תרגיל 4: עיבוד תלת-מימדי בביולוגיה מבנית

צבי מרמור, עמית יוסף, הדר אמסלם, שקד הרטל

1. שינוי השכבה שממנה נלקחים הembeddings משפיע על איכות הייצוג של הדאטה – במקרה שלנו הפפטידים, ולכן גם על הדיוק של המודל ועל ציון הAUC של הסיווג.
בחירה באחת השכבות נמוכות, למשל שכבה 9, לרוב תלכוד תכונות מבניות ותייצג פחות את המשמעות הביולוגית הכוללת של הפפטיד. בחירה בשכבות בינוניות, כמו 20 עשויות לאזן בין תכונות ״מקומיות״ ל״גלובליות״. בחירה באחת השכבות גבוהות, כמו 33, מכילות לרוב את הייצוגים העשירים ביותר, כלומר תכונות סמנטיות או מבניות כלליות שמכילות מידע על כל הרצף, ולכן נצפה שבחירה כזו תניב את התוצאות הטובות ביותר.
בפועל ראינו שAUC של הבייסליין משתפר כאשר בוחרים שכבה גבוהה יותר בהשוואה לבחירת שכבות נמוכות יותר.
2. לא, מודל גדול יותר לא בהכרח מוביל לתשובות טובות יותר, בגלל שככל שהמודל גדל ,כלומר ככל שיש לו יותר פרמטרים, כך הוא גם נהיה רגיש יותר לרעש בדאטה ועלול לבצע overfitting לדאטה. במצב כזה, המודל ילמד ״בעל פה״ את דוגמאות האימון, אבל לא יצליח כלל להכליל לדוגמאות חדשות.
בנוסף, מודלים גדולים דורשים יותר משאבי חישוב וזמן אימון, ואם זה לא נעשה כמו שצריך זה עלול לפגוע בביצועים במקום לשפר אותם.
3. הנוסחה הנתונה מזהה את הסיווג של הפפטיד בכך שהיא משווה בין המרחק שלו לדוגמאות השליליות ולדוגמאות החיוביות: אם הפפטיד קרוב יותר לחיוביים, אז הסיווג יהיה חיובי; ואם הוא קרוב יותר לשליליים, הסיווג יהיה שלילי.
(השימוש בlog1p ״מרכך״ הבדלים קיצוניים ונמנע מערכים לא תקינים (log0)).
4. הAUC הטוב ביותר שהושג: 0.92 עם ההיפרפרמטרים:
embedding_size=2560,
hidden_dim=256,
dropout=0.2,
batch_size=64,
epochs=100,
learning_rate=1e-3
5. א. קלטים נוספים שיכולים לשפר ביצועים יכולים להיות מאפיינים פיזיקלים וכימיים של הפפטידים שנוסיף לembedding של הפפטידים.
ב. נוכל להשתמש באלגוריתמים אחרים שנוצרו כדי לבצע קלסיפיקציות, כמו Random-Forest, SVM, KNN (או kmeans כמו שמומש בתרגיל).
6. גרף הtSNE הצבוע לפי התוויות האמיתיות מראה ששני הקלאסים (חיוביים ושליליים) לא נפרדים לחלוטין במרחב דו־ממדי ויש חפיפה רבה ביניהן. אך כן ניתן לראות שקיימת מגמה כללית של הפרדה חלקית, במיוחד בצד הימני של הגרף שבו נמצאים יותר פפטידים חיוביים ובצד השמאלי יותר פפטידים שליליים. לכן, למרות שההפרדה אינה חדה או ליניארית, קיים מבנה מסוים שאולי ניתן לזהות.
אלגוריתם Kmeans הצליח למצוא מבנה כלשהו המפריד בין הקלאסים, אך הוא לא לגמרי תואם את ההפרדה ״האמיתית״ בין הקלאסים.

7. .לפי התוצאות שקיבלנו pLDDT הוא מדד טוב יותר להפרדה בין הקבוצות:
עבור COM distance התקבל AUC של 0.47, לא מופרד בכלל.
עבור mean pLDDT: התקבל AUC של 0.76 , הפרדה בינונית-טובה (אם כי פחות טובה מההפרדה של הרשת).