Review 145: Textbooks Are All You Need II: phi-1.5 technical report, 13.09.23

https://arxiv.org/abs/2309.05463.pdf  
סקירה זו נכתבה על ידי עדן יביןֿ

**Paper: https://arxiv.org/abs/2309.05463v1**

בזמן האחרון עם שחרור מודלי שפה כגון Falcon-180b, שגודל המודל הוא האינדיקטור העיקרי לביצועי המודל, כלומר מודל גדול יותר שווה מודל טוב יותר. חוקרים ממיקרוסופט הראו שזה לא בהכרח נכון עם שחרור של מודל חדש קטן עם ביצועים מצוינים. איך הם עשו את זה? היום נגלה ב-#shorthebrewpapereviews.  
  
מבחינה טכנית אין משהו מיוחד בארכיטקטורה של המודל: טרנספורמר עם 24 שכבות ו-32 ראשי attention. הם השתמשו ב-flash attention לצורך אימון מהיר יותר (לא משהו מיוחד) והטוקנייזר של codegen-mono. אז מה כן מיוחד? ובכן העבודה העיקרית שלהם הייתה המידע לאימון. מבחינה טכנית אין משהו מיוחד בארכיטקטורה של המודל: טרנספורמר עם 24 שכבות ו-32 ראשי attention. הם השתמשו ב-flash attention לצורך אימון מהיר יותר (לא משהו מיוחד) וטוקנייזר של codegen-mono .

אז מה כן מיוחד? ובכן העבודה העיקרית שלהם הייתה המידע לאימון. המידע לאימון הורכב מ-7 מיליארד שהיו שייכים לדאטה סט של המודל הקודם שלהם – phi-1. אליו הם הוסיפו עוד 20 מיליארד טוקונים בצורת ספר לימוד(textbook) על 20 אלף תחומים שונים כגון מדע, תאורית המוח, פעילויות יומיומיות ועוד. לא הוצג כיצד ספרים אלו נוצרו אלא רק שהם לא נוצרו על ידי אדם.   
  
המטרה בספרים אלו הייתה לשפר את יכולת הסקת המסקנות והידע הכולל של המודל לגבי העולם. החוקרים ציינו שהם הוסיפו גם 6 מיליארד טוקונים של קוד שהיו היחידים בכל הדאטה סט (30 מיליארד טוקנים עד כה) שנכתבו על ידי אדם ולא על ידי מכונה. עם דאטה סט זה הם אימנו את phi-1.5 אך לא עצרו כאן. החוקרים רצו לבדוק את השפעת הוספת מידע שנאסף מהאינטרנט (כלומר נכתב על ידי אדם) ומחקו מידע לא חינוכי על ידי השיטה הבאה:

שימוש במודל שפה כדי להחליט עד כמה דגימה היא ״בעלת ערך חינוכי״

אימון מסווג שבהינתן אמבדינג של דגימה מסווג את הערך החינוכי

מחיקה של כל דגימות ללא ערך זה.

חשוב לציין שהחוקרים לא השתמשו בשום שיטה של אימון נוסף (fine-tune) על מעקב אחר הוראות או שיטות כגון RLHF, ובכל זאת הגיעו לביצועים מצוינים מול מודלים שכן עברו את אחד או כל תהליכים אלו. לטענתם, זה רק מוכיח עד כמה הדאטה הוא הדבר החשוב ביותר. סה״כ מאמר מאוד טכני ופחות תיאורטי עם פינות קטנות שמיקרוסופט לא חושפת אך עדיין אפשר ללמוד ממנו לא מעט על הפרטים הטכניים הקשורים לאימון מודלי שפה גדול אלו.