⚡️🚀המאמר היומי של מייק -18.10.24: ⚡️🚀  
SimCSE: Simple Contrastive Learning of Sentence Embeddings

סקירה קצרה מאוד על איך ניתן לעשות למידה ניגודית (contrastive learning) כדי לבנות ייצוג חזק של הטקסט. הרי כבר הסברנו בסקירה הקודמת שהמטרה של CL היא לאמן ייצוג של דאטה כך שייצוגים קרובים סמנטית יהיו קרובים במרחב הייצוג ואילו ייצוגים של דוגמאות לא דומות יהיו רחוקות שם. מאמנים ייצוג כזה בדרך כלל דרך מזעור היחס שבין ייצוגי פיסות דאטה דומות (זוג חיובי) לבין אלו של הלא דומות (שליליים).

השאלה איך לבנות את הייצוגים האלו (במיוחד הזוגות החיוביים)? זה בעצם נושא מחקר פעיל מלפני שנתיים-שלוש. המאמר המסוקר מציע לבנות זוגות חיובים דרך dropouts שונים של רשת הנוירונים (שאותה מאמנים לבנות את הייצוג). כלומר עבור אותו הטקסט זוג דוגמאות חיובי נבנה עם עם הפעלת הרשת עליו עם שני dropouts שונים. נזכיר dropout מבטל באקראי קשרים בין נוירונים ברשת ומהווה כלי ידוע לשיפור יכולת ההכללה של הרשת. הזוגות השליליים נבנים עם דוגמאות שנבחרו בצורה אקראית.

לדאטהסטים המכיל משפטים מתויגים כמו למשל NLI (למשפט נתון הדאטהסט מכיל משפט אחד עם אותה המשמעות(entailment), משפט אחד בעל משמעות דומה ומשפט אחד בעל משמעות הפוכה או סתירה - contrary). באופן לא מפתיע המאמר מציע לבחור בתור זוג שלילי את שני המשפטים בעלי משמעות הפוכה ובתור זוג חיובי שניים עם אותה משמעות.

בנוסף המשפט הזכיר לי לייצוג דאטה טוב יש 2 תכונות מהותיות: קרבה בין ייצוגי הדאטה הדומה והתפלגות יוניפורמית של כלל הייצוגים של הדאטה - זה חשוב.

https://arxiv.org/pdf/2104.08821