הסבר כללי על טכניקת LoRA ותאימותה לפרוייקט

LoRA הינה דרך חדשנית לעשות Fine tuning למודלים גדולים (נועד בשביל מודלי שפה אך יכול להיות גם עבור אחרים), כיוון דק עד היום של מודלי שפה גדולים בשביל שיהיו טובים למשימות ספציפיות שהם לא אומנו עבורם מראש דורש כיוונון ושינוי של כמות פרמטרים אסטרונומית, וזה דורש משאבים מאוד גדולים (בין אם זה כוח עיבוד ובין אם זה מאגר מידע גדול מאוד) לכן הגיעו אל LoRA על מנת לפתור את הבעיה הזאת.

הרעיון של LoRA – Low Rank Adaptation זה להשתמש ב2 מטריצות בעלות דירוג נמוך (Low Rank) של פרמטרים (כלומר במקום כלל הפרמטרים של המודל, ייבחרו רק חלק מהם, לרוב חלק קטן אף מאוד) שמתארות את כלל ההסתגלות שצריך (adaptation) בשביל המשימה החדשה שעבורה אנחנו מכוונים. בזמן האימון אך ורק המטריצות האלו משתנות ומחושבות מחדש ולא כלל הפרמטרים של המודל שנשארים ללא שינוי (מה שגורם להורדה משמעותית בכוח חישוב שצריך וגם בגודל מאגר המידע הנדרש). בגדול, הרעיון של LoRA הגיע מאלגברה לינארית שאומר שאם במטריצה יש עמודות או שורות שתלויות אחת בשנייה (ניתן להגיע אליה מחיבור כפל של אחת אחרת וכדומה) אז לא ניקח את התלויה אלא רק את זאת שמגיעה אליה כלומר ה2 מטריצות של LoRA יהוו אך ורק את העמודות/שורות שלא תלויות בשום עמודה/שורה אחרת בהתאמה.

לסיכום, טכניקת LoRA יכולה מאוד להתאים לפרוייקט הזה מכיוון שהוא מיועד להשתמש בלפחות מודל שפה גדול אחד, והמאגר מידע שלנו לא גדול מספיק וכוח החישוב שלנו (ככול הנראה) לא יהיה חזק מספיק על מנת לעשות למודל שפה גדול כיוון דק מסורתי, ולכן LoRA פותרת לנו את שני הבעיות גם של כוח החישוב וגם של גודל מאגר המידע הדרוש.