המאמר היומי של מייק - 12.03.25  
Transformers are Universal In-context Learner

היום נסקור קצרות מאמר תיאורטי כבד החוקר את יכולת האקספרסיביות של טרנספורמרים עמוקים. טרנספורמרים הם ארכיטקטורות עמוקות המגדירות "מיפויים הקשריים" (in-context mappings), אשר מאפשרים חיזוי של טוקנים חדשים בהתבסס על קבוצת טוקנים נתונה. שימו לב של-in-context כאן יש משמעות קצת שונה מאשר בלמידת in-context הקשור ליכולת של טרנספורמרים ללמוד משימות שלא אומן להם בהתבסס על כמה דוגמאות בפרומפט (לפחות למיטב הבנתי).

המחברים מוכיחים כי טרנספורמרים עמוקים (בעלי מספר רב של בלוקי הטרנספורמרים) הם מקרבים(approximators) אוניברסליים, כלומר, הם יכולים לקרב כל מיפוי הקשרי רציף מהתפלגויות טוקנים בכל דיוק. יתרה מכך, התוצאות תקפות הן עבור מנגנוני attention דו-כיווניים (כמו באנקודר) והן עבור מנגנוני attention סיבתיים (כמו בדקודרים), תוך שמירה על ממד אמבדינג קבוע שאינו תלוי במספר הטוקנים.

הגישה המוצע מבוססת על תורת המידה(סוף סוף מצאתי לה שימוש במאמרי DL), שבה רצפי סדרות מיוצגים כהתפלגויות הסתברותיות במרחב האמבדינגס. זה מאפשר שימוש בכלים מאנליזה פונקציונלית(פלאשבקים מלפני כמעט 30 שנה בתואר הראשון) ובתורת הטרנספורט האופטימלי (כתבתי על זה לא מעט בזמו בהקשר של Wasserstein GAN) על מנת להוכיח את יכולת הקירוב האוניברסלית של טרנספורמרים. תרומה טכנית מרכזית היא הגדרה מחדש של מנגנון ה-attention בטרנספורמרים כאופרטור על התפלגויות. זה מאפשר שימוש במשפט סטון-ויירשטראס(Stone–Weierstrass המהווה הכללה קשוחה של משפט Weierstrass הנלמד באינפי2 לדעתי) - תוצאה יסודית בתורת הקירוב על כך שניתן לקרב כל פונקציה ״נוחה״ על ידי משפחת פונקציות צפופות יחסית (המשפט באמת קשוח המגדיר פונקציות במרחבי האוסדורף וכאלו).

ייצוג מבוסס-מידה של למידה בהקשר

חידוש מרכזי במאמר הוא ייצוג של מנגנון ה-attention כאופרטור על התפלגויות במקום על סדרות טוקנים סופיות. דבר זה מאפשר ניתוח אחיד של למידת ההקשר (in-context learning), ללא תלות במספר הטוקנים בסדרה. במקום לעבוד עם קבוצות סופיות של האמבדינגס של הטוקנים, המחברים מגדירים מרחב של התפלגויות הסתברותיות על תת-קבוצה קומפקטית של מרחב אוקלידי (של האמבדינגס). התפלגות משייכת משקלים לאמבדינגס שונים של טוקנים, ובכך מייצגת את המעבר מלמידה על מספר טוקנים סופי לייצוג רציף ואינסופי.

באופן פורמלי, רצף של טוקנים ניתן לייצוג כהתפלגות הסתברות בדידה, המורכבת מסכום משוקלל של פונקציות דלתא דיראק, שכל אחת מהן ממוקמת על הטמעה של טוקן בודד. כאשר מספר הטוקנים גדל, התפלגויות אלה מתכנסות להתפלגויות רציפות. ניסוח זה מאפשר להוכיח תוצאות החלות על כל מספר אפשרי של טוקנים, כולל אינסוף.

הגדרת attention כאופרטור על מרחב מידות

שכבת טרנספורמר טיפוסית מורכבת משני רכיבים:

1. מנגנון attention רב-ראשי, האחראי על עדכון הייצוגים של הטוקנים על ידי חישוב יחסי הגומלין ביניהם.

2. שכבות FFN, המעבדות כל טוקן באופן עצמאי לאחר שלב ה-attention.

המחברים מנסחים מחדש את מנגנון ה-attention כמיפוי הפועל על התפלגויות של טוקנים. במקום לחבר סכום על קבוצת טוקנים סופית, ה-attention מוגדרת כאופרטור אינטגרלי על מרחב ההתפלגויות, מה שהופך את הטוקנים למבנה רציף. ניסוח זה חשוב במיוחד, מכיוון שהוא מאפשר להגדיר רציפות וחלקות של מיפויים בהקשר באמצעות מרחק וסרשטיין (מקרה פרטי שלו הוא earth mover distance), המודד את המרחק בין התפלגויות הסתברותיות. פונקציה היא רציפה במובן וסרשטיין אם שינויים קטנים בהתפלגות הקלט מובילים לשינויים קטנים בהתפלגות הפלט. תכונה זו מבטיחה שהמיפויים שיוצרים טרנספורמרים יציבים לשינויים בהקשר הלימודי.

הוכחת אוניברסליות: קירוב מיפויים הקשריים

התוצאות המרכזיות של המאמר מוכיחות כי טרנספורמרים הם מקרבים אוניברסליים למיפויים הקשריים. המחברים מראים כי עבור כל פונקציה רציפה הממפה התפלגויות טוקנים לפלטים, קיים טרנספורמר עמוק שיכול לקרב אותה בכל דיוק. חלק מרכזי בהוכחה הוא בנייה של פונקציות יסודיות בהקשר, המשמשות כיחידות הבסיס לקירוב כל פונקציה כללית במרחבים שהגדרנו קודם. פונקציות אלו הן גרסאות פשוטות יותר של שכבות טרנספורמר, אשר לוכדות את העקרונות המרכזיים של מנגנוני ה-attention.

פונקציה יסודית כזו מורכבת משלושה מרכיבים:

1. טרנספורמציה לינארית על הטמעת הטוקן (מיפוי אפיני).

2. אינטראקציה לא-ליניארית המתחשבת בהתפלגות של כלל הטוקנים.

3. התאמה תלוית-הקשר, המאפשרת למודל "ללמוד בהקשר".

פונקציות אלו פועלות באופן דומה למנגנון ה-attention בעל ראש בודד, אך הן קלות יותר לניתוח מתמטי. המחברים מוכיחים כי על ידי הרכבת מספר שכבות של פונקציות אלו, ניתן ליצור טרנספורמרים עמוקים המסוגלים לקרב כל פונקציה בהקשר.

שימוש במשפט סטון-ויירשטראס

כדי להוכיח אוניברסליות, המחברים מראים כי קבוצת הפונקציות היסודיות שהם הגדירו מקיימת את תנאי משפט סטון-ויירשטראס, שכאמור הוא משפט מרכזי באנליזה פונקציונלית. המחברים מוכיחים כי הפונקציות היסודיות שלהם מקיימות תנאים אלו, מה שמבטיח כי טרנספורמרים עמוקים יכולים לקרב כל מיפוי הקשרי.

סיכום:

המאמר מספק מסגרת מתמטית להוכחת האקספרסיביות של טרנספורמרים בלמידת  מיפוים הקשריים, תוך שימוש באנליזה פונקציונלית, תורת המידה ותורת הטרנספורט האופטימלי. התוצאות מראות כי טרנספורמרים עמוקים יכולים לקרב כל פונקציה תלויה-הקשר, ללא תלות במספר הטוקנים בחלון ההקשר.

https://arxiv.org/abs/2408.01367