דוח מעבדה טים בוכבינדר 313419814

מבוא:

המטרה של פרויקט זה היא לבנות מודל שמזהה פפטידים שמזוהים על ידי HLA\_A0201 בדיוק כמה שיותר גבוהה.

בדוח נחקור את הבעיה ונציע פתרונות אפשריים שיעזרו למקסם את הדיוק של המודל.

בכל שלב:

1) נציג את הביצועים הנוכחיים של המודל בעזרת וויזואליזציה של התוצאות

2) נעלה בעיה שלדעתינו כדאי להתמקד בה

3) נציע פתרון אפשרי.

4) נציג את התוצאות לאחר הפתרון.

**חקירת הדאטא:**

לפני שנתמקד ביצירת המודל עלינו להכיר את הדאטא איתו אנחנו עובדים,

ממבט קצר ניתן לראות כי 89% מהרצפים תויגו כשליליים ( לא זוהו על ידי ה HLA), כלומר הדאטא מאוד לא מאוזן.

המחשה קצרה להתפלגות הדאטא (לאחר חלוקה ל TRAIN ו TEST 90\10)

TRAIN

A picture containing text, orange, close

Description automatically generated

**מודל: (נסיון ראשון)**

כעת לאחר שלמדנו את הדאטא נתחיל מלבנות מודל מאוד פשוט, בעזרתו נוכל "להרגיש" את הבעיה, את התוצאות שלו נוכל לנתח ולגלות איפה כדאי להשתפר.

ישנן 20 חומצות אמינו שונות ולכן החלטנו לייצג כל חומצת אמינו כ ONEHOT VECTOR באורך 20 כאשר פפטיד אחד ייצגנו בתור שרשור של כל חומצות האמינו מהן הוא מורכב (כלומר כל פפטיד מורכב מוקטור באורך 180)

הגדרת המודל :

Text

Description automatically generated

ניתן לראות שהמודל מצפה לרצף באורך 180 תווים כאשר הפלט שלו יהיה סקלר (על הסקלר אנחנו נפעיל סיגמויד כדי שהערך יהיה בין 0 ל 1 והחיזוי יתבצע כ ROUND של הערך) .

|  |  |
| --- | --- |
| FIGURE 1 |  |
| BINARY CROSS ENTROPY | Loss function |
| SGD | optimizer |
| 0.001 | Learning rate |
| 30 | epochs |

הרצנו את המודל עם הפרמטרים שאפשר לראות ב FIGURE 1

אלו התוצאות :

הגרף מראה את ה LOSS בTRAIN וב TEST במשך האפוקים, הנתונים מראים את אחוז הדיוק על הTEST

Chart, line chart

Description automatically generated

|  |  |
| --- | --- |
| TRAIN |  |
| TEST |  |

מסקנות: אנחנו מתייחסים ל פפטידים ש HLA מזהה כחיוביים ולשאר כשליליים

אפשר לראות שהמודל מסווג את כל הפפטידים כשליליים , אמנם הוא משיג אחוז דיוק לא רע על הTEST ברור לנו שצריך לשפר את המודל.

ראינו קודם ש89% מהפפטידים שלנו הם חיוביים, זאת אחת הסיבות שהמודל לא מצליח ללמוד איך לסווג את השליליים.

כדי שהמודל יצליח ללמוד ולהכליל עבור הפפטידים השונים נרצה לאזן את הדאטא.

**מודל: (נסיון שני )**

השלב הבא אם כך הוא לאזן את הדאטא:

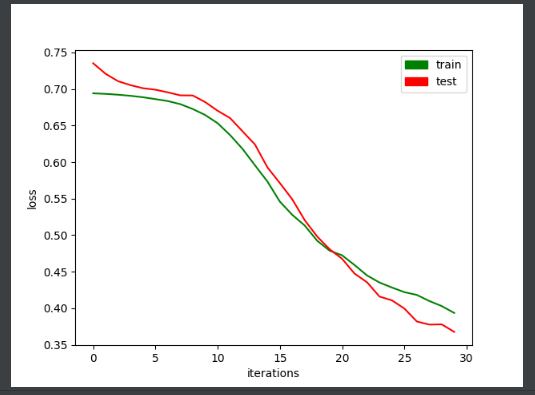
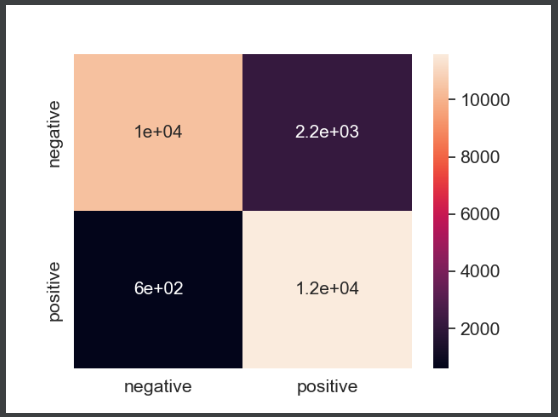
בנוסף ראינו שמדד ה ACCURACY לא נותן לנו מספיק מידע, שכן הרשת השיגה מדד דיוק גבוהה אבל תוצאות רעות, לכן הוספנו מטריקה נוספת למדידת ההצלחה שהיא

.F1 SCORE+CONFUSION MATRIX (confusion matrix תעזור לנו להבין כמה המודל טוב על כל אחת מהקטגוריות ו F1 ישקלל לנו את האקיורסי עם המדדים של כמה המודל טוב על כל קטגוריה, הסבר יותר מפורט ניתן למצוא בקישור למטה)

<https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/> הסבר על המטריקות

איזנו את הBATCHES בעזרת WEIGHTED SAMPLER

כל לייבל קיבל משקל 1 פחות אחוז ההופעות שלו בדאטא. הרצנו את אותו המודל על מאזן הדאטא החדש וקיבלנו:

|  |  |
| --- | --- |
| TRAIN |  |
| TEST |  |

מסקנות:

ניתן לראות כי אחוז הדיוק של המודל ירד קצת אך כעת הוא מצליח לסווג גם רצפים חיוביים וגם שליליים. כמו כן למודל אין אוברפיט משמעותי.

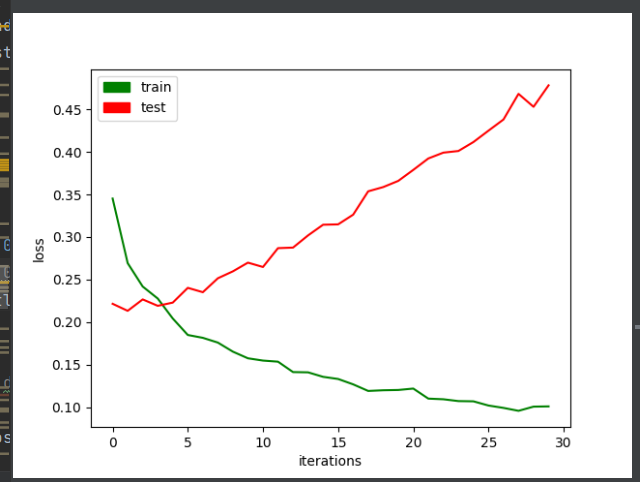
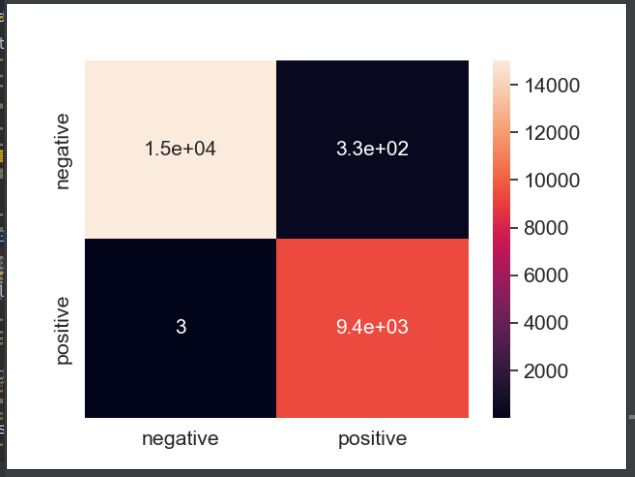
ננסה לאזן את הדאטא כך שהמודל יסווג את הרצפים החיוביים והשליליים בערך באותה המידה (הגדלנו קצת את המשקל של הדוגמאות השליליות לעומת החיוביות וקיבלנו את התוצאה הבאה )

|  |  |
| --- | --- |
| TRAIN |  |
| TEST |  |

**מודל: (נסיון 3)**

כעת ננסה לגרום לרשת לסווג את הפפטידים בצורה טובה יותר, נתחיל מלנסות לשנות אופטימייזר לADAM :

|  |  |
| --- | --- |
| FIGURE 2 |  |
| BINARY CROSS ENTROPY | Loss function |
| ADAM(betas = 0.9,0.999) | optimizer |
| 0.001 | Learning rate |
| 30 | epochs |

|  |  |
| --- | --- |
| TRAIN |  |
| TEST |  |

מסקנות:

ניתן לראות כי ההצלחה של המודל על הטריין עלתה משמעותית גם במדד הדיוק וגם במדד הF1 כלומר הוא מצליח להבין את הדאטא בצורה טובה יותר, לעומת זאת ב TEST אמנם מדד הדיוק עלה אך מדד ה F1 ירד , ואכן אפשר לראות כי המודל מכליל בצורה טובה עבור רצפים שליליים אבל מכליל בצורה הרבה פחות טובה עבור רצפים חיוביים. כלומר למודל יש אובר פיט על הדוגמאות החיוביות. אחת הסיבות האפשריות לתוצאה הזו היא שבדאטא של ה TRAIN שכפלנו את התוצאות החיוביות כדי לאזן את הדאטא ומכאן שהמודל אמנם רואה הרבה דוגמאות חיוביות אך הרבה מהן יהיו זהות ולכן הוא מצליח להכליל בצורה פחות טובה.

מחשבה לשיפור:

ננסה להפעיל רגולריזציה על המודל כדי לשפר את האוברפיט. הוספנו דרופאאוט (0.5) על השכבות השונות של הרשת ו RIDGE REGULARIZATION בנוסף הורדנו את כמות האפוקים ל 15 .

Chart, line chart

Description automatically generated Chart, treemap chart

Description automatically generated

* אפשר לראות כאן שהטעות על ה TRAIN גבוהה מהטעות על הTEST זה קורה מכיוון שהמודל יודע לזהות בצורה טובה יותר דוגמאות שליליות, הTRAIN מאוזן מבחינת יחס הדוגמאות השליליות והחיוביות, אך הTEST מורכב ברובו מדוגמאות שליליות.

|  |  |
| --- | --- |
| TRAIN |  |
| TEST |  |

**מודל (נסיון 4) :**

שיחקנו עם גודל הרשת ומספר הנוירונים בכל שכבה ( למרות שכרגע נראה שהרשת מספיק חזקה כדי ללמוד את הדאטא, אבל אולי היא אפילו חזקה מדי ולכן האוברפיט ) ועם ה LR .

(מכיוון שיש כמות עצומה של ניסויים שאפשר להציג היינו צריכים לסנן, בחרנו את הניסויים שהיו נראים לנו יותר מעניינים)

ומצאנו שהאכיטקטורה הנ"ל מחזירה לנו את רמת הדיוק הגבוהה ביותר בשקלול של המטריקות שבחרנו

הארכיטקטורה:

Text

Description automatically generated

הפרמטרים:

לADAM :

|  |  |
| --- | --- |
| FIGURE 2 |  |
| BINARY CROSS ENTROPY | Loss function |
| ADAM(betas = 0.9,0.999) | optimizer |
| 0.00065 | Learning rate |
| 15 | epochs |

התוצאות:

Chart, line chart

Description automatically generated Chart, treemap chart

Description automatically generated

|  |  |
| --- | --- |
| TRAIN |  |
| TEST |  |

תרגילים 6-7

|  |  |
| --- | --- |
| הפפטידים | הציון שלהם במודל |
| PLVDLPIGI | 0.97222 |
| ALLAGTITS | 0.93837 |
| LLTDEMIAQ | 0.90346 |
| GLTVLPPLL | 0.89040 |
| LLFNKVTLA | 0.88609 |

אלו הרצפים שהמודל מזהה בצורה הכי חד משמעית כ חיוביים.

מכיוון שהוספנו למודל סיגמואיד על הקלט, כל הפלטים שלו הם בין 0 ל 1 כאשר תוצאות שקטנות מ 0.5 יסווגו כשליליות ותוצאות שגדולות מ 0.5 יסווגו כחיוביות, לכן ככל שהתוצאה גבוהה יותר כך המודל "בטוח" יותר שהדוגמא חיובית.