# Review 5: Single Headed Attention RNN: Stop Thinking With Your Head

**Paper: https://arxiv.org/abs/1911.11423v2**

תחום מאמר: טרנספורמרים, multi-head attention

תקציר בשתי שורות: המאמר מצהיר שהוא מציע ארכיטקטורה פשוטה בהרבה ותופסת משמעותית פחות זיכרון מהטרנספורמר בעל ראשים מרובים המהווה חלק מהותי (כמעט כל מה שיש שם) ב BERT וצאצאיו.

תקציר מאמר: מכיוון שרובכם יודעים היטב מה זה טרנספורמר ו- Attention מודול בעל ראשים מרובים , לא ארחיב על זה כאן. מחבר המאמר מתרעם רבות על כך שהטרנספורמר בעל ראשים מרובים זה ארכיטקטורה מאוד כבדה, קשה לאימון ודורשת הרבה משאבי GPU/TPU. הוא מציין שדי לא ברור למה צריך ראשים מרובים וטוען שאפשר להשיג תוצאות דומות גם עם ראש אחד בלבד עם כל מיני תוספות ארכיטקטוניות נחמדות. דרך אגב קיימות עבודות המציעות שיטות לניצול יעיל יותר של הפלט של הראשים הרבים של הטרנספורמר. אז מה הוא בעצם מציע זה להחליף את הראשים הרבים של הטרנספורמר עי״ שילוב של attention מודל מנוון עם ראש אחד בלבד עם מה שנקרא AWD-LSTM, שזה בעצם LSTM עם תוספת של כמה טריקים נחמדים כמו DropConnect, Average SGD, אורך משתנה של סדרת backprop בזמן ועוד כמה. הוא קורא לארכיטקטורה הזו SHA-RNN

אז מה שעושה מחבר המאמר הוא פשוט מוסיף לארכיטקטורה הזו את מודול ה- attention מנוון (תיכף אסביר מה מנוון שם) עם ראש אחד ומודול שנקרא Boom. מודול ה- attention הינו מנוון כי רק האינפוט q מוכפל במטריצת משקלים נלמדים כאשר האינפוטים key ו value נכנסים כמו שהם ללא הכפלה במטריצה. עבור הארכיטקטורה הזו q הינו היציאה של LSTM אחרי שכבת נירמול (layer norm), כאשר k ו v הם הזיכרון של התא lstm (למיטב זכרוני הוא מסומן ב c בדרך כלל) ואותו זכרון מועבר דרך LN (לא ברור לי מה הטעם להעביר למודול הזה שני אינפוטים שהם שונים רק בממוצע ובסקייל). את הפלט מעבירים דרך המודול הנקרא boom שזה בעצם רשת דו שכבתית. בשכבה הראשונה מגדילים את המימד (מטריצה עם הרבה שורות), מעבירים דרך GELU, ואז מצמצמים מימד עיי חלוקה של וקטור לכמה וקטורים קטנים וסכימתם (סוג של average pooling).

הישגי מאמר: המחבר בחר להשוות את bpw שזה bit per word מול מודלים שונים של הטרנספורמרים. הוא הצליח להראות שהוא מצליח להפיק את אותו bpw עם ארכיטקטורה פשוטה יותר.צריך לציין שכמות המשקלים בארכיטקטורה שלו עולה על זה המתחרים (בעלי אותם ביצועים). דרך אגב אחד הווריאנטים של SHA-LSTM זה ארכיטקטורה עם 4 שכבות של LSTM כל אחד עם הראש שלו שזה רחוק מלהיות פשוט. בקיצור לא השתכנעתי.

דאטה סטים: enwik8, wikiText, wikiText103

נ.ב. המאמר לא הצליח להוכיח לי שהארכיטקטורה שהוא מציע עדיפה על הטרנספורמר בעל ראשים מרובים. לא הוצגו תוצאות על בנצ'מארקים ידועים (סוגי משימות שונות). השילוב של lstm עם ראש אחד של attention נחמד אך התועלת לא ברורה. בקיצור המלצת קריאה - רק אם יש לכם הרבה זמן פנוי.