# Review 140: [Short] One Wide Feedforward is All You Need, 07.09.2023 https://huggingface.co/papers/2309.01826

**Paper: https://arxiv.org/abs/2309.01826v2**

ארכיטקטורת הטרנספורמרים היא המלכה הבלתי מעורערת של עולם AI. רוב המודלים כמו מודלי שפה או מודלי דיפוזיה גנרטיביים המככבים היום בחדשות AI בנויים על הארכיטקטורה הזו. כמובן שיש לא מעט מחקר גם באקדמיה וגם בתעשיה על שיפור ביצועי הטרנספורמרים.   
  
אז היום ב-shorthebrewpapereviews נסקור מאמר שמנסה לשפר שני ההיבטים של הטרנספורמרים: נפח האחסון וכמות חישובים (בכיוון הקטנתם). קודם כל ניזכר שכל בלוק של טרנספורמר (שהוא גרעין של כל מודל המבוסס על הטרנספורמרים) בנוי ממנגנון של תשומת הלב (attention) ועוד שתי שכבות של fully-connected שאחת מהן עם ReLU והשנייה לינארית.   
  
לפי המאמר השכבות האלו מהוות 2/3 ממספר המשקלים (ב-BERT) וכמובן ״תורמים״ לעומס החישובי. המחברים שאולים מה יקרה עם נוותר על השכבות האלו או שנעשה אותם ״שיתופיים״ (shared) בין כל בלוקי הטרנספורמים של המודל. זה עתיד להקטין את כמות המשקלים במודל באופן משמעותי כי רוב המודלים מכילים עשרות רבות או מאות בלוקי הטרנספורמרים.   
  
המאמר גם מציע ״לשתף״ משקלים בין האנקודר לדקודר. אבל איך זה משפיע על ביצועים. המאמר מראה שהפגיעה בביצועים לא גדולה במיוחד (למרות שהם ביצעו מספר בדיקות די מצומצם והם בדקו זאת על מודלים די קטנים עם 6 בלוקי טרנספורמרים בלבד). הם גם השווה דמיון בין הייצוגים של משפטים עבור המודל המקורי והמודל ״הקל״ המוצע וגילו שהוא די גבוה. בנוסף הם השווה k משפטים הדומים ביותר מבחינת הייצוג לשני המודלים וגילו גם כאן דמיון רב. נראה מבטיח אך נדרשות בדיקות מקיפות יותר על מודלים רציניים יותר.