תרגיל 1 - קורס מתקדם בעיבוד שפה טבעית

אדיר ברק 207009739

4 במאי 2025

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

חלק 1 - שאלות פתוחות

שאלה 1

ראינו בהרצאה

1. SNLI-(Stanford-Natural-Language-Inference) : זוגות של משפטי מוצא והשערות אפשריות (Premise ו-Hypothesis). המשימה היא לקבוע אם ההשערה נובעת, סותרת או נייטרלית ביחס למוצא.

- הצלחה במשימה מעידה על היכולת של המודל להבין את הקשרים הסמנטיים בין טקסטים שונים, (הסקה, סתירה או נייטרליות) מה שמצביעה על הבנה בסיסית של קשרים לוגיים (אף סמויים, הנשענים על ידע כללי) בין משפטים.

2. מאגרי מידע של Coreference (כמו WSC) : משפטים בהם יש להבין את הייחוס של מילים או כינויים (כמו "הוא", "היא") ולזהות לאיזה ישות הם מתייחסים. לדוגמה, במשפט "הילד חובש כובע, והוא נראה טוב", השאלה היא האם "הוא" מתייחס לילד או לכובע.

- הצלחה במשימה מעידה על היכולת של המודל להבין הקשרים סמנטיים של ייחוס בין מילים ויישויות בטקסט, תהליך שדורש הבנה עמוקה של קשרים בין אובייקטים והיכולת לקשר בין כינויים והיישויות שאליהם הם מתייחסים.

3. מאגרי מידע של Entity-Linking (למשל Wikidata) : טקסטים עם מילים או כינויים שיש למפות אותם לישויות כמו אנשים, מקומות או אירועים. המשימה היא לזהות את הישות הנכונה מתוך מאגר גדול ולהבין לאיזה יישות מתייחסות המילים בטקסט.

- הצלחה במשימה מעידה על היכולת של המודל לזהות ולקשר בין ישויות בטקסטים בצורה מדויקת, ומסייעת לו להבין קשרים סמנטיים בין ישויות שונות. בנוסף על המודל לבצע ייחוס בין מידע טקסטואלי ומאגרי ידע "ולהוכיח" הבנה מעמיקה של הטקסט, הקשרים והיחסים הסמנטיים בו.

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 (0.87255) הייתה גם זו שהשיגה את ה-acc הכי גבוה ב-test (0.84638). הייתה זו "epoch_num_4_lr_0.0001_batch_size_64".

Configuration	eval_acc	test_acc
epoch_num_2_lr_0.0002_batch_size_64	0.85539	0.83130
epoch_num_3_lr_0.00015_batch_size_64	0.84804	0.83536
epoch_num_4_lr_0.0001_batch_size_64	**0.87255**	**0.84638**

- 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 - משפטים כמעט זהים מבחינה מילולית אך שונים בכמה מילות מפתח או מכילים תוספות קטנות שמשנות את המשמעות - למשל לא הצליח להבדיל בין "משפט_א_ארוך" לבין "משפט_א_ארוך, אבל משפט_ב_קצר".

  – המודל הגרוע לא הסתמך מספיק על הפרטים ולעיתים הסתפק רק במשמעות הכללית - למשל לא הצליח לדייק ולהבדיל בין "עליית מחירים ל-$50" אל מול "עליית מחירים ל-$18".