

הקדמה

כחלק מעבודת התזה שלי התעמקתי לאחרונה בארכיטקטורת הטרנספורמרים. כשקראתי את המאמרים הראשוניים בתחום (Attention is all You Need, ViT) התקשיתי לעבד את הרעיונות הקונקרטיים שעליהם מבוססים הטרנספורמרים לרעיונות מופשטים. כתוצאה מכך צללתי לנושא, ניתחתי קוד ממספר מקורות באינטרנט לצד קריאת המאמרים, ולבסוף ניהלתי עם ChatGPT סיבוב שאלות-תשובות כדי לחדד את ההבנה. לאחר שבוע וחצי של מחקר מאומץ, הגעתי למסקנה שאני רוצה לשתף את הידע שצברתי עם העולם. ולכן, החלטתי לכתוב סדרת פוסטים שאוכל להביע דרכה את ההבנה שלי, ואולי לקצר תהליך למי שאין לו הזמן שהיה לי להשקיע בלמידה הזו. כחלק מהכתיבה, אסקור את ההיסטוריה של התחום, מבנה הרשת, יתרונות וחסרונות, וכיצד מאמנים אותה. בשלב השני אסקור מאמרים בתחום הראייה הממוחשבת המציגים את השימוש ברשת למשימות כגון סגמנטציה, הדבקת תמונות (Image Matting), העתקת מנח גוף (Gait Transfer), סיווג וכדומה. בכל פוסט אוסיף קישור לקובץ מתעדכן המכיל את כל המידע המתקצר בפוסטים.

פרט אחרון, במשך זמן מה התלבטתי האם לכתוב את התוכן הזה בעברית או באנגלית. לבסוף בחרתי בעברית מכיוון שישנו המון חומר בנושא באנגלית, ומי שירצה יכול בחיפוש קצר בגוגל למצוא אותו. לעומת זאת בעברית אני מצליח להביא את הקול הייחודי שלי. מאחל לכם קריאה מהנה, ולי הצלחה.

סקירה היסטורית

טרנספורמים הוצעו בהתחלה כפתרון לבעיית ניתוח טקסט. המאמר שבו הטרנספורמרים הופיעו לראשונה נקרא Attention is all You Need, הכותבים השתמשו בארכיטקטורה זו למשימת תרגום מאנגלית לצרפתית. מאמר זה למעשה פתח את הסכר למהפכת ה-NLP שאנחנו רואים היום. טרנספורמרים כללו מספר רעיונות חדשים בתחום ניתוח השפה. השניים המרכזיים, שזורים אחד בשני בארכיטקטורה, ומהווים את אבני הבניין הקונספטואלים של הרשת (בנוסף לעוד חידושים שהמאמר הציג). הראשון היה עיבוד מידע מקבילי, שהוביל ליעילות חישובית באימון המודל ביחס למודלים קודמים (RNN\LSTM\GRU), ואפשר בפעם הראשונה לפרוץ את מחסום הלמידה התלויה בזמן של קלט סדרתי. כלומר, ניתן ללמוד במקביל תלויות קצרות וארוכות טווח בקלט סדרתי. נציין שגם טרנספורמרים מוגבלים ביכולתם לעבד קלט מקבילי, אולם זוהי מגבלה התלויה במשאבי חישוב כדוגמת זיכרון זמין ויחידות עיבוד (כרגע מספר הטוקנים המקסימלי הוא

בנוסף, הרשת הציגה שני מנגנוני תשומת הלב (attention). הראשון הוא תשומת לב-עצמית (self-attention) שאפשרה למודל להתמקד במידע החשוב ביותר באופן סלקטיבי. המנגנון השני הינו תשומת לב מוצלבת (cross-attention) שאפשרה מידול תלויות וקשרים בין חלקי הקלט והפלט השונים. תכונות אלו הכרחיות במשימות כדוגמת תרגום ומענה על שאלות סיכום טקסט, שדורשות בחירה מודעת בחלקים החשובים ביותר של הקלט והפלט. רעיונות אלו הלהיבו אותי, וגרמו לי לרצות להבין מה הוביל להתפתחותם. על מנת לעשות זאת התחלתי לסקור את הארכיטקטורות שקדמו לטרנספורמים, כיצד הם עבדו, ומדוע לא צלחו במשימה שהטרנספורמים כן הצליחו בה.