תרגיל 1 - קורס מתקדם בעיבוד שפה טבעית

אדיר ברק 207009739

12 במאי 2025

גיטהאב התרגיל: https://github.com/adir-barak/anlp-ex1

חלק 1 - שאלות פתוחות

שאלה 1

1. NarrativeQA: דאטהסט של שאלות המבוססות על ספרים ותסריטים שלמים (של סרטים).

- מודל שמפגין ביצועים גבוהים יעיד על היכולת שלו להבין מסמכים ארוכים במיוחד, ולהסיק מידע מהקשר עצום.

2. DocVQA: דאטהסט של תמונות של מסמכים סרוקים.

- מודל שמפגין ביצועים גבוהים יעיד על היכולת שלו לפרסר ולהבין מבנה של מסמכים, ולעבד יחד מידע טקסטואלי, ו-ויזואלי.

3. ReCoRD-(Reading-Comprehension-with-Commonsense-Reasoning-Dataset): דאטהסט של טקסטים עם חלקים מוסתרים שעל המודל להסיק ולהשלים ע"פ ההקשר.

- מודל שמפגין ביצועים גבוהים יעיד על היכולת שלו להשתמש 'בידע הכללי' שלו, על 'היגיון בריא' ומעקב אחר ישויות.

שאלה a2

Chain-of-Thought-Reasoning:

- תיאור: המודל בונה שרשרת של צעדי חשיבה לקראת הפתרון. הוא יכול לתכנן (Planning) מראש את הצעדים, לזהות טעויות ולחזור לאחור (Backtracking), ולבצע הערכה עצמית (Self-Evaluation) של התשובות כדי לשפר את האיכות.
- יתרונות: מאפשר התמודדות עם בעיות מורכבות ורב-שלביות, ומשפר את הדיוק על ידי בדיקה עצמית ותיקון שגיאות.
- צוואר בקבוק חישובי: החישוב אורך זמן רב בגלל תכנון, ניסוי וטעייה, וביצוע הערכה על תוצאות רבות. אורך 'השרשרת' הינו פקטור משמעותי בהקשרי העלות.
- האם ניתן למקבל את התהליך?: אולי חלקית. ניתן לחשב רישאות של שרשראות שונות במקביל, ולבחור באילו להמשיך.

Self-Consistency:

- תיאור: יצירת מספר 'נתיבי' Chain-of-Thought או היסק רגיל (עם שימוש ב-temprature או רנדומיזציה אחרת כלשהי), ובחירת התשובה שניתנת על ידי הרוב.
- יתרונות: שיפור הדיוק והאמינות של התשובה הסופית, במיוחד במשימות הדורשות מספר שלבים לפתרון.
- צוואר בקבוק חישובי: דורש יצירת מספר רב של תשובות, ועל כן מייקר משמעותית את תהליך וזמן ה-inference.
- האם ניתן למקבל את התהליך?: כן. ניתן לחשב את כל השרשראות במקביל.

Verifiers:

- תיאור: אימות אוטומטי של תשובות בעזרת כללים או מודלים נפרדים (לדוג' regex,unit-tests,etc').
- יתרונות: מאפשר סינון תשובות לא נכונות, שיפור הביצועים והאמינות מבלי להגדיל את המודל באופן ישיר.
- צוואר בקבוק חישובי: דורש הפעלה של מודלים או כלים נוספים על כל תשובה שנוצרת.
- האם ניתן למקבל את התהליך?: כן. אפשר לבדוק מספר תשובות במקביל (ושוב, גם Best-of-N-/-Rejection-Sampling).

שאלה b2

הייתי בוחר ב-CoT-Reasoning : השיטה מאפשרת חשיבה מסודרת ואפקטיבית (תכנון, backtracking והערכה עצמית) מה שמשפר משמעותית את יכולות המודל לפתור משימות מדעיות מורכבות הדורשות צעדים לוגיים וביקורתיות.
עם GPU יחיד ומספיק זיכרון, 'נשלם' בזמן חישוב ארוך יותר בתמורה לשיפור משמעותי ואיכותי בפתרון.

- Self-Consistency : ע"פ אופי המשימה המתוארת, נשמע שהשקעה גדולה בזמן החישוב (הרבה נתיבי חישוב) לא בהכרח תניב תוצאה מספקת - סביר להניח שרוב ההיסקים לא יהיו מתוחכמים מספיק על מנת לבצע את המשימה באופן הדרוש.
- Verifiers : מחייבים פיתוח או שימוש בכללי אימות מותאמים. לא בהכרח קיימים כאלו לבעיה המתוארת, וגם אם כן, בדר"כ נצטרך לפתח אותם או לכל הפחות לעשות אדפטציה לכלים קיימים.

חלק 2 - קוד וניתוח

שאלה 2

- מבין 3 הקונפיגורציות שבדקתי, הקונפיגורציה שהשיגה את ה-eval_acc הגבוה ביותר באימון (0.85874) הייתה גם זו שהשיגה את ה-acc הכי גבוה ב-test (0.82550). הייתה זו "epoch_num_3_lr_0.00015_batch_size_32".

Configuration	eval_acc	test_acc
epoch_num_2_lr_0.00015_batch_size_32	0.85294	0.82376
epoch_num_3_lr_0.00015_batch_size_32	**0.85874**	**0.82550**
epoch_num_4_lr_0.0001_batch_size_32	0.85539	0.82318

- סקירה של הדוגמאות בהן המודל הטוב ביותר צדק והמודל הגרוע ביותר טעה (comparison.txt), מעלה דפוסים אפשריים:

  - המודל הגרוע ביותר הסתמך יותר מדי על דמיון מילולי ולא תפס הבדלים קלים ששומרים על המשמעות. למשל:

    ```
    ------------------
    Validation Sample 29
    Sentence 1: Ricky Clemons ' brief , troubled Missouri basketball career is over .
    Sentence 2: Missouri kicked Ricky Clemons off its team , ending his troubled career there .
    True Label: 1
    ------------------
    ```

  - המודל הגרוע ביותר התקשה לבצע Coreference/Entity-Resolution באופן מספק. למשל:

    ```
    ------------------
    Validation Sample 91
    Sentence 1: They were at Raffles Hospital over the weekend for further evaluation .
    Sentence 2: They underwent more tests over the weekend , and are now warded at Raffles Hospital .
    True Label: 1
    ------------------
    ```

  - המודל הגרוע ביותר לא הצליח לעקוב אחרי פרטים קטנים אך משמעותיים המשנים למעשה את המשמעות. למשל:

    ```
    ------------------
    Validation Sample 229
    Sentence 1: Klarman was arrested by FBI agents in the Hamptons , an exclusive summer resort enclave east of New York City .
    Sentence 2: Klarman was arrested by FBI agents Monday morning at his home in New York .
    True Label: 0
    ------------------
    ```

    ```
    ------------------
    Validation Sample 133
    Sentence 1: By Sunday night , the fires had blackened 277,000 acres , hundreds of miles apart .
    Sentence 2: Major fires had burned 264,000 acres by early last night .
    True Label: 0
    ------------------
    ```

  - המודל הגרוע ביותר הראה יכולות פחות טובות במה שקשור ל-'ידע כללי'. למשל:

    ```
    ------------------
    Validation Sample 212
    Sentence 1: The driver of the truck escaped and is now being sought by the police , Supoyo said .
    Sentence 2: But police say the driver of the truck has not been found and is wanted for questioning .
    True Label: 1
    ------------------
    ```