# Review 168: [Short] VERA: VECTOR-BASED RANDOM MATRIX ADAPTATION

**Paper: https://arxiv.org/abs/2310.11454v2**

https://huggingface.co/papers/2310.11454

כבר סקרנו השבוע מאמר שמציע שיפור ל-LoRA שיטת טיוב(finetune) מודלי שפה חסכונית מבחינת משאבי חישוב הנדרשים. היום ב-#shorthebrewpapereviews נסקור שיפור נוסף ל-LoRA המאפשר להקטין את כמות המשאבים הנדרשים לטיוב עוד יותר.

הסקירה של היום הולכת להיות קצרה וקלילה. אז הקטע ב-LoRA המקורי היה לא לכייל את כל הפרמטרים של מודל שפה (אחרי אימון מקדים) אלא לאמן תוספת לפרמטרים של המודל. כלומר לוקחים את כל הפרמטרים של המודל אחרי המאומן ומאמנים תוספת אליהם שהיא מוגדרת בצורה A\*B כאשר A ו- B הם מטריצות בעלות דרגה (ראנק) נמוך (קטנות יותר).

לאחר שמאמנים מודל שפה (מוצאים ערכים אופטימליים של A ו-B) על דאטהסט ולידציה ומבצעים קווינטוט של המטריצה המקורית ושל התוספת. אז המאמר המסוקר מציע להקטין עוד יותר את מספר הפרמטרים במטריצת התוספת ולהציג אותה כמכפלה של bAdB כאשר מטריצות A ו-B הן קבועות לכל השכבות(נגדמות מהתפלגות נורמלית) וקטורים (לא מטריצות!) b ו-d נלמדות פר שכבה. כך מספר הפרמטרים המנלמדים יורד בצורה משמעות בלי לפגוע בביצועי המודל. בקיצור מודיפיקציה נחמדה של LoRA.