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:

PreTrained Image Processing Transformer

פינת הסוקר:

המלצת קריאה ממייק: רק עם קשה לכם להירדם בלילה (שווה לאלו שמתעסקים במשימות low-level בתחום עיבוד תמונה).

בהירות כתיבה: בינוני מינוס.

רמת היכרות עם כלים מתמטיים וטכניקות של ML/DL הנדרשים להבנת מאמר: היכרות עם מושגי יסוד של DL.

יישומים פרקטיים אפשריים: הגישה המוצעת במאמר יכולה לשמש כשיטת אימון למשימות כמו סופר-רזולוציה, ניקוי רעש רגיל או הסרת רעש גשם (deraining) עבור דאטהסטים קטנים.

פרטי מאמר:

לינק למאמר: זמין להורדה.

לינק לקוד: לא הצלחתי לאתר.

פורסם בתאריך: 03.12.20, בארקיב.

הוצג בכנס: לא מצאתי מידע על כך.

תחומי מאמר:

- למידה עם משימות מרובות (multi-task learning – MLT).

- למידה מנוגדת (contrastive learning – CL).

כלים מתמטיים, טכניקות, מושגים וסימונים:

- טרנספורמר ויזואלי (הפועל על פאצ'ים של תמונות).
- לוס מנוגד (contrastive loss).
- משימות low-level של הראייה הממוחשבת כמו סופר-רזולוציה ,ניקוי רעשים וכדומה.

לינקים להסברים טובים על מושגי יסוד במאמר:

- אימון מקדים של רשתות.
- טרנספורמר1 ,טרנספורמר2.
- למידה מנוגדת.(contrastive learning)

מבוא והסבר כללי על תחום המאמר:

רשתות נוירונים הפכו לכלי הנפוץ ביותר עבור מגוון רחב של משימות בראייה הממוחשבת החל ממשימות high-level כמו סיווג, סגמנטציה, זיהוי אובייקטים וכדומה וכלה במשימות low-level כמו ניקוי רעש, סופר-רזולוציה, שחזור חלקים פגומים של תמונה (inpainting) ועוד. עקב דמיון בין משימות low-level רבות ניתן לצפות שמודל (או הייצוג שנבנה באמצעותו) שאומן על דאטהסט מסוים יהיה שימושי גם עבור דאטהסטים אחרים. אז איך ניתן לנצל את הדמיון הזה? מאמנים מודל אחד על דאטהסט גדול (pretraining) ובדרך כלל (אך לא תמיד) על אותה משימה (!!) ולאחר מכן מכיילים את המודל המאומן (fine-tuning) על דאטהסט אחר (נקרא לו דאטהסט מטרה) שיכול להיות קטן בהרבה. די ברור שככל הדומיינים של הדאטהסטים דומים יותר, היעילות של האימון המקדים עולה.

גישה זו טומנת בעצמה שתי אתגרים עיקריים:

- לא תמיד יש דאטהסטים זמינים לאימון מקדים (למשל בדומיין הרפואי או בדומיין של תמונות לווין) למשימה נתונה.
- לא תמיד ניתן לדעת לאיזו משימה יאומן מודל בתהליך הכיול שמקשה על בחירה של דאטהסט לאימון מקדים (נקרא לדאטהסט זה דאטהסט מקור).

תמצית מאמר:

במטרה להתמודד עם אתגרים אלו, המאמר מציע שיטת אימון מקדים למשימות מרובות בדומיין הויזואלי. הם קראו לגישה שלהם IPT- Image Processing Transformer (כמו שאתם יכולים לנחש הארכיטקטורה שלהם מבוסת על הטרנספורמרים). IPT מורכב מארבעה מרכיבים (רשתות) עיקריים שמאומנים לכמה משימות במקביל (!!):

- רשתות "ראשים" (heads): מספר ראשים שווה למספר משימות שעליהן IPT מאומן ,(ראש פר משימה). כל ראש הוא למעשה רשת קונוולוציונית שיעדה להפיק מהקלט פיצ'רים רלוונטים למשימה שעליה אחראי הראש הזה.
- מקודד (encoder): הפלט של כל ראש מוזן למקודד הסטנדרטי של הטרנספורמר.
- מפענח (decoder): הפלט של המקודד עובר למפענח די סטנדרטי של הטרנספורמר עם שינוי קטן (יפורט בפרק הבא).
- רשתות זנבות (tails): מספר "זנבות" שווה למספר משימות (כמו ראשים) והם מיועדים בשביל ליצור קלט עבור כל משימה (שעשוי להיות במימד שונה לכל משימה).

תוספת של האימון המנוגד (contrastive training) ל- IPT: נציין שקיים מגוון רחב של משימות בעלות אופייניים שונים ובדומיינים שונים, שלא ניתן לאמן את כולם במהלך אימון מקדים. אז בשביל לשפר את עוצמת הייצוג של תמונה, המופק עי" IPT, המאמר מציע לאמן אותו בשיטת הלמידה המנוגדת (עם הלוס המנוגד הקלאסי) בנוסף לאימון על מספר משימות low-level של עיבוד תמונה.

הסבר של רעיונות בסיסיים:

תחילה בואו נבין איך בונים קלט למקודד. נתחיל מזה שהארכיטקטורה שלו זהה לזו של הטרנספורמר. הקלט למקודד של הטרנספורמר הסטנדרטי (למשימות NLP) הוא האמבדינגס (embeddings) של טוקנים במשפט שמתווסף אליהם קידוד מיקומי (positional encoding) שמטרתו "להעביר" למקודד את המיקום של המילה במשפט. ב- IPT עושים משהו דומה רק שבמקום טוקנים יש לנו פאטצ'ים של תמונה (בגודל של 48×48). נציין שלהבדיל מהטרנספורמר הקלאסי, הקידודים המיקומיים כאן הינם נלמדים (זה מה שעשו במאמר המפורסם An Image is Worth 16×16 Words). נציין כי להבדיל מהטרנספורמר המקורי יש למקודד שכבת self-attention אחת ולא 2.

הפלט של מקודד נכנס למפענח שהוא מאוד דומה לזה של הטרנספורמר המקורי עם שני הבדלים. הבדל הראשון הוא בנוסף לפלט של המקודד גם קידודים מיקומיים נלמדים פר משימה (!!) מוזנים למפענח (הסכום שלהם). המחברים טוענים כי תוספת זו תרמה רבות לביצועי המודל. ההבדל השני הוא העדר שכבה attention מקודד-מפענח (encoder-decoder) והוחלפה בשכבת self-attention.

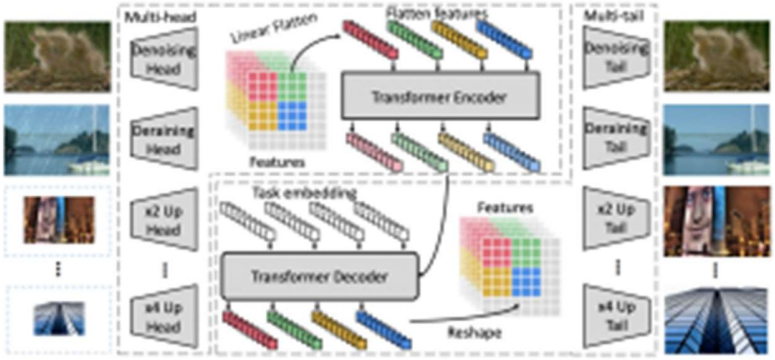

Figure 2. The diagram of the proposed image processing transformer (IPT). The IPT model consists of multi-head and multi-tail for different tasks and a shared transformer body including encoder and decoder. The input images are first converted to visual features and then divided into patches as visual words for subsequent processing. The resulting images with high visual quality are reconstructed by ensembling output patches.

<u>הערה על ארכיטקטורת המפענח</u>: העדר שכבה attention מקודד-מפענח במפענח של IPT נראה לי קצת לא הגיוני. מבנה המשימות של עיבוד תמונה low-level (כמו ניקוי רעש או סופר-רזולוציה) דומה לזה של התרגום האוטומטי. כידוע הטרנספורמר הציג ביצועים טובים מאוד במשימות תרגום כאשר שכבת attention מקודד-מפענח משחקת תפקיד מאוד חשוב בהבנת/כימות קשרים בין המשפט המקורי לתרגום שלו.

כמו שאמרנו בנוסף לאימון של IPT על מספר משימות low-level בו-זמנית, המאמר מציע לבצע אימון מנוגד במטרה לשפר את הייצוג התמונה. אז בואו קודם כל נרענן מה זה שיטת אימון (למידה מנוגדת).

עקרונות הלמידה המנוגדת: העקרון החשוב של גישה זו מניח שייצוגים של דוגמאות דומות צריכים להיות קרובים, כאשר ייצוגים של דוגמאות לא דומות צריכים להיות רחוקים. פונקצית המטרה בלימדה המנוגדת מנסה למקסם את היחס בין אקספוננט של דמיון של זוג דוגמאות קרובות לסכום הדמיונות בינו לבין דוגמאות רנדומליות (זוגות שליליים)

המאמר משתמש בגרסה הסטנדרטית של הלוס המנוגד, כאשר הדמיון בין ייצוגים מוגדר כדמיון קוסינוס (cosine similarity). דוגמאות קרובות כאן זה למעשה פאטצ'ים של אותה תמונה, כאשר כל זוג של פאטצ'ים מתמונות שונות מוגדר כשלילי.

<u>איך מאמנים **IPT**:</u>

- <u>אימון מקדים:</u> לוקחים את הדאטהסט של ImageNet ויוצרים ממנו דוגמאות למשימות עליהן מאומן IPT. למשל למשימה של ניקוי רעש רגיל הם יוצרים תמונות מורעשות עי" הוספה של רעש לבן לתמונה כאשר המשימה היא לשחזר את התמונה המקורית.
- מאמנים את IPT למשימות low-level שונות כאשר כל באטץ' מכיל דוגמאות למשימה אחת בלבד (בשביל לחסוך זמן חישוב עקב שימוש בשכבת ראש ובשכבת זנב אחת בלבד).
- מבצעים למידה מנוגדת על Imagenet. המאמר לא מציין מה סדר בין אימון למשימות low-level ולבין האימון המנוגד. אני משער כי הם מתבצעים ביחד (נגיד באטץ' אחד עבור משימת low-level ובאטץ' אחד עבור הלמידה המנוגדת).

- כיול של IPT מתבצע על דאטהסט מטרה.

לבסוף נציין כי פונקצית לוס למשימות low-level הינה 1L.

הישגי מאמר:

המאמר מצליח להראות שיפור בביצועים עבור מספר משימות עיבוד תמונה low-level כמו סופר-רזולוציה, ניקוי רעש לבן והסרת רעש גשם (deraining) עבור כמה דאטהסטים. עבור כל משימה לוקחים IPT מאומן ומכיילים אותו על דאטהסט נתון.

gure 1. Comparison on the performance of the proposed IP
e state-of-the-art image processing models on different tas

דאטהסטים:DIV2K, Urban100, B100, Set14 , Set5

נ.ב. מאמר מציע שיטה לאימון מקדים של רשת נוירונים עבור משימות low-level מרובות. הארכיטקטורה שלהם כוללת מקודד ומפנעח של הטרנספורמר הסטנדרטי ורשתות המפיקות פיצ'רים ייעודיים לכל משימה. המאמר מראה שיפור על מספר רב של שיטות SOTA והגישה נראית די מבטיחה. הייתי רוצה לראות הוכחת עליוניות טיפה יותר מבוססת על דאטהסטים מדומיינים מגוונים יותר. גם הקוד לא שותף שזה תמיד מאכזב. בקיצור מאמר מעניין אך נראה קצת לא מבושל למרות שהרעיון שהוא מציע נראה די חדשני ומעניין.

#deepnightlearners