המאמר היומי של מייק - 26.03.25  
DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining

המאמר שנסקור היום שייך לתחום שלא הכרתי אז יש סיכוי שיהיו שגיאות בסקירה למרות מאמציי הכבירים. המאמר דן באופטימיזצית אימון של מודלי שפה כאשר יש ברשותנו דאטהסטים מדומיינים שונים. בנימה יותר מתמטית המחברים מציעים שיטה למשקול של הדאטהסטים השונים במהלך האימון. כלומר אם יש לנו d דאטהסטים המטרה היא למצוא וקטור d-מימדי α של מספרים אי שליליים המסתכמים ל-1 כאשר α\_k היא ההסתברות לדגום דוגמא מדאטהסט D\_k. כלומר אנו מרכיבים את סט האימון שלנו בשני שלבים: בשלב הראשון בוחרים דאטהסט עם דגימה מ- α ולאחר מכן בוחרים דוגמא הדאטהסט הנבחר.

דרך אחת פשוטה היא לבחור את α\_i פרופורציונלית לגודל הדאטהסט D\_i כלומר ככל שדאטהסט גדול מדי הוא ייבחר יותר פעמיים. אפשר לדגום גם בצורה יוניפורמית כאשר כל דאטהסט ייבחר בהסתברות 1/d כאשר d הוא מספר הדאטהסטים. יש שיטות שבוחרות α לפי איכות הדאטהסט ומעדיפים דאטהסטים איכותיים יותר על פני אלו שפחות איכותיים.

אבל איך לבחור את α בצורה שתמקסם את ביצועי המודל המאומן? זו השאלה שעליה מנסה המאמר לענות. אחת השיטות היא לנסות כל מיני ערכים של α ועבור כל אחד לאמן את המודל (brute-force). עבור מודלים גדולים ומספר גבוה של דאטהסטים d המחיר החישובי (= עלות) עלול להיות עצום. השאלה האם ניתן לעשות משהו חכם מזה?

התשובה על השאלה הזו היא כן וזה מה שהמחברים מציעים. בשלב הראשון המחברים מציעים לאמן מודל M\_ref קטן עם α f כלשהו (נגיד יוניפורמי). החברים מציעים להשתמש בשיטת (distributionally robust language modeling (DRO-LM שמאמנת מודל קטן הממזער את השגיאה המקסימלית ביחס מעל כל וקטורי α (החוקיים) יחסית לשגיאה של M\_ref (הפרש השגיאות בין M\_ref למודל המאומן). השגיאה במקרה הזה היא לוג של הנראות של הטוקן הנכון (עבור כל מודל ממצעים עבור כל הטוקנים עבור כל דאטהסט בנפרד).

אם ראיתם כאן בעיית minimax, אתם צודקים. בצורה איטריבית ממקסמים(כלומר עושים מעלה הגרדיאנט או gradient ascent) את הפרש השגיאות (עבור באצ'ים של דוגמאות) מעל α ולאחר מכן ממזערים את הפרש השגיאות מעל משקלי המודל המאומן (כלומר gradient ascent). וקטור המשקול α הסופי שנבחר על ידי מיצוע של כל וקטורי α עבור כל האיטרציות של בעיית המינימקס הזו. מעניין שהבאצ'ים נדגמים באקראי עבור כל האיטרציות. בשלב האחרון מאמנים מודל גדול עם α שמצאנו בצורה הזו.

מקווה שהצלחתי להסביר את זה ברור…

https://arxiv.org/abs/2305.10429