המאמר היומי של אביב ומייק: 18.06.25  
Harnessing the Universal Geometry of Embeddings

לפני כחודש הופיע המאמר, וישר היכה גלים. הוא בונה על מאמר אחר, The Platonic Representation Hypothesis, שגם-כן היכה גלים בזמנו, ומתיימר לחזקו משמעותית, במידה מפתיעה. ועל הדרך הדרך גם מדגים איך למנף את ההישג המסקרן-תיאורטית לכדי פריצה משמעותית של חילוץ מידע. גובה הגלים היה תוצר של כל אלו - בשילוב עם כתיבה שמעודדת קריאה בומבסטית \*מדי\* של מה שהמאמר בעצם מראה. בואו נעשה קצת סדר.

דאטה, בין עם טקסטואלי, תמונתי, או אחר, מגיע בסופו של דבר מהתהליך הזה שנקרא המציאות. ככל שהמודלים הגדלים מאומנים על יותר דאטה, יותר מגוון, עבור משימות רבות ומגוונות - הייצוגים שלהם ייטו יותר ויותר להתכנס אל אותה המציאות המשותפת שמאחורי הדברים, אל המרחב הלטנטי ה"אמיתי" שמאפשר את ההיסקים האופטימליים. זו הייתה הטענה אותה המאמר המקורי קידם - והשתדל להדגים, באמצעות מגוון מדדים והשוואות. עד כאן אז.

המאמר החדש בא לטעון טענה לכאורה חזקה יותר, ו"קונסטרוקטיבית":

ניתן ללמוד את המרחב הלטנטי האוניברסלי הזה עבור ייצוגי טקסט, ולמעשה לרתום אותו על-מנת "לתרגם" ממרחב ייצוג אחד לשני - ללא דאטה המאפשר הצלבה (כלומר שקיבלנו את הקידודים שלו משני הצדדים), ולמעשה ללא גישה או שום ידע על אחד מן המודלים, רק לדוגמיות הקידוד שלו.

או אז, בהינתן היכולת לתרגם מהמרחב של מודל לא ידוע אל אחד שבשליטתנו, ניתן לגלות בזה האחרון תכונות על מידע שקודד במודל הלא-ידוע, ואף לשחזרו באמצעות טכניקות "היפוך-שיכון" קיימות.

אז איך כל זה עובד?

ראשית כל, בואו נחדד את שתי הנקודות בהן המאמר חוטא בניסוחי "הבטחת-יתר" (שבתקווה יתעדנו כשיתנגשו בקיר ביקורת העמיתים האקדמית):

המאמר המקורי דיבר על מרחב משותף, \*יחיד\*, שמאחד בין כל המודלים, המציאות שמאחורי מגוון ההשתקפויות שלה. זהו לב העניין התיאורטי שהצדיק רפרנס פילוסופי שהולך אחורה 2400 שנה. המאמר בו עסקינן עכשיו, לעומת זאת, \*לא\* משיג ייצוג יחיד שכזה. שיטת למידת הייצוג שהוא מאמן, כפי שנראה בהמשך, מגשרת רק בין כל \*צמד מודלים ספציפי\*. זה בהחלט בכיוון, אבל עוד לא ממש שם.

המאמר המקורי דיבר על התאמה (alignment) בין modalities שונים (ועוד), כמו למשל בין שמות אובייקטים (במודלי טקסט) לתמונות שלהם (במודלים ויזואלים) (ראו תמונה). זה הרבה יותר עמוק ומשמעותי מהתאמה פשוט בין מודלי טקסט שונים - בהם עוסק המאמר החדש. (הוא אמנם נוגע גם ב-CLIP, אך זהו בפרט מודל ייצוג לטקסט, שעל-פי בנייתו כבר מראש בא מותאם גם מול דאטה תמונתי, אין באמת מה ללמוד מכך בהקשר שלנו.)

ועכשיו אחרי שכיבינו את להבת ההייפ המוגזם, בואו נצלול אל הפרטים והדברים שיוצאים מהם, שכן מעניינים בפני עצמם. אז כפי שאמרנו, המאמר בונה מיפויים ממרחב אמבדינג X למרחב אמבדינג Y. הוא עושה זאת באמצעות חמישה מיפויים, המיוצגים באמצעות מודלים מאומנים:

מיפויים A1 ו-A2 הממפים מ-X ו-Y למרחב אמבדינג משותף Z בהתאמה

מיפוי T מיישר את האמבדינגס אחרי A1 ו-A2 לייצוג לטנטי משותף Z\_m

מיפויים B1 ו-B2 המחזירים את האמבדינגס מ-Z\_m ל X ו- Y בהתאמה

על גבי אלו ניתן להגדיר גם את:

מיפויי "תרגום" F1 = B2 ◦ T ◦ A1, F2 = B1 ◦ T ◦ A2 - הממפים מרחב אמבדינג מקורי X למרחב השני Y ובכיוון ההפוך בהתאמה

מיפויים עצמיים R1 = B1 ◦ T ◦ A1, R2 = B2 ◦ T ◦ A2 הממפים את X ו- Y לתוך עצמם (X ו-Y) דרך מרחב אמבדינג משותף באמצעות T.

אחרי שהגדרנו את שורת המיפויים הארוכה, נסביר את מבנה פונקציית הלוס המלאה. היא מורכבת מכמה לוסים. הלוס הראשון מנסה לכפות על התפלגות המיפוים מ-X הממופים דרך F1 ל-Y להיראות כמו המיפויים מ-Y עצמו. למטרה זו משתמשים בגאן (אלו ששלטו ללא עוררין בתחום גנרוט התמונות לפני מודלי דיפוזיה). הגאנים (GANs) מאמנים שני מודלים בו זמנית עם לוסים מנוגדים: המודל המגנרט (F1) מאומן לעשות את המיפויים מ-X מאוד דומים לאלו מ-Y, כאשר המודל השני (D1) מאומן להבחין בין מיפויים מ-X לאלו אחרי F1. בסוף (אם התהליך מתכנס) מקבלים מודל גנרטיבי חזק (F1) המסוגל "לתת פייט" למודל דיסקרימינטור חזק (D1), וכך עושים גם ל-F2 עם D2. ד"א המאמר משתמש בגישה הקלאסית לגאן מהמאמר של גודפלואו מ-2014.

בנוסף מאמנים עוד שני גאנים לייצוגים מהמרחב הלטנטי הממופים מ-X וגם מ-Y. כלומר המודל הגנרטיבי במקרה הזה הוא T◦A1 שמאומן באופן שמתואר בפסקה הקודמת. גאן נוסף מאומן עבור T◦A2. בסוף, ארבעת הלוסים שתארנו מהווים את החלק הראשון של פונקציית לוס המלאה.

החלק השני של הלוס הוא לוס "שחזור" המוודא שכל אמבדינג המועבר מ-X (לחילופין מ-Y) למרחב המשותף יודע לחזור אל עצמו אחרי מיפוי B1 (מיפוי B2). החלק השלישי של הלוס הוא לוס "עקביות התרגום" (cycle consistency loss) שדואג שהייצוג שהגיע מ-X ל-Y (לחילופין מ-Y ל-X) עם F1 (עם F2) חוזר לתוך עצמו אם מפעילים עליו את התרגום ההפוך F2 (ו-F1 בהתאמה). החלק האחרון של הלוס דואג שיחסים בין זוגות אמבדינגס שונים מ-X (לחילופין מ-Y) יישמרו בתרגום ל-Y (ל-X).

בסופו של דבר, הלוס כולל את כל הלוסים המתוארים לעיל. כך אנחנו לומדים את הייצוג המשותף (שוב, ללא שום דוגמאות שנתונים לנו קידודיהן משני הצדדים!). באופן מרשים, המיפוי הנלמד למעשה מכליל גם להתפלגויות טקסטים מאוד אחרות, כך שהגישור שהושג כאן הוא די כללי. אבל לפרטים אלה כמו גם האפליקציה של הטכניקה לגילוי מידע, נשאיר משהו למאמר עצמו :)

https://arxiv.org/abs/2505.12540