המאמר היומי של יניב ומייק: 03.06.25  
Reinforcement Learning for Reasoning in Small LLMs: What Works and What Doesn’t

מודלי Reasoning כמו GPT-o1 של OpenAI חוללו מהפכה ביכולת ההסקה לטובת פתרון בעיות מתמטיות מורכבות ועד כתיבת קוד אלגנטי הודות למשאבי חישוב עצומים ומאגרי נתונים עצומים שאומנו באמצעות למידה מחיזוקים לבצע חשיבה לוגית. אך האם מודלים קטנים וזולים יותר יכולים להגיע להישגים דומים? מאמר חדש מאת עוסק בדיוק בשאלה הזו.

למה זה חשוב?

מודלים גדולים מספקים ביצועי הסקה מרשימים, אך דורשים משאבים כבדים והם יקרים לשימוש נרחב. מנגד, מודלים קטנים (בסדר גודל של 1–2 מיליארד פרמטרים) זולים ונוחים לפריסה, אך לרוב נופלים מאחור במשימות הסקה מורכבות. מטרת המחקר של דאנג ונו היא שאפתנית אך מעשית: לשפר את ביצועי ההסקה של מודלים קטנים תוך שימוש מינימלי במשאבים.

השיטה: Group Relative Policy Optimization (GRPO) על דאטה באיכות גבוהה

החוקרים בחרו במודל DeepSeek-R1-Distill-Qwen-1.5B (בגודל 1.5 מיליארד פרמטרים) ואימנו אותו באמצעות GRPO: אלגוריתם למידה מחיזוקים שהציגו DeepSeek והוכיח את עצמו במודלים גדולים, אך יושם כאן בקנה מידה קטן משמעותית.  
כדי לשמור על עלויות נמוכות, האימון היה מוגבל מאוד. מבחינת חומרה, היה שימוש ב 4 כרטיסי מסך NVIDIA A40 בלבד. משך האימון הוגבל ל 24 שעות. הדאטסט שנוצר היה מורכב מ - 7,000 שאלות מתמטיות שנבחרו בקפידה.

תוצאות מפתיעות

למרות המגבלות החמורות, השיפור בביצועים היה יוצא דופן:  
דיוק בבנצ'מרק AMC23 קפץ מ־63% ל־80%. בנצ'מרק AIME24 התקבל ציון של 46.7%, שעוקף את o1-preview של OpenAI שעומד על 44.6%. עלות האימון הכוללת: כ־42 דולר בלבד שזה סדרי גודל זול יותר ממתודות עכשוויות מובילות.

מה בדיוק נעשה?

החוקרים ביצעו שלושה ניסויים:

ניסוי 1: אימון עם שאלות מתמטיקה קשות ואיכותיות. שיפור מהיר אך ירידה חדה בביצועים עקב חוסר יציבות ו"סטיית שפה".

ניסוי 2: שילוב שאלות קלות עם קשות. השיג יציבות התחלתית גבוהה יותר ותוצאות שיא מרשימות, אך גם כאן חלה הידרדרות לבסוף.

ניסוי 3: שימוש ב־cosine reward לעידוד תשובות קצרות שיפר את היציבות וביצועים. ממצאים אלה עולים בקנה אחד עם מאמר "DR GRPO", שעלה לארכיב ימים ספורים קודם ומצא נטיה מובנית לתשובות ארוכות ב GRPO.

מגבלות ושאלות פתוחות

למה זה עבד כל כך טוב? התוצאה המפתיעה ביותר היא היכולת של מודל קטן, שאומן במהירות ועל מעט דאטה, להציג ביצועים כל כך מרשימים. החוקרים אינם מספקים הסבר אינטואיטיבי, דבר שמשאיר שאלות על עמידות התוצאות למשימות שונות שאינן מתמטיות.

סטיית שפה: עם התמשכות האימון הבסיס הרב-לשוני של המודל הוביל לתשובות שאינן באנגלית, מה שגרם לאי־יציבות בכל גרסאות המודל.

ספציפיות לתחום: ההערכה התמקדה רק בהסקה מתמטית. לא ברור אם הגישה תעבוד גם במדעים, קוד או תחומים אחרים.

מה הלאה?

המחקר מוכיח שלא חייבים מודלים ענקיים ויקרים כדי להגיע לביצועים טובים בהסקה. ההצלחה של המודל הקטן מצביעה על פוטנציאל לבדוק גרסאות נוספות של GRPO כמו DR GRPO תחת מגבלות משאבים, ולבחון ביצועים על מגוון רחב יותר של משימות. כיוון נוסף הוא כיוונון היפר-פרמטרים וייתכן מאוד ש־KL loss גבוה יותר ישפר את היציבות.

בשורה התחתונה

המאמר מתווה כיוון חדש: מודלים לשוניים קטנים, זולים ובעלי יכולת הסקה משמעותית. גם אם הסיבה לכך עדיין לא ברורה לחלוטין, ההשלכות המעשיות ברורות; פוטנציאל דמוקרטיזציה של יכולות AI מתקדמות גם למעבדות קטנות ולחוקרים עצמאיים.