🚀המאמר היומי של מייק 29.09.24: ⚡️🚀  
ASR Error Correction using Large Language Models

ממשיך לסקור מאמרים בדומיין אודיו. הפעם נדבר על מאמר המציע שיטה לשיפור איכות של פענוח אות דיבור ניתן להשתמש בה במערכות ל-Automatic Speech Recognition או בקצרה ASR. המטרה בכל הסיפור הזה היא לתמלל אות קולי או במילים פשוטות להבין מה נאמר שם.

בד"כ הקלט ל- ASR הוא כמה פלטים של המודול שנקרא Error Correction או EC שמטרתו היא ליצור כמה וריאנטים של תמלול Z (בעלי "סבירות גבוהה ביותר") עבור אות דיבור נתון. למעשה מטרתו של ה- EC היא לבנות את התמלול הסופי בהינתן Z.

בעידננו של מודלי שפה עוצמתיים ניתן למנף את יכולתם למשימה הזו בצורה די ישרה. כלומר אנו מזינים ל-LLM את הוריאנטים השונים של התמלול ומבקשים מ-LLM לבחור את התמלול הגיוני ביותר מבחינה סמנטית (עם פרומפט מתאים). המאמר בחר LLM לא סטנדרטי המורכב מאנקודר ומדקודר (כמו במאמר המקורי של הטרנספורמרים) למשימה זו וזה עבד לא רע. אם יש לנו דאטהסט המכיל את התמלולים מה-ASR והתמלול הנכון, ניתן לבצע פיינטיון.

האם ניתן לעשות יותר טוב? מתברר שכן אם בנוסף לתמלולים אנו מזינים למודל שפה גם את תכונות אות הדיבור עצמו (למשל ייצוגו אחרי האנקודר או מטה-דאטה שלו) ניתן לשפר את הביצועים של ה-EC. המחברים מציעים לבנות את התוצאה באמצעות מקסום של סכום משוקלל של הנראויות (log-likelihoods) מהסעיף הקודם (בהינתן התמלולים מהסעיף הקודם) והנראות של התמלול בהינתן התכונת של סיגנל הדיבור עצמו. באופן לא מפתיע זה משפר את הביצועים כי המודל מקבל יותר מידע רלוונטי.

עוד שכלול אחד הוא תוספת ההתחשבות במרחק Levenstein מינימלי בין הפלט הסופי של EC לבין הפלטים של ASR (המוזנים ל-EC). מרחק לבּינשטיין הוא מדד הבודק את מספר השינויים המינימלי הנדרש כדי להפוך מחרוזת אחת לאחרת. כלומר אנו בוחרים את התיקון הקרוב ביותר (מבחינת LD) לאחד הפלטים של ה-ASR.

מקווה שלא פספסתי שום דבר…

arxiv.org/pdf/2409.09554