⚡️🚀המאמר היומי של מייק -28.11.24: ⚡️🚀  
Parameter-Efficient Fine-Tuning with Discrete Fourier Transform

רקע: PeFT:

נתחיל את הסקירה ברענון קצרצר לגבי שיטות טיוב (fine-tuning) חסכוניות של מודלי שפה. PeFT הינה משפחה של שיטות המאפשרות טיוב של מודלים גדולים (בפרט מודל שפה) תוך שימוש במספר מצומצם של פרמטרים, מה שחוסך משמעותית במשאבי חישוב וזיכרון.

רקע: LoRA:

אחת השיטות הפופולריות ביותר ב-PeFT, הנקראת LoRA, מקפיאה את משקולות המודל ומאמנת מטריצות תוספת לכל שכבה של הטרנספורמטורים. כל מטריצת תוספת נלמדת הינה בעלת בדרגה נמוכה (low-rank), כך שניתן לייצגה על ידי מכפלה של שתי מטריצות קטנות (במימד האמצעי של המכפלה).

היתרון המרכזי של LoRA הוא שהיא מאפשרת להתאים מודלים גדולים למשימות ספציפיות תוך אימון של חלק קטן (נגיד 1% מכלל הפרמטרים שלו), מה שהופך אותה ליעילה במיוחד. שיטה זו הוכיחה את עצמה כאפקטיבית במיוחד בהתאמת מודלי שפה גדולים למשימות ספציפיות. בנוסף, LoRA מאפשרת החלפה מהירה בין גרסאות שונות של המודל המטויב, מכיוון שניתן לשמור את המטריצות הקטנות בנפרד מהמודל המקורי.

שיטה מוצעת:

הרעיון המרכזי הוא להסתכל על שינויי המשקולות של רשת הנוירונים כמו על תמונה או אות, ולייצג אותם בציר התדר במקום ערכים ישירים. כשאנחנו רוצים לטייב את המודל, במקום לשנות את כל המשקולות באופן ישיר (שדורש המון פרמטרים), אנחנו:

1. מגדירים מראש כמה נקודות דגימה במרחב התדרים שבהן נרצה להתמקד. זה כמו לבחור אילו תדרים אנחנו רוצים לשמור בייצוג הדחוס שלנו. זה נעשה על ידי בחירת מטריצת תדרים קבועה (לא נלמדת) E בגודל 2xn המשמשת לבניית ייצוג של מטריצת תוספת. מטריצה זו היא קבועה לכל השכבות של הטרנספורמרים.

2. לומדים וקטור c בגודל n (לכל שכבה) כאשר דרך שילובו עם E בונים את מטריצת התוספות בתחום התדר F (הסבר לאיך זה נבנה לא נראה ברור במאמר)

3. מעבירים את F דרך Gaussian bandpass filter (כלומר דוגמים בעיקר תדרים נמוכים, הנמצאים קרוב למרכז המטריצה).

4. מעבירים את מטריצת F לתחום הזמן (הרגיל) ומשתמשים בה בדיוק כמו ב-LoRA

יתרונות השיטה המוצעת:

היתרון הגדול הוא שתדרים הם דרך מאוד יעילה לייצג מידע (צריך 2n+ Ln משקלים כאשר L מספר השכבות במודל). בדיוק כמו שאפשר לדחוס תמונה או מוזיקה על ידי שמירת התדרים החשובים ביותר, כאן אנחנו יכולים לייצג שינויים מורכבים במשקולות באמצעות מספר קטן מאוד של תדרים.

זה עובד טוב(כנראה):

- שינויים במשקולות נוטים להיות "חלקים" יחסית, כלומר יש בהם מבנה שאפשר לתפוס טוב עם תדרים

- הבסיס המתמטי של פורייה הוא אורתוגונלי, מה שאומר שכל תדר מוסיף מידע ייחודי

- אנחנו יכולים לבחור מראש כמה תדרים אנחנו רוצים לשמור, ובכך לשלוט ישירות בכמות הפרמטרים

סיכום:

בניגוד לשיטות אחרות שמנסות להקטין את כמות הפרמטרים על ידי הגבלת הדרגה של המטריצות (כמו LoRA), הגישה הזו מסתכלת על הבעיה מזווית שונה - דרך עדשת התדרים, ומצליחה להשיג דחיסה משמעותית יותר.

https://arxiv.org/abs/2405.03003