⚡️🚀המאמר היומי של מייק -19.10.24: ⚡️🚀  
DiffCSE: Difference-based Contrastive Learning for Sentence Embeddings

סקירה קצרה ואחרונה(כנראה) במיני-סדרה על איך לבנות ייצוג דאטה באמצעות שיטות למידה ניגודית. כבר הסברתי על הלמידה הניגודית בשתי בסקירות הקודמות. בקצרה, מאמנים מודל הבונה אמבדינג לדאטה המקרב ייצוגים של פיסות דאטה דומות ולהרחיק פיסות דאטה לא דומות. וכאמור הוצעו עשרות שיטות לעשות זאת לדאטה מדומיינים שונים.

המאמר מציע שיטת CL העושה זאת בצורה מתוחכמת יותר (לטעמי). הרי אחת המטרות של בניית ייצוג הדאטה היא שהוא ישקף את התכונות האינהרנטיות של הדאטה והמחברים הציעו דרך ״לאכוף״ את זה על הייצוג. הם מאמנים מודל לבניית ייצוג טקסט כך שהמודל ״יבדיל בין מה אמור ומה לא אמור להיות בתוך הטקסט״.

איך הם עשו זאת? הם מיסכו כמה טוקנים בטקסט, ביקשו ממודל אחר לחזות את הטוקן הזה ואז אימנו את ייצוג כך שבעזרתו יהיה ניתן להבדיל בין הטוקנים שנחזו ואלו שלא. כלומר בנוסף למודל החיזוי (לא אומן) ומודל לבניית אמבדינג הם אימנו עוד מודל לסיווג בינארי שמטרתו להגיד האם טוקן נחזה או לא. וייצוג הטקסט מוזן למודל הסיווג הזה.

דרך אגב פונקציית הלוס למודל הסיווג דומה לזו של GAN אבל אין באמת קשר בין שני הדברים (זה טיפה בלבל אותי בהתחלה)....

https://arxiv.org/pdf/2204.10298