המאמר היומי של מייק - 21.01.25  
Time-MoE: Billion-Scale Time Series Foundation Models with Mixture of Experts

המאמר משך את תשומת ליבי למרות הידע הרדוד שאני מחזיק לגבי תחום הסדרות העתיות (time-series). בגדול הסיבה העיקרית לכך שבשמו מופיע צמד מילים "Foundational Models" שזה חיה די נדירה בתחום הסדרות העתיות להבדיל מתחום מודלי שפה. הסיבה לכך (כנראה) היא מגוון עשיר הרבה יותר של סדרות עתיות השונות יחסית לשפה טבעית.

האמת לא מצאתי ב- Time-MoE, המבוססת כמובן על הטרנספורמרים, מציאות ארכיטקטוניות מאוד מעניינות ועם זאת יש בו כמה דברים שונים מאלו שאנו רגילים לראות ב-LLMs. למשל במקום שכבת טוקניזציה ואמבדינג, מבוססים על מילון טוקנים, שיש לנו ב-LLMs במודל המוצע יש כל טוקן (שזו נקודה בסדרה) עובר טרנספורמציה לא לינאריות עם אקטיבציה מסוג SwiGLU וכמה טרנספורמציות לינאריות.

בנוגע לשכבת הטרנספורמרים, המחברים לוקחים ארכיטקטורת MoE די סטנדרטית. השוני היחיד שמשך את עיניי הוא שימוש בשיטת נרמול RMSNorm שלא הכרתי. פרט לכך יש את כל השכבות הרגילות של הטרנספורמרים כולל כמובן שכבות residual.

השכבה האחרונה של Time-MoE היא קצת שונה ממה שאנו רגילים לראות בטרנספורמרים. מכיוון שלהבדיל ממודלי שפה אנו צריכים מודל בעולם של TS אנו צריכים לחזות במספר נקודות זמן שונה (נגיד שניה, דקה או יום קדימה), המחברים משתמשים בכמה ראשים בשכבה האחרונה. כל ראש אחראי על חיזוי באופק מסוים (כמות דגימות קדימה). באימון משלבים את הלוסים מכל הראשים.

גם פונקציות לוס במאמר הן די סטנדרטיות: פונקצית הובר שהיא הגרסה הרובסטית של L2 (הלא נותנת לא להגיע לערכים גבוהים מאוד). בנוסף יש איבר רגולריזציה שמנסה להפעיל את כל המומחים ב-MoE בצורה אחידה. וכמובן אימנו את המודל על דאטהסטים ענקיים ומגוונים.

זהו וזה - סקירה קצרה, ובתקווה גם ברורה….

https://arxiv.org/pdf/2409.16040