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:

DETReg: Unsupervised Pretraining with Region Priors for Object Detection

---

פינת הסוקר:

המלצת קריאה ממייק: חובה לעוסקים בזיהוי אובייקטים בתמונות.

בהירות כתיבה: גבוהה.

רמת היכרות עם כלים מתמטיים וטכניקות של ML/DL הנדרשים להבנת מאמר: נדרשת היכרות עם DeTR, שיטות למידת ייצוג בצורה unsupervised וטרנספורמרים.

יישומים פרקטיים אפשריים: pretraining של מודל לזיהוי אובייקטים בדומיינים עם כמות מועטה של דאטה מתויג.

פרטי מאמר:

לינק למאמר: זמין להורדה.

לינק לקוד: זמין כאן.

פורסם בתאריך: 08.06.21, בארקיב.

הוצג בכנס: טרם ידוע.

תחומי מאמר:

- זיהוי אובייקטים בתמונה (Object Detection)

ידע מוקדם:

- למידת ייצוג של דאטה לא מתויג (representation learning)
- Region proposals
- (Detection with transformers (DETR
- SwaV1 (סקירה שלי בעברית), SwaV2 (סקירה של רחל שלום באנגלית)
- טרנספורמרים למשימות הראייה הממוחשבת (בפרט למשימות זיהוי אובייקטים)
- Selective Search
- אלגוריתם התאמת הזוגות ההונגרי (Hungarian bipartite matching algorithm)

מבוא:

זיהוי אובייקטים בתמונה הינה משימת ראייה ממוחשבת קלאסית שמטרתה איתור מיקום של אובייקטים בתמונה בנוסף לזיהוי הקטגוריה של כל אובייקט. בדרך כלל נדרש דאטהסט מתויג גדול כדי לאמן רשת לזיהוי אובייקטים בתמונה בדיוק גבוה (הדיוק מתייחס גם למיקום וגם לקטגוריה של האובייקטים). דאטהסט מתויג למשימת זיהוי אובייקטים מכיל תמונות עם bounding boxes(BB) לכל אובייקט והקטגוריה שלו כאשר מספר האובייקטים בתמונה עשוי להיות די גדול. בניית דאטהסטים כאלו יכולה להיות עסק די יקר. עקב כך נוצר צורך בבניית ייצוגיים "טובים" של תמונה שניתן ללמוד אותם בצורה unsupervised (קרי ללא דאטה מתויג) לצורך pretraining של מודל לזיהוי אובייקטים. ייצוג "טוב" של תמונה בהקשר זה יאפשר להקטין בצורה משמעותית גודל דאטהסט הנדרש לאימון (למעשה לכיול - fine-tuning) של מודל לזיהוי אובייקטים.

בשנים האחרונות יצאו מחקרים רבים המציעים שיטות ללמידת ייצוג של תמונה ללא דאטהסט מתויג. לטענת המאמר שיטות אלו לא מצליחות לבנות ייצוג של תמונה שהוא "מטורגט" למשימה של זיהוי אובייקטים. כלומר ייצוגי תמונות הנבנים באמצעות השיטות הקיימות לא מצליחים "לדחוף" כמות מספקת של "מידע רלוונטי לזיהוי אובייקטים בתמונה לוקטור הייצוג של התמונה. ייתכן שהסיבה לכך היא שוני גדול בין אופיים של תהליכי למידת (אימון) ייצוג unsupervised של תמונה לבין תכונות הנחוצות (במאמר קוראים להם תכונות objectness) עבור משימת זיהוי אובייקטים. בגדול המטרה של שיטות אימון של ייצוג unsupervised הקיימות היא "לקרב ייצוגים של תמונות דומות ולהרחיק ייצוגים של תמונות לא דומות". כנראה ייצוג בעל תכונה זו לא מכיל מספיק מידע רלוונטי למשימת זיהוי אובייקטים.

למיטב ידיעתי, לא קיימת שיטה לבניית ייצוג של תמונה בצורה unsupervised, המאומנת על משימה "דומה" לזיהוי אובייקטים.

תמצית מאמר:

המאמר הנסקר מציע שיטה, הנקראת DETReg לבניית ייצוג של תמונה כך שהוא יכיל מידע רלוונטי למשימת זיהוי אובייקטים (כלומר מידע על מיקום וסוג האובייקט). השיטה המוצעת היא למעשה זיהוי אובייקטים בתמונה. אבל איך ניתן לבנות משימה כזו כאשר אין ברשותנו דאטהסט מתויג? המחברים השתמשו בשיטה קלאסית (שהוצעה עוד ב-2013) לזיהוי אובייקטים בתמונה הנקראת Selective

Search או בקיצור SS. המאמר מציע לנצל BB-ים (ללא קטגוריה של אובייקט) המחושבים באמצעות SS למטרת pretraining של DETReg.

(a) Def. DETR [59] w/ SwAV [6] (b) UP-DETR [12] (c) DETReg (Ours)

Figure 1: **Prediction examples of unsupervised pretraining approaches**. Recent methods, shown in (a) and (b), do not learn "objectness" during the pretraining stage. In contrast, our method DETReg (c) learns to localize objects more accurately in its pretraining. The included prediction examples were obtained after pretraining and before finetuning with annotated data.

אבל זה לא מספיק בשביל לאמן ייצוג חזק לזיהוי אובייקטים! צריך לזכור כי המטרה של אימון DETReg היא לבנות ייצוג של תמונה המכיל אינפורמציה על מיקומים ועל סוגים של האובייקטים בתמונה (שזו למעשה המטרה של משימת זיהוי אובייקטים). מידע על מיקום האובייקטים מועבר באמצעות BB-ים המסופקים באמצעות SS. כעת נשאלת השאלה איך "להעביר מידע **על סוג האובייקטים** לייצוג התמונה" במהלך pretraining? המאמר מציע מנסה לכפות על ייצוגים של BB-ים (שהם למעשה פאטצ'ים של תמונה), הנבנים באמצעות DETReg להיות קרובים לייצוגים של BB-ים המוצעים ע"י SS.

אבל באיזה ייצוג נשתמש כדי "להעביר" ל-DETReg את האינפורמציה על סוג האובייקט בכל BB? נזכור כי שיטות unsupervised מצליחות להפיק ייצוג של תמונה המכיל מידע על סוגי האובייקטים בתמונה. למעשה ייצוגים של תמונות עם אותו סוג של אובייקטים (שייכים לאותה קטגוריה) "קרובים" במרחב הייצוג כאשר אלו של התמונות מקטגוריות שונות רחוקים יותר. למעשה ייצוגים של תמונות מאותה קטגוריה מהווים קלאסטרים במרחב הייצוג והקלאסטרים של קטגוריות שונות "מופרדים" זה מזה.

המאמר בחר בשיטה הנקראת SwaV לייצוג של BB-ים המופקים באמצעות SS. ד"א, סקרתי מאמר זה בעבר ובנוסף יש סקירה מעולה של Rachel Shalom באנגלית למי שרוצה להבין את השיטה המעניינת הזו(SwaV) לעומק. זאת אומרת "התיוגים" שעליהם מאומן DETReg הם:

1. BB-ים המחושבים באמצעות SS.
2. ייצוגי SwaV של BB-ים אלו.

לב הרעיון של DETReg הוא ללמוד ייצוג של אובייקטים תמונה כאשר מטרת האימון היא:

1. להפיק BB-ים דומים לאלו המופקים באמצעות Selective Search.
2. לכפות על ייצוגי SwaV של BB-ים "מתאימים" (יפורט בהמשך) של SS ו-DERTeg להיות קרובים.

תקציר המאמר:

לאחר שהבנו את הרעיון העיקרי של המאמר הנסקר, נתבונן כעת בפרטי האימון של DETReg. למעשה תהליך האימון מורכב משני שלבים:

1. הפעלת אלגוריתם SS על תמונות מהדאטהסט (לא מתויג).

נציין כי SS פולט מספר רב של BB-ים כאשר רובם מכילים רק חלק מאובייקט או לא מכילים אובייקטים כלל. עקב כך המאמר מציע לאחד את האיזורים המוצעים (region proposals) על סמך הדמיון ביניהם. דמיון זה תלוי בקרבה בין המאפיינים שלהם (כגון צבע, טקסטורה, צורת האיחוד ביניהם וכדומה). המאמר מציע מספר אסטרטגיות לבחירה של מועמדים לאיחוד (Top-K, k-random) ואחת שמבצעת importance sampling בהתבסס על הציונים של האיזורים).

הערה: ראה פרק "הסבר על מושגי היסוד" להסבר קצר על SS.

2. אימון של DETReg על BB-s שהתקבלו בשלב 1.

לאחר שבנינו "דאטהסט מתויג", נותר לנו "רק" לאמן את הרשת עליו.

למעשה נותר לנו לתאר רק את הארכיטקטורה ואת פונקציית הלוס של DETReg. המחברים בחרו להשתמש בארכיטקטורה שהוצעה במאמר Deformable DETR, שהיא שכלול של מאמר מפורסם של קבוצת מחקר מ-Facebook AI, הנקרא DETR (לתיאור קצר של הגישה של DETR ראה פרק "הסבר על מושגי היסוד"). Deformable DETR מציע להקטין את הסיבוכיות החישובית של DETR באמצעות חישוב משקלי ה-cross ו-self-attention באופן יותר לוקאלי שבפועל מקטין את מספר החישובים באנקודר ובדקודר של הטרנספורמר. כאמור DETReg משתמש בארכיטקטורה של Deformable DETR ל-pretraining כאשר הדאטהסט הוא הפלטים של SS לאחר האיחוד.

Figure 2: The **DETReg** pretext task and model. We pretrain a Deformable DETR [59] based detector to predict region proposals and their corresponding object embeddings in the pretraining stage.

הדבר האחרון שנשאר לנו לדבר עליו זו פונקציית לוס של DETReg. היא מורכבת מסכום משוקלל של שלושה לוסים הבאים:

- **לוס על המיקום של BB:**

משתמשים ב- Generalized Intersection Over Union (GIoU) - די סטנדרטי בסך הכל :) .

- **לוס על ייצוג SwaV של הפאטץ' של התמונה המוגדר באמצעות BB:**

מחושב כנורמה L1 של ההפרש בין ייצוגי SwaV של DETReg לבין אלו של אלו של ה-ground truth (שהופקו באמצעות SS).

- **לוס על קטגוריה של אובייקט:**

כאן נראה שיש לנו בעיה. הרי SS לא מוציא לנו קטגוריה אלא רק BB-ים. המחברים מצאו פתרון אלגנטי לסוגיה הזו. הם הניחו כי מספר ה-BB-ים של DETReg פולט (נסמן אותו ב-N) הוא יותר גדול ממספר ה-BB-ים המופקים באמצעות SS (נסמן אותו ב-M). אז המחברים הוסיפו N-M פסאודו BB-ים ל-SS ותייגו אותם עם קטגוריה 0, כאשר ה-BB-ים האמיתיים קיבלו לייבל 1. כעת DETReg מנסה לחזות הסתברויות לשתי קטגוריות בלבד - BB המכיל אובייקט אמיתי (לייבל 1) ו-BB עם הרקע (לייבל 0). באופן זה המשימה של "זיהוי קטגוריה" הופכת לבעיית סיווג בינארית כאשר פונקצית לוס עבורה היא Focal Loss.

הערה: DETR המקורי משתמש בקטגוריה של רקע ל-BB-ים ללא אובייקט בתוכם (מספר BB-ים בפלט של DETR הוא קבוע).

הסבר על מושגי היסוד במאמר:

הסבר על Selective Search:

שיטה זו מאתרת "אזורים החשודים להימצאות אובייקטים בהם". אזורים אלו מחושבים באמצעות תהליך איטרטיבי שמקבץ באופן היררכי אזורים קטנים יותר על סמך הדמיון והקרבה שלהם. SS לא דורש אימון ולא נדרשת התערבות אנושית כדי להפעיל אותו (כמובן קיים מימוש בפייטון). SS מוציא גם ציונים לכל BB שמודד סבירות של הימצאות האובייקט שם (למעשה האלגוריתם ממיין את האיזורים לפי הצפי של הימצאות אובייקט בו).

תיאור קצר של DETR:

המאמר המקורי DETR מציע להשתמש בטרנספורמרים (כולל אנקודר דקודר) לבניית מודל לזיהוי אובייקטים בתמונה (בצורה unsupervised). נזכיר הפלט של DETR הוא סט S_mod של BB-ים עם יחד עם התפלגות מעל הקטגוריות של אובייקט בתוך BB. לאחר מכן DETR משתמש באלגוריתם התאמת הזוגות ההונגרי (Hungarian bipartite matching algorithm) שמחפש "התאמה מקסימלית" (מבחינת מיקום וקטגוריה) בין סט S_gt של BB-ים (עם הקטגוריה) האמיתיים (ground truth) לבין סט של BB-ים שזוהו באמצעות המודל. כלומר המטרה היא לבנות את זוגות ה-BB-ים הדומים ביותר מאיברי S_mod ו-S_gt. לאחר שזוגות אלו אותרו, מחשבים פונקצית לוס שהיא סכום ה"מרחקים" של הזוגות שנבנו (מרחק של זוג מודד את מידת השוני בין מיקום של BB-ים ולבין הקטגוריות של איברי הזוג). ליותר פרטים על DETR אתם מוזמנים לקרוא את הסקירה המעולה של אברהם רביב.

הישגי מאמר:

המאמר מראה כי DETReg שעבר pretraining על ImageNET מציג ביצועים טובים יותר משיטות pretraining אחרות (כמו SwaV ו-MOCOv2). בנוסף DETReg (לאחר pretraining) הציג ביצועים טובים יותר מהמתחרים כאשר הוא מכויל על חלק קטן של דאטהסטים מתויגים לזיהוי אובייקטים.

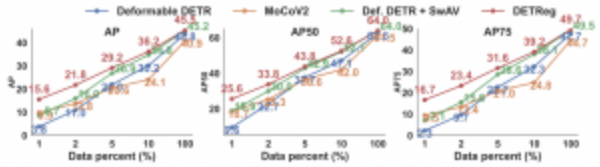

Figure 3: Object detection results finetuned on MS COCO `train2017` and evaluated on `val2017`. DETReg consistently outperforms previous pretraining approaches by a large margin. When fine-tuning with 1% data, DETReg improves 5 points in AP over prior methods.

נ.ב.

מאמר עם רעיון מגניב המאפשר pretraining של מודל לזיהוי אובייקטים בתמונות ללא דאטה מתויג. הגישה המוצעת הצליחה לשפר בצורה משמעותית את ביצועי המודל לאחר כיול (גם על דאטהסטים קטנים).

#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.