⚡️🚀המאמר היומי של מייק 09.07.24: ⚡️🚀

Learning to (Learn at Test Time): RNNs with Expressive Hidden States

המאמר הזה המצהיר שהוא לומד ב״זמן טסט״ משך את עיניי היום. המאמר מציע ארכיטקטורה חדשה ומעניינת לעיבוד דאטה סדרתי. בעיקרון הרשת די דומה ל-RNN מבחינת המהות אבל יש כמה הבדלים מהותיים.

ֿאז מה יש לנו בארכיטקטורה הזו? בדומה ל-RNN אנו מחשבים את הייצוג עבור יחידת דאטה בזמן t (נגיד טוקן t) אבל כאן עושים זאת בשיטה שונה. לפי המאמר במקום לחשב את הייצוג עצמו אנו מחשבים את וקטור המשקלים שיאפשר לנו לחשב את ייצוגו של יחידת דאטה t. כלומר אנו מעדכנים את משקלות מודל בתנועה בהתאם לדאטה כלומר הרשת מתאפטמת ומתאימה את עצמה לדאטה שעליה היא מופעלת. זה נעשה באמצעות הזזה של המקשלים בכיוון הנגדי של הגרדיאנט של פונקציית לוס l.

מה זה בעצם פונקציית l ואיך מאמנים אותה? נניח שהייצוג של איבר דאטה t מחושב על ידי פונקציית f. במקרה הזה פונקציית l יכולה להיות (למשל) נורמה של הפרש ריבוע של ייצוג דאטה z (המחושב עם f) מהדאטה עצמו. כלומר אנו מאמנים את וקטור הייצוג להיות מסוגל לשחזר (כלומר לזכור) את הדאטה עצמו x\_t. כמובן שאין בזה הרבה משמעות אבל אם נאמן רשת עם קלט מורעש ונשווה את ייצוג עם הדאטה האמיתי נקבל סוג של רשת denoising שהרשת לומדת להפיק ייצוג המאפשר לזכור את הפיצ'רים המהותיים של דאטה הנחוצים לשחזור.

דרך אחרת המוצעת במאמר לאמן את רשת לשחזר הטלה למימד נמוך של דאטה להטלה אחרת כאשר שתי ההטלות נלמדות גם כן. הייצוג של דאטה במקרה הזה מחושב עם הטלה נלמדת שלישית (עם פונקציית f). כלומר המטרה כאן ללמוד את ייצוג של דאטה כאשר המשקלים מחושבים עם GD מהמשקלים הקודמים.

הארכיטקטורה קיבלה שם ttt וניתן לשלב אותם על שכבות אחרות (כמו טרנספורמרים או SSM). רעיון מגניב שבינתיים לא הפנמתי אותו עד הסוף…

https://arxiv.org/pdf/2407.04620