המאמר היומי של מייק: 28.08.25

Memento: Fine-tuning LLM Agents without Fine-tuning LLMs

המאמר מציע פרדיגמה חדשה לבניית סוכנים שיכולים ללמוד ולהסתגל מניסיון, כל זאת ללא העלות המשתקת של פיינטיון של ה-LLM עצמו. זהו שילוב מעניין של AI קלאסי ולמידה עם חיזוקים מודרנית שמרגיש כמו צעד אמיתי קדימה.

המאמר שואל את השאלה הבאה: הכיצד נוכל לבנות סוכנים שלומדים באופן רציף מסביבתם ללא העלות הגבוהה של פיינטיון למשימה ספציפית? במקום לשנות את הידע הפנימי והפרמטרי של ה-LLM, השיטה המוצעת מצמידה לסוכן את זיכרון אדפטיבית. הסוכן מכוונן את היכולת שלו להשתמש בזיכרון, בלי לעשות פיינטיון ל-LLM עצמו וללא ניהול ידני של הזיכרון.

### החידוש: הזיכרון כפוליסי

החידוש המרכזי בממנטו הוא למסגר מחדש את תהליך הלמידה של הסוכן. במקום ללמד את ה-LLM טריקים חדשים, המטרה היא ללמד את הסוכן להפוך למומחה בהתייחסות לחוויות העבר שלו, הן הצלחות והן כישלונות. זה מושג על ידי שילוב של פריימוורק מתמטית מדויק עם תהליך הסקה מעוגן פסיכולוגית (ככה טוענים במאמר).

תרומתם הראשונה של המחברים היא למדל באופן פורמלי את עולמו של הסוכן כתהליך החלטה מרקובי מוגבר-זיכרון (M-MDP). וזה עמוק יותר ממה שזה נשמע. תהליך החלטה מרקובי (MDP) סטנדרטי מגדיר כיצד סוכן צריך לבחור פעולה בהתבסס על המצב הנוכחי שלו. ה-M-MDP מוסיף משתנה חדש וחיוני למשוואה: הזיכרון של הסוכן. כעת, הפעולה האופטימלית תלויה לא רק במצב הנוכחי, אלא בכל מאגר חוויות העבר שהסוכן צבר.

פורמליזם זה הופך את הרעיון המעורפל של "למידה מניסיון" לבעיית אופטימיזציה פתירה. התנהגות הסוכן אינה עוד רק פונקציה של המצב הנוכחי שלו, אלא מדיניות המותנית במפורש בזיכרונו.

### המנוע: למידת חיזוקית מבוססת-מקרים

ה-M-MDP הוא ה"מה", אבל הסקה מבוססת-מקרים (CBR) היא ה"איך". ממנטו מיישם פוליסי CBR שבה, בכל שלב, הסוכן מבצע תהליך דו-שלבי:

שליפה: תחילה הוא מתייעץ עם הזיכרון שלו, "בנק מקרים" הולך וגדל של מסלולי עבר, ובוחר מקרה עבר רלוונטי. מקרה הוא שלשה פשוטה: המצב שבו היה, הפעולה שנקט, והתגמול שקיבל.

שימוש חוזר והתאמה: לאחר מכן המקרה שנשלף מוזן למתכנן ה-LLM, יחד עם המשימה הנוכחית. תפקידו של ה-LLM הוא להתאים את הפתרון מהמקרה הישן לבעיה החדשה.

החדשנות של המערכת טמונה בשלב השליפה. כיצד הוא לומד איזה מקרה לשלוף? גישה נאיבית עשויה פשוט למצוא את חווית העבר הדומה ביותר מבחינה סמנטית. אבל ממנטו מתוחכם הרבה יותר. הוא לומד פוליסי שליפת-מקרים באמצעות גישות מעולם למידה עם חיזוקים (כלומר RL).

ה"פעולה" אינה קריאה לכלי או שורת קוד; זוהי פעולת בחירת הזיכרון. המערכת לומדת, באמצעות ניסוי וטעייה, פונקציית ערך החוזה עד כמה מקרה עבר מסוים יהיה שימושי לפתרון הבעיה הנוכחית. זה מושג באמצעות למידת Q רכה (soft Q-learning), שבה הסוכן מתוגמל על בחירת מקרים המובילים לתוצאות מוצלחות. החלק ה"רך" מעודד חקירה, ומונע מהסוכן להיתקע בשליפת אותם זיכרונות ספורים שוב ושוב.

פונקציית ה-Q הנלמדת הזו היא החלק מה"כוונון העדין" של הסוכן. חשוב לציין כי זוהי רשת נוירונים שבאמצעותה היא מחושבת הינה רדודה וקטנה, לא LLM עם עשרות מיליארדי פרמטרים. המוח של הסוכן (ה-LLM) נשאר קפוא, בעוד שהמיומנות שלו בגישה לניסיון שלו (מדיניות שליפת הזיכרון) משתפרת ללא הרף.

### זיכרון פרמטרי מול זיכרון לא-פרמטרי

ממנטו מיישם מדיניות שליפה זו בשני אופנים:

זיכרון לא-פרמטרי: זוהי גרסת הבסיס הפשוטה יותר, שבה מקרים נשלפים על בסיס דמיון קוסינוס. זה עובד, אבל זה "טיפש", ומתייחס לכל חוויות העבר הדומות כשוות ערך.

זיכרון פרמטרי: זוהי הגישה המלאה והנלמדת. כאן, פונקציית Q נוירונית קטנה מאומנת אונליין כדי לחזות את התועלת שבשליפת מקרה נתון עבור המצב הנוכחי. בכל פעם שהסוכן משלים משימה, הוא לא רק שומר את החוויה; הוא משתמש בתוצאה זו כדי לעדכן את פונקציית ה-Q שלו, ובכך מחדד בעדינות את הבנתו אילו זיכרונות הם בעלי הערך הרב ביותר. גישה פרמטרית זו משיגה ביצועים טובים יותר באופן עקבי מהגישה הלא-פרמטרית, ומוכיחה שללמוד כיצד לשלוף הוא מנגנון חזק יותר מאשר פשוט למצוא דברים דומים.

ביצועים ומדוע זה חשוב

התוצאות מדברות בעד עצמן. ממנטו השיג ביצועי טופ-1 במערך האימות של מבחן GAIA והדגים שיפורים משמעותיים ועקביים במגוון רחב של מבחנים אחרים כמו DeepResearcher ו-SimpleQA. אך התוצאות החשובות ביותר מגיעות ממחקרי האבלציה (ablation studies), אשר מפרקים באופן שיטתי את ממנטו כדי להוכיח מהיכן נובע הקסם:

תכנון הוא חיוני: ארכיטקטורת המתכנן-מבצע (planner-executor) הבסיסית מספקת שיפור עצום לעומת LLM פשוט המשתמש בכלים, ומאשרת שפירוק משימות הוא המפתח.

CBR מספק דחיפה נוספת: הוספת הזיכרון מבוסס-המקרים מעל המתכנן מניבה קפיצה עקבית נוספת בביצועים בכל המשימות. זה מוכיח שמערכת הזיכרון אינה רק גימיק, אלא תורם מרכזי להצלחת הסוכן.

הוא מכליל: כאשר אומן על קבוצת משימות אחת ונבדק על דאטהסטים שהם לחלוטין מחוץ להתפלגות (OOD), ממנטו הראה שיפורים אבסולוטיים בביצועים של עד 9.6%. זהו נתון מכריע: למידה מניסיון מאפשרת לסוכן לפתח אסטרטגיות פתרון בעיות מוכללות המועברות למצבים חדשים.

ממנטו מציע תוכנית-אב חדשה ומשכנעת לבניית סוכני LLM. על ידי ניתוק בין מנוע ההסקה היציב והמרכזי של הסוכן (ה-LLM) לבין הניסיון הדינמי והמתפתח שלו (זיכרון המקרים האדפטיבי), הוא מספק נתיב אפשרי מבחינה חישובית ליצירת סוכנים המסוגלים ללמידה אמיתית לאורך החיים (lifelong learning). זוהי מסגרת עקרונית שמתקדמת מעבר לטריקים של הנחיה (prompting) אד-הוק, לעבר מדע חזק ומבוסס מתמטית של תכנון סוכנים.

http://arxiv.org/abs/2508.16153

האנטרופיה של הביטחון: לגרום למודל להאמין בעצמו