⚡️🚀המאמר היומי של מייק -11.11.24: ⚡️🚀

Stealing Part of a Production Language Model

מזמן לא סקרתי מאמר על איזה ניתן לפרוץ למודלים עמוקים. יש תחום שלם שנקרא adversarial learning שבו חוקרים מפתחים מנגנוני הגנה נגד התקפות שמנסות לגנוב משהו מהמודל או דרך המודל (למשל דאטה שהוא אומן עליו). המאמר שנסקור היום מציע שיטה שבאמצעותה ניתן לזהות המימד הפנימי (החבוי) של המודל (מימד ייצוגי הטוקנים) וגם את המטריצה בשכבה האחרונה של המודל. שכבה זו הממפה את האמבדינגס של כל הטוקנים ללוגיטים שלאחר מכן מוזנים לסופטמקס שממנו יוצרים ״ההסתברויות של הטוקנים.

נתחיל מכך שמימד המטריצה W בשכבה האחרונה הוא N\_voc x N\_emb ,כאשר N\_emb זה המימד הפנימי של המודל (אלפים בודדים) ו- N\_voc הוא מספר הטוקנים במילון (בד״כ כמה עשרות אלפים ולפעמים מגיע מעל 100K). כלומר N\_voc > N\_emb וזה בדיוק מה שמחברי המאמר מנצלים. מכיוון שהראנק של מטריצה W הוא N\_emb כל המכפלות בה ממפות את הוקטור לתת מרחב במימד N\_emb של מחרב הלוגיטים שהוא בעל מימד N\_voc. כלומר אם ניקח מספר וקטורי לוגיטים ונשים אותם לעמודות של המטריצה (נקרא לה V) המספר המקסימלי לי וקטורים בלתי תלוים שיהיה לנו יהיה בדיוק N\_emb.

זה בדיוק מה שמחברי המאמר עשו. אולם מכיוון שהחישובים בטרנספורמרים הם לא בדיוק המלאה (FP16 גג) אז קשה לתפוס מתי העמודות הופכות להיות בלתי תלויות. במקום זה הם חישבו את הערכים הסינגולריים(ע״ס) של V (דרך מה שנקרא SVD - מי שלא מכיר ממליץ לקרוא על זה) ומסתכלים מתי היחס של ע״ס העוקבים (הם ממוינים) צונח משמעתית.

למה זה חשוב? כי במקרה האידאלי ע״ס של V צריכים להתאפס אחר שעברנו את הראנק של או N\_emb. אז בגלל אי דיוקים נומריים במודל כמובן שלא נראה ממש אפסים שם אלא ערכים מאוד נמוכים ואיפה שזה מתחיל לקרות זה בדיוק במימד N\_emb + 1. אז עושים את הטריק הזה על הרבה מאוד דאטה ומגלים את המימד החבוי של המודל שלכם.

כמובן שבעולם האמיתי אין לכם גישה לכל הלוגיטים אלא רק ל-topK ואז המאמר מנצל את העובדה שניתן לקנפג חלק מהמודל להוסיף מרג'ין לטוקן נתון במילון. ואחרי מספיק משחקים מקבלים את כל הלוגיטים (זה די יקר חישובית).

מימד של W זה נחמד אבל מה עם מטריצה W עצמה. המאמר מציע התקפה כדי לגלות אותה (סוג של) גם. בכללי המאמר מלא ברעיונות יפים להתקפות על המודלים ומי שמתעניין מוזמן להעיף מבט.

https://arxiv.org/abs/2403.06634