המאמר היומי של מייק: 01.08.25

Hierarchical Reasoning Model

האם מודל ההיגיון ההיררכי הוא הצעד הראשון לקראת AI שהיא לא רק סימולציה של תבונה (אני חושד שלא)?

חקר AI נדמה לעיתים קרובות כמסע בלתי פוסק של הגדלת קנה מידה. מודלים גדולים יותר, יותר נתונים, יותר כוח חישובי. הפרדיגמה השלטת להסקת מסקנות ב-LLMs הייתה (Chain-of-Thought או CoT), טכניקה חכמה המשדלת מודלים "לחשוב בקול רם" על ידי יצירת הצדקות טקסטואליות שלב אחר שלב. אבל עד כמה ש-CoT יכולה להיות יעילה, היא תמיד הרגישה כמו קבּיים דרך לפצות על חסרון ארכיטקטוני. היא שבירה, תאבת נתונים ויקרה מבחינה חישובית, ומחצינה את תהליך המחשבה המורכב אל תוך הערוץ הצר של השפה.

אבל מה אם מודל היה יכול להסיק מסקנות באופן פנימי, שקט ויעיל, בדומה למוח האנושי? המאמר המסוקר מציג ארכיטקטורה חדשנית שאינה שיפור הדרגתי אלא חשיבה מחודשת מיסודה על האופן שבו אנו עשויים לבנות מכונות המסוגלות לחשיבה לוגית. זה לא עוד מודל; זוהי תוכנית אב מעניינת, בהשראת המוח, המדגימה יכולות חזקות (לכאורה) תוך שימוש במעט מאוד מהמשאבים. בואו נצלול לעומק החידושים המרכזיים של העבודה המלהיבה הזו.

רעיון הליבה: חשיבה סמויה (Latent Reasoning) במערכת דו-שכבתית

החידוש המרכזי של מודל ההיגיון ההיררכי (HRM) הוא נטישת המבנה השטוח והמונוליתי של מודלי טרנספורמר סטנדרטיים. בהשראת האופן שבו המוח מארגן חישובים באזורים שונים ובמהירויות שונות, HRM הוא ארכיטקטורה רקורסיבית הבנויה על שני מודולים התלויים זה בזה:

מודול ברמה גבוהה (H - High-Level): מודול זה פועל בסקאלת זמן איטית יותר. חשבו עליו כעל המתכנן האסטרטגי או על החשיבה המודעת והשקולה. הוא אינו מסתבך בפרטים הקטנים, אלא אחראי על יצירת תוכניות מופשטות והנחיית מסלול פתרון הבעיות הכולל.

מודול ברמה נמוכה (L - Low-Level): מודול זה הוא "סוס העבודה" המהיר. הוא מקבל את התוכנית המופשטת ממודול ה-H ומבצע חישובים וחיפושים מהירים ומפורטים.

התהליך כולו מתרחש במרחב לטנטי. במקום לייצר טוקנים (מילים), המודל מתפעל ומעדן וקטורים ממדיים גבוהים – מצב "המחשבה" הפנימי שלו. המצב של מודול H מספק הקשר מנחה, ובתוך אותו הקשר יציב, מודול L מבצע איטרציות מהירות כדי לחקור פתרונות. זהו שינוי תפיסתי עמוק. הוא מרמז שהשפה נועדה לתקשורת, ולא מהווה את המצע למחשבה עצמה – השקפה המהדהדת את מדעי המוח המודרניים.

השגת עומק חישובי אמיתי באמצעות "התכנסות היררכית"

כל מי שעבד עם רשתות נוירונים רקורסיביות (RNNs) סטנדרטיות מכיר את המלכודות שלהן. הן נוטות להתכנס לפתרון מהר מדי, ובכך עוצרות את החישוב ומגבילות את "עומק" המחשבה שלהן, או שהן סובלות מחוסר יציבות כמו דעיכה או התפוצצות של גרדיאנטים. מודל ה-HRM עוקף בעיה זו באמצעות קונספט אלגנטי שהמחברים מכנים התכנסות היררכית (Hierarchical Convergence).

זו האינטואיציה:

בהינתן הקשר אסטרטגי שנקבע על ידי מודול H האיטי, מודול L המהיר רץ למספר קבוע של צעדים ומבצע את החיפוש המפורט שלו. RNN זה יתחיל באופן טבעי להתייצב סביב שיווי משקל מקומי – מצב פנימי יציב.

בדיוק כשהאנרגיה החישובית שלו (כלומר השתנות) עומדת לדעוך, המחזור מסתיים. המצב הסופי של מודול L מוזן בחזרה למודול H.

מודול H מטמיע את התוצאה הזו ומבצע עדכון איטי משלו, ובכך קובע הקשר חדש ברמה הגבוהה.

הקשר החדש הזה למעשה "מאפס" את מודול L, ופותח שלב חדש של חישוב לקראת שיווי משקל מקומי אחר.

כפי שמודגם בניתוח במאמר של forward residuals (מדד לפעילות חישובית), תהליך זה מאפשר לפעילות של מודול L לזנק שוב ושוב, בעוד מודול H מתכנס ביציבות לעבר הפתרון. מבנה חישובי זה מאפשר למודל לבצע רצף של חישובים נפרדים, יציבים ועמוקים, תוך הימנעות מהתשישות המוקדמת של מודלים רקורסיביים סטנדרטיים.

אימון חכם יותר, לא קשה יותר: עקיפת Backpropagation-Through-Time או BPTT

אימון RNNs תמיד היה כאב ראש בשל עלויות הזיכרון והחישוב של BPTT. לעומת זאת HRM מציג שיטת אימון יעילה יותר, ומתקבלת יותר על הדעת מבחינה ביולוגית, המבוססת על קירוב גרדיאנט בצעד אחד.

גישה זו, המבוססת על התיאוריה של מודלי שיווי משקל עמוקים (DEQ), עוקפת את הצורך לפרוס את כל היסטוריית החישובים. היא מחשבת את הגרדיאנטים הנחוצים באמצעות המצב הסופי של כל מודול בלבד, ומתייחסת למצבי הביניים כאל קבועים. קיצור דרך חכם זה שומר על צריכת זיכרון קבועה עבור backprop ללא קשר למספר הצעדים הרקורסיביים שהמודל מבצע. יעילות זו מועצמת עוד יותר על ידי מנגנון "השגחה עמוקה" (Deep Supervision), שבו המודל מקבל משוב מתקן לאחר כל מעבר קדמי מלא (או "סגמנט"), מה שמייצב את האימון ומשמש כצורה חזקה של רגולריזציה.

חשיבה לפי דרישה: זמן חישוב מסתגל (ACT)

לא כל הבעיות דורשות את אותה כמות מחשבה. בהשראת יכולתו של המוח לעבור בין חשיבה מהירה ואוטומטית ("מערכת 1") לבין חשיבה איטית ושקולה ("מערכת 2"), HRM משלב מנגנון של זמן חישוב מסתגל (Adaptive Computational Time - ACT).

באמצעות אלגוריתם Q-learning, המודל לומד מדיניות להחליט אם "לעצור" ולהפיק תשובה או "להמשיך" ולבצע סגמנט נוסף של חישוב. הדבר מאפשר ל-HRM להקצות באופן דינמי את התקציב החישובי שלו, "לחשוב" יותר על בעיות קשות יותר תוך פתרון מהיר של בעיות קלות. התוצאה היא מערכת שמשיגה כמעט את אותם ביצועים כמו מודל עם מספר קבוע וגדול של צעדים חישוביים, אך ביעילות גדולה משמעותית.

החותם המתהווה של אינטליגנציה: היררכיה של ממדיות

אולי הממצא העמוק ביותר במאמר אינו רק ש-HRM עובד, אלא כיצד הוא מארגן את עצמו. החוקרים ניתחו את "הממדיות האפקטיבית" של הייצוגים בכל מודול באמצעות מדד הנקרא יחס השתתפות (Participation Ratio - PR). יחס PR גבוה יותר פירושו שייצוג הוא מורכב יותר ומפוזר על פני יותר ממדים.

התוצאות די חזקות:

לאחר האימון, מודול H לומד באופן אוטונומי לפעול במרחב בממד גבוה משמעותית יותר מאשר מודול L הנמוך.

היררכיה זו משקפת את מה שמדעני מוח צופים בקורטקס של יונקים, שם אזורים קוגניטיביים מסדר גבוה מציגים פעילות עצבית בממדיות גבוהה יותר כדי לתמוך במשימות גמישות ותלויות-הקשר.

מבנה זה אינו קיים ברשת לא מאומנת; זוהי תכונה נלמדת שמתהווה ככל שהמודל רוכש מיומנות בחשיבה מורכבת.

ממצא זה מצביע על כך ש-HRM לא רק אומן לפתור משימה; הוא גילה עיקרון ארגוני בסיסי לחישוב חזק וגמיש. הוא לומד לחלק את מרחב העבודה הפנימי שלו למרחב מופשט בעל קיבולת גבוהה לתכנון, ולמרחב ייעודי יותר, בממדיות נמוכה, לביצוע.

ומה השורה התחתונה? קריאת כיוון חדשה לביצועים (שדי מפתיעים)

החידושים הארכיטקטוניים והאימוניים של HRM מתורגמים לביצועים יוצאי דופן באמת. עם 27 מיליון פרמטרים בלבד, ולאחר אימון על כ-1,000 דוגמאות בלבד לכל משימה (ללא אימון-מוקדם), HRM משיג תוצאות המאפילות על מודלים גדולים ותאבי-נתונים בהרבה:

במבחן Abstraction and Reasoning Corpus (ARC-AGI), מבחן מפתח לאינטליגנציה פלואידית, HRM מתעלה על מודלים מובילים מבוססי CoT כמו Claude 3.7 ו-03-mini-high.

בחידות סודוקו קשות במיוחד ובמשימות מציאת נתיב במבוכים בגודל 30x30 – בעיות הדורשות חיפוש נרחב וחזרה לאחור (backtracking) – HRM משיג דיוק כמעט מושלם, בעוד שמודלי LLMs מתקדמים המשתמשים ב-CoT נכשלים לחלוטין.

תוצאות אלו מאתגרות לכאורה את המנטרה של "גודל זה כל מה שצריך". הן מראות שהארכיטקטורה הנכונה, כזו עם עומק חישובי מספיק והטיות אינדוקטיביות בהשראת המוח – יכולה להיות יעילה ורבת עוצמה בסדרי גודל עבור חשיבה מורכבת.

כמובן, נותרו שאלות פתוחות. עד כמה הארכיטקטורה הזו יכולה לגדול? האם ניתן לשלב את מנוע החשיבה השקט והעוצמתי שלה עם ידע העולם העשיר והשטף הלשוני של מודלי LLM? המחברים מבהירים שעבודתם היא צעד לקראת מסגרת יסוד לחישוב אוניברסלי, ולא המילה האחרונה.

HRM הוא תזכורת לכך שההשראה לדור הבא של AI עשויה שלא להגיע מהוספת עוד טריליון פרמטרים, אלא מהתבוננות בעקרונות החישוביים האלגנטיים והיעילים של מכונת החשיבה המוכחת היחידה שאנו מכירים: המוח האנושי. זהו מסע שיתופי, והמאמר הזה מספק מפה חדשה, מרתקת ומבטיחה.

https://arxiv.org/abs/2506.21734