המאמר היומי של מייק - 28.03.25  
UniMax: Fairer and More Effective Language Sampling for Large-Scale Multilingual Pretraining

אני ממשיך לסקור מאמרים בנושא של אופטימיזצית בחירת דאטה לאימון מודלים בפרט עבור מודלי שפה שאנו כה אוהבים. הבעיה ניתנת לניסוח די פשוט: יש לכם כמה דאטהסטים {D = {D1,..., D\_n. המטרה שלכם לבחור שילוב ״אופטימלי״ של דאטהסטים אלו לאימון המודל כאשר יש לנו תקציב B של מספר הטוקנים/סימבולים (במאמרים משתמשים בסימבולים) הכולל שאנו שהמודל ״רואה במהלך האימון״ (ניתן לתרגם את זה ל-FLOPs בהינתן ארכיטקטורה של המודל). במילים פשטות אנו מנסים להבין את ניתן לדגום דאטה מ-D כדי לקבל מודל בעל ביצועים הטובים ביותר אחרי האימון תחת תקציב B. כמובן ש״הטוב ביותר״ ניתן להגדיר במספר מובנים אבל אנו פחות נתמקד בנקודה הזו ונדון בעיקר איך ניתן לאזן בין דוגמאות ממקורות (דאטהסטים) שונים באימון המודל.

המאמר דן בתרחיש של אימון מודלים מולטי-שפתיים כאשר יש ברשותנו דאטה בכמה שפות. הגישה הפשוטה היא להעניק תקציב שווה לכל דאטהסט (=שפה) כלומר דאטהסטים בגודל שונה יאומנו מספר אפוקים (epoch) שונה כאשר דאטהסטים קטנים(שפות לא נפוצות) יותר יאומנו במספר אפוקים גבוה יחסית לדאטהסטים גדולים (שפות פופולריות יותר). המאמר טוען שחוסר התאמה זה עלול לגרום לביצועים ירודים של המודל. המחברים מציעים שיטה מאוד אינטואיטיבית ופשוטה לאיזון של מספר האפוקים לדאטהסטים שונים תחת תקציב B.

המאמר קובע מספר מקסימלי לאפוקים N שיינתן לכל דאטהסט. התהליך מתחיל בדאטהסט הקטן ביותר (שפה הכי פחות נפוצה) המחשב לפי מספר הסימבולים C\_i בשפה את מספר האפוקים לדאטהסט זה בהתאם לתקציב ממוצע פר שפה(B מחולק ב-|D|). אם מספר האפוקים E\_i עבור השפה הנוכחית עולה על N, הוא(מס' אפוקים) נקבע להיות N. לאחר מכן מחסרים מ-B את תקציב לדאטהסט הנוכחי ומחשבים תקציב ממוצע u\_i פר שפה (נותרו כרגע 1 - |D| שפות). אז ממשיכים את התהליך עבור כל השפות בדאטהסט. המחברים מציינים שהם לא משתמשים במספר הטוקנים כדי לאמוד ״גודל אפקטיבי״ של כל שפה בדאטהסט עקב מורכבות של טוקניזציה של דאטהסט מולטי-שפתי.

בשלב האחרון מנרמלים את כל התקציבים u\_i עם הסופטמקס כדי לקבל ההתפלגות p שממנו דוגמים דאטה עבור כל השפות. ניתן להשתמש ב-p עם טמפרטורה τ כדי להחליק או להבליט את ההתפלגות (מעלים p\_i בחזקה τ/1) ולדגום ממנה את השפות (=דאטהסטים).

זה וזה - מחר עוד מאמר בנושא…

https://arxiv.org/abs/2304.09151