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:

Unsupervised Learning of Visual Features by Contrasting Cluster Assignments

---

פינת הסוקר:

המלצת קריאה ממייק: חובה בטח לאוהבי למידת הייצוג.

בהירות כתיבה: בינונית פלוס.

רמת היכרות עם כלים מתמטיים וטכניקות של ML/DL הנדרשים להבנת מאמר: הבנה טובה בעקרונות הלוס המנוגד וידע טוב באופטימיזציה.

יישומים פרקטיים אפשריים: למידה ייצוגים חזקים על דאטהסטים לא מתויגים עם תקציב חישוב מצומצם.

פרטי מאמר:

לינק למאמר: זמין להורדה.

לינק לקוד: זמין כאן.

פורסם בתאריך: 08.01.21, בארקיב.

הוצג בכנס: NeurIPS 2020.

תחומי מאמר:

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

כלים מתמטיים, מושגים וסימונים:

- מולטי-קרופ - טכניקת אוגמנטציה המבוססת על לקיחת פאצ'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 דאטהסט לא מתויג. רוב גישות המודרניות בתחום הזה (SSRL) מורכבות משני מרכיבים עיקריים:

- הלוס המנוגד (contrastive loss - CL): מסתמך על ההנחה שייצוגים של דוגמאות קרובות צריכים להיות קרובים, בזמן שייצוגים של דוגמאות לא קשורות (נבחרות רנדומלית בד"כ) צריכים להיות רחוקים.
- שיטה ליצירה של דוגמאות "דומות", קרי אוגמנטציה: בדרך כלל זוג דוגמאות קרובות (אקרא לזוגות האלו בהמשך זוגות חיוביים או זוגות קרובים) נוצר 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 שמצריך כמות גדולה של זכרון ומשאבי עיבוד משמעותיים. דרישות אלו מקשות על יישום של שיטות אלו בצורת אונליין (לטענת המאמר רוב שיטות SSRL היום מיושמות בצורת אונליין שדי הפתיע אותי). אז בואו נדבר על החידושים שהמאמר הזה מציע:

- שיטת אימון SwaV: 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 למועמד טוב למימוש בצורת אונליין.
- שיטת אוגמנטציה מולטי-קרופ: המאמר מציע שיטת אוגמנטציה הנקראת מולטי-קרופ שמתחילה מהחישוב של שני "קרופים סטנדרטיים" x_cr1 ו- x_cr2 של תמונה x. לאחר מכן לוקחים "קרופים קטנים יותר" של x_cr1 ו- x_cr2 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 שיטת SwaV. פונקציית L מוגדרת באופן הבא (לכל דוגמא בבאטץ'):

- בונים מספר אוגמנטציות לדוגמא x עם מולטי קרופ או כל גישה אחרת.
- מרכיבים מאוגמנטציות אלו זוגות של דוגמאות.
- בונים וקטורי ייצוג z_i לכל הדוגמאות שבנינו.
- לכל זוג וקטורי ייצוג (z_1, z_2) מחשבים את הקודים שלהם q_1 ו- q_2.
- מחשבים את סכום הדמיונות l_s בין z_1 ו- q_2 ובין z_2 ל q_1.
- מחשבים את הסכום L_x של l_s של כל הזוגות של הדוגמאות החיוביות של דוגמא x.

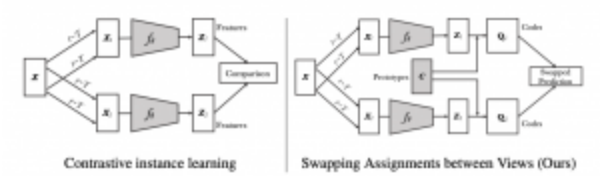

Figure 1: **Contrastive instance learning (left) *vs.* SwAV (right).** In contrastive learning methods applied to instance classification, the features from different transformations of the same images are compared directly to each other. In SwAV, we first obtain "codes" by assigning features to prototype vectors. We then solve a "swapped" prediction problem wherein the codes obtained from one data augmented view are predicted using the other view. Thus, SwAV does not directly compare image features. Prototype vectors are learned along with the ConvNet parameters by backpropagation.

פינת האינטואיציה:

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?

מבנה של פונקציה לוס בין ייצוג z לקוד q (של דוגמאות קרובות): 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 לוקטור מרחקים מנורמל שחישבנו. את הפונקציה זו אנו ממקסמים ביחס ל ייצוגים z וביחס לפרוטוטייפים c.

פינת האינטואיציה:

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 ייצוגים של דוגמאות חיוביות להיות רחוקים בצורה כמה שיותר דומה מכל הקלסטרים השליליים וקרובים באותה מידה מהקלסטרים החיוביים. לדעתי זה הנקודה הכי משמעותית במאמר (!!).

הסבר על בניית קוד q של ייצוג z: 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 באטץ' בודד בלבד(!!). אפשר להגיד שהקוד q מייצג את ההסתברויות שיוך של וקטור הייצוג z של הדוגמא נתונה לקלסטרים המיוצגים ע"י 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 הקלסטרים השונים). שימו לב שבעיית אופטימיזציה זו מזכירה בצורתה את בעיית הטרנספורט האופטימלי בין מידות הסתברות דיסקרטיות (האחידות) המוגדרות על שני דאטהסטים. את התפקיד של דאטהסטים 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 (j,k) שלה מגדיר את המסה ההסתברותית המועברות מווקטור z_k לוקטור c_j, כלומר הסתברות השיוך של z_k לקלסטר של c_j. מכיוון שאנו רוצים שאותו מספר דוגמאות "ישויך" לכל קלסטר, מוסיפים אילוץ על סכום השורות וסכום העמודת של Q. בעיה זו פותרים בעזרת אלגוריתם איטרטיבי של סינקהורן-קנופ.

הסבר על מושגים חשובים במאמר:

שיטות אימון של גישות SSRL המודרניות: 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 בשיטת SimCLR כל באטצי 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 באטץ' הינה סכום של פונקציות המטרה של כל 2N דוגמאות של הבאטץ'.

בנק של ייצוגי דוגמאות שליליות: ידוע שהגדלת מספר הזוגות השליליים לכל זוג חיובי באימון תורמת לעוצמת הייצוג של הדאטה. כתוצאה מכך משתמשים בבאטצ'ים מאוד גדולים (עשרות אלפי דוגמאות) שדורש משאבי זכרון גדולים, כח עיבוד רב (צריך לחשב את הייצוג של עשרות אלפי דוגמאות מהבאטץ'). 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 ומוסיפים את זה לייצוגים השליליים מהבאטצי 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 עבור מספר 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 דאטהסטים VOC07 iNaturalist2018, ו- Places205 מהייצוגים הנבנים על ImageNet מתויג (!!) גם על משימת סיווג ועל משימת זיהוי אובייקטים. בנוסף הם הראו שהייצוגים שלהם משיגים ביצועים יותר טובים מבחינת Top1/Top5 (לוקחים 1/5 דוגמאות הכי קרובות מבחינת הייצוג ומחשבים כמה מתוכם שייכים לאותה קטגוריה) מ- SimCLR ו- MoCov2. נזכיר שלהבדיל מ-MoCov2 , 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.

Method	Arch.	Param.	Top1
Supervised	R50	24	76.5
Colorization [65]	R50	24	39.6
Jigsaw [46]	R50	24	45.7
NPID [58]	R50	24	54.0
BigBiGAN [15]	R50	24	56.6
LA [68]	R50	24	58.8
NPID++ [44]	R50	24	59.0
MoCo [24]	R50	24	60.6
SeLa [2]	R50	24	61.5
PIRL [44]	R50	24	63.6
CPC v2 [28]	R50	24	63.8
PCL [37]	R50	24	65.9
SimCLR [10]	R50	24	70.0
MoCov2 [11]	R50	24	71.1
SwAV	R50	24	**75.3**

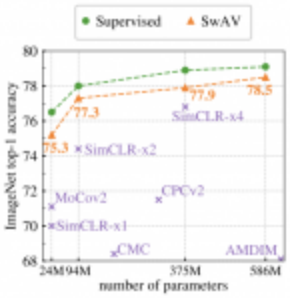

Figure 2: **Linear classification on ImageNet.** Top-1 accuracy for linear models trained on frozen features from different self-supervised methods. **(left)** Performance with a standard ResNet-50. **(right)** Performance as we multiply the width of a ResNet-50 by a factor ×2, ×4, and ×5.

Table 1: **Semi-supervised learning on ImageNet with a ResNet-50.** We finetune the model with 1% and 10% labels and report top-1 and top-5 accuracies. *: *uses RandAugment [12].*

	Method	1% labels Top-1	1% labels Top-5	10% labels Top-1	10% labels Top-5
	Supervised	25.4	48.4	56.4	80.4
Methods using label-propagation	UDA [60]	-	-	68.8*	88.5*
	FixMatch [51]	-	-	**71.5***	89.1*
Methods using self-supervision only	PIRL [44]	30.7	57.2	60.4	83.8
	PCL [37]	-	75.6	-	86.2
	SimCLR [10]	48.3	75.5	65.6	87.8
	SwAV	**53.9**	**78.5**	70.2	**89.9**

נ.ב.

מאמר ממש מגניב עם רעיון מסקרן 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

הפוסט נכתב על ידי מיכאל (מייק) ארליכסון, Michael Erlihson, 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.