המאמר היומי של מייק - 30.03.25  
Efficient Online Data Mixing For Language Model Pre-Training

ממשיך בלסקור את קו המחקר בנושא אופטימיזציה של אימון מודלים (בפרט מודלי שפה) כאשר יש בידינו כמה דאטהסטים שונים. מכיוון שכבר הגדרתי את הבעיה בסקירות מ-26.03 ומ-28.03 לא אעשה זאת בסקירה זו ומיד אתחיל מהסבר הרעיון העיקרי של המאמר.

המאמר ניגש לבעיה בצורה שונה משני המאמרים הקודמים שסקרתי אך לדעתי (למרות המורכבות מתמטית מסוימת) הגישה המוצעת במאמר היא מאוד אינטואיטיבית. המחברים מנסים לפתור בעיית בניית דאטהסט D לאימון מודל שפה מהדאטהסטים D\_1,..., D\_n עם מה שנקרא Multi-Arm Bandits או MAB. אזכיר בעיית MBA מוגדרת באופן הבא: יש לנו כמה n מכונות מזל עם הסתברויות זכייה p1,..., p\_n שלא ידועות לנו מראש. המטרה היא למצוא אסטרטגית בחירת מכונה הממקסמת את הזכייה (נגיד, התוחלת שלה) כאשר יש לנו N נסיונות.

שימו לב שבעיית אופטימיזצית האימון שלנו די דומה ל-MBA - גם פה אנו צריכים למצוא את אסטרטגית בחירת דאטהסטים לאימון בלי שאנו יודעים מה ״ההשפעה״ של כל דאטהסט לתוצאת האימון הסופית. בלי להיכנס יותר מדי עמוק למתמטיקה (תהליך החלטה מרקובי, התפלגות גיבס וכדומה) המטרה למצוא התפלגות p1,..., p\_n על הדאטהסטים שלנו למקסום ביצועי המודל המאומן. הקאץ' כאן שהתפלגות זו משתנה עם האיטרציות כאשר איטרציה במקרה הזה היא צעד אחד (או מספר כלשהו אך קבוע מראש) על הדאטה מהדאטהסט D\_i שנבחר באיטרציה זו.

כלומר כל פעם אנו בוחרים דאטהסט עם ההתפלגות הנוכחית p, מאמנים את המודל על הדאטה מהדאטהסט הנבחר ומעדכנים את p בהתבסס על תוצאות האימון. כמובן נשאלת השאלה איך ניתן לקבוע p עבור איטרציה הבא על סמך התוצאות של האיטרציה(בחירת דאטהסט) הקודמת. וכאן אנו מגיעים למה שנקרא תגמול (reward) שהוא משקף את ״ההצלחה״ בבחירת הדאטהסט d\_i באיטרציה זו. אם האימון על d\_i היה מוצלח, אנו רוצה להגדיל את ההסתברות שלו (על חשבון האחרים) כאשר אם הוא פחות מוצלח אז צריך להקטין אותה.

אוקיי, אז מה זה בעצם התגמול כאן? התגמול כאן היא המידה שהמודל ירוויח מהדאטה מדאטהסט D\_i כלומר ילמד יותר סוג של information gain או IG. המאמר מחשב את IG בתור פרפלקסיטי (שזה אקספוננט של הלוס) על הדאטה של דאטהסט d\_i. לוס זה משוערך על סמך באץ' מהדאטהסט. בנוסף יש גם עניין של exploration כי אנו לא רוצים ״להקטין דרסטית״ את הסתברות בחירה של דאטהסט מסוים על סמך מעט באצ'ים ואז מגדילים (כמו שמקובל ב-MBA ובשיטות אחרות של RL) כל הסתברות p\_i במספר קטן ε\_t שיורד עם האיטרציות.

אז האלגוריתם הסופי מכיל 3 שלבים:

עדכון הסתברויות בחירה p1,..., p\_n

דגימת דאטה מהדאטהסטים D\_1,..., D\_n לפי הסתברויות אלה ואימון מודל על דאטה

עדכון נוסף של ההסתברויות בהתבסס על המודל המאומן בשלב 2

מאמר מומלץ - נהניתי לצלול אליו

https://arxiv.org/pdf/2312.02406