבה בשנים האחרונות. לאחרונה שיטות אלו הצליחו לבנות ייצוגים מאוד חזקים שלא נופלים באיכות מאלו שנבנו במהלך אימון של מודלים עמוקים עם דאטהסטים מתויגים. יתרה מזו, יישום גישה זו בשילוב של דאטהסט מתויג לא גדול (למידה semi-supervised) הצליח להזניק את הביצועים במגוון של משימות כמו סיווג, סגמנטציה, זיהוי אובייקטים ואחרות מעבר לאלו של המודלים שנבנו עבור דאטהסטים מתויגים גדולים.

מרבית שיטות אלו הוצעו במקור עבור דומיין של התמונות (SimCLR, BYoL, MoCo ועוד). נציין שממש לאחרונה פורסמו עבודות המציעות גישות מבוססות על הלמידה המנוגדת גם בדומיינים אחרים כמו למידה מנוגדת לסדרות זמן ו- למידה מנוגדת לדומיין הווידאו. לעומת זאת רוב השיטות העכשווית לזיהוי כתב יד הינן שיטות למידה supervised קלאסיות. הסיבה לכך טמונה בעובדה שגישות למידה מנוגדת מצטיינות ביצירת ייצוגים חזקים למשימות כמו סיווג או זיהוי שבהם כל דוגמא הינה "פיסה אטומית" (במילים אחרות הדוגמא "מהווה מקשה אחת"). במקרה כזה דוגמאות "חיוביות" (קרובות) לדוגמא נתונה **x** נוצרות עי" הפעלת אוגמנטציות שונות על **x**, כאשר כל שאר הדוגמאות הינן "שליליות" ל **x** (רחוקות) באופן אינהרנטי. המצב במשימות זיהוי של כתב יד שונה כי תהליך החיזוי של מילה כתובה הינו סדרתי באופן טבעי (כמו כל זיהוי של טקסט). כל מילה כתובה מורכבת מאותיות ונראה שהכי הגיוני למדל אותה כסדרה של פאטצ'ים (נקרא לזה גם פריימים) סמוכים של תמונה שכל אחד מהם מייצג תת-מילה/אות/חלק מאות. עקב כך הצורה הסטנדרטית של הלמידה המנוגדת אינה ישימה עבור תמונות המכילות כתב יד (למעשה אוגמנטציה של תמונה עם מילה כתובה **x** עלול להפוך אותה ל"לא קרובה" ל- **x** מאחר והיא משנה את ה"סדר הטבעי" של האותיות.

הסבר על מושגים חשובים במאמר:

עקרונות הלמידה המנוגדת: גישה זו מסתמכת על ההנחה שייצוגים של דוגמאות קרובות צריכים להיות קרובים, בזמן שייצוגים של דוגמאות לא קשורות (נקראות שליליות) צריכים להיות רחוקים. בשביל לבנות פונקצית מטרה לשיטת למידה מנוגדת לוקחים זוג של דוגמאות קרובות (למשל שתי אוגמנטציות של אותה דוגמא) ומספר דוגמאות רנדומליות ומנסים למקסם את היחס בין אקספוננט של דמיון של הזוג הקרוב לסכום הדמיונות בינו לבין דוגמאות רנדומליות.

תמצית מאמר:

המאמר מציע גישה של למידה מנוגדת, הנקראת SeqCLR (הגרסה הסדרתית של SimCLR) המותאמת למשימות בעלות אופי סדרתי כגון זיהוי של טקסט בתמונה. בגדול הרעיון הוא ליצור מתמונה **סדרה(!!)** של אובייקטים כאשר כל אובייקט מורכב ממספר פאטצ'ים סמוכים וללמוד ייצוג כל אובייקט כזה באמצעות למידה מנוגדת. נדגיש שהסדר בין האובייקטים נשמר וזו הסיבה שהשתמשתי במונח "סדרה"(!!) ולא במונח "סט"(!!) בתיאור. כמובן ניתן ליצור מספר שונה של אובייקטים לכל תמונה. אז איך בעצם עובדת הלמידה המנוגדת כאן, כלומר איך בונים זוגות "חיוביים" וזוגות "שליליים" של דוגמאות(שזה הבסיס של כל גישת הלמידה המנוגדת). אז התהליך נראה כך.

- מפעילים שתי אוגמנטציות על תמונה **x** ויוצרים תמונות **x_1** ו- **x_2.** שימו לב שלא כל שיטת אוגמנטציה שמשתמשים בה בלמידה המנוגדת הסטנדרטית הדומיין התמונות, מתאימה כאן. למשל אוגמנטציות מסוג הזזה אופקית והיפוך אינם מתאימות כאן כי הן עלולות "לשנות את הסדר בין פריימים (הסיבה לכך מפורטת באחד הסעיפים הבאים).

- מחלקים את **x_1** ו- **x_2** לסדרות של אובייקטים **{x_2i}** ו- **{x_1i}** (הסדרות הן באותו אורך). נזכיר כל אובייקט מורכב מפריימים (פאטצ'ים) סמוכים.

- בונים ייצוגים ם **{z_2i}** ו- **{z_1i}** לכל האובייקטים שבנינו עי" העברתם דרך כמה רשתות נוירונים עוקבות (הייצוגים נבנים בצורה זהה עבור **x_1** ו- **x_2**).

- כל זוג חיובי בנוי מהאובייקט נבנה מייצוג **z_1k** ו- **z_2k** עבור k-ים שונים כאשר ייצוגים **z_1i** ו- **z_2j** עם אינדקסים שונים וייצוגים של האובייקטים של הדוגמאות האחרות מהבאטץ', מהווים זוגות של דוגמאות שליליות. שימו לב שככל שמספר האובייקטים, נוצרים מתמונה בודדת, גבוה יותר, צריך באטץ' יותר קטן בשביל ליצור אותה כמות של זוגות של דוגמאות שליליות לדוגמא **x**. בנוסף יש לי תחושה שהדוגמאות השליליות הנוצרות מאותה תמונה "מאלצות" את המודל ללמוד להבחין בין אובייקטים שונים בתוכן (אותיות שונות) אך דומים מבחינת הסגנון שתורם חיובית לעוצמת הייצוג . בנייה זו ממחישה למה ניתן להשתמש רק באוגמנטציות שלא משנות את הסדר של האובייקטים(פריימים) בתמונה.

הסבר של רעיונות בסיסיים:

אחרי שמבינים את עיקרי SeqCLR, נותר רק לפרט איך נבנים הייצוגים **{z_i}** מסדרת הפריימים של תמונה **x.** בנייה זו נעשית דרך שימוש בכמה רשתות נוירונים עוקבות שמפורטות בסעיף הבא:

תהליך בניית ייצוגים עבור תמונה x:

- <u>בניית ייצוגים התחלתיים של פריימים:</u> קודם כל מחלקים תמונה ל- T פריימים. אחר כך יש שתי אופציות: הראשונה היא להעביר כל פריים דרך רשת ייצוג סטנדרטית לתמונות ולהשתמש בתוצאה בתור ייצוג של כל פריים. האפשרות השנייה היא לבנות ייצוגים המנצלים את הקשרים בין הייצוגים של הפריימים השונים מהשלב הראשון (ייצוג קונטקסטואלי או הקשרי). המאמר מציע לעשות זאת עם LSTM דו כיווני (כאן ברוח הזמן הייתי מציע להשתמש באיזו גרסה "קלה חישובית" של הטרנספורמר). ניתן כמובן לשלב את שני הייצוגים האלו כקלט לשלב הבא.

- <u>בניית ייצוגים של אובייקטים:</u> המטרה בשלב זה הינה ליצור קלט לחישוב של הלוס המנוגד. המאמר מציע 3 דרכים לבנות אותם מהייצוגים של הפריימים שהתקבלו מהשלב הראשון:
  1. למצע את כל ייצוגים ולקבל וקטור ייצוג אחד **z** עבור התמונה ולקבל אובייקט אחד לכל תמונה. במקרה כזה זוגות שליליים נבנים רק בין דוגמאות שונות בבאטץ'.
  2. לבנות T_p אובייקטים מ-T פריימים כאשר T > T_p (פולינג) המאפשר בנייה של זוגות דוגמאות שליליות גם מאובייקטים מאותה תמונה.
  3. כל פריים הופך לאובייקט.

- <u>חישוב של לוס מנוגד</u> מהייצוגים שהתקבלו בסעיף הקודם.

הישגי מאמר: המאמר הצליח להוכיח שהייצוגים שנבנו באמצעות SeqCLR מסוגלים להביא לשיפור ביצועים במשימות סיווג על מגוון דאטהסטים. הבדיקות נעשו בצורה סטנדרטית עבור שיטות מהסוג הזה: מקפיאים את משקלי המודל(שבונה את ייצוג הדאטה), מוסיפים שכבת לינארית למודל ומאמנים אותה על דאטה מתויג. SeqCLR הגיע לביצועים יותר טובים גם עבור משימות semi-supervised (כלומר האימון של השכבה הלינארית מתבצע רק על אחוז מסוים של דוגמאות מדאטהסט מתויג).

דאטהסטים:

- כתב יד באנגלית: IAM, CVL .
- כתב יד בצרפתית RIMES.
- זיהוי טקסט בתמונה: SyntText, IIT5K, ICO3.

נ.ב. המאמר מכליל את שיטת הלמידה המנוגדת לדיאטה בעלת אופי סדרתי כמו טקטס בתמונה. השיטה מאוד אינטואיטיבית וקל להשתכנע שהיא אכן מגיעה לתוצאות טובות. השיטה נבדקה על לא מעט דאטהסטים בשני דומיינים שונים (כתב יד וטקסט בתמונה - OCR) ונמצאה עדיפה על מתחריה. החלק היחיד שחסר לי בשביל להשתכנע זה הקוד שלא נמצא באקריב - בתקווה שיוסף בקרוב.

#deepnightlearners

הפוסט נכתב על ידי מיכאל (מייק) ארליכסון, PhD.

מיכאל עובד בחברת סייבר Salt Security בתור Principal Data Scientist. מיכאל חוקר ופועל בתחום הלמידה העמוקה, ולצד זאת מרצה ומנגיש את החומרים המדעיים לקהל הרחב.