⚡️🚀המאמר היומי של מייק -05.12.24: ⚡️🚀  
Memory3: Language Modeling with Explicit Memory

#### א. רעיון כללי:

המאמר מציע זיכרון מפורש (explicit memory או EP) כתוספת לארכיטקטורה של מודלים לשוניים. בניגוד לאופי הסטטי של פרמטרי המודל או הזיכרון הזמני (משקלי K ו- V), הזיכרון המפורש פועל כמחסן ידע מובנה ודינמי, הניתן לאחזור מחוץ למודל שפה.

זיכרון מפורש מיועד ל״שיפור טרייד-אוף״ בין גודל של LLMs לבין ביצועיהם. באמצעות החצנת ידע פחות מופשט (כמו עובדות, נתונים, חוקים ספציפיים לתחום) אל תוך EM, המודל נמנע מהגדלה משמעותית של פרמטרי המודל, תוך שמירה או אף שיפור של ביצועים. חידוש זה לא רק משפר את היעילות החישובית, אלא גם הופך את המערכת למודולרית. עדכוני ידע אינם מחייבים אימון מחדש של כל המודל, מה שמדמה תהליך למידה אנושי שבו מידע חדש נשמר מבלי לשנות את הפונקציות הקוגניטיביות הבסיסיות.

#### ב. היררכיית זיכרון מוצעת

היררכיית הזיכרון שהוצעה במאמר שואבת השראה ממערכות קוגניטיביות אנושיות, שבהן הזיכרון לטווח ארוך מסווג לפי נגישות ותדירות שימוש. המחברים מעצבים מסגרת זו כדי להקצות ידע אסטרטגית ב- 3 רמות:

טקסט פשוט (עלויות קריאה גבוהות, עלויות כתיבה נמוכות):

מתאים למידע שניגשים אליו באופן נדיר, אחסון טקסט פשוט שומר על קלילות המערכת הכוללת. אחזור מזיכרון זה פחות יעיל אך משמש כגיבוי לשאילתות נדירות.

זיכרון מפורש (עלויות מאוזנות):

ידע הנמצא בשימוש תדיר יותר אך לא קריטי (כמו ידע מופשט על השפה) נשמר ב-EM, המאזן בין מהירות האחזור(retrieval) לעלויות האחסון. האינטגרציה שלו עם מנגנוני attention דלילים מבטיחה שרק חלקי הזיכרון הרלוונטיים ביותר יופעלו, מה שמשפר את יעילות האינפרנס.

פרמטרי מודל (עלויות קריאה נמוכות, עלויות כתיבה גבוהות):

שמור לידע מופשט המהווה ליבה ליכולות האינפרנס הבסיסיות של המודל. עדכונים בשכבה זו מתבצעים באימון, מה שהופך אותם ליקרים חישובית.

היררכיה זו מאפשרת ל-Memory3 לתעדף הקצאת משאבים בצורה דינמית, ומבטיחה שהעלויות החישוביות יישארו ניתנות לניהול תוך שמירה על ביצועים גבוהים. עיצוב זה רלוונטי במיוחד ליישומים הדורשים עדכוני ידע בזמן אמת, כגון מערכות תמיכת לקוחות או בוטים מותאמים לתחום ספציפי.

#### ג. ארכיטקטורה

ארכיטקטורת Memory3 היא אבולוציה משמעותית של מודלים סטנדרטיים מבוססי טרנספורמרים, תוך שילוב זיכרון מפורש באופן חלק.

חידושים עיקריים:

מנגנוני attention דלילים:

באמצעות שילוב הזיכרון המפורש במנגנון attention, הגישה המוצעת נמנעת מעסקיילינג הריבועי של attention (היו בעבר טרנספורמרים שעשו משהו דומה). attention דליל מפחית כמות חישובים על ידי התמקדות רק בתת-קבוצות של זיכרון הרלוונטיות ביותר לשאילתה.

אחזור זיכרון יעיל:

המודל משתמש בחיפוש מבוסס דמיון קוסינוס כדי לאחזר זוגות מפתח-ערך(KV) רלוונטיים. אמבדינגס של חלקי הזיכרון הרלוונטיים מחושבים מראש שמבטיח אחזור מהיר וסקיילבילי, כך שמהירות האינפרנס לא נפגעת גם כשהזיכרון גדל.

דילול(sparsification) זיכרון:

כדי לשמור על יעילות הזיכרון, המחברים מציעים טכניקות כמו בחירת טוקנים מדורגת (Top-k), שבהם נשמרים רק הטוקנים האינפורמטיביים ביותר. זאת בשילוב עם קוונטיזציה של וקטורים, שמכווצת את אמבדינגס של הזיכרון מבלי לאבד משמעותית מכוח הייצוג שלהן.

גמישות בעדכוני ידע:

בניגוד לאחסון מבוסס פרמטרים, זיכרון מפורש מאפשר עדכונים מודולריים. לדוגמה, הוספת ידע חדש כרוכה רק בהוספת זוגות KV במקום אימון מחדש של המודל, מה שהופך את Memory3 למותאם ומתאים לעתיד.

#### ד. פרדיגמת האימון

המחברים מאמצים פרדיגמת אימון בשני שלבים אשר מותאמת לשילוב זיכרון מפורש:

שלב אימון warm-up:

המודל עובר אימון בסיסי ללא EM. שלב זה מבטיח פיתוח של יכולות הפשטה חזקות והבנה לשונית בסיסית. שלב זה דומה לאימון מקדים במודלים טרנספורמריים מסורתיים.

שלב אימון continual:

המודל לומד לכתוב ולקרוא מ-EM. מטרות האימון מתרחבות כדי לכלול משימות ספציפיות לזיכרון כמו:

כתיבת זיכרון: אופטימיזציה של אחסון ידע בתור זוגות KV.

אחזור זיכרון: שיפור היכולת לאחזר מידע רלוונטי באופן יעיל ומדויק במהלך האינפרנס.

### סיכום:

שילוב EM ב-Memory3 ממחיש דרך חדשנית לבניית מודלים לשוניים יעילים, ניתנים להתאמה ומודולריים. הגישה הזו עשויה (למרות שב-5 החודשים מאז יציאת המאמר לא ראיתי ניצנים לכך) להוות בסיס לדור הבא של LLMs, במיוחד בתחומים הדורשים עדכונים שוטפים של ידע ו-interpretability גבוה (בגלל שיש זיכרון מפורש).

https://arxiv.org/abs/2407.01178