סקירה זו נכתבה בשיתוף עם אברהם רביב.

סקירה זו היא חלק מפינה קבועה בה שותפיי ואנוכי סוקרים מאמרים חשובים בתחום ה-ML/DL, וכותב גרסה פשוטה וברורה יותר שלהם בעברית. במידה ותרצו לקרוא את המאמרים הנוספים שסיכמנו שותפיי ואנוכי, 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 בחרנו לסקירה את המאמר שנקרא:

PIX2SEQ: A LANGUAGE MODELING FRAMEWORK FOR OBJECT DETECTION

פינת הסוקרים:

המלצת קריאה מאברהם וממייק: מומלץ מאוד לחובבי לחובבי תחום זיהוי האובייקטים

בהירות כתיבה: גבוהה

ידע מוקדם:

- יסודות של מודלי שפה
- יסודות של שיטות מבוססות רשתות נוירונים לזיהוי אובייקטים

יישומים פרקטיים:

- ניתן ליישם אותה לבניית מודלים לזיהוי אובייקטים בתמונות.

פרטי מאמר:

מאמר: כאן

קוד: כאן

פורסם בתאריך: ארקיב,27.05.2022

הוצג בכנס: ICLR 2022

תחומי מאמר:

- יסודות של זיהוי אובייקטים בתמונות
- מודלי שפה אוטורגרסיביים

כלים מתמטיים, מושגים וסימונים:

- טרנספורמרים (אנקודר ודקודר)
- Bounding Boxes

מבוא:

זיהוי אובייקטים היא משימה מאוד נפוצה בעולם של ראייה ממוחשבת, ויש לה מספר רב של יישומים מגוונים. ניקח למשל מכונית אוטונומית, שבכל רגע צריכה לזהות את האובייקטים שבסביבתה ולקבל תמונת מצב עדכנית על המתרחש, או לחילופין מצלמה של טלפון נייד שיודעת לזהות פנים של בנאדם בכדי לבצע עליהם פוקוס או לתקן רעשי רקע, ועוד המון יישומים במגוון תחומים. משימת זיהוי אובייקטים מורכבת משתי תתי משימות – מציאת המיקום של האובייקט (Localization/Regression) וסיווג האובייקט לקטגוריה (Class) הנכון (Classification). כמובן שניתן להכליל את משימת Object Detection גם למספר אובייקטים, כאשר במקרה זה על המודל לספק מספר Bounding Boxes(BB) ועבור כל אחד מהם לזהות את הקטגוריה שלו.

כאמור הפלט של מודל לזיהוי אובייקטים מורכב ממיקומי האובייקטים, המתואר באמצעות מלבן (BB) המכיל אותו, והקטגוריה של כל אחד מהם. כל מלבן כזה מתואר באמצעות ערכי הקודקוד הימני העליון (x_{min}, y_{min}) וערכי הקודקוד הימני התחתון (x_{max}, y_{max}). בנוסף, לכל אובייקט יש ערך נוסף המייצג של המחלקה אליה הוא שייך, ובסך הכל כל אובייקט מתואר באמצעות tuple של חמישה ערכים: {x_{min}, y_{min} ,x_{max}, y_{max}, Category}. כדי לאמן מודל לזיהוי אובייקטים צריך לבנות פונקצית לוס המורכבת משני חלקים:

1. **לוס רגרסיה** שמטרתו לשפר את דיוק ה-BB שהמודל מספק. איבר זה בפונקציית הלוס "יעניש" את המודל ככל שערכי ה-BB של הפלט יהיו רחוקים מערכי ה-BB האמיתיים (ground-truth).
2. **לוס סיווג** הבודק האם המודל סיווג את האובייקט לקטגוריה הנכונה. בדרך כלל מקובל להשתמש ב-cross-entropy loss, באופן דומה לשימוש בו במשימות סיווג רגילות.

בעבודות אחרות, למשל DETR (סקרנו אותה בעבר), פונקצית המחיר שאמורה **לשפר את הרגרסיה(למה רק רגרסיה)** על ה-BB מבוססת על Hungarian Loss. כך גם בעבודות המשך, כמו למשל DETReg (שגם סקרנו בעבר), הבנויות על קונספט דומה ומצליחות לשפר את הביצועים של DETR.

הסבר על הרעיון העיקרי של המאמר:

בכל הגישות שתוארו עד כה הפלט היה מיוצג באמצעות ה-tuple שתיארנו קודם, המחזיק ערכים מספריים המייצגים את מיקום ה-BB-ים ואיבר נוסף המייצג את המחלקה של האובייקט. במאמר הנסקר המחברים יישמו פרדיגמה שונה לחלוטין לייצוג של האובייקטים וגם כן למשימת הזיהוי של מיקומיהם וסיווגם. הרעיון העיקרי של המאמר הוא ייצוג של מיקום האובייקט ומחלקתו באמצעות סדרה (רצף) של מספרים. כל אובייקט בתמונה יתואר באמצעות 5 טוקנים (המיוצגים באמצעות ערכים מספריים) ובסך הכל כל האובייקטים בתמונה יוצגו באמצעות סדרת טוקנים באורך של 5 x מספר האובייקטים + טוקן נוסף לסימון של סיום הסדרה, כלומר:

(a, a, a, a, a, b, b, b, b, b, …, EOS)

(a, a, a, a, a, b, b, b, b, b, …, EOS)

הטוקנים המייצגים את ערכי הקודקודים הם למעשה הערכים המספריים של הקואורדינטה אותה הם מייצגים (כלומר - x_{min}, y_{min}, x_{max}, y_{max}) והקטגוריה מיוצגת באמצעות טוקן נוסף **שגם הוא מספר שלם**. כדי לייצג את סוף הסדרה/הרצף מוסיפים טוקן של End Of Sentence (EOS) שערכו המספרי הוא פשוט 0.

נציין שגודל המילון, המכיל את כל הטוקנים האפשריים, של משימה זו הוא קטן משמעותית מזה של מודלי שפה גדולים. למשל לתמונה ברזולוציה 1024x1024 ו- 100 קטגוריות אנו צריכים בסך הכל 1024+100 = 1124 טוקנים.

Source

כיוון שהייצוג של האובייקטים הוא רצף ולא כפי שהיה נהוג עד כה, גם החיזוי של המודל נעשה באופן שונה (ואולי זו בעצם המוטיבציה הכי משמעותית לניסוח זה של בעיית זיהוי האובייקטים). כל המודלים המוכרים מקבלים כקלט תמונה והפלט הוא רשימה של האובייקטים והסיווג שלהם, המתקבל **כמקשה אחת ובבת אחת**. אמנם קיימים מודלים המבצעים זיהוי אובייקטים באמצעות שני ראשים: ראש הרגרסיה (למיקומים של אובייקטים) וראש סיווג (לזיהוי קטגוריה) ויש מודלים שמבצעים את הרגרסיה והסיווג יחד, אך המשותף להם הוא שהפלט מתקבל בו זמנית וכמקשה אחת עבור כל האובייקטים. בעבודה הנסקרת לעומת זאת החיזוי נעשה באופן אוטורגרסיבי, בדומה לאופן בו נעשה במשימות של גנרוט סדרות כמו תרגום או יצירת טקסט. המשמעות היא שהפלט לא נבנה בבת אחת אלא כל פעם המודל מייצר איבר פלט נוסף המבוסס גם על הקלט וגם על איברי הפלט שנוצרו לפניו.

המודל המוצע בכל פעם מספק טוקן אחד בלבד המסמן קואורדינטה של אובייקט או קטגוריה. כל איבר פלט מתבסס גם על הקלט אך גם על הפלטים שנחזו קודם לכן. לכן כל טוקן שנוצר מתבסס גם על המידע החדש (קרי הטוקנים שנוצרו לפניו). באופן הזה החיזוי של הקואורדינטות נעשה יותר מדויק כיוון שהוא משתמש במידע עדכני כל הזמן. גם חיזוי של קטגוריה מתבסס לא רק על הקלט אלא גם על הקואורדינטות הקודמות שהתקבלו, מה שיכול לתרום בדיוק שלו. **חשוב לשים לב שבמשך האימון טוקן בעל ההסתברות הגבוהה ביותר (מותנית בהינתן הטוקנים הקודמים) נבחר בתור טוקן הבא בסדרה.**

אתגרים המוזכרים במאמר:

הגישה המוצעת להתמודדות עם משימת זיהוי אובייקטים העלתה כמובן כמה אתגרים, כאשר לחלקם המחברים התייחסו באופן ישיר ואף ביצעו ablation study מקיף.

האתגר הראשון קשור לטוקן end-of-sentence (EOS) המסמן את סוף הפלט (כאשר המודל תיאור את כל האובייקטים בתמונה). בסט האימון לכל תמונה יש סדרה המתארת את האובייקטים הנמצאים בה, ובסוף הסדרה

יש את הטוקן EOS (שערכו המספרי 0) שמסמן שכאן נגמר הפלט. ב-inference המודל בעצמו נדרש להוציא סדרה של מספרים ובנוסף להוציא את הטוקן EOS בסוף הסדרה אחרי שהוא סיפ תיאור של כל האובייקטים שהוא חושב שיש בתמונה. המחברים שמו לב כי לעיתים קרובות המודל מוציא את הטוקן הזה מוקדם מדי, ובכך מפספס אובייקטים בתמונה, והם ניסו כמה גישות בכדי להתמודד עם סוגיה זו.

לפני שנראה את דרכי ההתמודדות ננסח בצורה מעט שונה את הבעיה, בטרמינולוגיה דומה לאיך שהצגנו את אופי הפעולה של המודל האוטורגרסיבי. מודל כזה בכל פעם מוציא טוקן שיש לו את ההסתברות המותנית הגבוהה ביותר, **ומה שקרה המודל אמנם זיהה את כל האובייקטים, אך לטוקן של EOS היתה הסתברות גבוהה יותר מאשר לשאר האובייקטים**, ולכן בסדרת הפלט הוא נבחר לפניהם. במקרה זה הסדרה למעשה נעצרת, וממילא אנו לא רואים את החיזוי עבור יתר האובייקטים בתמונה. תופעה זו גרמה לירידה ב-recall עקב הופעתם של הרבה False negatives (אובייקטים שקיימים בתמונה אך המודל לא זיהה אותם). אחת הדרכים להתמודד עם סוגיה זו היא "להקטין באופן ידני" את ה-likelihood של הטוקן EOS, אך זה יגרום לעלייה ב-False positive (אובייקטים שלא קיימים אך המודל חזה אותם), מה שכמובן מקטין את ה-precision. אמנם הטרייד-אוף של precision-recall אינו ייחודי רק לעבודה הזו, אך פרדיגמה שונה שהמחברים אימצו לפתרון משימה זו, אפשרה להם להגיע לביצועים משופרים מבחינת היחס של precision-recall באמצעות 2 טריקים נחמדים של תהליך האימון:

1. **הוספה של BB רנדומליים**
המחברים הוסיפו של BB רנדומליים כאשר כל BB תויג עם קטגוריה לא קיימת ("n/a"). המודל מאומן לזהות BB כאלו עם הקטגוריה הלא קיימת.
2. **עיוות של BB קיימים**
כאן במקום ליצור BB באופן רנדומלי, המחברים הציעו "לעוות" (להזיז או להקטין/להגדיל) BB קיימים ולתייג אותם עם אותה קטגוריה לא קיימת.

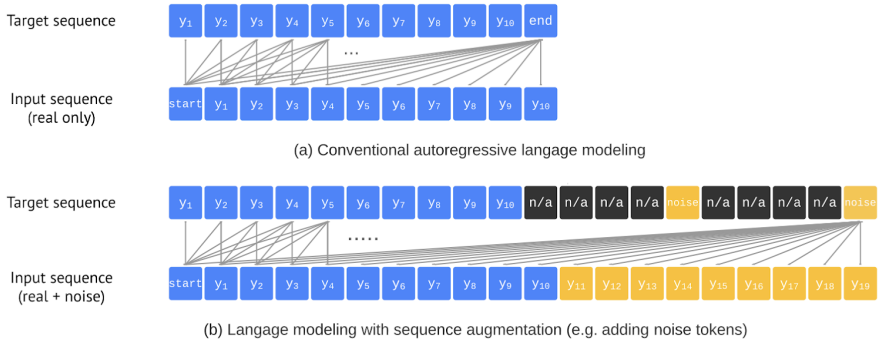

Figure 5: Illustration of language modeling with / without sequence augmentation. With sequence augmentation, input tokens are constructed to include both real objects (blue) and synthetic noise objects (orange). For the noise objects, the model is trained to identify them as the "noise" class, and we set the loss weight of "n/a" tokens (corresponding to coordinates of noise objects) to zero since we do not want the model to mimic them.

נציין כי במהלך האימון הלוס על BB-ים לא אמיתיים מתווסף לפונקצית לוס הרגילה כאשר המודל לא נקנס על זיהוי שגוי של הקואורדינטות של BB-ים אלה.

טריקים אלו מאפשרים לאמן את המודל עם מספר טוקנים קבוע וגבוה מספיק בשביל לזהות את כל האובייקטים בתמונה. בזמן ה-inference לכל אובייקט נבחרת מקטגוריה אמיתית בעלת הסתברות הגבוהה ביותר.

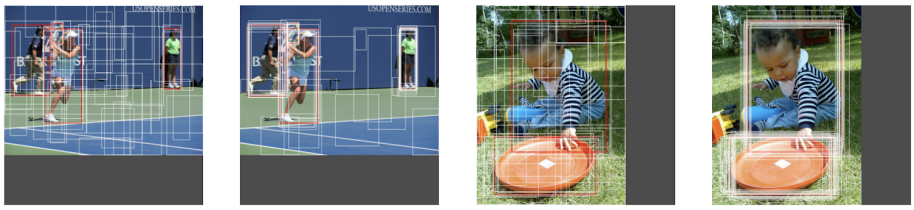

Figure 6: Illustrations of randomly sampled noise objects (in white), vs. ground-truth objects (in red).

בעיה נוספת שעולה מהייצוג של האובייקטים כסדרה היא הסדר בין האובייקטים. מצד האמת, אין משמעות לסדר של האובייקטים ואין אובייקט אחד שנמצא בתמונה בוודאות גדולה יותר מאשר אובייקט אחר. מודל שמוציא זיהוי של כל האובייקטים ביחד באמת אינו מעדיף אחד על אחר. ייצוג של האובייקטים כסדרה לעומת זאת הוא בעייתי, כיוון שהוא אומר שככל שאובייקט נמצא מוקדם יותר בסדרה ככה ה-likelihood שלו גבוה יותר. המשמעות של זה היא שבזמן האימון אנו מכריחים את המודל ללמוד סדר מסוים בין האובייקטים ולספק כפלט אובייקט אחד לפני האחר, למרות שאין באמת יחס סדר כלשהו. כך למשל אם יש בתמונה כוס ופיל ובסדרה המייצגת אותם הכוס מופיעה לפני הפיל, אז אם המודל יחזה את הפיל לפני הכוס - תיווצר חוסר התאמה בין ה-ground-truth לבין ה-detection, וממילא בתהליך האימון המודל "יענש" על חיזוי כזה למרות שהוא נכון לחלוטין.

המחברים בחנו 4 אפשרויות לסידור האובייקטים:

1. סדר רנדומלי
2. מיון לפי גודל האובייקט
3. מיון לפי מרחק אובייקט ממרכז התמונה
4. מיון לפי סדר הקטגוריות

הביצועים הטובים ביותר הושגו כאשר האובייקטים מסודרים בסדר רנדומלי (1). כפי שציינו המחברים, אין הגיון לאלץ את המודל לחזות את האובייקטים דווקא בסדר מסוים, ולכן זה לא מפתיע שאף אחת מאפשרויות המיון לא היתה טובה מהאחרות או מסידור רנדומלי, אם כי עדיין חשוב לציין שגם בסדר הרנדומלי ישנה בעייתיות מובנית.

הישגי המאמר:

המחברים השוו את ביצועי Pix2Seq עם כמה שיטות פופולריות לזיהוי אובייקטים כמו Faster R-CNN+, Faster R-CNN ו- DETR על דאטהסט קלאסי לזיהוי אובייקטים CoCo. השיטה המוצעות הצליחה להשיג ביצועים ברי השוואה לשיטות הנ״ל מבחינת דיוק ממוצע (average precision) עבור כמה ספים (thresholds) שונים. בנוסף המאמר בחן את ביצועי הגישה המוצעת בשתי קונפיגורציות של אימון: אימון מהתחלה ו-pretraining על דאטסט גדול Object365. עם זאת לא ראייתי השוואה עם שיטות יותר עדכניות כמו DETReg.

Method	Backbone	#params	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster R-CNN	R50-FPN	42M	40.2	61.0	43.8	24.2	43.5	52.0
Faster R-CNN+	R50-FPN	42M	42.0	62.1	45.5	26.6	45.4	53.4
DETR	R50	41M	42.0	62.4	44.2	20.5	45.8	61.1
Pix2seq (Ours)	R50	37M	**43.0**	61.0	45.6	25.1	46.9	59.4
Faster R-CNN	R101-FPN	60M	42.0	62.5	45.9	25.2	45.6	54.6
Faster R-CNN+	R101-FPN	60M	44.0	63.9	47.8	27.2	48.1	56.0
DETR	R101	60M	43.5	63.8	46.4	21.9	48.0	61.8
Pix2seq (Ours)	R101	56M	**44.5**	62.8	47.5	26.0	48.2	60.3
Faster R-CNN	R50-DC5	166M	39.0	60.5	42.3	21.4	43.5	52.5
Faster R-CNN+	R50-DC5	166M	41.1	61.4	44.3	22.9	45.9	55.0
DETR	R50-DC5	41M	**43.3**	63.1	45.9	22.5	47.3	61.1
Pix2seq (Ours)	R50-DC5	38M	**43.2**	61.0	46.1	26.6	47.0	58.6
DETR	R101-DC5	60M	**44.9**	64.7	47.7	23.7	49.5	62.3
Pix2seq (Ours)	R101-DC5	57M	**45.0**	63.2	48.6	28.2	48.9	60.4

נ.ב.

המאמר הציע פרדיגמה חדשנית ומעניינת לבניית מודל לזיהוי אובייקטים בתמונה שהפלט שלה מיוצר באופן אוטורגרסיבי. הביצועים של הגישה די דומים לשיטות קלאסיות לזיהוי אובייקטים. אני מניח שבקרוב יצאו מאמרים המשכללים גישה זו ומצליחים להשיג ביצועים טובים יותר משיטות SOTA.

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. מרצה ומנגיש בעברית חומרים בתחום הבינה המלאכותית, ומחבר ספר על למידת מכונה ולמידה עמוקה בעברית.

ברצוננו להודות לעדו בן יאיר על העזרה בהגהת הסקירה.