מבוא ללמידה עמוקה – תרגיל 2:

חלק מעשי:

במהלך התרגיל, לשם השוואת פרמטרים לרשתות אותן הרצנו, השתמשנו בספריית Weights and biases הנלמדה בתרגול. לאורך התרגיל נציג את הגרפים, שקיבלנו מהספרייה, המשווים בין פרמטרים אלו.

מטרתנו בכל המודלים הייתה למקסמם את ה-Accuracy אשר חושב לפי מספר ההצלחות של המודל ב-Testing data.

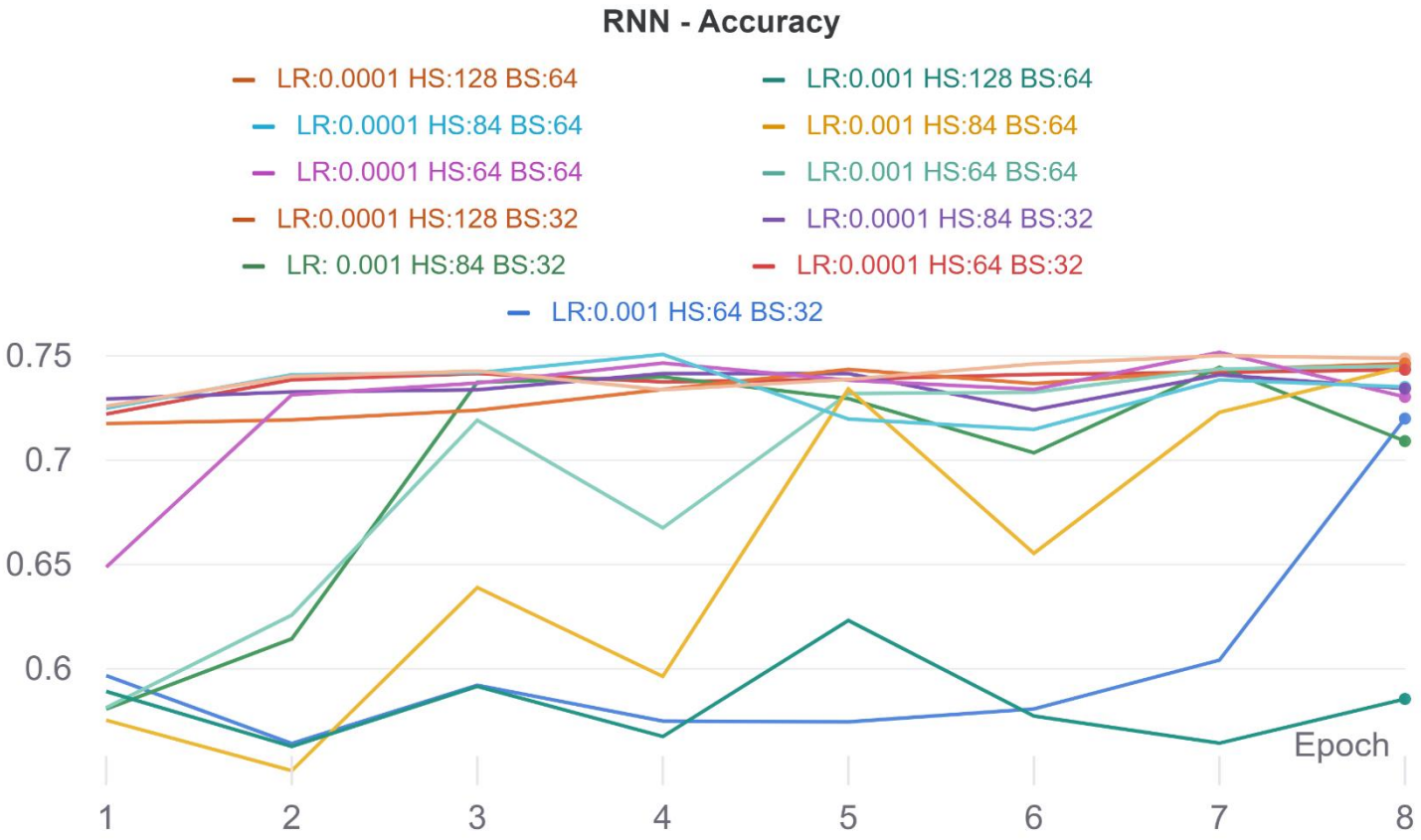
כמו כן, נציין כי:

- הגדלנו את גודל ה-Training data ל45,000 דגימות (90% מהנתונים).
- מספר ה-Epoch-ים נע בין 7-8 (תלוי במודל).

1) RNN:

בבניית רשת זו, כללנו שתי מטריצות נלמדות: אחת המייצגת את פלט הרשת והשנייה המייצגת את "זיכרון" הרשת.
בחרנו מספר פרמטרים אותם רצינו לבדוק כדי להשיג ביצועים אופטימליים לרשת זו.
הפרמטרים הם: Learning rate, Hidden size, Batch size, ובחנו את הרשת לפי הדיוק (Accuracy) אותה השיגה.

נציג את תוצאות ההרצות –

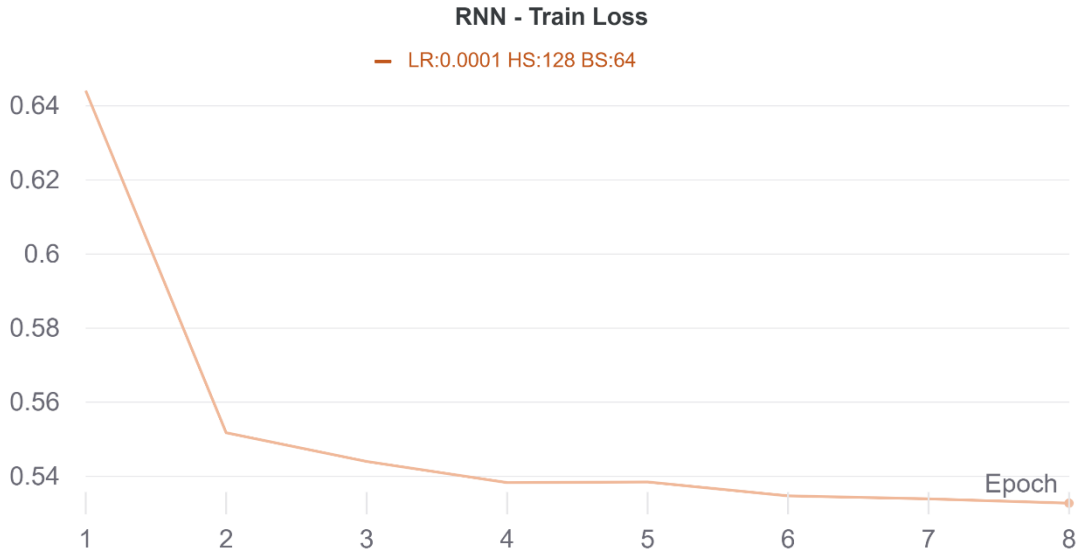


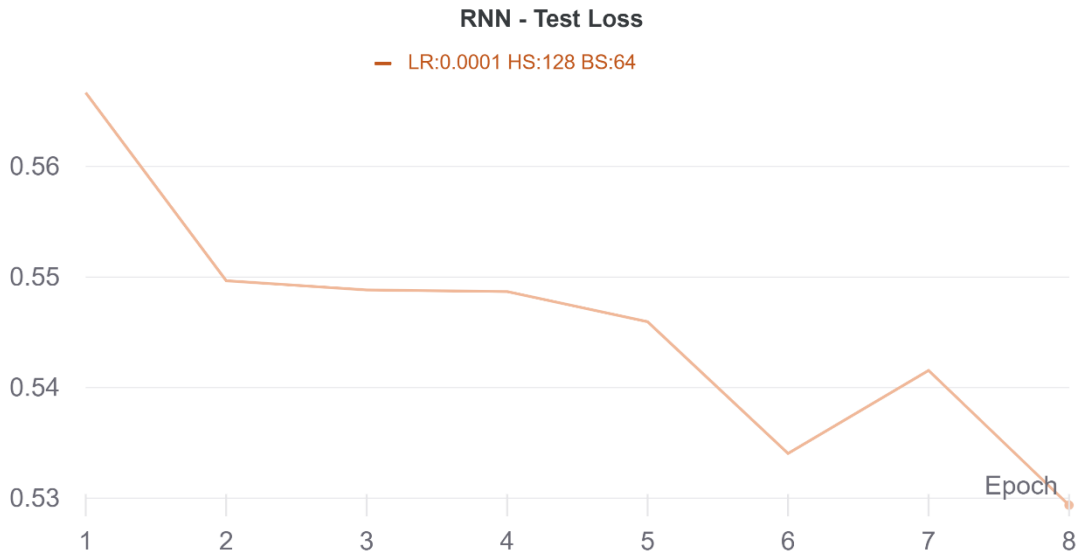
נשים לב, כי הפרמטרים אשר ממקסמים את הדיוק (מגיעים ל-Accuracy 0.7488) הם:

Learning rate = 0.0001
Hidden Size = 128
Batch Size = 64

באופן כללי, במודל זה ניתן לראות התאמה בין Accuracy גבוהה ל-Learning rate נמוך.

נציג את הגרפים של ה-Train Loss וה-Test Loss-



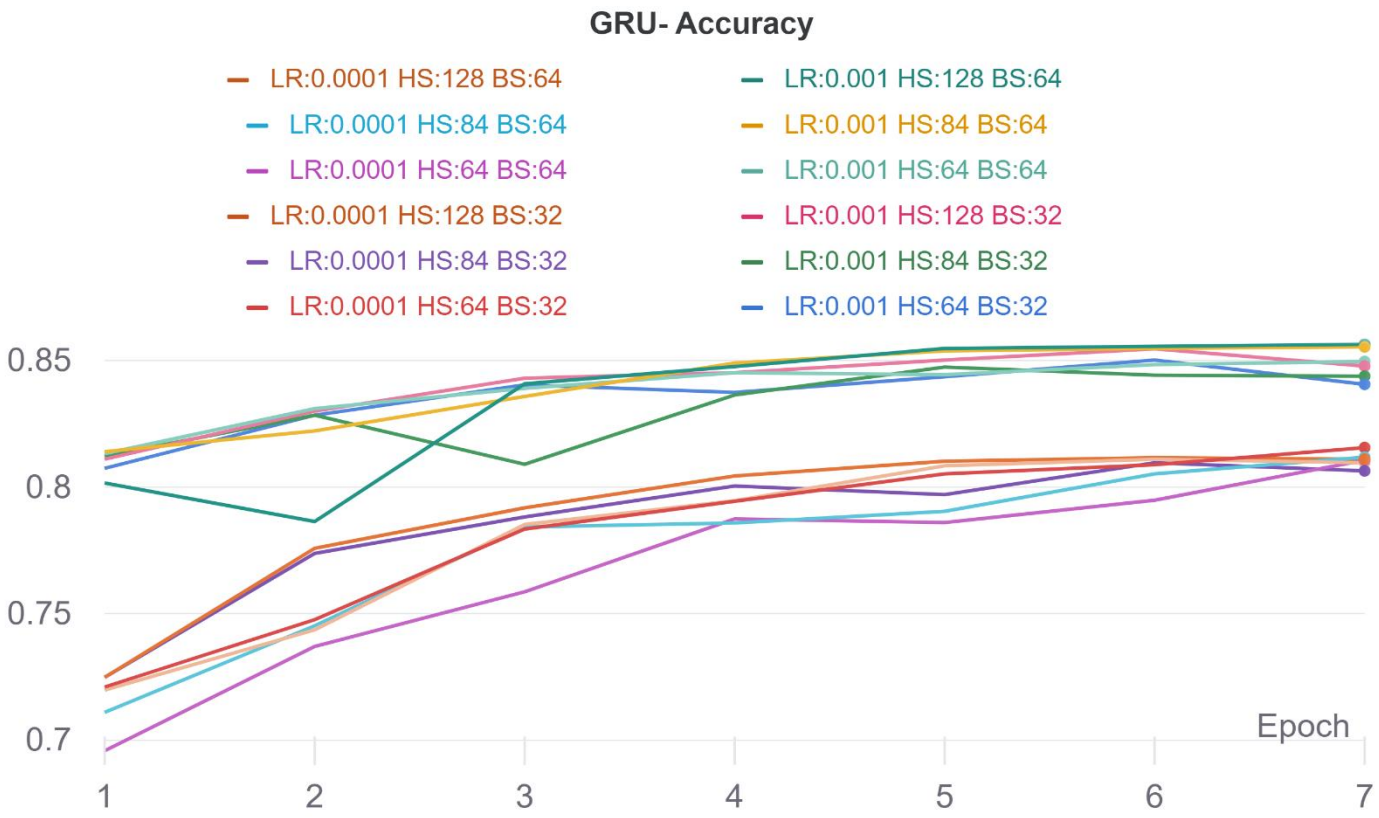


ניתן לראות, כי ה-Train Loss מתכנס וכי לא נגרם Overfit.
לבסוף ה-Loss אותו קיבלנו למודל הוא:

Train Loss: 0.5327
Test Loss: 0.5294

:GRU

בבניית רשת זו, כללנו ארבע מטריצות נלמדות (כפי שנלמד בכיתה).
בחרנו להשוות את אותם הפרמטרים אשר השוונו ב-RNN ורצינו למקסם את דיוק הרשת.
נציג את תוצאות ההרצות –

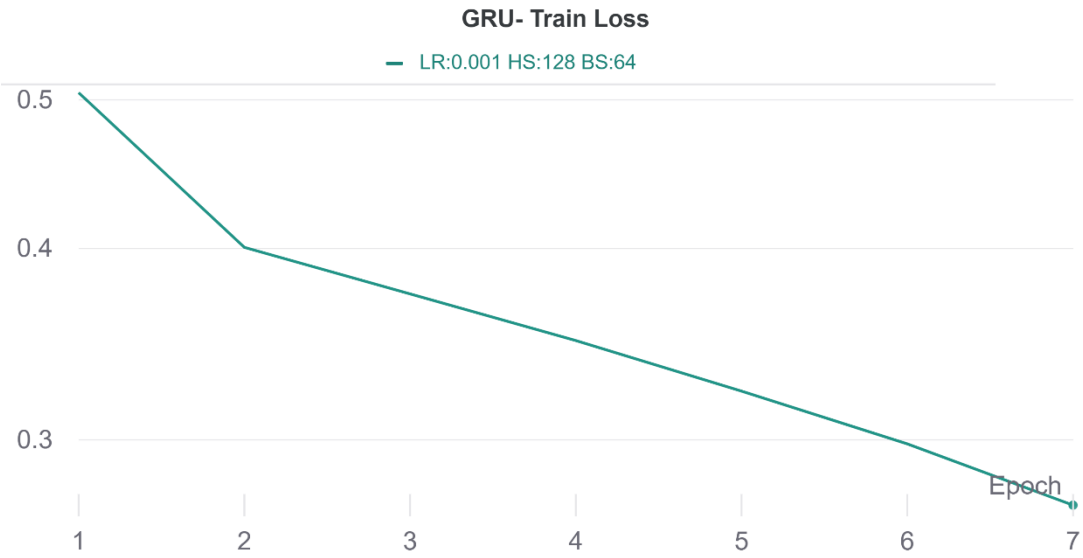


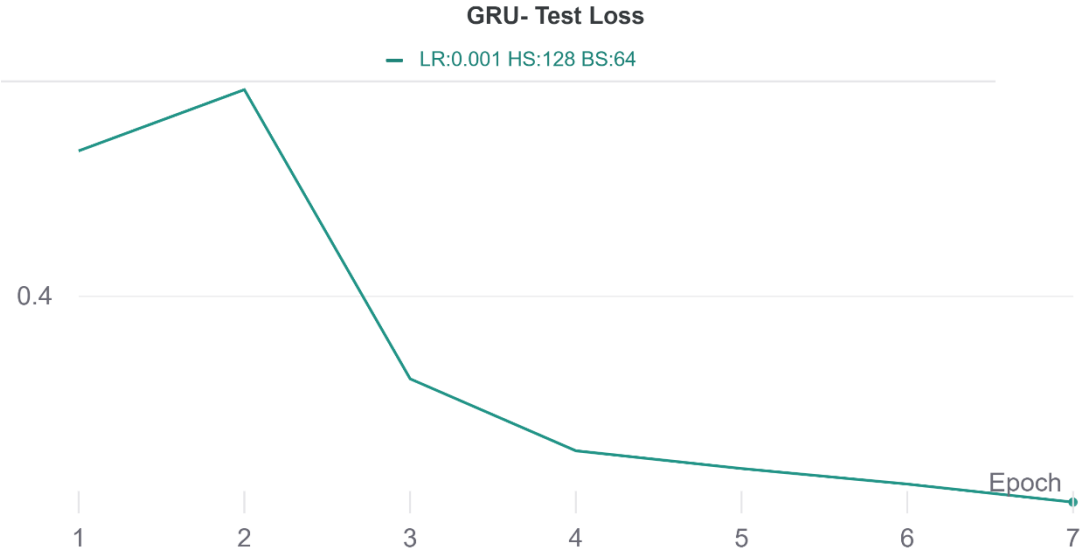
נשים לב, כי הפרמטרים אשר ממקסמים את הדיוק (מגיעים ל-Accuracy 0.8564) הם:

Learning rate = 0.001
Hidden Size = 128
Batch Size = 64

באופן כללי, במודל זה (לעומת ה-RNN) ישנה התאמה בין Accuracy גבוהה ל-Learning rate נמוך יותר.
ניתן לראות, כי אכן מודל זה הינו בעל ביצועים טובים משמעותית מה-RNN שכן מגיע לרמת דיוק גבוהה ב-10%.

נציג את הגרפים של ה-Train Loss וה-Test Loss-





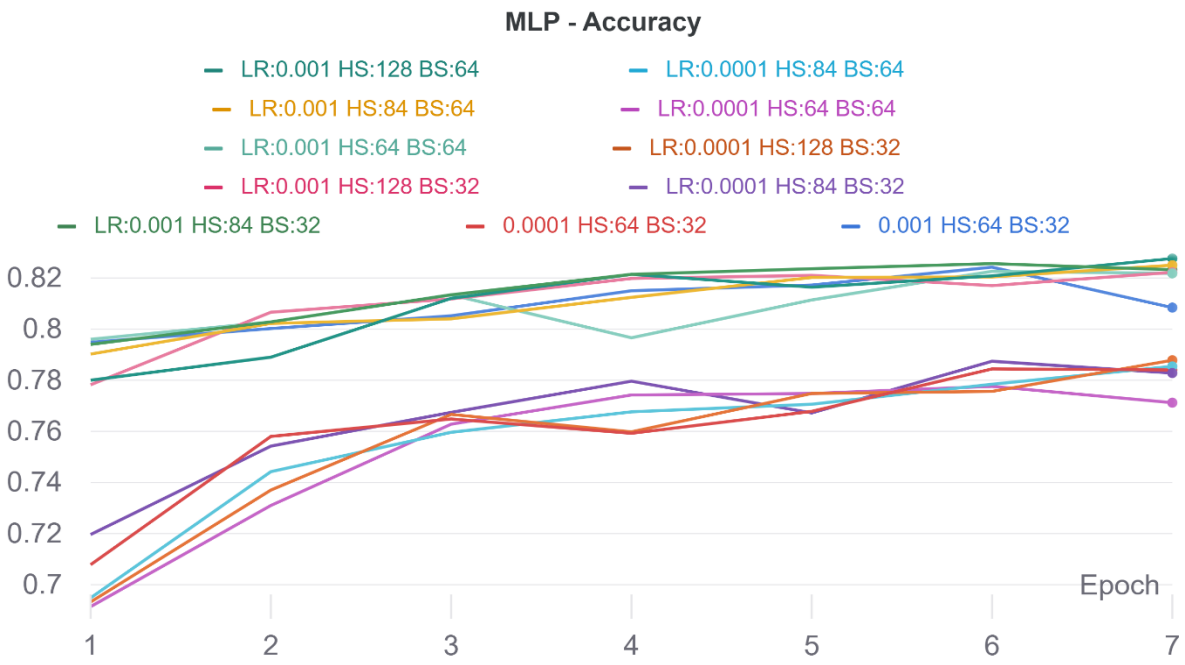
ניתן לראות, כי ה-Train Loss מתכנס וכי לא נגרם Overfit.
לבסוף ה-Loss אותו קיבלנו למודל הוא:

Train Loss: 0.272
Test Loss: 0.3441

שוב, גם במדד זה המודל בעל תוצאות טובות בהרבה מאשר ה-RNN.
השיפור המשמעותי בין המודלים, נובע מכך שמודל ה-GRU פותר את בעיית ה-Vanishing Gradient עליה למדנו בקורס.

2) MLP:

בבניית רשת זו, לא כללנו מספר רב של שכבות נלמדות, כי רצינו להבחין ביתרון אשר נקבל מתוספת שכבת ה-Attention במשימה הבאה.
כללנו 3 מטריצות ברשת זו ופונקציית האקטיבציה Relu בין כל שכבה.
הפרמטרים אותם רצינו ללמוד כדי למקסם את דיוק המודל זהים לפרמטרים הקודמים.
נציג את תוצאות ההרצות –

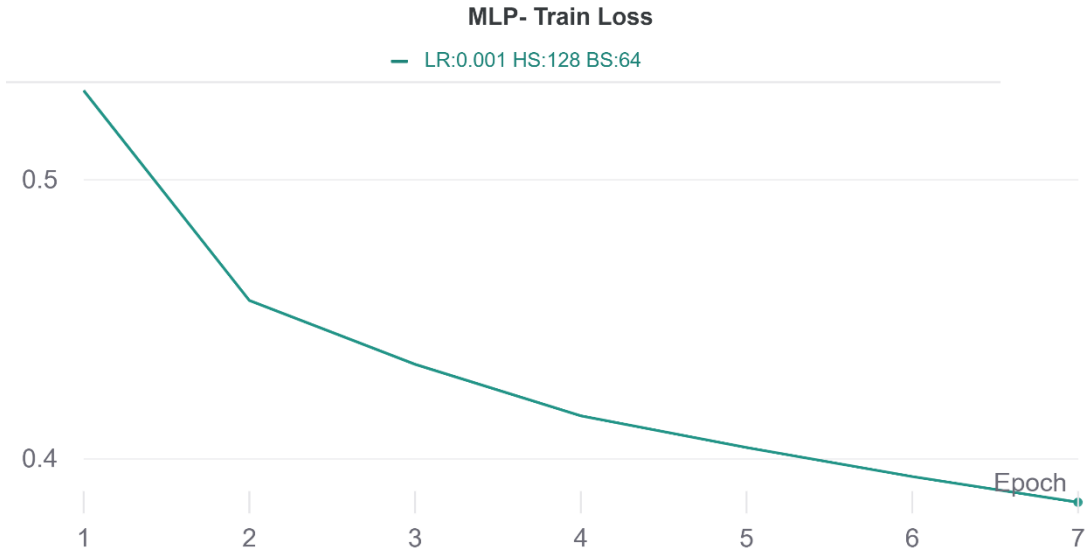


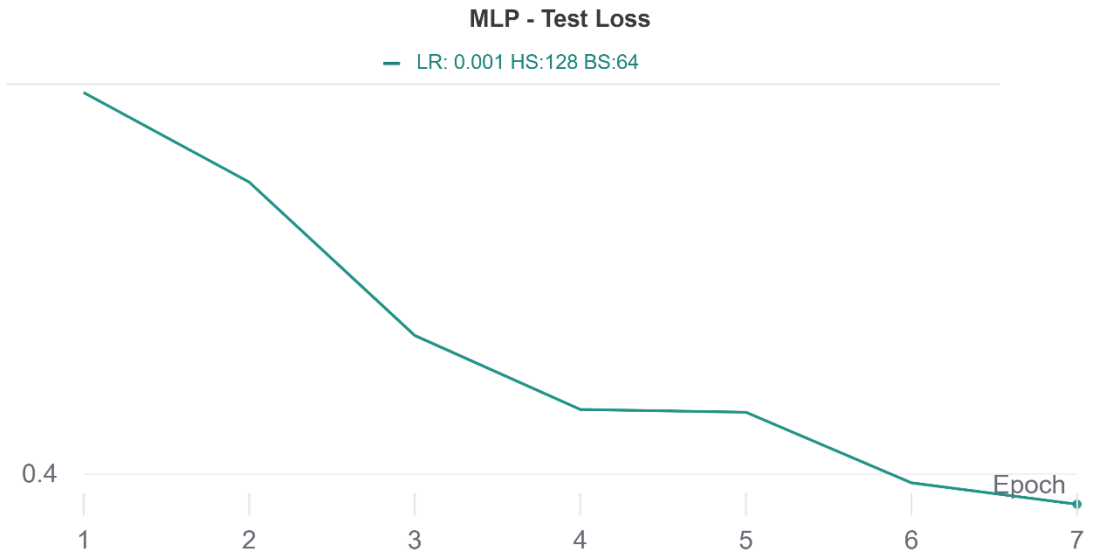
בדומה ל-RNN, ניתן לראות בבירור כי הפרמטר המשפיע ביותר הוא ה- Learning rate. ישנה התאמה בין Learning rate נמוך ל-Accuracy גבוה.

נשים לב, כי הפרמטרים אשר ממקסמים את הדיוק (מגיעים ל-Accuracy 0.8276) הם:

Learning rate = 0.001
Hidden Size = 128
Batch Size = 64

נציג את הגרפים של ה-Train Loss וה-Test Loss-





ניתן לראות, כי ה-Train Loss מתכנס וכי לא נגרם Overfit.
לבסוף ה-Loss אותו קיבלנו למודל הוא:

Train Loss: 0.3864
Test Loss: 0.3947

כעת נציג שתי ביקורות ואת תוצאות המודל עליהן (לכל מילה בנפרד ולביקורת כולה):

```
Model: MLP
Review: great movie the best ever saw the actors were amazing if you like action movies you have to go see this movie classic
With the label:    positive
Was labeld as    positive
word: great    with sub-scores: [23.280099868774414, -23.92298698425293] was labeled as: positive
word: movie    with sub-scores: [-2.3427553176879883, 1.1890686750411987] was labeled as: negative
word: the    with sub-scores: [0.08670783787965775, -0.020285658538341522] was labeled as: positive
word: best    with sub-scores: [22.167131423950195, -22.590938568115234] was labeled as: positive
word: ever    with sub-scores: [4.448002338409424, -3.987426519393921] was labeled as: positive
word: saw    with sub-scores: [2.9792354106903076, -3.295314311981201] was labeled as: positive
word: the    with sub-scores: [0.08670783787965775, -0.020285658538341522] was labeled as: positive
word: actors    with sub-scores: [-11.738157272338867, 9.782042503356934] was labeled as: negative
word: were    with sub-scores: [-3.429141044616699, 4.649611473083496] was labeled as: negative
word: amazing    with sub-scores: [45.759910583496094, -45.584503173828125] was labeled as: positive
word: if    with sub-scores: [-3.5256032943725586, 4.134836673736572] was labeled as: negative
word: you    with sub-scores: [4.158991813659668, -4.094771385192871] was labeled as: positive
word: like    with sub-scores: [-2.6577186584472656, 2.6851723194122314] was labeled as: negative
word: action    with sub-scores: [0.9762357473373413, -0.8259413242340088] was labeled as: positive
word: movies    with sub-scores: [0.2952585816383362, -1.6163772344589233] was labeled as: positive
word: you    with sub-scores: [4.158991813659668, -4.094771385192871] was labeled as: positive
word: have    with sub-scores: [-3.56712007522583, 4.112929821014404] was labeled as: negative
word: to    with sub-scores: [-0.2185843139886856, 0.11393018811941147] was labeled as: negative
word: go    with sub-scores: [-2.367783308029175, 2.506488800048828] was labeled as: negative
word: see    with sub-scores: [2.3574788570404053, -2.9372901916503906] was labeled as: positive
word: this    with sub-scores: [-0.8611427545547485, 0.043915409594774246] was labeled as: negative
word: movie    with sub-scores: [-2.3427553176879883, 1.1890686750411987] was labeled as: negative
word: classic    with sub-scores: [13.839287757873535, -15.184500694274902] was labeled as: positive

Model: MLP
Review: not bad you should watch it liked it
With the label:    positive
Was labeld as    negative
word: not    with sub-scores: [-4.606049537658691, 5.322415351867676] was labeled as: negative
word: bad    with sub-scores: [-38.599605560302734, 39.24014663696289] was labeled as: negative
word: you    with sub-scores: [4.158991813659668, -4.094771385192871] was labeled as: positive
word: should    with sub-scores: [-4.25764799118042, 4.362118721008301] was labeled as: negative
word: watch    with sub-scores: [0.5240914225578308, -1.4072353839874268] was labeled as: positive
word: it    with sub-scores: [3.1495680809020996, -4.620395183563232] was labeled as: positive
word: liked    with sub-scores: [10.897064208984375, -11.573066711425781] was labeled as: positive
word: it    with sub-scores: [3.1495680809020996, -4.620395183563232] was labeled as: positive
```

ניתן לראות, כי על הביקורת הראשונה, המודל אכן צודק. הוא קיבל ביקורת חיובית, לא מבלבלת (מבחינת הקשר מילים) והצליח לזהות כי מדובר בביקורת חיובית.

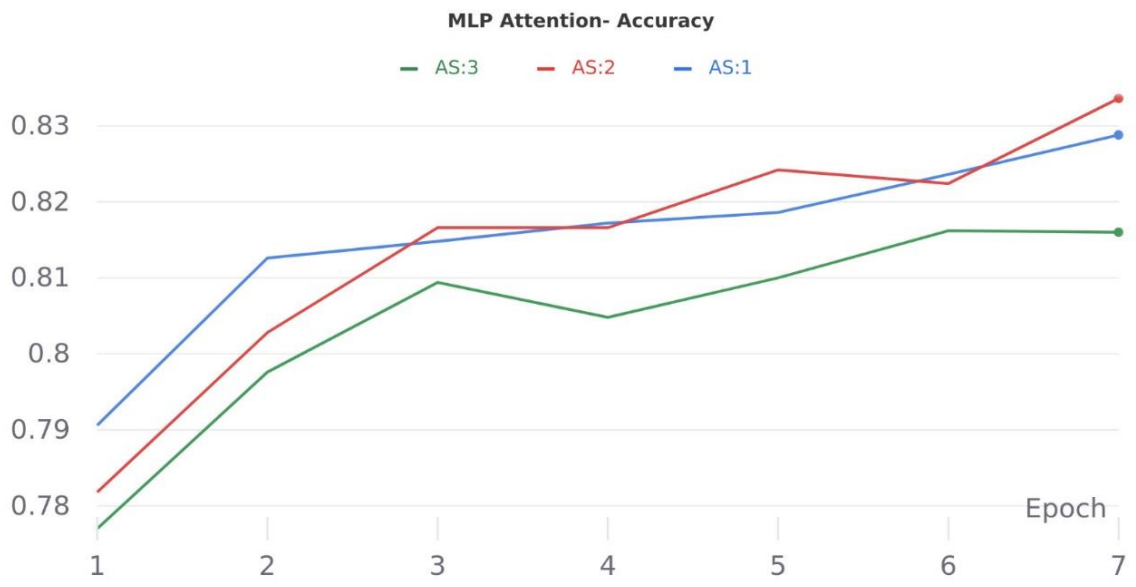
המודל אכן זיהה מילים חיוביות כמו great, best, amazing ותייג אותן עם ערכים גבוהים מאוד אשר השפיעו על הפלט הסופי.

אמנם, על הביקורת השנייה, המודל טועה. המילה "bad" משליכה על תיוג המשפט והמודל לא מזהה / מקשר את המילה "not" המגיעה לפניה. ניתן לראות, כי התיוג של המילה bad כה נמוך, עד כדי כך שהמשך משפט כמו: I liked it לא משפיע על התיוג הסופי.

3) <u>MLP Attention:</u>

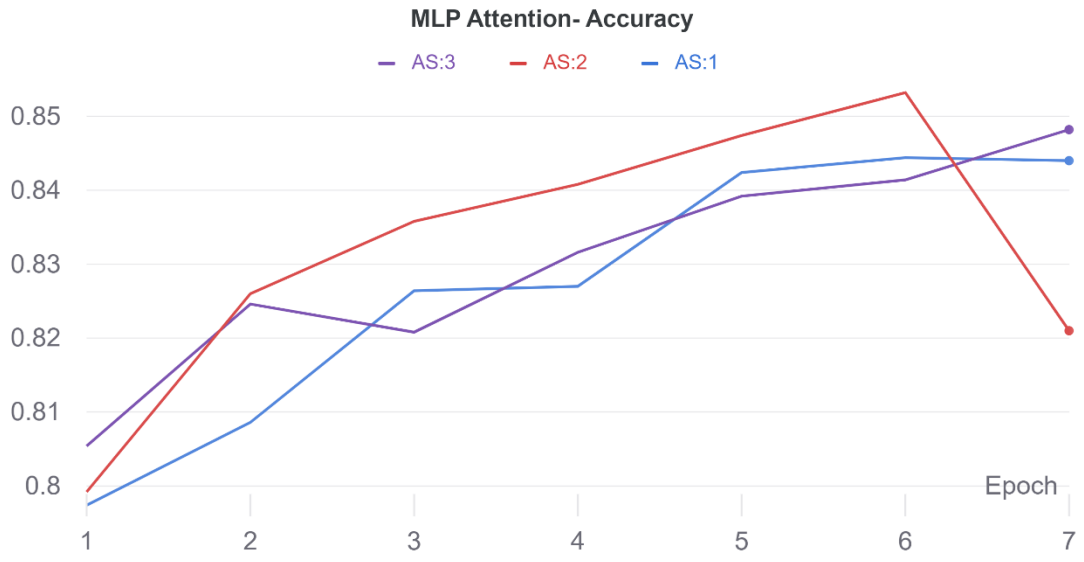
תחילה נציין, כי את שכבת ה-Attention הוספנו למודל ה-MLP המוצלח ביותר אותו בנינו בסעיף הקודם, שכן רצינו לראות את ההבדל שהשכבה מוסיפה.
למודל זה ביצענו 2 בדיקות, הראשונה הייתה מיקום השכבה בתוך הרשת. בחנו את האופציה למקם את השכבה בתחילת המודל ובאמצע המודל.
הבדיקה השנייה הייתה גודל ה-Attention size, בחנו אפשרויות של Attention size בגדלים 1-3 (שמנו לב כי החל מגודל 4 ביצועי הרשת נפגעים).
נציג את דיוק המודלים לפי הבדיקות.

תחילה נציג את גרף המודל בו השכבה הייתה ממוקמת בתחילת הרשת:

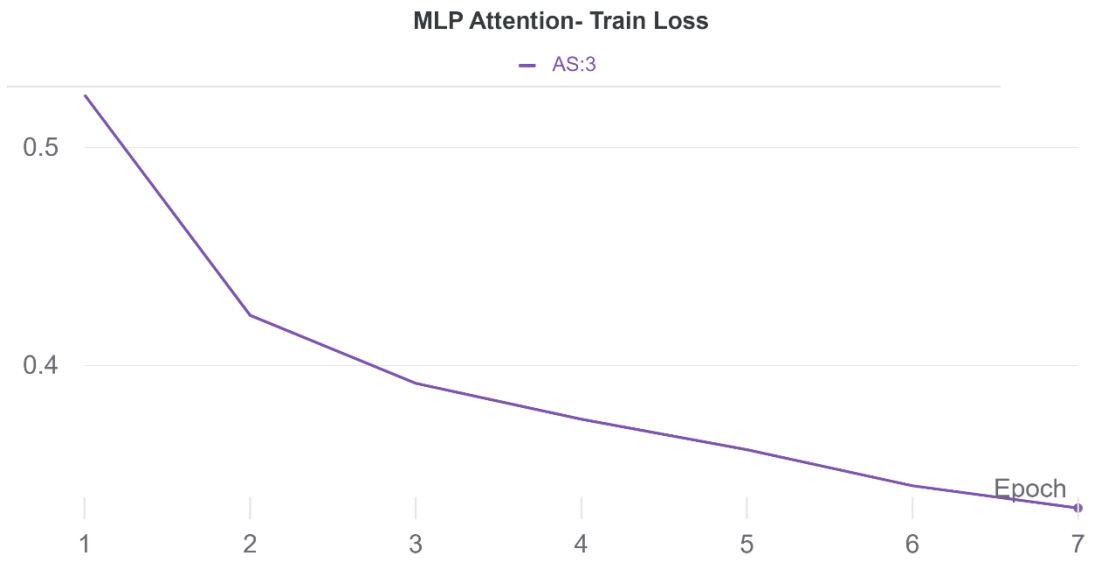


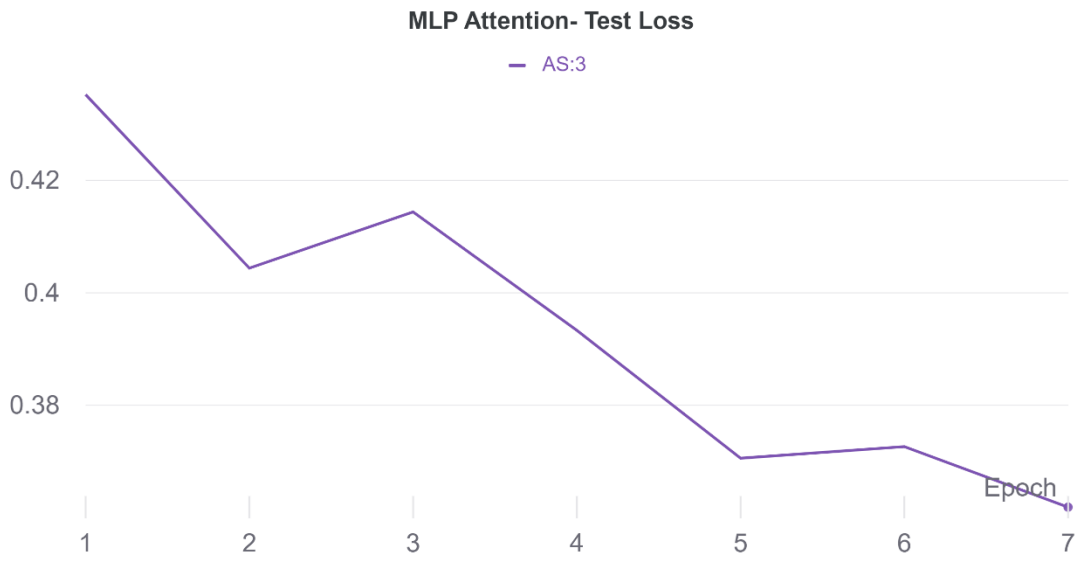
הדיוק הגבוה ביותר אותו הרשת השיגה הוא 0.8336, וזאת כאשר הרשת כללה Attention size בגודל 2.

כעת נציג את גרף הדיוק כאשר מיקמנו את שכבת ה-Attention באמצע ה-MLP
(לאחר הפעלת השכבה הראשונה ופונקציית Relu):



ניתן לראות כי כעת, הדיוק הגבוה ביותר אותו הרשת השיגה הוא 0.8488, וזאת כאשר הרשת כללה Attention size בגודל 3.
באופן כללי, גילינו כי מיקום שכבת ה-Attention באמצע הרשת, משפר את ביצועיה.
בנוסף, מודל זה, השיג שיפור של מעל ל-2% מרשת ה-MLP הרגילה.
נציג כעת את ה-Train Loss וה-Test Loss של המודל-





ניתן לראות, כי ה-Train Loss מתכנס וכי לא נגרם Overfit.
לבסוף ה-Loss אותו קיבלנו למודל הוא:

Train Loss: 0.3456
Test Loss: 0.3619

אכן, שכבת ה-Attention עוזרת למודל ונדגים זאת כעת על אותם 2 משפטים שבחנו את מודל ה-MLP עליהם ונראה את ההבדלים:

```
Model: MLP_atten
Review: great movie the best ever saw the actors were amazing if you like action movies you have to go see this movie classic
With the label:    positive
Was labeld as    positive
word: great    with sub-scores: [9.198075294494629, -15.057229042053223] was labeled as: positive
word: movie    with sub-scores: [10.768322944641113, -17.415424346923828] was labeled as: positive
word: the    with sub-scores: [11.625757217407227, -18.349143981933594] was labeled as: positive
word: best    with sub-scores: [11.47944450378418, -17.975053787231445] was labeled as: positive
word: ever    with sub-scores: [7.932000160217285, -12.915009498596191] was labeled as: positive
word: saw    with sub-scores: [4.400181770324707, -7.463406562805176] was labeled as: positive
word: the    with sub-scores: [12.09378719329834, -19.201013565063477] was labeled as: positive
word: actors    with sub-scores: [5.523909568786621, -7.327303886413574] was labeled as: positive
word: were    with sub-scores: [4.763569355010986, -5.830705642700195] was labeled as: positive
word: amazing    with sub-scores: [3.461582899093628, -2.5658316612243652] was labeled as: positive
word: if    with sub-scores: [3.938483238220215, -2.283785343170166] was labeled as: positive
word: you    with sub-scores: [3.332749843597412, -4.403890609741211] was labeled as: positive
word: like    with sub-scores: [4.765642166137695, -8.530525207519531] was labeled as: positive
word: action    with sub-scores: [0.20870034396648407, 1.2862675189971924] was labeled as: negative
word: movies    with sub-scores: [0.20206326246261597, 0.021390853449702263] was labeled as: positive
word: you    with sub-scores: [0.2707158923149109, 0.9155313968658447] was labeled as: negative
word: have    with sub-scores: [0.29231181740760803, -0.6671969890594482] was labeled as: positive
word: to    with sub-scores: [0.6072117686271667, -0.9501044154167175] was labeled as: positive
word: go    with sub-scores: [1.5161367654800415, -0.6970281004905701] was labeled as: positive
word: see    with sub-scores: [2.5651609897613525, -4.6389946937561035] was labeled as: positive
word: this    with sub-scores: [3.9950215816497803, -7.193515777587891] was labeled as: positive
word: movie    with sub-scores: [4.825226306915283, -8.377191543579102] was labeled as: positive
word: classic    with sub-scores: [4.95777702331543, -8.979754447937012] was labeled as: positive

Model: MLP_atten
Review: not bad you should watch it liked it
With the label:    positive
Was labeld as    positive
word: not    with sub-scores: [7.869879245758057, 12.281719207763672] was labeled as: negative
word: bad    with sub-scores: [6.430748462677002, 12.522133827209473] was labeled as: negative
word: you    with sub-scores: [5.029680252075195, 11.920699119567871] was labeled as: negative
word: should    with sub-scores: [2.8423798084259033, 7.463054180145264] was labeled as: negative
word: watch    with sub-scores: [-0.1892348825931549, 7.391153812408447] was labeled as: negative
word: it    with sub-scores: [7.152981758117676, -11.651902198791504] was labeled as: positive
word: liked    with sub-scores: [6.5604095458984375, -10.592193603515625] was labeled as: positive
word: it    with sub-scores: [7.495264053344727, -12.092079162597656] was labeled as: positive
```

ניתן לראות, כי מודל זה צודק על שתי הביקורת (הזהות לביקורת אותן בחנו את מודל ה-MLP שלנו).
ההבדל המשמעותי הגרם לרשת זו, לזהות כי הביקורת השנייה חיובית היה השימוש בשכבת ה-Attention. ניתן לראות כי כעת, המילה "bad" עדיין מקבלת תיוג שלילי, אך ערכה נמוך בהרבה והרשת מבינה את הקשר למילה "not" המופיעה לפניה.

הבדל זה, מדגיש את החיזוק ששכבת ה-Attention תרמה למודל הקודם, ואת חשיבותה לניתוח משימות מסוג זה.

מבוא ללמידה עמוקה - 67822

תרגיל 2

מגישים: אור בוגו 208489732

שי כהן 314997388

1) א. ~~בשאלה~~ ניתוח גיבור -

בבעיה זו נרצה לתרגם אורכו לטקסט. נשים לב כי זו בעיה מסוג many to many שבן לאו נרצה את מספר המילים באורכו.

נפנה לבעיה זו באופן הבא -

נבנה מודל אשר בראשיתו יכיל ~~שלמה~~ רשת קונבולוציה. מאחר וזהו קובץ אורכו, המידע אי קלט, זהו מידע אשר מקיים את תכונת ה- LTI ולכן הכניסה ~~בשלב~~ ברשת קונבולוציה.

מבחינה אינטואיטיבית, משמעות מיקום השכבה תהיה לרפרזנט בין מילים וכן לתת לכן יחוס שונה. לאחר מכן, המודל יכיל שכבת Attention שתוביל לרשת GRU. בחירה זו ~~ב~~ נובעת מכך שיש חשיבות לרקשר המילה כי שנוכל לתרגם את האורכו לטקסט המתאים. הקשר חשוב במקרה זה, שכן יש מילים שונות שנשמעות אותו דבר. (לדוגמה- (hight, knight).

כאמור, לבסוף נניח את כלל ה- Attention ל- GRU מודל זה יבלע לפעור בעיות many to many בעיקר מתכונת ה- GRU לא יכיל את בעיות ה- Vanishing Gradient.

ב. תשובות לשאלות

את בעיה זו פתרו באמצעות רשת Transformer
בדיה זו אוטוריגרסיבית & היא בעלת מבנה many to many
שמן לא נמצא את אורך השפה ומן את אורך
בתשובה.

מובן ה-Transformer מתאים לבעיה מסוג
זה, בעצם הוא יכול לעצם טוב את השאלה
ואת הקשר בין המילים השונות. עצם באמצעות
שכבת ה-Multi Headed Attention אותה הוא מכיל.

ג. סווג משפט-

באופן דומה לסעיף הקודם, גם כאן אנחנו
יכולים מהו אורך הקלט. נמרר שוב להשתמש
ברשת Transformer. רשת זו תדע הוצב את
המשפט, יבין קשר בין המילים השונות ואמורה
לסווג מן סיווגים שבחרים.

ד. סיווג תמונה-

זו משימה קלאסית לרשת קונבולוציה. קלט מסוג
תמונה הינו LTI. רשת קונבולוציה תבין ב
באמצעות קונבולוציות נלמדות מצבעת פינות, פרטים
וצבעים בתמונה ולסווגה לפיהם.

ה. תירגום מילה-

בבעיה זו אורך הקלט קבוע, לרשת אין
מנגנון בנוכחות ובעיות Attention, לכן נשתמש
לבעיה זו ב בעזרת MLP רגיל.

2) אם A מטריצה ו- x וקטור משתנים.

נתון שמן-

$$x=\begin{bmatrix}x_1\\ \vdots\\ x_n\end{bmatrix} \qquad A=\begin{bmatrix}a_1^1 & \cdots & a_1^n\\ \vdots & & \\ a_m^1 & \cdots & a_m^n\end{bmatrix} \qquad n,m\in\mathbb{N}$$

נחשב את $J(Ax)$.

$$\forall_{\substack{i\in[n]\\ j\in[m]}} \quad J_{i,j}(Ax)=\frac{\partial (Ax)_j}{\partial x_j}=$$

$$=\frac{\partial\left(\sum_{k=1}^{n} a_j^k x_k\right)}{\partial (x_j)}=a_j^i$$

לכן סה"כ $\boxed{J(Ax)=A}$.

ב. יהיו $x,b\in\mathbb{R}^n$ נתון-

$$x=\begin{bmatrix}x_1\\ \vdots\\ x_n\end{bmatrix},\quad b=\begin{bmatrix}b_1\\ \vdots\\ b_n\end{bmatrix}$$

נחשב את $J(x+b)$-

לכל $i,j\in[n]$

$$J_{i,j}(x+b)=\frac{\partial (x+b)_i}{\partial x_j}=\begin{cases}1 & i=k\\ 0 & \text{else}\end{cases}$$

לכן סה"כ $\boxed{J(x+b)=I_n}$

ג. יהי x וקטור ו- f דרגל דיפרנציאבילי.

ידוע בהרצאה כי לכל פ"ג דיפרנציאבילי בין x ל- f נגזרת של x במטריצה סיבוכיות המורכבת מהפונקציות עבור f. סה"כ נוכל לסמן את הפעולה הדיפרנציאלית כך- $f*x=Fx$ עבור F המטריצה הסיבוכיות המתאימה. מאחר והגזרה כי נהיה נוכל למצוא את $J(fx)$ במקום.

לפי סע' 1 מתקיים. $\boxed{J(f*x)=J(Fx)=F}$

3) אינטואיציה ארכיטקטורה טלית אשר בה נמזן משפט תפלוט

תמונה משתמש קבוצת תמונות סבירה.

נפנה לפליה נ ~~[illegible]~~ באופן הבא: נבנה המודל

שלנו יהיה Transformer אשר מקבל תהיה

לעמת את המשפט ולייצג בייצוג מתאים לפי

הקשרי המילים. שנבה נ תהיה מזין Encoder

למשפט.

את פלט ה- Transformer נכניס לרשת GRU

אשר מקבלת תהיה לייצר קידוד נכון לתמונת

הפלט, (אם הייתה שינתן לקודד ולפענח תמונות).

את הרשת נאמן כדי לקודד תמונות ממשפטים

בצורה הטובה ביותר. מוקל ה- GRU ושתמש

בתכונת הזכרון שלו ויש לבקש עמ בצורה טובה.

ב. נלמד מאמר והתמונה מקודדת ל-4 איברים נכנס

לפענוח ל- GRU שבתוך Attention. נצירה

באופן הבא:

הי queries יהיו הדברים, וישלוי לאחר אשר

הם המילה הנוכחית משפחה על איבר נה.

ה- keys וה- values שלו יהיו פלט ה- hidden state

של ~~[illegible]~~ ה- GRU.

באופן נה נקבל ייצוג ~~[illegible]~~ תכון לפרטים ונוכל

לחלקם לאחר מכן איברי התמונה.

ע"ן אז בכל שכבה ברשת הקונבולוציה, אם נסמן
את גודל הפלטר ב- m ואת גודל ה- stride
m, מתקיים שאחרי הפעלת הקונבולוציה, גודל
הפלט יהיה- $m-(n-1) = \underline{m+1-n}$

לכן לאחר השכבה הראשונה נקבל-

$128+1-3 = 126$

בשנייה- $126+1-3 = 124$

באופן דומה בשכבת ה- maxpool-

$124+1-2 = 123$

בשלישית- $123+1-3 = 121$

ולבסוף- $121+1-5 = 117$

סה"כ גודל פלט הרשת יהיה $\boxed{117\times117}$

ב. נחשב את גודל ה- receptive field
נחשב זאת מהסוף להתחלה. בשכבה האחרונה, גודל
הפלטר הינו 5x5 ולכן לכל פיקסל בתשובה 5x5
פיקסלים מעליו.
בשכבה שמעליו, גודל הקרנל הינו 3x3 ולכן
מאחר וכדאי לספר בפיקסלים אותם הוא כבר לקח
בחשבון, התוספת היחידה מגיעה מהקרנל של
מוסף מלון "מסביב" ברוחב 1 לכל כיוון.
לכן גודל ה- receptive field של- 7x7.
באופן דומה שכבת ה- max pool תגדל את
השדה ל- 8x8 ושתי השכבות הראשונות יגדלו
את ה- receptive field ל- $\boxed{12\times12}$.

ג. נראה שיטה להורדת חשיבות של ט איטר בתמונה לחיצי הסופי של רשת הר נאט.

בנוסף, הראנו בסעיף הקודם, ט ט (נירון) בפלט מושפע מאיטר של 12x12 פיקסלים בתמונה המקורית. כדי להוריד מבחן השפעה של ט פיקסל נשתמש בגרדיאנט של השכבה האחרונה.

מחשב גרדיאנט זה ושמור עבור כל נירון את גודל הגרדיאנט שלו. כמה לט, בהינתן פיקסל במקור נגיש אותו ודגים אולו נירונים מושפעים ממנו, נסנן את הגרדיאנטים של הנירונים המתאימים וכך נוריד את מבחן חשיבותו עבור השכבה הספציפית.

באופן טבעי, בהינתן איטר בתמונה המקורית, נחזור על פעולה זו לט פיקסל באיטר ונסכום את הגרדיאנטים כמתואר וכך נוריד את חשיבותו.