# Review 156: Short : LONGLORA: EFFICIENT FINE-TUNING OF LONG CONTEXT LARGE LANGUAGE MODELS, 28.09.2023

**Paper: https://arxiv.org/abs/2309.12307v3**

https://arxiv.org/abs/2309.12307.pdf

כל מי שעוסק במודלי שפה בטח שמע על סוגיית אורך ההקשר (context length). אנו רוצים שהמודלים שלנו יהיו מסוגלים ״להחזיק בבטן״ כמות כמה שיותר גדולה של טקסט. אולם הקשר ארוך דורש כמות עצומה של משאבים לאימון ולאינפרנס.   
  
אז היום ב-#shorthebrewpapereviews אנו סוקרים מאמר שמציע גישת טיוב(fine-tuning) שמגדילה את אורך הקשר של מודל שפה. כלומר אם מודל שפה היה מאומן באימון מקדים (pretraining) עם אורך הקשר של 2048, השיטה המוצעת מאפשרת להאריכו פי 4 ל-8192. כמו שאתם רואים השם של השיטה מכיל את המילה LoRA שהיא שיטה מאוד פופולרית לפיין-טיון של מודלי שפה.

במקום לכייל (לשנות) את כל המשקלים של מודל השפה המכויל LoRA מעדכנת רק את התוספת למשקלי המודל (כמו ResNet). בנוסף התוספת למשקלים מיוצגת על ידי מטריצה עם רנק נמוך שניתן לתאר אותה על ידי מכפלה של מטריצות בעלות מימד נמוך יחסית.   
  
אז מה מציע LongLoRA בנוסף? כדי להגדיל את אורך הקשר נגיד מ- 2048 ל-8192 היא מחלקת את 8192 טוקנים ל-4 קבוצות בעלות 2048 טוקנים כל אחת שעבור כל אחת מהם ציוני ה-attention מחושבים בנפרד (חיסכון פי 16 בחישובים). את זה עושים בחצי מהראשים. בשאר הראשים פשוט מזיזים את הקבוצות האלו בחצי גודל כלומר הקבוצה הראשונה תכילי טוקנים מ-1024 עד 3071, השניה מ 3072 ל 5195 וכדומה. טריק פשוט מאוד אבל מביא תוצאות לא רעות בכלל.