לילה טוב חברים, היום אנחנו שוב בפינתנו deepnightlearners עם סקירה של מאמר בתחום הלמידה העמוקה
היום בחרתי לסקירה את המאמר שנקרא:
Unsupervised Learning of Visual Features by Contrasting Cluster Assignments
שיצא לפני בערך חודשיים

הוצג בכנס: NeurIPS 2020

תחומי מאמר:

- למידת ייצוג ללא דאטה סט מתויג (SSRL -self-supervised representation learning)
- SSRL מבוסס על טכניקות קליסטור (Clustering for deep representation learning)

כלים מתמטיים, מושגים וסימונים:

- מולטי-קרופ - טכניקת אוגמנטציה המבוסס על לקיחת פאטצ'ים קטנים של תמונה ברזולוציות נמוכות שונות
- האלגוריתם של סינקהורן קנופ (Sinkhorn-Knopp) לפתרון בעיית הטרנספורט האופטימלי למידות הסתברות דיסקרטיות

תמצית מאמר:
המאמר מציע שיטת למידת ייצוג על דאטה סט לא מתויג. רוב גישות המודרניות בתחום הזה (SSRL) מורכבות משני מרכיבים עיקריים:

<u>הלוס המנוגד (:contrastive loss - CL)</u> מסתמך על ההנחה שייצוגים של דוגמאות קרובות צריכים להיות קרובים, בזמן שייצוגים של דוגמאות לא קשורות (נבחרות רנדומלית בד"כ) צריכים להיות רחוקים.
<u>שיטה ליצירה של דוגמאות "דומות", קרי אוגמנטציה:</u> בדרך כלל זוג דוגמאות קרובות(אקרא לזוגות האלו בהמשך זוגות חיוביים או זוגות קרו נוצרות עי" ההפעלה של שתי אוגמנטציות שונות על אותה דוגמא.

נציין כי גישות SSRL המודרניות מסתמכות של השוואה של מספר גבוה מאוד של זוגות ייצוגים של דוגמאות שדורש כמות גדולה של זכרון ומשאבי עיבוד משמעותיים. דרישות אלו מקשות על יישום של שיטות אלו בצורה אונליין (לטענת המאמר רוב שיטות SSRL היום מיושמות בצורת אונליין שדי הפתיע אותי). אז בואו נדבר על החידושים שהמאמר הזה מציע:
<u>שיטת אימון SwaV</u>: המאמר הנסקר מציע שיטה חדשה SSLR (הנקראת SwaV) העשויה להוריד גם את כמות החישובים וגם לצמצם את כמות הזכרון הנדרשות. הרעיון העיקרי של המאמר הינו שינוי "ההגדרה של מושג הדמיון בין ייצוגי דוגמאות". למעשה המאמר "מאלץ" זוגות של הדוגמאות הקרובים "להשתייך" לאותם הקלאסטרים במרחב הייצוג במקום להשוות את הייצוגים בצורה מפורשת (שיוך זה המיוצג עי" הקוד של דוגמא המחושב על סמך הבאטץ' שלו - אופן בנייתו יפורט בהמשך). נציין ש SwaV אינו דורש לשמור בנק של דוגמאות שליליות שהופך אותו <u>למועמד טוב למימוש בצורה אונליין</u>.
<u>שיטת אוגמנטציה מולטי-קרופ:</u> המאמר מציע שיטת אוגמנטציה הנקראת מולטי-קרופ שמתחילה מהחישוב של שני "קרופים סטנדרטיים" 1x_cr ו- 2x_cr של תמונה x. לאחר מכן לוקחים "קרופים קטנים יותר" של 1x_cr ו- 2x_cr במגוון רזולוציות נמוכות ובונים מהם סט דוגמאות חיוביות עבור תמונה x. לטענת המאמר שיטה זו מקטינה את כמות החישובים הנדרשת תוך שמירה על הביצועים

הסבר של רעיונות בסיסיים:
עכשיו ננסה להבין מה פונקצית המטרה L שבליבה של SwaV. פונקציית L מוגדרת באופן הבא (לכל דוגמא בבאטץ':)

- בונים מספר אוגמנטציות שלה לדוגמא x עם מולטי קרופ או כל גישה אחרת

- מרכיבים מאוגמנטציות אלו זוגות של דוגמאות
- בונים וקטורי ייצוג z לכל הדוגמאות שבנינו
- לכל זוג וקטורי ייצוג (2_z_1, z) מחשבים את הקודים שלהם 1_q ו- 2_q
- מחשבים את סכום הדמיונות s_l בין 1_z ו- 2_q ובין 2_z ל 1_q
- מחשבים את הסכום L_x של s_l של כל הזוגות של הדוגמאות החיוביות של דוגמא x

<u>אינטואיציה:</u> למעשה תהליך אימון זה "מאלץ" וקטורי ייצוגי של דוגמא להכיל מידע על הקוד של הדוגמאות הקרובות. בצורה לא פורמלית ניתן לומר שאנו מנסים למקסם את "המידע הדדי" בין הייצוגים של הדוגמאות שזה המטרה העיקרית של האימון עם הלוס המונגד CL.

דרך אגב השם של השיטה נובע מהפעולה שחלוף (swap) שמבצעים בין הייצוגים ובין הקודים של דוגמאות קרובות באימון.

השאלה האחרונה שטרם התייחסנו אליה הינה מבנה של פונקציית לוס בין ייצוג z לקוד q.

<u>מבנה של פונקציה לוס בין ייצוג z לקוד q (של דוגמאות קרובות):</u> אם אתם זוכרים הקוד q ניתן לפרש כווקטור הסתברויות שיוך לקלסטרים. למעשה אנו רוצים שהקוד q ישקף בצורה כמה שיותר טובה את המרחקים של z מהפרוטוטייפים c_i שניתן לראות אותם בתור מרכזים (סנטרואידים) של קלסטרים של ייצוגים. אז קודם כל אנו בונים את וקטור המרחקים המנורמלים מ- z לכל c_i. מרחק זה מחושב כאקספוננט של המכפלה הפנימית בין z ל c_i. בסוף לוקחים את וקטור המרחקים ומנרמלים אותו. לאחר מכן מחשבים את קרוס אנטרופי בין q לווקטור מרחקים מנורמל שחישבנו. את הפונקציה זו אנו ממקסמים ביחס ל ייצוגים z וביחס לפרוטוטייפים c.

<u>אינטואיציה:</u> שימו לב על הדמיון של המרחק בין וקטור הייצוג z ל- c_i לביטוי של החוב המנוגד CL. וזה לא מקרי - אתם זוכרים שלהבדיל משיטות מבוססות CL קלאסי, אין לנו כאן דוגמאות שליליות בצורה מפורשת. אז מה שמשחק כאן את תפקיד "הדוגמאות השליליות" זה מרכזי הקלסטרים שרחוקים מ z. כלומר הם מאלצים <u>ייצוגים של דוגמאות חיוביות להיות רחוקים בצורה כמה שיותר דומה מכל הקלסטרים השליליים וקרובים באותה מידה מהקלסטרים החיוביים</u>. לדעתי זה הנקודה הכי משמעותית במאמר(!!).

<u>הסבר על בניית קוד q של ייצוג z:</u> הקוד q של וקטור ייצוג z מתאר את "רמת קרבתו" של z ל K וקטורי פרוטוטייפ c_i. וקטור c_i "מייצג" את הקלסטר i. קוד של דוגמא (וגם של כל האוגמנטציות שלה) מחושב על סמך הבאטץ' בלבד(!!). אפשר להגיד שהקוד q מייצג את ההסתברויות שיוך של וקטור הייצוג z של הדוגמא לקלסטרים המיוצגים עי" וקטורי c_k.

מטריצה Q המכילה את הקודים של כל הדוגמאות מהבאטץ' הינה פתרון של בעיית אופטימיזציה לינארית עם איבר רגולריזציה השווה לאנטרופיה הכוללת של Q עם מקדם קטן. פונקציה מטרה זו מנסה למקסם את הדמיון הכולל בין וקטורי ייצוג של הדוגמאות בבאטץ' לפרוטוטייפים c_i(כלומר לפזר את הקודים בצורה המשקפת את את יחס המרחקים בין ייצוג הדוגמא למרכזי הקלאסטרים השונים). שימו לב שבעיית אופטימיזציה זו מזכירה בצורתה את בעיית הטרנספורט האופטימלי בין מידות הסתברות דיסקרטיות (האחידות) המוגדרות על שני דאטה סטים. את התפקיד של דאטה סטים כאן משחקים הפרוטוטייפים c_i וקטורי הייצוג z_i של כל הדוגמאות בבאטץ'. המטרה כאן זה למצוא את האופן האופטימלי שבו ניתן "להעביר את המסה ההסתברותית מווקטורי z_i לוקטורי c_i (נציין שפונקצית המרחק שיש בהגדרה של הטרנספורט האופטימלי הינה פרופורציונלית במקרה שלנו למרחק בין z ל c). למעשה אנו מנסים למצוא מטריצה Q האי שלילית, שאיבר jk שלה מגדיר את המסה ההסתברותית המועברות מווקטור z_k לווקטור c_j כלומר הסתברות השיוך של z_k לקלסטר של c_j. מכיוון שאנו רוצים שאותו מספר דוגמאות "ישויך" לכל קלאסטר, מוסיפים אילוץ על סכום השורות וסכום העמודת של Q. בעיה זו פותרים בעזרת אלגוריתם איטרטיבי של סינקהורן-קנופ .

הסבר על מושגים חשובים במאמר:

<u>שיטות אימון של גישות SSRL המודרניות:</u> בדרך כלל בזמן האימון של SSLR לכל זוג של דוגמאות קרובות בונים מספר גדול של זוגות רנדומליים (אקרא לזוגות כאלו זוגות רחוקים או זוגות שליליים). כאן פונקציית המטרה F_ob

(שממקסמים אותה) הינה יחס בין אקספוננט של דמיון של "הזוג הקרוב" (בין הייצוגים שלהם) לסכום הדמיונות בינו לבין כל הזוגות שליליים. למשל שיטת SimCLR כל באטצ' מורכב מ- N זוגות של דוגמאות קרובות (אוגמנטציה של אותה הדוגמא) המהווים את הזוגות החייבים כאשר עבור דוגמא נתונה, כל הדוגמאות פרט ל"בת הזוג" שלה נחשבת לדוגמא שלילית עבורה. פונקציה המטרה לכל באטצ' הינה סכום של פונקציות המטרה של כל N2 דוגמאות של הבאטצ'.

בנק של ייצוגי דוגמאות שליליות: ידוע שהגדלת מספר הזוגות השליליים לכל זוג חיובי באימון תורמת לעוצמת הייצוג של הדאטה. כתוצאה מכך משתמשים בבאטצ'ם מאוד גדולים (עשרות אלפי דוגמאות) שדורש משאבי זכרון גדולים, כח עיבוד רב (צריך לחשב את הייצוג של עשרות אלפי דוגמאות מהבאטצ'). כדי להקטין את כוח העיבוד הנדרש הוצע (MOCO) "בנק הדוגמאות השליליות" מהבאטצים הקודמים המכיל את הייצוגים של הדוגמאות מכמה הבאטצים הקודמים. כל פעם דוגמים משם ייצוגים של דוגמאות שליליות ומוספים את זה לייצוגיים השליליים מהבאטצ' הנוכחי. צריך לזכור שגישה זו כרוכה בהקצאת משאבי אחסון נוספים לשמירת בנק זה.

הישגי מאמר: המאמר מראה ש SwaV משולב עם מולטי-קרופ מצליח לייצר ייצוגים יותר חזקים משיטות בניית ייצוג רבות במספר משימות. ההשוואה בוצעה בדרך הסטנדרטית: הוספה של שכבה לינארית לרשת הבונה ייצוג(עם משקלים מוקפאים) ובחינת ביצועיה על משימה מסוימת. קודם כל הם הראה שייצוג שנבנה באמצעות SwaV מציג ביצועים יותר טובים על דאטה סטים iNaturalist 2018VOC07 ו- 205Places מהייצוגים הנבנים על ImageNet מתויג (!!) גם על משימת סיווג ועל משימת זיהוי אובייקטים. בנוסף הם הראו שהייצוגים שלהם משיגים ביצועים יותר טובים מבחינת 5Top1/Top (לוקחים 5/1 דוגמאות הכי קרובות מבחינת הייצוג ומחשבים כמה מתוכם שייכים לאותה קטגוריה) משיטות כמו SimCLR ו- 2MoCov. נזכיר שלהבדיל מ 2MoCov , אין צורך בשמירה של בנק דוגמאות שליליות ב -SwaV. הם גם הראה את עליונותה של SwaV במשימות semi-supervised על שיטות כמו UDA ו- FixMatch. וזה רק חלק קטן מכל השוואות שהם עשו - הם באמת עשו עבודה מרשימה בהיבט הזה.

לינק למאמר: https://arxiv.org/pdf/2006.09882.pdf

לינק לקוד: **https://github.com/facebookresearch/swav**

נ.ב. מאמר ממש מגניב עם רעיון מגניב המשלב תובנות רבות ממגוון שיטת SSRL. הם גם טרחו להשוות את הביצועים של השיטה שלהם מול מגוון רחב של אלגוריתמים, משימות, דאטה סטים וקונפיגורציות שזה בהחלט מרשים. בקיצור המלצת קריאה לוהטת ממני:)

#deepnightlearners