המאמר היומי של מייק - 02.03.25  
An Empirical Model of Large-Batch Training

מאמר מלפני 6 שנים של חוקרי OpenAI אך מצאתי אותו די מעניין לסקירה קצרה. המאמר חוקר גדול באץ' אופטימלי עבור אימון Mini-Batch Gradient Descent או MBGD. מה זה אופטימלי כאן? כזה שימעזר את מספר הדוגמאות ש-MiGD משתמש בהם כדי להביא את המודל לערך יעד של הלוס. כמובן שניתן ״להריץ״ את אותו הדוגמא כמה פעמים במהלך MBGD.

למי שכח MBGD שייך למשפחת שיטות המבוססות על מורד הגרדיאנט. עם MBGD אנו מחלקים את הדאטהסט למיני-באצ'ים שכל באץ' מורכב מכמה דוגמאות. עבור כל באץ' אנו מבצעים עדכון אחד של משקלי מודל כאשר הגרדיאנט מחושב בתור ממוצע של כל ערכי הגרדיאנטים עבור כל הדוגמאות בבאץ'. למעשה ממוצע זה הינו משערך של הגרדיאנט הממוצע של המודל עבור כל הדוגמאות מהדאטהסט. נזכיר שכל עדכון הוא הזזה (לינארית) של משקולות המודל בכיוון ההפוך לכיוון הגרדיאנט. כל עדכון כזה תלוי בקצב למידה שקובע את גודל עדכון המשקולות (מוכפל בגרדיאנט ממוצע).

המאמר מציע שיטה למציאת גודל באץ' אופטימלי (לפי ההגדרה שנתתי קודם) שעבור קצב למידה אופטימלי (הממזער את הלוס בכל איטרציה). די ברור כי גודל באץ' אופטימלי צריך להיות תלוי בפרמטרי המודל - למשל בצורת משטח הלוס וגם בערכי הגרדיאנט. המאמר טוען כי גודל באץ' אופטימלי ניתן לחשב בתור הטרייס (trace, סכום הערכים העצמיים) של המכפלה של מטריצת קווריאנס של גרדיאנט הלוס וההיסאין H של פונקציית לוס מחולקת ב G^T)HG) כאשר G הוא הממוצע של וקטור הגרדיאנט.

תוצאה זו התקבלה דרך פיתוח טיילור מסדר שני (בכיוון הגרדיאנט)ֿ, מציאה גודל קצב למידה אופטימלי והצבתו לנוסחה כדי לחשב את גודל הבאץ' שעבורו מתקבל ירידה מקסימלית של הלוס. לאחר מכן משווים את הירידה המקסימלית עם זו עבור גודל באץ' נתון B.

המאמר מדגיש שגודל אופטימלי של באץ' אינו תלוי בגודל שלה דאטהסט וכמובן משתנה במהלך האימון כי גם ההסיאן H וגם הגרדיאנט הממוצע H וגם מטריצת קווריאנס של גרדיאנט הלוס לא נשארים קבועים (בד״כ). המחברים מציינים מקרה פרטי די מעניין (לא קורה במציאות אמנם) שבו ההיסאין H שווה למטריצה היחידה I. במקרה הזה גודל באץ' אופטימלי שווה לסכום השונויות של כל רכיבי הגרדיאנט.

המאמר כתוב בצורה מאוד מובנת וניתן לקריאה קלילה יחסית…

https://arxiv.org/abs/1812.06162