המאמר היומי של טדי מייק 29.06.2025

In-Context Symbolic Regression: Leveraging Large Language Models for Function Discovery

היום יש לנו מאמר טיפה ישן (בן שנה) אבל שהזדקן ממש טוב בינתיים.

המאמר מציג את (In-Context Symbolic Regression (ICSR, גישה חדשנית המשתמשת ב-LLMs לפתרון בעיות רגרסיה סימבולית (SR - symbolic regression). רגרסיה סימבולית היא בעיה שבה מתקבל דאטה (טבלאי בדרך כלל) ואנחנו מתבקשים להחזיר משוואה אנליטית המתארת דאטה זה. הבעיה בעצם מרחיבה את בעיית הרגרסיה הקלאסית, כמו רגרסיה לינארית לדוגמה, בכך שהיא לא רק מוצאת את המקדמים הכי מתאימים אלא גם את מבנה הפונקציה עצמה. בהמשך לדוגמא של פונקציה לינארית, שם אנחנו מניחים שמבנה הפונקציה הוא, ובכן, לינארי. במקום לבנות מודלים ייעודיים, ICSR ממנפת את יכולות הלמידה בתוך הקשר של LLMs כדי להציע ולשפר צורות פונקציונליות באופן איטרטיבי. לדוגמא במחקרים פיזיקלים, ה LLM יודע שצריכים להתחשב ביחידות מידה באופן אימפליסי ולכן ידע לא להציע פונקציות שאינן פולינומיות למשתני הקלט. חידוש זה מאפשר למצוא משוואות פשוטות ומדויקות יותר בהשוואה לשיטות קיימות, ואף להכליל אותן טוב יותר לנתונים חדשים. הגישה גמישה מאוד בזכות הידע המקדימים שמגיע מתהליך האימון מבוסס על כמויות ענקיות של טקסטים עם משוואות, התיאור שלהן, והדאטה ששומש כדי ליצור אותם, ומשתפרת עם התקדמות ה-LLMs ללא צורך באימון נוסף של SR. כלומר, בהקשר הזה, LLMs משתמש בתור meta-learner של משוואות אנליטיות מדאטה טבלאי.

השיטה פועלת בשני שלבים עיקריים. בהתחלה, נוצרות פונקציות התחלתיות (Seed Functions). בשלב הראשון, ה-LLM מקבל קבוצה של תצפיות (נקודות דאטה) ומתבקש לייצר אוכלוסייה ראשונית של פונקציות מועמדות. במקום להסתמך על פונקציות מוגדרת מראש כמו סינוס וכאלה, ה-LLM מייצר את הפונקציות בעצמו, מה שמוביל בדרך כלל למגוון רחב ומורכב יותר של פונקציות. התהליך הזה מבוצע מספר פעמים כדי להתמודד עם פונקציות לא מוגדרות עבור נקודות קלט מסוימות (נניח ln לערכים שליליים).

בשלב השני , המחברים משתמשים בלולאת אופטימיזציה הבאה. בכל איטרציה קוראים כמה דברים. תחילה, הזנת הקשר (In-Context Learning): ה-LLM מקבל כקלט "מטא-פרומפט" (meta-prompt) המכיל את התצפיות כלומר זוגות (X,Y), וכן רשימה של הפונקציות המועמדות הטובות ביותר מהאיטרציות הקודמות יחד עם ציוני ההתאמה (fitness scores) שלהן. ההנחה היא שה-LLM יכול להסיק דפוסים מהדוגמאות הללו ולהציע פונקציה חדשה וטובה יותר.

לאחר מכן ה-LLM מייצר רק את הצורה הפונקציונלית ("שלד") של הפונקציה (למשל, "ax + bx^2 + c"), מבלי לקבוע את המקדמים המספריים. בסוף, המקדמים הבלתי ידועים של הצורה הפונקציונלית שהוצעה על ידי ה-LLM מותאמים לנתונים באמצעות אופטימיזציה חיצונית של בסגנון ריבועים פחותים לא לינאריים (Non-linear Least Squares - NLS). למה לא מבקשים מה LLM גם את הערכים האלה? כי הוא מפשל! לעומת זאת, שימוש באופטימייזר חיצוני מבטיח ערכי מקדמים טובים יותר ומאפשר חקירה יעילה יותר של מרחב הפונקציות. התהליך חוזר על עצמו עד שהשגיאה נמוכה מספיק או עד שהתקציב החישובי נגמר כי בואו נודה באמת, כמה אתם מוכנים למצוא משווה בסופו של יום.

גישת ICSR נבדלת משיטות SR מבוססות טרנספורמר אחרות בכך שהיא לא דורשת אימון מוקדם על דאטהסטים SR סינתטיים גדולים, אלא מסתמכת על הידע המתמטי הקיים ב-LLM המאומן מראש. בנוסף, ICSR מציגה ממשק בשפה טבעית, מה שמאפשר לה לחקור מגוון רחב יותר של פונקציות.

התוצאות של ICSR הן בעלות חשיבות עצומה שכן הן מציגות פריצת דרך בגישת הרגרסיה הסימבולית. הן מוכיחות ש-LLMs, שאינם אומנו במיוחד למטרה זו, יש את היכולת לזהות ולנסח פונקציות מתמטיות לא רק בדיוק גבוה, אלא גם בפשטות אלגנטית, תוך שמירה על יכולת הכללה יוצאת דופן לנתונים שטרם נצפו. יתרון זה של פונקציות פשוטות אך כלליות הוא קריטי ביישומים מדעיים והנדסיים, שכן הוא מאפשר הבנה עמוקה יותר של התופעות הנחקרות ומניעת התאמת יתר לנתוני האימון. כי בואו נודה באמת, אתם שמחים אם יבנו את המטוס הבא שלכם בעזרת איזה ML\DL אבל בסוף אתם רוצים שהמנדס יבין את הפיזיקה של הכלי שהוא בנה ופה SR נותן עבודה חבל על הזמן.

מה גם, שבעידן ש LLM כבר כותבים מאמרים חצי בעצמו, SR סוגר פינה חשובה בהקשר הזה כי SR יכול לשמש ככלי עוצמתי לגילוי חוקים פיזיקליים חדשים, ניסוח משוואות כימיות המתארות תהליכים מורכבים, בניית מודלים ביולוגיים, או זיהוי קשרים כלכליים שדורשים הסבר.

למי שמתעניין בלייצר משוואות בעזרת תיאור וקצת דאטא, שווה קריאה:

https://arxiv.org/pdf/2404.19094