סקירת המאמר:

VirTex: Learning Visual Representations from Textual Annotations

שיצא לפני כמה חודשים. נכתב על ידי Justin Johnson ו-Karan Desai.

תחומי מאמר:

Pre-training model, transfer learning, Image Captioning

כלים מתמטיים, מושגים וסימונים:

כלים סטנדרטיים של רשתות עמוקות – רשתות קונבולוציה, טרנספורמרים.

בהירות כתיבה: מובן מאוד.

רמת היכרות עם כלים מתמטיים וטכניקות של ML/DL הנדרשים להבנת מאמר: בינוני

תמצית מאמר:

(כיוון שהמאמר מתעסק בעיבוד תמונה, אתמקד בדומיין הזה).

עבור משימות של עיבוד תמונה, כמו למשל: סיווג, זיהוי אובייקטים, תיאור מילולי של התמונה ועוד, כמעט תמיד משתמשים במודל שאומן כבר על דאטא-סט מסוים, ועל גביו מוסיפים שכבות עבור משימה ספציפית. התהליך הזה נקרא transfer-learning – לקחת מודל שאומן עבור משימה מסוימת ולהשתמש בו עבור משימה אחרת, תוך כדי התאמת הרשת למשימה החדשה. המודל שאומן כבר נקרא pretrained model. הדבר הנפוץ הוא לקחת מודל שאומן על ImageNet, כיוון שזה דאטא-סט שמכיל הרבה תמונות באופן יחסי. לאחרונה התפרסמו עבודות שבנו unsupervised pretraining models בעזרת דאטא-סטים לא מתויגים גדולים בהרבה מ-ImageNet. המודל המאומן, שבדרך כלל מורכב מרשתות קונבולוציה, נקרא backbone.

במאמר הנוכחי המחברים מציעים דרך אחרת ליצירת backbone, על ידי אימון רשת על תמונות שה-labels שלהן הוא תיאור מילולי (Image Captioning). לדוג' – יש תמונה של חתול כתום-לבן ליד עוגה, אז ה-label הוא:

"An orange and white cat near a plate and a white cake."

המחברים מצליחים להראות שעל ידי אימון ה-backbone על תמונות שמתויגות בעזרת תיאור מילולי, ניתן לבצע בהצלחה transfer למשימות אחרות של עיבוד תמונה, כאשר האימון של ה-backbone דורש הרבה פחות תמונות בסט האימון (פי עשרה פחות). בנוסף, המחברים מראים שה-backbone שלהם מתאים גם למשימות supervised וגם למשימות unsupervised.

backbones המבוססים על ImageNet למעשה נבנים על בסיס משימת סיווג – לוקחים דאטא-סט גדול של תמונות מתויגות ומאמנים את הרשת. במאמר הנוכחי משימת האימון ליצירת ה-backbone מבוססת על משימת Visual Representation – נתינת תיאור מילולי לתמונה, וזה לטענת המחברים מקור ההצלחה של השיטה. היכולת ליצור backbone טוב שישמש בסיס גם למשימות אחרות, נובעת לטענתם מהעובדה שאיכות הלמידה היא גבוהה, כיוון שכל label מכיל יותר מידע מאשר label בודד כמו ב-ImageNet. למעשה יש פה טרייד-אוף של כמות מול איכות – ב-ImageNet יש המון תמונות בעלות label בודד (וב-unsupervised pretrained models בכלל אין labels), ואילו במאמר יש מאגר עם הרבה פחות של תמונות אך עם label בעל יותר מידע.

הסבר של רעיונות בסיסיים:

הרעיון הבסיסי של המאמר הוא יחסית פשוט – לייצר מודל מאומן שיאפשר transfer learning למשימות אחרות, כאשר האימון של המודל הוא על משימת Visual Representation. משימת האימון הזו מבוססת על שני חלקים: רשת קונבולוציה סטנדרטית (ResNet50) ולאחריה שני טרנספורמרים מקבילים היוצרים textual head. שני הטרנספורמרים מייצרים יחד את אותו מודל שפה, אך פועלים בכיוונים הפוכים – טרנספורמר אחד מנסה לנחש בכל פעם

את המילה הבאה המשפט, ואילו הטרנספורמר השני פועל בכיוון ההפוך – הוא מקבל חלק מסוף המשפט ומנסה לנחש את המילה שלפני חלק זה.

חיבור שני החלקים יוצר רשת המסוגלת לתת ייצוג מילולי לתמונה – רשת הקונבולוציה מנסה ללמוד את הפיצ'רים הנמצאים במרחב התמונה, ואילו החלק של הטרנספורמרים מנסה לתרגם את הפיצ'רים הויזואלים למשפט המתאר בצורה הטובה ביותר את התמונה. כיוון שבסופו של דבר יש פה למידה של תמונה על ידי התרגום שלה לשפה אנושית, יש אפשרות ליצור מודל שיוכל להיות backbone גם עבור משימות אחרות בעיבוד תמונה.

כיוון שהטרנספורמרים נועדו בשביל לתרגם את התמונה לתיאור מילולי, עבור משימות אחרות יותר פשוטות לא צריך אותן. לכן למעשה ה-backbone הוא רק רשת הקונבולוציה המאומנת – אותה ניתן לקחת גם עבור משימות אחרות.

תוצאות המאמר:

בעזרת הארכיטקטורה של ה-backbone ביצעו transfer learning למגוון משימות – סיווג וזיהוי אובייקטים וכמובן גם בחנו את הרשת עבור משימה של נתינת תיאור מילולי לתמונה. עבור המשימה האחרונה התוצאות היו טובות, אבל זה לא מפתיע, כיוון שזו המשימה עליה אומן המודל. עבור משימות אחרות ה-transfer learning מציג תוצאות יפות מאוד, כאשר הדגש הוא מספר התמונות הקטן יחסית שנדרש בשביל לייצר backbone איכותי.

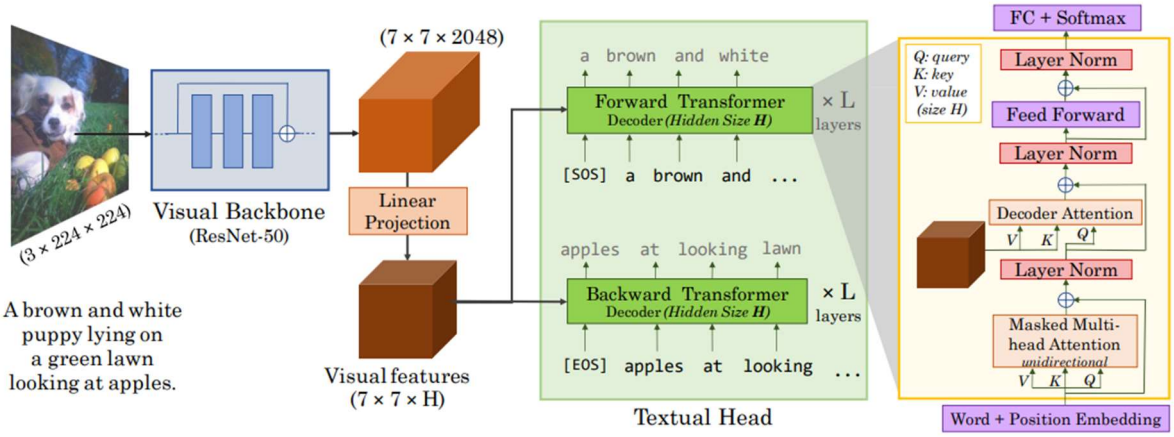
הישגי מאמר: המחברים אומרים בפירוש שהשאיפה של VirTex היא לא בהכרח להגיע לתוצאות SOTA במגוון משימות (אם כי התוצאות שלהם טובות למדי). החידוש שלהם הוא בחשיבה מקורית מנוגדת לטרנד – במקום לשפר את ה-backbone בעזרת אימון על דאטא-סט לא מתויג יותר גדול, הם דווקא לקחו דאטא-סט מתויג, אך בחרו משימת אימון של Visual Representation. בזכות "איכות" התוויות של הדאטא-סט – תהליך האימון יוצר מודל איכותי, ממנו ניתן לבצע בהצלחה מגוון משימות אחרות בעזרת transfer learninig. היופי במבנה שהם יצרו נובע מכך שהוסיפו את הטרנספורמרים לצורך אימון המודל, אך לאחר מכן במשימת אחרות ניתן לקחת רק את רשת הקונבולוציה בתור pre-trained model בלי להזדקק יותר לטרנספורמים.

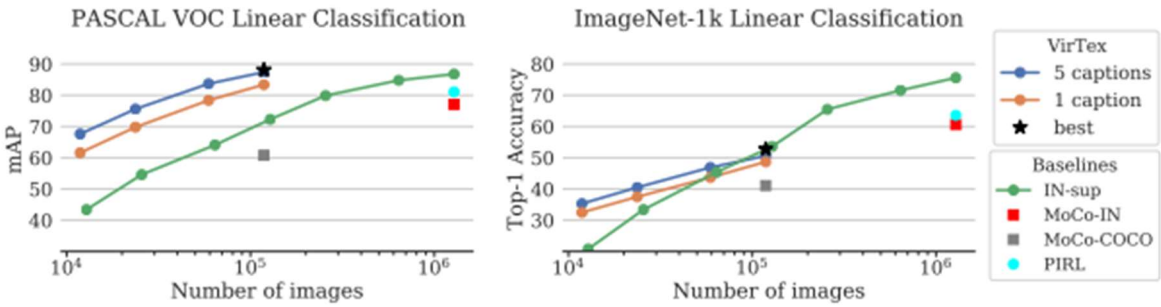
לינק למאמר: https://arxiv.org/abs/2006.06666

לינק לקוד: https://github.com/kdexd/virtex

יש ביוטיוב סרטון הסבר של המחבר (ג'סטין ג'ונסון) וסרטון הסבר של יאניק.

נ.ב. מאמר מקורי ומעניין, המסתיים בקריאה: "Finally, the usage of captions opens a clear pathway to scaling our approach to web-scale image-text pairs, which are orders of magnitude larger, albeit more noisy than COCO.". נמתין ונראה אם אכן יהיו עבודות המשך.





#deepnightlearners