המאמר היומי של מייק - 22.03.25  
Physics in Next-token Prediction

המאמר הזה לא רגיל. זה מתחיל מהשם שלו: הרי איך חיזוי של טוקן הבא (NTP) יכול להיות קשור לפיזיקה. מתברר שהקשר הזה קיים והוא עובר דרך תורת המידע (information theory). מי שמכיר אותי יודע שאני מאוד מתעניין בהיבט מידעי (אינפורמציוני) שקיים בתחום למידת מכונה, בדחיסת מידע על ידי המודלים, על איך הידע נשמר במודלים מאומנים וכדומה. והמאמר הזה מדבר בדיוק על הנושאים האלו ולמרות שאין בו מתמטיקה יותר מדי מורכבת הוא די עמוק (בספק הצלחתי להפנים אותו אני במלאו:) ).

נתחיל בלציין שלפי חוק שנון (עם טוויסט קטן) הטוען כי כדי להעביר מילה {x\_{t+1 לאחר שהעברנו t מילים הינה שווה לאנטרופיה מותנית H של {x\_{t+1 בהינתן x\_1, …x\_t או מידע עצמי I. האנטרופיה H שווה במקרה הזה ל-logprob של הסתברות מותנית של {x\_{t+1 בהינתן x\_1, …x\_t. נובע מכך (די בקלות) שמספר הביטים הנדרש כדי להעביר את כל המילים מדאטהסט D כלשהו הינו סכום של האנטרופיות המותנות אלו עבור t מ- 1 עד |D|.

עכשיו נניח שיש לנו מודל שאימנו אותו לחזות טוקן הבא בהינתן ההקשר (כלומר הטוקנים הקודמים), למשל מודל שפה. מספר הביטים הנדרש להעביר את אותן המילים מדאטהסט D מחושב לפי אותה הנוסחה, כלומר סכום של אנטרופיה מותנות של {x\_{t+1 בהינתן x\_1, …x\_t. אבל הפעם, כאשר המודל משמש לחיזוי אנטרופיה זו (כלומר הסתברות מותנית) כנראה שנצטרך פחות ביטים להעברת אותו דאטהסט D. למה זה בעצם קורה? איפה נעלם ההפרש בין מספר הביטים שצריך כדי להעביר את D בלי המודל ועם המודל?

מכיוון שאינפורמציה לא יכולה ללכת לאיבוד ההנחה היא שהמודל צבר אותו (למד). המאמר קורא למידע זה השמור בתוך המודל מידע אפקטיבי של המודל על דאטהסט D (או משימה). המאמר גם מגדיר η שהיא הקיבולת של המודל בתור היחס בין המידע האפקטיבי של המודל למספר הפרמטרים של המודל (בביטים). בנוסף נציין משהו מאוד מעניין: כי מספר הביטים שצריך כדי להעביר את D עם המודל הוא לוס (cross-entropy) של המודל עבור D מוכפל ב |D|.

אז אם נקשר את כל הגדלים שהגדרנו קודם נקבל את חוק הקיבול המידעי הראשון שמוגדר במאמר: (ηN = |D| (H − L, כאשר N זה מספר הפרמטרים של המודל, L זה קרוס-אנטרופי לוס של המודל על הדאטהסט D, ו-H היא האנטרופיה ההתחלתית של D. במהלך האימון H ו-N נשארים קבועים ו- |D| הוא מספר הטוקנים שהמודל ״ראה״ במהלך האימון. כלומר האימון הוא תהליך של דחיסת דאטהסט D והעברותו מידע ממנו למודל המאומן.

המאמר גם מגדיר את החוק הקיבול המידעי השני המתאר את האנרגיה המינימלית הנדרשת להעברת אינפורמציה מ-D למודל. היא פרופורציונלית ל N ו-ל η ומופיע בו גם טמפרטורה T (לא לבלבל עם הטמפרטורה של LLMs) וגם קבוע בולצמן k- מודה שלא הצלחתי להבין את המשמעות של השניים האחרונים (T ו- k).

בהתבסס על תורה זו המבחרים מגיעים למסקנות מעניינות לגבי אימון המודל וגם משווים את החוקים שניסחו עם חוקי סקיילינג של מודלי שפה. מי שמתעניין בזה, מוזמן לצלול - מאמר מרתק.

https://arxiv.org/abs/2411.00660