סקירת המאמר:

DETR: End-to-end Object Detection with Transformers

שיצא לפני כחצי שנה, על ידי Facebook AI

**הוצג בכנס: ECCV 2020**

**תחומי מאמר:**

זיהוי אובייקטים בתמונה באמצעות טרנספורמרים.

יצירת masking עבור תמונה באמצעות טרנספורמרים.

**כלים מתמטיים, מושגים וסימונים:**

שום כלי מלבד טרנספורמר רגיל. אין משוואות במאמר.

**תמצית מאמר:**

ניתן לתקצר את המאמר במשפט אחד: ניתן לקחת תמונה ולהפוך אותה לסדרה (reshape לסדרה במימד אחד), לבצע attention ו-self-attention, ולהבדיל בקלות בין העצמים שבתמונה.

ויותר בהרחבה: זיהוי אובייקטים בתמונה הוא נושא שתמיד מתעסקים בו, ויש כמה אלגוריתמים של SOTA, בעיקר אלו המשתמשים ברשתות Faster R-CNN ו-ResNet). במאמר זה מתמודדים עם האתגר בעזרת טרנספורמרים (שילוב של encoder-decoder), באופן כזה שהופכים את התמונה לסדרה ממימד אחד ואז מפעילים עליו את הטרנספורמר.

**הסבר של רעיונות בסיסיים**:

קצת רקע על attention: ניתן לקחת כל סדרה, ובכל פעם לקחת איבר אחד ממנה ולבצע מכפלה פנימית עם שאר האיברים בסדרה. איברים דומים בסדרה יתנו ערכים גבוהים ואיברים שונים בסדרה יתנו ערכים נמוכים. כשאומרים איברים "דומים" או "שונים" – הכוונה שיש ביניהם קשר מסוים. ב-NLP זה יכול להיות מילים שסביר שיופיעו בסמיכות, ובתמונה זה יכול להיות פיקסלים דומים.

כאשר לוקחים איבר מסוים ומכפילים בכל שאר האיברים (נקרא לכל תוצאה: מקדם a_i) – ניתן לסכם את מכפלת כל המקדמים באיברים המקוריים, וככה לקבל ייצוג חדש לאיבר המקורי, והייצוג הזה קושר אותו לאיברים דומים בסדרה. זה בעצם הרעיון הבסיסי של attention – לקבל סדרה של דאטא, וליצור סדרה חדשה, שכל איבר בסדרה החדשה מייצג איבר בסדרה המקורית, כאשר בייצוג החדש כלול גם מידע על הקשרים שבין האיברים בסדרה המקורית.

(כמובן שבשלב יצירת המקדמים a_i, מעבירים את כל התוצאות ב-softmax על מנת לנרמל את התוצאה).

זהו בעצם self-attention – ביצוע מניפולציה על סדרה מסוימת בכדי לקבל ייצוג מחודש שלה, הכולל מידע על קשרים בין האיברים. לצורך המחשה, אם יש תמונה של פנים ויש בתמונה גם רקע מסוים מסביב לפנים שפחות מעניין, ניתן על ידי ההכפלה הזו לגלות איפה בתמונה נמצאים הפנים ואיפה הרקע, ובכך להתייחס רק לאזורים המעניינים. איך זה קורה? עושים מכפלה פנימית של הפיקסלים עם עצמם, וכך מגלים אזורים הדומים אחד לשני.

אם עושים את אותה פעולה, אך במקום להכפיל את איברי הסדרה בעצמם מביאים query חיצוני, אז מקבלים הפעם ייצוג חדש, הכולל מידע על היחס בין האיברים בסדרה לבין ה-query. אם למשל מכפילים תמונה של פיל באיברים של תמונה מסוימת, אז הפיקסלים המכילים חלקים של פיל יתנו תוצאה גבוהה, ושאר המפכלות ישאפו ל-0. ואם נחזור לדוגמה הקודמת – הפרדת פנים מהרקע – לאחר שביצענו self-attention וגילינו אזורים דומים, נכפיל את איברי התמונה ב-query של פנים, וכך נגלה איפה הפנים בתמונה ונדע שכל השאר הוא רקע.

שילוב של self-attention (הנקרא גם encoder) לבין attention (הנקרא גם decoder), מכונה בשם transformer. הטרנספורמרים הגיעו לעולם לפני כמה שנים, וההצלחה שלהם הייתה מאוד גדולה, והם נמצאים בכל אזור שמכיל רשתות. לכן זה היה טבעי לקחת טרנספורמר ולהפעיל אותו גם על תמונה.

כל מה שהיה צריך זה רק להבין שניתן להפוך תמונה לסדרה, ואז: א. לבצע על הסדרה self-attention בכדי להבדיל בין אזורים שונים בתמונה. ב. לבצע attention עם query חיצוני בכדי לסווג כל אזור בתמונה ל-label מסוים.

ההבדל העיקרי בעיני בין שיטה זו לשיטות קודמות (כמו Resnet), והוא גם היופי במאמר – ניתן בבת אחת לבצע ניתוח של כל התמונה. כמובן שגם כאן יש חלוקה לפיקסלים, אבל ברגע שהופכים את התמונה לדאטא בצורה של מערך, כל הלמידה נעשית על כל הסדרה יחד.

אז לסיכום – מעבירים תמונה ב-CNN רגיל על מנת לחלק את התמונה לאזורים שונים, לאחר מכן מייצרים מהתמונה סדרה, ועליו מפעילים טרנספורמר ומקבלים את הדרוש.

כמובן שנעשתה התאמה של הרשת לזיהוי אובייקטים – ה-loss שנבחר הוא Hungarian loss, וה-metric הינו IoU.

ניתן לראות שעיקר ההסבר הוא על טרנספורמר, ומי שמכיר את זה, המאמר הזה הוא פשוט יישום טרנספורמר על תמונה, ולמעשה כמעט לא התווסף כלום מעבר לטרנספורמר המוכר. החידוש העיקרי הוא בהתאמת תמונה לרשת של טרנספורמר, וכאשר עניין זה מובן, אין צורך להוסיף עוד הסברים והתוצאות הן המשך ישיר של החיבור בין טרנספורמר לבין תמונה שעברה reshape.

**תוצאות המאמר:**

השיטה נוסתה על המאגר COCO וניתן לראות תוצאות טובות מאוד, הן בזיהוי אובייקטים והן יצירת masking לתמונה שלמה. עבודות המשך שנעשו כבר הראו תוצאות על זיהוי אובייקטים בווידאו, וניתן לראות איך האובייקטים מזוהים בצורה מדויקת באפס זמן.

המחברים מציינים את העובדה שביחס לשיטות קודמות, הרשת של DETR מאוד פשוטה למימוש, ובעלת הרבה פחות פרמטרים. בנוסף, כל הפרמטרים של ה-decoder נלמדים גם הם, כלומר, ה-querys מאותחלים בערכים אקראיים ומהר מאוד נלמדים. זאת אומרת שאין צורך לומר בהתחלה מהו פיל, אלא עם הזמן הרשת לומדת מהו הייצוג של פיל, ואיפה הוא בתמונה. זה בעיני מרשים מאוד.

עובדה מעניינת נוספת: המחברים מראים הצלחה בזיהוי אובייקטים גם עבור אובייקטים שלא הופיעו הרבה ב-training set. כלומר, הרשת לומדת במהירות ייצוג של אובייקט, ומסוגלת לזהות גם אוגמנטציות שלו בקלות ובהצלחה רבה.

אחד הדברים היפים זה הוויזואליזציה שעשו למאמר. בסרטון הבא ניתן לראות בהמחשה יפה מאוד כיצד בפועל פועל הטרנספורמר על תמונה – הן בזיהוי האובייקטים והן בסיווג שלהם:

https://www.youtube.com/watch?v=utxbUlo9CyY&ab_channel=NicolasCarion

**הישגי מאמר:** המאמר מציג תוצאות של SOTA על COCO במשימת זיהוי אובייקטים וביצירת masking, כאשר הקוד פשוט יותר מהשיטות הקודמות, עם הרבה פחות פרמטרים. בנוסף, ניתן בקלות להוסיף לייבלים ולהכליל למשימות אחרות.

**לינק למאמר:** https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123460205.pdf

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

**נ.ב.** מאמר מאוד מעניין שהיופי בו הוא הפשטות והקלות של המימוש והדרך להשיג את התוצאות. במאמר יש בסך הכל 3 שורות של משוואות, וגם הן זה בסה"כ הloss וה-metric, שהותאמו למשימה הספציפית של זיהוי אובייקטים בעזרת טרנספורמר. ניתן להניח שיהיו עבודות המשך שייקחו את הרעיון וישפרו אותו, אז שווה לעקוב.

#deepnightlearners