⚡️🚀המאמר היומי של מייק -24.10.24: ⚡️🚀  
HOW MANY VAN GOGHS DOES IT TAKE TO VAN GOGH? FINDING THE IMITATION THRESHOLD

מאמר מעניין שנטלו בו חלק חוקרים ישראלים מאוניברסיטת בר-אילן. הם חקרו נושא די חשוב שקשור להפרת זכויות יוצרים אפשרית על ידי מודלים גנרטיביים לתמונות. הרי יש מודלים שאומנו בחלקם על דאטה שהוא פרטי, מוגן על ידי זכויות יוצרים ואם המודל יתחיל לגנרט לנו תמונות דומות מדי להם זה עלול להוות עבירה על החוק. אבל איך להבטיח (או לפחות לתת הערכה כלשהי) לכך שזה לא יקרה?

המאמר בחר בגישה די אינטואיטיבית לכך. הרי כישורי העתקה של קונספט מסוים על ידי המודל קשורים קשר סיבתי (אמנם לא ב 100% מובן כרגע) במספר פיסות דאטה (= תמונות) המוכלות בדאטהסט שהמודל אומן עליו. אבל איך נדע זאת? הרי אז נצטרך לאמן הרבה מודלים כדי לבדוק מתי התמונות המגונרטות על ידי המודל יהיו דומות מדי קונספט T מסוים (עם פרומפט מתאים).

כמובן שזה לא בר עשייה והמאמר מציע שיטה יחסית פשוטה לעשות את זה כאשר הוא מניח הנחה מהותית אחת: מספר התמונות המכיל קונספט T מספיק לכך שהמודל יהיה מסוגל להעתיקו איננו תלוי ב-T. אני מניח שזה נכון בגבולות הסביר זאת אומרת המספר הזה נע באינטרוול יחסית צר לכל הסגנונות. יש עוד הנחה שניה (גם חשובה) שאין איזה confounded בין מספר התמונות ליכולת המודל להעתקה (גם די סביר).

עם הנחה כזו המאמר מציע לאמן מודל על הדאטהסט שיש בו שונות גדולה בין כמות ההופעות של כל קונספט. לאחר מכן המאמר מגנרט תמונות מכל T שהופיע בטקסט ובודק כמה מהם קרובים סמנטית (משווים אמבדינגס) ל T. זה נעשה עם הסף שנקבע דרך השוואה בין דמיון האמבדינגס של תמונות שונות של אותו הקונספט מול תמונות של מכילות את הקונספט הזה (כדי למזער FP יחד FN).

לאחר מכן מגנרטים תמונות עבור כל הקונספטים T שיש בדאטהסט ומחשבים כמה מהם (היחס) מכילים את T. זה נקרא imitation score. בסוף אנו נקבל imitation score עבור כל קונספט T ובגלל שיש לנו שונות גדולה בין הופעה של כל קונספט בדאטהסט ניתן לזהות איפה יש עלייה מובהקת ב- score הזה מבחינת מספר ההופעות של קונספט T בתמונה. זה קצת דומה לזיהוי elbow ב-k-means ויש אלגוריתמים מעולים (כמו PELT) שיודעים לעשות זאת. ככה נקבל את הסף של מספר ההופעות של קונספט בדאטהסט שממנו המודל יידע להעתיקו ופוטנציאלית לגרום לתביעות.

אהבתי - המאמר גם כתוב יפה וברור.

https://arxiv.org/pdf/2410.15002