# Review 205: PanGu-π: Enhancing Language Model Architectures via Nonlinearity Compensation

**Paper: https://arxiv.org/abs/2505.21411v2**

https://arxiv.org/abs/2312.17276

היום סוקרים מאמר המציע שדרוג לארכיטקטורת הטרנספורמר. כמו שאתם בטח יודעים בלוק של טרנספורמר מורכב משני החלקים העיקריים (פרט לשכבות נרמול):

מנגנון תשומת לב עצמי בעל ראשים מרובים (multi-head self-attention or MSA)

שכבת (MLP (fully connected המורכבת משכבה לינארית עם פונקציה אקטיבציה לא לינארית ולאחר מהן שכבה לינארית נוספת (ללא פונקציית אקטיבציה)

כמו שאתם זוכרים מטרת בלוק הטרנספורמר היא להפיק ייצוגים תלויי הקשר של טוקני הקלט. כלומר כל ייצוג של כל טוקן לוקח בחשבון את הטוקנים בתוך ההקשר. המחברים מנתחים את התכונות של ייצוגי טוקנים תלוי הקשר הנוצרים על ידי הטרנספורמרים עלי ידי השוואתם עם ייצוגי הטוקנים המוזנים לבלוק הראשון של הטרנספורמר (כלומר ייצוגי הטוקנים ממטריצת embeddings של מודל שפה). השיפורים המוצעים במאמר באים למנוע מצב שבו ייצוגי תלוי ההקשר של טוקנים יהיו דומים מאוד אחד לשני.

תופעה דומה למתוארת בפסקה הקודמת נקראת over-smoothing ברשתות נוירונים גרפיות (GNN). זה קורה יש מספר גבוה מדי של שכבות MSA שמוביל לייצוגים דומים למדי של הקודקוד העלולים לגרום ל״קריסה״ של הייצוגים לתת-מחרב קטן של מרחב הקלט. ד״א מטריצת משקלי ה-attention בטרנספורמרים ניתן לראות בתור מטריצה שכנויות מנורמלת של גרף שלם.

אבל איך נמדוד את מידת שוני (diversity) בין ייצוגי הטוקנים? המאמר מגדיר את שוני של מטריצה M(= קבוצה של וקטורים) בתור מינימום נורמת פרובניוס של ההפרש של M-A מעל כל המטריצות A בעלות רנק 1(כל הוקטורים במטריצה תלויים לינארית).

המחברים מראים כי עבור מודל המורכב מ l בלוקי MSA מוערמים (stacked) בלבד (ללא MLP) השוני של ייצוגי הפלט ניתן לחסום על מכפלה של הערכים הסינגולריים (הכללה של ערכים עצמיים למטריצות לא ריבועיות) המקסימליים של מטריצות משקלים השונות במנגנון MSA ובשוני של ייצוג הקלט (ממטריצת אמבדינגס של מודל השפה). ללא שכבות MLP ייצוגים אלו נוטים להתנוון ולהפוך להיות תלויים לינארית ככל מספר הבלוק l גדל. זו הסיבה ל״הימצאות״ של MLP בטרנספורמרים.

בנוסף עבור המודל המורכבים מבלוק MLP מוערמים המאמר מוכיח כי השוני של ייצוג הפלט הינם מכפלה של שוני ייצוג הקלט, הערכים הסינגולריים המקסימליים של מטריצות המשקלים וקבועי ליפשיץ של פונקציות האקטיבציה של MLP.

במטרה לשפר את תכונות ייצוגי הטוקנים בפלט של הטרנספורמר המאמר מציע שני שדרוגים, אחד ל MSA והשני ל-MLP. זוכרים בבלוק הטרנספורמר יש לנו יש לנו חיבור שארי (residual or shortcut according to the paper) - כלומר הפלט של MSA מחובר לייצוגי הקלט ל-MSA, המחברים מציעים לפתוח חיבורי ״קיצור דרך״ נוספים. כל חיבור כזה הוא למעשה שכבה לינארית עם מטריצה נלמדת ופונקצית אקטיבציה לא לינארית. כדי לא להכביד מדי על העומס החישובי המתווסף בעקבות כך(מטריצות המשקלים בחיבורי קיצור דרך אלו יכולות להיות 4096x4096 וזה די הרבה עם רוצה להשתמש בכמה חיבורים כאלו) משתמשים במטריצות בעלי רנק נמוך. המחברים מוכיחים שהוספה של שכבות לבלוקי הטרנספורמרים המקוריים אלו תורם להקטנת הפגיעה בשוני של ייצוגי פלט.

בנוסף המאמר מציע לשדרג פונקציה אקטיבציה שזה החלק המהותי של מנגנון הטרנספורמרים בנוסף ל-MSA. במקום להשתמש בפונקציה אקטיבציה רגיל (כמו סיגמאויד או ReLU) המאמר מציע לשלב (חיבורית) n פונקציות אקטיבציה בצורה הבא:

כאשר a\_i ו- b\_i הם פרמטרים נלמדים. כמובן שיש הוכחה שהחלפה כזו תורמת להגדלת השוני בין ייצוגי הפלט.

בנוסף השיפצורים המוצעים נבדקו על מספר בנצ'מארקים והראה ביצועים לא רעים.