# Review 165: [Short] LoftQ: LoRA-Fine-Tuning-Aware Quantization for Large Language Models

**Paper: https://arxiv.org/abs/2310.08659v4**

https://huggingface.co/papers/2310.08659

כולכם מכירים את LoRA (Low Rank Adaptation) – שיטה מאוד פופולרית לטיוב מודל שפה. יצאו כבר כמה מאמרים שמשכללים את השיטה הזו והיום ב-#shorthebrewpapereviews נסקור את אחד השכלולים האלו. קודם כל נרענן מה זה LoRA.

כאמור LoRA היא שיטה לטיוב(fine-tuning) מודלי שפה שבמקום לאפטם את המשקלים של המודל על דאטהסט נתון מנסה למצוא את התוספת למטריצת המשקלים (שמכילה את כל משקלי המודל W אחרי אימון מקדים) שממזערת את הלוס על דאטהסט זה. מטריצה תוספת זו היא מטריצה low-rank שניתן לתאר אותה כמכפלה של שתי מטריצות מרנק נמוך A ו- B (מלבניות וקטנות יחסית).

כך מספר המשקלים הנלמדים במטריצת התוספת הזו נשמר יחסית נמוך ויותר קל לאמן אותו. בסוף מקוונטטים את המטריצה שיצא אחר הפיין-טיון(FT): ניתן לתאר קווינטטי על ידי מכפלה של סכום של W ו-AB במטריצת קווינטוט X שניתן לחשבה בקלות. המאמר מציע שני חידושים:

מתחילים את FT עם מטריצות Q, A ו-B כשכל אחת מהם מטריצה מקוונטטת (8-ביט, למשל) כאשר A ו- B הן מטריצות בעלות רנק נמוך. מטריצות אלו מאותחלות כך שנורמת פרובניוס (שורש מסכום הריבועים של מטריצה) של W-Q-AB יהיה מינימלי.

מחשבים את W-Q ואז מוצאים מטריצות A ו-B על ידי הפעלת טרנספורמציית SVD של W-Q.

חוזרים ל-2 T מספר איטרציות נתון T

מאוד פשוט ואלגנטי וגם הביצועים לא רעים