

הקדמה

להלן מבוזת התזה שלי התעמקתי לאחרונה ארכיטקטורת הטרנספורמרים. כשקרהתי את המאמרים הראשונים בתחום (Attention is all You Need, ViT) הת קישיתי לעבד את הרעיון ה konkretiyim שליהם מושגים הטרנספורמרים לרעיונות מופשטים. כתוצאה לכך צלתי לנושא, ניתחתי קוד מספר מקורות באינטרנט לצד קריית המאמרים, ולבסוף ניהلت עם ChatGPT סיבוב שאלות-תשובות כדי לחזק את הבנה. לאחר שבוע וחצי של מחקר מאומץ, הגעת למסקנה שאין רציה לשטף את הידע שצברתי עם העולם. ولكن, החלטתי לכתוב סדרת פוסטים שאוכל להביע דרכה את ההבנה שלי, ואולי לפחות תחילך למי שאין לו זמן שהיה לי להשקיע בלמידה זו. חלק מהכתיבה, אסקור את ההיסטוריה של התחום, מבנה הרשות, יתרונות וחסרונות, כיצד מאמנים אותה. בשלב השני אסקור מאמרים בתחום הראייה הממוחשבת המציגים את השימוש ברשות למשימות כגון סגמנטציה, הדבקת תמונות (Image Matting), העתקת מנה גוף (Gait Transfer), סיוג וכדומה. בכל פוסט אוסיף קישור לקובץ מתעדכן המכיל את כל המידע המתווך בפוסטים.

פרט אחרון, במשך זמן מה התלבטתי האם לכתוב את התוכן זהה בעברית או אנגלית. לבסוף בחרתי בעברית מכיוון שישנו המון חומר בנושא אנגלית,ומי שירצה יכול בחיפוש קצר ב.google למצוא אותו. לעומת זאת בעברית אני מצליח להביא את הקול הייחודי שלו. מחלל לכם קריאה מהנה, ולucky.

קירה היסטורית

טרנספורמרים הוצעו בהתחלה כפתרון לבניית ניתוח טקסטי. המאמר שבו הטרנספורמרים הופיעו לראשונה נקרא *Attention is all You Need*, הכותבים השתמשו ארכיטקטורה זו למשימת תרגום מאנגלית לצרפתית. מאמר זה למעשה פתח את הסבר למהפכת ה-NLP שאנו רואים היום. טרנספורמרים כללו מספר רעיונות חדשים בתחום ניתוח השפה. השניים המרכזיים, שזרים אחד בשני ארכיטקטורה, ומהווים את אבני הבניין הkonceptualים של הרשות (בנוסף לפחות חידושים שהמאמר הציג). הראשון היה עיבוד מידע מקבילי, שהוביל ליעילות חישובית באימון המודל ביחס למודלים קודמים (GRU&LSTM&NN), ואפשר בכך הרשונה לפזר את מחסום הלמידה התלויה בזמן של קלט סדרתי. כלומר, ניתן ללמידה במקביל תלויות קצחות וארוכות טווח בקלט סדרתי. נציין שגם טרנספורמרים מוגבלים ביכולתם לעבד קלט מקבילי, אולם זהי מוגבלת התלויה במשאבי חישוב כדוגמת זיכרון זמני ויחידות עיבוד (כרגע מספר הטוקנים המקסימלי הוא

בנוסף, הרשות הצינה שני מנגנון 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 חלקו הקלט והפלט השונים. תוכנות אלו הכרחיות במשימות כדוגמת תרגום ומענה על שאלות סיקום טקסט, שדורשות בחירה מודעת בחלקים החשובים ביותר של הקלט והפלט. רעיון זה הוביל אותנו לירצחות להבין מה הוביל להתקפותיהם. על מנת לעשות זאת תחילה לסקור את הארכיטקטורות שקדמו לטרנספורמים, כיצד הם עבדו, ומדוע לא צלחו במשימה שהטרנספורמים כן הצליחו בה.