⚡️🚀המאמר היומי של מייק 17.08.24: ⚡️🚀

Faster Machine Unlearning via Natural Gradient Descent

היום סוקרים מאמר כחול לבן בנושא מעניין הנקרא unlearning. בדרך כלל אנו מעוניינים שהמודל שלנו ילמד מהדאטה אבל כאן אנו רוצים שהמודל ישכח דאטה מסוים. הנושא די חשוב לחברות שרוצה להיות compliant עם הדרישות של תקנים סטייל GDPR כאשר יוזר או קבוצת יוזרים מקבשים למחוק את הדאטה שלו באופן מוחלט. כמובן שבנוסף למחיקת הדאטה עצמו צריך ״למחוק״ אותו מה״מוח״ כלומר המשקלים של המודלים שאומנו (בפרט) על הדאטה הזה.

אחת השיטות הנאיביות לעשות unlearning היא למחוק את הדאטה ולאמן מודל מחדש. אבל זה יכול להיות די יקר ולא יעיל במיוחד למודלים גדולים. האם קיימת שיטה אחרת לעשות את זה?

אכן יש לא מעט מחקר בנושא של unlearning ואחת הגישות הפופולריות היא לקחת מודל מאומן ולמזער את הפרש של הביצועים על הדאטה שנותר והדאטה שאמור להימחק. כלומר אנו רוצים למזער את הלוס על הדאטה הנותר ולמקסם אותו על הדאטה שנמחק. ככה ״נמחק״ מהמוח(אולי הזכרון) של המודל את הדאטה המיועד למחיקה.

כמובן ששיטה נוספת ״למחוק״ את הדאטה מהמודל היא פשוט למקסם את הלוס על הדאטה המיועד למחיקה.

כמובן שיש שיטות רבות לעשות את זה באמצעות וריאציות שונות של מורד הגרדיאנט (SGD - stochastic gradient descent). המאמר מציע לעשות את זה עם מה שנקרא natural gradient או NG. זה קונספט פחות ידוע ואני אסביר אותו בקצרה. אתם בטח זוכרים מה זה קצב למידה ב-SGD, נכון? זה פרמטר קריטי לתהליך הלמידה וקיימות לא מעט שכלולים של SGD כמו ADAM ו-RMSProp שבפועל (בצורה לא מפורשת) קובעים את קצב הלמידה האופטימלי כתלות בפונקציית לוס.

יש כמובן דרך נוספת לבחור את קצב הלמידה בצורה אופטימלית וזה מה שעושה שיטת ניוטון קלאסית (נראה לי שזה השם) לאופטימיזציה. במקום להשתמש בקצב למידה סקלרי משתמשים בהופכית של ההסיאן של פונקציית לוס (מטריצה של נגזרות שניות). זה אופטימלי מבחינת ההתכנסות (כי משתמשים בקירוב טיילור מסדר שני של פונקציית לוס). אבל כמובן לא ניתן לעשות זאת לרשתות (יש קירובים אמנם) כי קשה מאוד להפוך מטריצה בגדול מיליארד על מיליארד.

המאמר מציע להחליף את ההיסאן ב- FIM או Fisher Information Matrix. למעשה FIM היא תוחלת של המכפלה הוקטורית של הגרדיאנט הלוג של הנראות (likelihood) של הדאטה המקורב על ידי המודל עם עצמו. למעשה FIM מודד עד כמה שינוי בפרמטרים של המודל משפיע על הנראות של הדאטה באמצעות המודל (עם המשקלים הנוכחיים). זה בעצם מצביע לנו עד כמה הנראות של הדאטה רגישה לשינוי בערכי המודל.

יש ל-NG הרבה יתרונות (למשל הוא חסין לרפרמטריזציה של המודל) אבל כמו ההסיאן עדיין מאוד קשה לחשב אותו עבור מודלים ענקיים. כמובן שקיימות שיטות המחשבות אותו באופן מקורב באמצעות שילוב עם פונקצית רגולריזציה ״נוחה״.

בנוסף לעדכון הרגיל של הגרדיאנט כמו ב-SGD עם FIM (כלומר בהופכית שלו) המאמר משתמש במה שנקרא proximal operator כדי לתקן את משקלי המודל אחרי שעודכנו עם SGD. למעשה PO מתקן את משקלי המודל אחרי העדכון ולא מאפשר להם להתרחק יותר ממשקלי המודל לפני עדכון כאשר ה״מרחק״ כאן מנורמל עם ההופכית של FIM תוך כדי לקיחה בחשבון של פונקציית רגולריזציה (שלא תתפוצץ).

המאמר די קשוח מתמטית ומקווה שהצלחתי לשפוך קצת אור עליו…

https://arxiv.org/abs/2407.08169