⚡🚀המאמר היומי של מייק 03.06.24: ⚡🚀

Better & Faster Large Language Models via Multi-token Prediction

אתם בטח שיודעים אנו רגילים לאמן מודל שפה גנרטיביים באמצעות חיזוי טוקן הבא בהינתם הטוקנים הקודמים (הקשר או קונטקסט). המאמר הזה (שקיבל די הרבה pr כשיצא) מציע לחזות כמה טוקנים עוקבים בו זמנית בהינתן הקשר. המחברים הראו שזה יכול לשפר את ביצועי המודל - זה לא מפתיע(לתחושתי) כי משימת חיזוי טוקנים מרובים דורשת מהמודל הבנה יותר מעמיקה של השפה. השיטה גם עשויה לתרום להאצת זמן ריצה והרווחים גדלים עם גודל המודל.

מאמר: https://arxiv.org/pdf/2404.19737

טלגרם: https://t.me/MathyAIwithMike/69