ניר חביב ובתאל ששון

**דו"ח פרוייקט סיום – קורס עיבוד שפה טבעית**

**המשימה:**

Named entity recognition

**הטכנולוגיה:**

PyTorch

**המאגר:** שימוש בקורפוס מתוייג של kaggle

https://www.kaggle.com/abhinavwalia95/entity-annotated-corpus/data

ממה שקראנו על מנת לבצע את המשימה בצורה הטובה ביותר כולל שימוש בהקשר של המשפט , ארכיטקטורת הרשת המומלצת היא bi directional lstm with crf אז התרכזנו בלממש מודל כזה בעצמנו. בארכיטקטורת הרשת שבנינו משתמשים בLSTM כהוצאת features ובשכבת CRF נוספת כדי למצוא את "המסלול הכי טוב" לתיוג בהקשר של המשפט.בנוסף, הוספנו תהליך של pre processing של בניית 2 מילונים : המילים והתגים שבכל הקורפוס. פונקציית האופטימיזציה המומלצת על פי המאמרים היא adam אז השתמשנו בה.

לאחר בניית המודל בנינו את פונקציית האימון. בסיום אימון על כל הEPOCH הוספנו גם פונקציית VAKIDATE שקיבלה 20 אחוז מהDATA של הEPOCH על מנת לבדוק את המודל המאומן לפני שמריצים אותו על קבוצת הTEST.

בנוסף, הוספנו שמירה של המודל המאומן לקובץ באמצעות checkpoint וזאת על מנת שלא נצטרך להתחיל כל פעם מחדש אלא נוכל להמשיך מנקודה קיימת.

באימון נתקלנו במס' בעיות :

* OVERFFITING לתג הפופולרי ביותר באמת בקורפוס שהוא Object.
* אימון על כמויות גדולות לא התאפשר בגלל כוח העיבוד המצומצם שלנו במחשבים האישיים.

הנחנו כי בגלל שגם לא יכולנו לאמן על הרבה DATA זה גרם לOVERFITTING.

כפתרון לזה, הוספנו dropout ויחקנו עם הערכים שלו כדי להגיע לתוצאות טובות יותר.

בסיום כל אימון מעבר לשמירת המודל המאומן הוספנו גם תצוגה גרפית של התוצאות :

(שמנו לב לפני ההגנה כי בטעות התבלבלנו במקרא והחלפנו בקוד את זה אך לא הספקנו לאמן מחדש כדי להוציא תוצאה)

