המאמר היומי של מייק: 04.08.25  
Rethinking Transformers Through the Lens of Physics: The Rise of Energy-Based Models

פיזיקה פוגשת AI: כך מודל חדש לומד שפה בלי לחזות אפילו טוקן אחד

במשך שנים, הפרדיגמה הדומיננטית לאימון LLMs הייתה פשוטה באופן מטעה: ללמד אותם לחזות את המילה הבאה. גישה אוטורגרסיבית זו, המבוססת על נראות, זכתה להצלחה אדירה, אך יש לה מגבלות אינהרנטיות. מודלים שאומנו כך חושבים באופן מקומי, טוקן אחר טוקן. הם עלולים לאבד את הקוהרנטיות הגלובלית, להתקשות עם תלויות ארוכות טווח, ולהתקשות במילוי אילוצים מורכבים והוליסטיים.

אבל מה אם במקום ללמד מודל לחזות את הצעד הבא, נוכל ללמד אותו לזהות תוצאה טובה כשהוא רואה אחת? מאמר של צוות חוקרים מסטנפורד מציע בדיוק את זה, על ידי הגדרה מחדש של הטרנספורמר לא כמודל חיזוי סדרתי, אלא כמודל מבוסס אנרגיה (Energy-Based Model - EBM). זו אינה רק ארכיטקטורה חדשה; זוהי פילוסופיה חדשה, כזו שממירה את ההיגיון המקומי של הסבירות באינטואיציה הגלובלית של מערכת פיזיקלית.

#### הרעיון המרכזי: מחיזוי טוקנים לניקוד רצפים

בבסיסו, מודל מבוסס אנרגיה אינו מחשב את ההסתברות של פיסת מידע באופן ישיר. במקום זאת, הוא מקצה ערך סקלרי, אנרגיה, לכל תצורה אפשרית. העיקרון המרכזי פשוט: תצורות בעלות אנרגיה נמוכה הן סבירות יותר, יציבות יותר, ו"נכונות" יותר. תצורות בעלות אנרגיה גבוהה אינן סבירות. מחברי המאמר מיישמים תפיסה זו על שפה. הטרנספורמר מבוסס האנרגיה (EBT) שלהם אינו חוזה טוקנים. הוא קורא טקסט שלם ומוציא מספר בודד: האנרגיה שלו. משפט בנוי היטב, קוהרנטי והגיוני יקבל ציון אנרגיה נמוך מאוד. משפט משובש או חסר פשר יקבל ציון גבוה.

זהו שינוי יסודי. בניגוד למודל GPT סטנדרטי, שהוא חד-כיווני ומעבד טקסט טוקן אחר טוקן, ה-EBT הוא דו-כיווני לחלוטין. הוא יכול להעריך את הקוהרנטיות הגלובלית של משפט על ידי התבוננות בכל חלקיו בו-זמנית, בדומה לאופן שבו קורא אנושי היה עושה זאת.

אז איך מאמנים מודל כזה? אם אי אפשר למקסם את הסבירות של הטוקן הבא, מהי מטרת האופטימיזציה? התשובה היא למידה ניגודית (contrastive learning).

תהליך האימון הוא אלגנטי:

מציגים למודל דוגמה "חיובית", משפט אמיתי מנתוני האימון, ומלמדים אותו להקצות למשפט זה ציון אנרגיה נמוך.

לאחר מכן, מציגים לו דוגמה "שלילית" – גרסה משובשת של המשפט, אולי עם כמה מילים שהוחלפו באקראי. מלמדים את המודל להקצות למשפט חסר פשר זה ציון אנרגיה גבוה.

על ידי חזרה על תהליך זה מיליוני פעמים, ה-EBT לומד לבנות "מדטח אנרגיה" (energy landscape) עבור כל מרחב המשפטים האפשריים. שפה תקינה שוכנת בעמקי האנרגיה הנמוכה, בעוד שכל השאר נדחף אל הרי האנרגיה הגבוהה.

#### חשיבה ויצירה באמצעות גרדיאנט

הפרספקטיבה הגלובלית הזו היא מה שמשחרר את ה"הוגה" (thinker) שבכותרת המודל. מכיוון שה-EBT יוצר את הטקסט כולו, הוא מצטיין במשימות הדורשות חשיבה הוליסטית ועמידה באילוצים, תחומים שבהם מודלים אוטו-רגרסיביים נוטים להיכשל.

יצירת טקסט (generation), לעומת זאת, היא סיפור אחר. אי אפשר פשוט לדגום מתוך משטי אנרגיה באופן ישיר. במקום זאת, המודל צריך למצוא את עמקי האנרגיה הנמוכה. המחברים משתמשים בטכניקה איטרטיבית בהשראת הפיזיקה שנקראת Langevin dynamics, סוג של דגימת MCMC (בערך). התהליך נראה כך:

מתחילים עם סדרה של רעש טהור (טוקנים אקראיים).

מחשבים את האנרגיה של הסדרה הזבל הזה.

דוחפים קלות את הטוקנים בכיוון שמפחית את האנרגיה במידה המרבית (כלומר, יורדים במורד הגרדיאנט של פונקציית האנרגיה).

חוזרים על תהליך זה מאות פעמים.

באטיות, באופן איטרטיבי, הסדרה האקראית מעודנת, מתייצבת מההרים הגבוהים של האנרגיה מטה אל עמק נמוך-אנרגיה, ומתגבש למשפט קוהרנטי ובנוי היטב. אמנם תהליך זה איטי יותר מיצירה אוטו-רגרסיבית סטנדרטית, אך הוא מאפשר צורת יצירה מבוקרת ומודעת-גלובלית הרבה יותר.

#### מדוע הוא "לומד וחושב סקיילאבילי"?

המאמר מספק ראיות חזקות לכך שגישה זו הינה סקיילאבילית (ניתנת להרחבה). ככל שהמודלים גדלים, יכולתם להבחין בין רצפים טובים לרעים משתפרת, ואיכות הדגימות שהם יוצרים עולה. חשוב מכך, המסגרת מבוססת האנרגיה היא גמישה להפליא. אינך כבול עוד לחיזוי הטוקן הבא. רוצה מודל שמייצר ביקורות קולנוע חיוביות? פשוט הוסף "מקדם אנרגיה" נוסף למטרת האימון שמעניש סנטימנט שלילי. מודולריות זו הופכת את ה-EBT לכלי רב עוצמה ליצירה נשלטת.

עבודה זו מאלצת אותנו לבחון מחדש את יסודות המודלים הנוכחיים שלנו. היא מציעה שהדרך לבינה מלאכותית חזקה, קוהרנטית ונשלטת יותר עשויה שלא להיות טמונה רק בהגדלה אינסופית של יכולת חיזוי הטוקן הבא, אלא בבניית מודלים שמבינים שפה ברמה הוליסטית ופיזיקלית יותר.

https://arxiv.org/abs/2507.02092

זה לא רק מה אתם כותבים בפרומפט, אלא איפה