בס"ד

שירה יאיר 31538975
אסנת אקרמן 315747204

Report1-experiment-assimgnet3:

במשימה זו נדרשנו לאמן מודל אשר יצליח לזהות את השפה הבאה:

```
good_regex = r'[1-9]+a+[1-9]+b+[1-9]+c+[1-9]+d+[1-9]+
bad_regex = r'[1-9]+a+[1-9]+c+[1-9]+b+[1-9]+d+[1-9]+'
```

כדי לממש את זה בצורה הטובה ביותר פעלנו כך:

1. טענו את הדאטה והמרנו את כל הרצפים של אותיות לרצפים עם מספרים. כל אות הוחלפה במספר הASCII שלה.
2. הכנסנו את הרצפים למודל כאשר האורכים של הרצפים שונים זה מזה, השתמשנו בספרייה המובנית של pytorch בשביל לרפד אותם לאורך אחיד.
3. בנינו טבלת Embedding באורך מס' תווי הASCII.
4. לאחר מכן הכנסנו את הפלט מEmbedding כקלט לLSTM ומהפלט של הLSTM לקחנו את הlast state שלו .
5. את ה last state שקיבלנו מהLSTM הכנסנו לשכבה לינארית בעלת שכבת hidden layer נוספת.
6. תוצאת השכבה הלינארית זו הפרדיקציה שלנו.

היפר- פרמטרים

לאחר מספר ניסיונות ובדיקת פרמטרים שונים הגענו למסקנה שהפרמטרים הבאים הם האידיאלים בשביל לקבל את אחוזי הדיוק הגבוהים ביותר.

גילינו שהתוצאות הטובות ביותר שקיבלנו מהמודל שלנו כאשר:

EMBEDDING_VOCAB = 126

EMBEDDING_LENGTH = 30

HIDDEN_LAYER_LSTM = 50

HIDDEN_LAYER = 50

LR = 0.001

BATCH_SIZE = 50

BATCH_SIZE_DEV = 500

10000:SAMPLE_SIZE_TRAIN

2000:SAMPLE_SIZE_DEV

- הערה- ניסינו לשחק עם מספר הדוגמאות שייצרנו. בהתחלה ייצרנו 500 דוגמאות של negative_example ו500 דוגמאות של positive_example וראינו שהאחוזים נמוכים והמודל לא מצליח ללמוד את השפה. שיחקנו עם הגדלים של הדאטה בdev ובtrain עד שהגענו לתוצאה שתביא לנו את האחוזים הכי גבוהים- 100%.

בס"ד

שירה יאיר 31538975
אסנת אקרמן 315747204

לקחנו לנו EPOCH 5 עד שהגענו ל100%

וה-TIME שלקח באיטרציה הראשונה זה:Epoch: 01 | Epoch Time: 0m 37s

ובאיטרציה האחרונה:Epoch: 05 | Epoch Time: 0m 35s
(במהלך האיטרציות הזמן עלה וירד בין 37 ל35 שניות.)
-ניסינו להשתמש בDROPOUT (0.5=P ו0.2=P) אך, ההוספה שלנו לא עזרה למודל ללמוד והביאה לאחוזים נמוכים.

ולכן סה"כ,

בסופו של דבר קיבלנו 100% הצלחה על סט הבדיקה.

*הערה: גילינו במהלך אימון הדאטה שבחלק מהמקרים האחוזים בTRAIN היו נמוכים משמעותית מהDEV ובאיטרציות הבאות הפער היה פחות משמעותי אך, המשיך להיות נמוך ממנו.