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

Improving Self-supervised Learning with Automated Unsupervised Outlier Arbitration

---

פינת הסוקר:

המלצת קריאה ממייק ואברהם: שווה קריאה לחובבי למידה ייצוגית (unsupervised learning)

בהירות כתיבה: בינונית

ידע מוקדם:

- עקרונות בסיסיים של למידה ייצוגית (representation learning)
- שיטות אימון ניגודיות (contrastive learning)
- importance sampling

יישומים פרקטיים אפשריים: שיפור באיכות הייצוג של דאטה המופק באמצעות מגוון שיטות של למידה ייצוגית

פרטי מאמר:

מאמר: זמין להורדה.

קוד: כאן

פורסם בתאריך: 15.12.21, בארקיב.

הוצג בכנס: NeurIPS 2021 Poster

תחומי מאמר:

- self-supervised learning

כלים מתמטיים, מושגים וסימונים:

- Importance sampling

- Radon-Nykodim derivative

מבוא:

אחת הדרכים המקובלות ביותר להגדיל את מאגר הדאטה לאימון של מודל היא באמצעות אוגמנטציות - לוקחים דוגמאות מהמאגר שיש לנו ומשנים אותן במגוון צורות. למשל, נניח ויש לנו תמונה של חתול, אז גם סיבוב של התמונה, חיתוך שלה בזוויות שונות, שינוי הגודל ועוד כמה טרנספורמציות (אוגמנטציות) נוספות, אינן אמורות לשנות את העובדה שבתמונה יש חתול. על ידי פעולות פשוטות אלו ניתן לייצר מתמונה אחת עוד תמונות רבות השייכות לאותה קטגוריה. העובדה שאוגמנטציות מסוימות של הדוגמה המקורית שייכות לאותה הקטגוריה כמו הדוגמה המקורית הינה בעלת ערך מוסף, כיוון שלא רק הגדלנו את הדאטהסט אלא הוספנו דוגמאות מתויגות. גם כאשר אין לנו כלל דאטה מתויג ואנו רוצים להפיק ייצוג חזק של דאטה שעשוי לשמש למשימות downstream, אוגמנטציות יכולות להועיל מאוד לה כדי "לרמוז" למודל לגבי פיצ'רים סמנטיים חשובים של הדאטה.

ההנחה המסתתרת מאחורי השימוש באוגמנטציות הינה פשוטה והגיונית, אך כפי שנראה, היא לא תמיד מתקיימת. בעצם אנו מניחים כי הדוגמאות הן אינווריאנטיות לאוגמנטציות, כלומר טרנספורמציות מסוימות אינן משנות את המאפיינים העיקריים של הדוגמה: התמונה לאחר אוגמנטציה עדיין שייכת לאותה קטגוריה, והתוכן הסמנטי שלה נשמר. במאמר הנסקר מערערים על הנחה זו ומראים שהיא עלולה להיות בעייתית. המחברים מראים דוגמאות שעבורן אוגמנטציות מסוימות משנות את התוכן הסמנטי של התוצאה.

המחברים דנים באימון מודלים על דאטה לא מתויג, הנקראת למידת ייצוג (representation learning), ומראים כי אוגמנטציות עלולות לפגוע ביכולת ההכללה של המודל המאומן, ומציעים פתרון אלגנטי להתמודדות עם בעיה זו.

הרעיון הבסיסי:

הסבר על הבעיה:

כאמור המאמר מנסה לטפל בבעית דוגמאות חיוביות "שקריות" שעלולה לעלות במהלך למידת ייצוג (representation learning) בפרט כאלה שמבוססות על שיטות ניגודיות (contrastive methods). בשיטות למידה ניגודית זוג דוגמאות חיוביות מוגדר בנוי משתי אוגמנטציות של אותה דוגמא. ההנחה המרכזית בשיטות אלו אומרת כי דוגמאות אלו הן בעלות אותו תוכן סמנטי. בדומיין הויזואלי (קרי תמונות) זוג של דוגמאות חיוביות יכול להיות שני פאצ'ים שונים (או crops) של אותה תמונה או שתי טרנספורמציות של אותה תמונה (למשל שני סיבובים שונים).

אך גישה פשטנית זו טומנת בחובה מוקש פוטנציאלי: למשל אם בתמונה יש שתי חיות, חתול וכלב, זוג של דוגמאות חיובית עלול להכיל פאץ' בו מוצג כלב ופאץ' שני בו מוצג חתול. פונקציית הלוס הניגודי תנסה לקרב את הייצוגים של שני הפאצ'ים, על אף שהתוכן הסמנטי שלהם מאוד שונה.

נזכיר כי ההנחה המרכזית מאחורי שיטת למידה ניגודית אומרת כי ייצוגים של דוגמאות "דומות" (= חיוביות) צריכים להיות קרובים במרחב הייצוג והייצוגים של דוגמאות לא דומות צריכים להיות רחוקים.

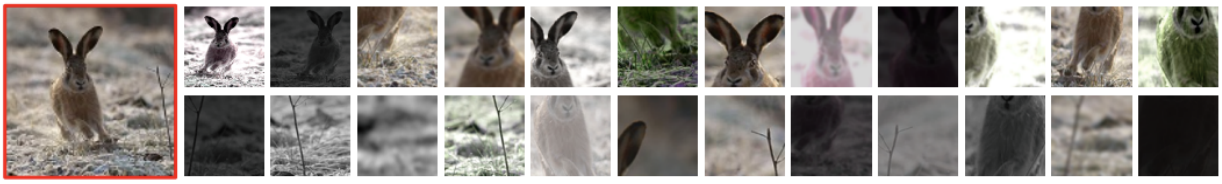

Figure 1: Illustration of OOD samples sampled from distribution $\widetilde{p}$. Figures are ranked according to a descent order of its associated $w_{i,j}$. The biggest image in the red box is the original instance input $\boldsymbol{x}_i$.

<u>הערה</u>: ניתן להשתמש בשיטה המוצעת במאמר גם עבור שיטות למידת ייצוג שלא משתמשות באופן מפורש בהנחת היסוד של הגישה הניגודית (קירוב ייצוגיים של דוגמאות חיוביות והרחקה של אלו עבור דוגמאות שליליות). בין השיטות הללו נמנות בין השאר SwaV ו- ByOL (אשר משתמש בזוגות חיוביים בלבד). בהמשך נדון בעיקר בשימוש בשיטה המוצעת עבור גישות ניגודיות קלאסיות.

המאמר הנסקר מנסה לתת מענה לבעיית הזוגות החיוביים ה-"שקריים" באמצעות משקול אדפטיבי כאשר תרומה של כל זוג חיובי תהיה פרופורציאונלית למידת "אמיתיות" שלו. כלומר, עבור זוג דוגמאות חיוביות נשערך את הסבירות שהן אכן מכילים את אותו התוכן הסמנטי. ככל שהסבירות הזאת גבוהה יותר, הציון שהזוג מקבל הוא גבוה יותר והתרומה שלו לפונקציית הלוס גם היא גבוהה יותר. גישה זו דומה ל-Importance Sampling, טכניקה לדגימה מהתפלגויות מורכבות הממשקלת באמצעות דגימה והתפלגות פשוטה יותר ומשקול של דגימות אלו.

פרטים נוספים:

כעת נבין את העקרונות המתמטיים של הגישה המוצעת במאמר. קודם כל ניזכר במבנה של פונקציית הלוס של שיטת למידת ייצוג כלשהי. פונקציית לוס זו היא ממוצע של ערכי הלוס על פני מספר מיני-באצ'ים בנויים מזוגות חיוביים (ולפעמים מיני-באץ' כולל גם דוגמאות שליליות, הנבחרות בדרך כלל רנדומלית). ערך פונקציית לוס של מיני-באץ' הוא סכום ערכי פונקציית L עבור הזוגות המרכיבים מיני-באץ'. אבל מה היא L?

פונקציית L מקבלת ייצוגים של דוגמאות חיוביות (ולפעמים שליליות) ומודדת "מידת התאמה של ייצוגים אלו להנחת היסוד של השיטה".

דוגמאות של פונקציה עבור כמה שיטות למידת הייצוג:

- **MoCO** - פונקציה L היא יחס מרחקים בין ייצוגים של זוגות שליליים לבין אלה של הזוגות החיוביים.
- **SWaV** - פונקציה L היא מרחק בין קלאסטרים של דוגמאות חיוביות.
- **ByOL** -פונקציה L היא מרחק L2 בין ייצוגי דוגמאות חיוביות (אין דוגמאות שליליות).

לאחר מכן, מגרילים סט דוגמאות חיוביות (ולפעמים גם שליליות) ומחשבים את הערך של L על הסט הזה.

עכשיו נשאלת השאלה מאיזו התפלגות אנו דוגמים טרנספורמציות (אוגמנטציות) המניבות לנו זוגות חיוביים. קודם כל, רצוי להשתמש בזוגות של טרנספורמציות שייצרו תמונות בעלות תוכן סמנטי זהה, או לפחות דומה, בסבירות גבוהה. אולם חוק התפלגות (נסמן אותו ב-P) של זוגות טרנספורמציות כאלו אינו ידוע ובנוסף הטרנספורמציות עשויות להיות תלויות בדוגמה עצמה. כלומר, שני קרופים זהים יכולים להכיל תוכן סמנטי זהה לתמונה אחת (למשל שני חתולים) ותוכן שונה לגמרי (למשל חתול ודשא) עבור תמונה אחרת.

במטרה להתגבר על המכשול הזה ולדגום מההתפלגות הטרנספורמציות הזו, נשתמש בשיטה הנקראת importance sampling (IS). כאמור IS מאפשרת לדגום מהתפלגות P, אשר ניסיון לדגום ממנה באופן ישיר יצריך דגימות רבות, באמצעות התפלגות אחרת Q, ממנה ניתן לדגום יותר בקלות. המשקל של דגימה x הוא היחס P(x)/Q(x). כלומר, דגימות בעלות יחס גבוה מקבלות משקל גדול יותר ולהיפך. במקרה שלנו, ההתפלגות Q תהיה ההתפלגות שממנה אנו דוגמים אוגמנטציות לבנייה של זוגות חיוביים של דוגמאות.

אבל יש לנו בעיה כאן. אנחנו לא יודעים לחשב את P(x) באופן מפורש, כלומר אין לנו דרך לדעת מהי ההסתברות שזוג אוגמנטציות יוביל לתמונות בעלות אותו תוכן סמנטי. למזלנו אנחנו כן יכולים לחשב בקירוב את היחס של P ו-Q וזה למעשה הרעיון העיקרי של המאמר. הקירוב מתבסס על כך שרוב האוגמנטציות שבדרך כלל משתמשים בהם בלמידת ייצוג כן שומרות על אותו תוכן סמנטי ויש יחסית מעט כאלו שלא מקיימות את התכונה הזו. התובנה העיקרית היא שניתן לזהות אוגמנטציות שלא מקיימות את התכונה הזאת באמצעות ניתוח של מרחק מהייצוג הממוצע z_{mean} המתקבל מתמונה נתונה לאחר הפעלת מספר אוגמנטציות (מסומן בתור M במאמר). כלומר אם ייצוג של תמונה לאחר אוגמנטציה A רחוק מדי מ- z_{mean} (גם השונות נלקחת בחשבון כאן) אז ניתן להסיק כי A אינה שומרת על התוכן הסמנטי.

אם כך, המשקל w של אוגמנטציה A מוגדר כמרחק ייצוג הדוגמה לאחר A מהייצוג הממוצע z_{mean} (במונחי שונות המחושבת על פני כל הדוגמאות במיני-באץ'):

$$w_{i,j} = \exp[-(\boldsymbol{z}_{i,j} - \boldsymbol{\mu}_i)^T (\tau \boldsymbol{\Sigma})^{-1} (\boldsymbol{z}_{i,j} - \boldsymbol{\mu}_i)],$$

$$\boldsymbol{\mu}_i = \frac{1}{M} \sum_j^M \boldsymbol{z}_{i,j}, \qquad \boldsymbol{\Sigma} = \frac{1}{NM} \sum_i^N \sum_j^M (\boldsymbol{z}_{i,j} - \boldsymbol{\mu}_i)(\boldsymbol{z}_{i,j} - \boldsymbol{\mu}_i)^T.$$

כאשר הפרמטר τ שולט ב"יחס התלות" של משקל w במרחק הנ"ל, וככל שהוא נמוך יותר המשקל יהיה יותר תלוי במרחק מהממוצע.

אופן מעט יותר פורמלי: נסמן את ההתפלגות של האוגמנטציות ה"רצויה" ב-P. התפלגות זו אינה ידועה לנו, ונדגום ממנה באמצעות היוריסטיקה מהתפלגות Q, המכילה את כל האוגמנטציות "הרגילות" שבדרך כלל שומרות על התוכן הסמנטי כמו קרופ או סיבוב. ההתפלגות Q כן ידועה לנו ונרצה להיעזר בה על מנת לדגום מ- P באמצעות שיטת importance sampling(IS). כעבור זוג דוגמאות x_i ו-x_j (שהן התמונה המקורית וגרסתה לאחר הפעלת אוגמנטציה A_j), נרצה לחשב את $P(A_j)/Q(A_j)$. כיוון שאנחנו לא יכולים לחשב את היחס באופן מפורש, כפי שאיננו יודעים את ההתפלגות הרצויה של אוגמנטציות P, נרצה לשערך את היחס הזה (נסמן אותו בתור w_{ij}). משקל w_{ij} ייכנס לפונקציית הלוס, וינסה לתת לאוגמנטציה הזו את המשקול "ההולם" עבורה ביחס לדוגמה המקורית. ככל שהאוגמנטציה שומרת יותר על התוכן הסמנטי של התמונה כך המשקל שלה יהיה גבוה יותר ואילו משקל האוגמנטציות ה"לא טובות" יהיה קרוב ל-0. ניתן לשערך את הערך של w_{ij} באמצעות מרחק (מאוד דומה למרחק Mahalanobis) של ייצוג התמונה לאחר אוגמנטציה A_j מהייצוג הממוצע z_{mean} של M אוגמנטציות שונות של התמונה.

הישגי מאמר:

המאמר השווה אספקטים רבים של ביצועי הגישה המוצעת עם שיטות SOTA רבות, אך החשובה העיקריות ביניהם היא ההשוואת ייצוגים המופקים באמצעות הטכניקה המוצעת עם אלו של שיטות unsupervised אחרות עבור משימות סיווג וזיהוי אובייקטים. השיטה המוצעת מצליחה להשיג את הביצועים הטובים ביותר, אך השיפור על פני אלגוריתמים אחרים לא גדול.

Table 6: Accuracy of linear classification model on ImageNet1K. Bold numbers are the best performance among models trained for 200 epochs. Numbers **(+x%)** denotes additional gain compared to the baseline model (i.e., SwAV here in the table) without UOTA approach. † denotes results represented from [3, 9]. ‡ means results of our reproduced reproduced based on SwAV official code.

Method	Epochs	Batch size	Top-1 (%)	Top-5 (%)
CPC v2 [21]	200	512	63.8	85.3
CMC [39]	240	/	64.8	86.1
MoCo [18]	200	256	60.6†	83.1†
MoCo v2 [8]	200	256	67.6†	88.0†
JCL [3]	200	256	68.7	89.0
SimCLR [7]	1000	4096	69.3	89.0
SimSiam [9]	200	256	70.0	/
InfoMin Aug. [40]	200	256	70.1	89.4
BYOL [15]	200	4096	70.6†	/
Barlow Twin [47]	1000	2048	73.2	91.0
SwAV [5]	200	256	72.7‡	91.5‡
SwAV+UOTA (Ours)	200	256	**73.5 (+0.8%)**	**91.8 (+0.3%)**

Table 7: Performance on downstream tasks: object detection [35] (left), instance segmentation [19] (middle) and keypoint detection [19] (right). Accuracy in %. All models pretrained 200 epochs and finetuned on MS COCO with 1 × schedule.

Model	*Faster R-CNN + R50-FPN*			*Mask R-CNN + R50-FPN*			*Keypoint R-CNN + R50-FPN*		
	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}	AP^{kp}	AP^{kp}_{50}	AP^{kp}_{75}
random	30.1	48.6	31.9	28.5	46.8	30.4	63.5	85.3	69.3
supervised	38.2	59.1	41.5	35.4	56.5	38.1	65.4	87.0	71.0
MoCo-v1[18]	37.1	57.4	40.2	35.1	55.9	37.7	65.6	87.1	71.3
MoCo-v2 [8]	37.6	57.9	40.8	35.3	55.9	37.9	66.0	87.2	71.4
JCL [3]	38.1	58.3	41.3	35.6	56.2	38.3	66.2	87.2	**72.3**
InfoMin Aug. [40]	/	/	/	**36.7**	57.7	**39.4**	/	/	/
SwAV	38.5	60.5	41.4	36.3	57.7	38.9	65.6	86.9	71.6
SwAV+UOTA	**39.0**	**61.0**	**42.0**	**36.7**	**58.4**	**39.4**	**66.3**	**87.4**	**72.3**

נ.ב.

לסיכום, מדובר במאמר מאוד מעניין המציע דרך להתמודד עם אוגמנטציות שלא משמרות את התוכן הסמנטי של תמונה במהלך למידה ייצוגית על דאטהסטים לא מתויגים.

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

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.

אברהם סטודנט לתואר שלישי בתחום של למידת מכונה באוניברסיטת בר-אילן ועובד בחברת סמסונג. מתעניין בלמידה עמוקה ומגוון יישומים כולל ראייה ממוחשבת ועיבוד שפה טבעית.

מיכאל ואברהם כתבו יחד ספר על למידת מכונה ולמידה עמוקה בעברית.

ברצוננו להודות לעדו בן-יאיר על עזרה בהגהה ובחידוד של נקודות חשובות.