דו"ח מעבדה – הגשה 2

מגישים:

- אלמוג בן סימון 206571135
- באנא סעדי 206611477

לינק לגיט: https://github.com/almog2065/Rome-Digits.git

מטרה:

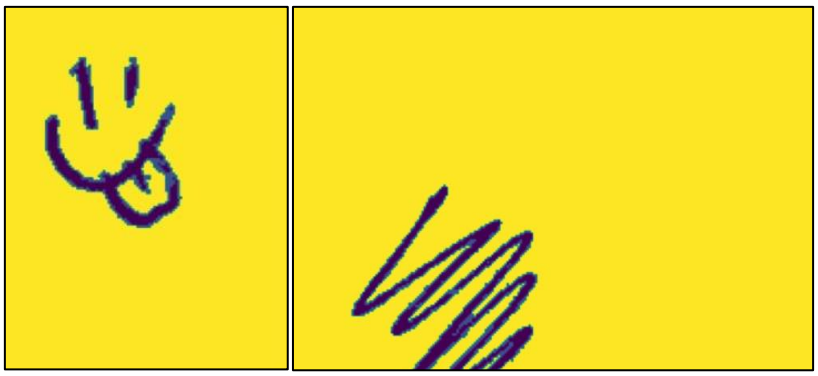
מטרתנו בתרגיל זה היא ליצור סט אימון המהווה מדגם מייצג ומאוזן לספרות רומיות. הסט יהיה מייצג במובן הזה שיכלול מגוון גרסאות אפשריות לתמונות שייתכן ונתקל בהם בשלב הבדיקה, והוא יהיה מאוזן במובן הזה שלכל לייבל יש מספר כמעט זהה של תמונות בסט באימון.

תיאור תהליך עבודה:

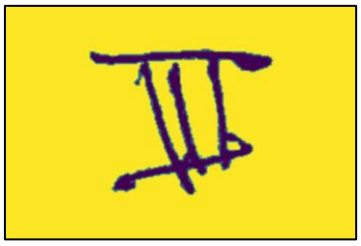
במהלך בניית מסד הנתונים אנו מניחים שהתמונות בסט הבדיקה (test set) מגיעות מאותה התפלגות כמו התמונות שקיבלנו בסט האימון. לכן מעבר ראשוני על התמונות היה חיוני ובכך התחלנו הן בשביל ניקוי וסידור הדאטה והן בשביל לקבל רעיון כללי על אופי התמונות בכדי שנוכל לבחור אוגמנטציות בצורה יעילה.

-Data Cleaning and relabeling:

עברנו על כל התמונות בכל אחת מהתיקיות ומחקנו כאלה שלא תואמות אף תיוג באופן מובהק כגון:

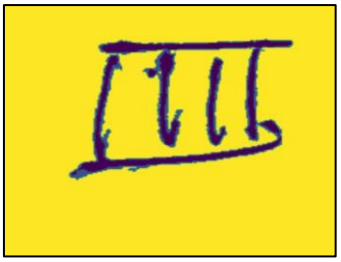


בנוסף, תייגנו מחדש חלק מהתמונות שלא התאימו לתיוג המקורי, למשל התמונה הבאה הייתה תחת התיקייה vii כאשר התיוג הנכון הוא iii:

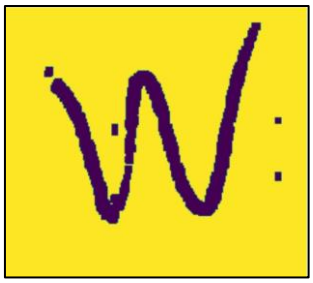


נציין שהתיוג מחדש היה תהליך מורכב יחסית. חלק מהתמונות היה קשה לשייך לקטוגריה מסויימת, אך השתדלנו לתייג את כל התמונות באופן עקבי ולפי שיקולים זהים. ניתן שתי דוגמאות לתמונות כאלה ונסביר את השיקולים מאחורי התיוג שבחרנו.

דוגמה 1: התמונה הבאה תויגה iv על אף שהיא לא תואמת אף אות רומית. רצינו שהמודל יזהה גם טעויות נפוצות של משתמשים חסרי בקיאות בספרות הרומיות ובכל זאת יצליח לזהות את כוונתם לכן, מתוך הבנה שהמשתמש כנראה התכוון למספר 4 השארנו את התמונה.



דוגמה 2: התמונה הבאה גם סווגה כ iv. שמנו לב שבהרבה מהתמונות הספרות כתובות בצורה מחוברת כך שלמרות שאיכות תמונות כאלה נמוכה ויותר קשה להבחין בספרה הרצויה זה דפוס נפוץ שנרצה שהמודל ילמד.



לאחר שסיימנו שלב זה ישנן 1942 תמונות בקובץ ה train כאשר מחקנו 126 תמונות. נציין שתהליך זה לא נעשה ידנית באופן מלא, כתבנו קוד אשר עבר על כל תמונה וסידר את התמונות בקבצים לפי הקלט שהבאנו לנו או לחילופין מחק את התמונה במידה והזנו d.

להלן חלוקת התמונות בכל קובץ:

```
The directory 'hw2_094295/data/train/i' contains 229 files.
The directory 'hw2_094295/data/train/ii' contains 146 files.
The directory 'hw2_094295/data/train/iii' contains 162 files.
The directory 'hw2_094295/data/train/iv' contains 271 files.
The directory 'hw2_094295/data/train/ix' contains 214 files.
The directory 'hw2_094295/data/train/v' contains 185 files.
The directory 'hw2_094295/data/train/vi' contains 184 files.
The directory 'hw2_094295/data/train/vii' contains 186 files.
The directory 'hw2_094295/data/train/viii' contains 180 files.
The directory 'hw2_094295/data/train/x' contains 185 files.
```

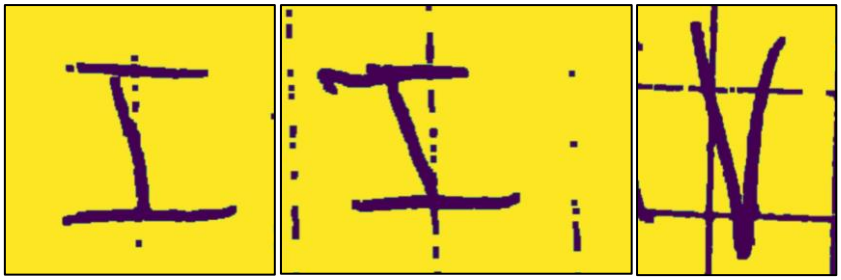
- **אוגמנטציות:**

סיימנו את השלב הקודם עם יחסית מעט תמונות. בשלב זה רצינו להעשיר את סט הנתונים באמצעות אוגמנטציות שונות על הדאטה. האוגמנטציות נבחרו לפי דפוסים שצפינו בהם בסט הנתון בנוסף לקחנו בחשבון שיקולים הגיוניים המלווים תהליכי כתיבה וסריקת ספרות.

בחרנו ליצור תמונות חדשות באופנים הבאים:

1. **הוספת רעש:**

שיטה זו תעזור לנו לסווג תמונות גם במקרה שיש סביב הספרות סימנים לא רצויים. בהשארת דוגמאות הקיימות בקבצי האימון (ניתן לראות דוגמאות למטה), בחרנו בשני סוגי הוספת רעש - הוספת נקודות והוספת קווים.

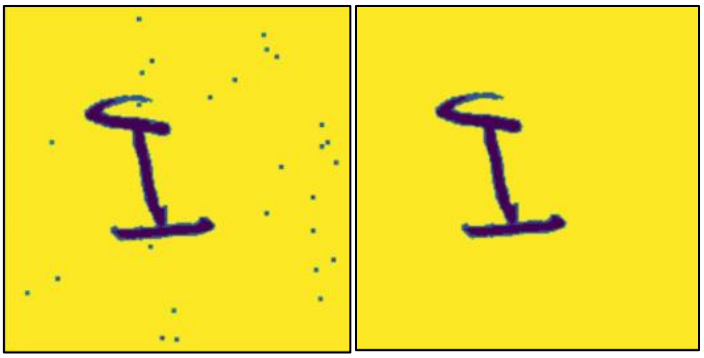


בתמונות מעל אנחנו רואים קווים אופקיים ואנכיים, קווים מקווקים ונקודות מעל הספרות הרומיות. הוספנו את הרעשים באופן שתדמה כמה שיותר דברים שראינו בדוגמאות.

נסביר על אופן הוספת הרעש:

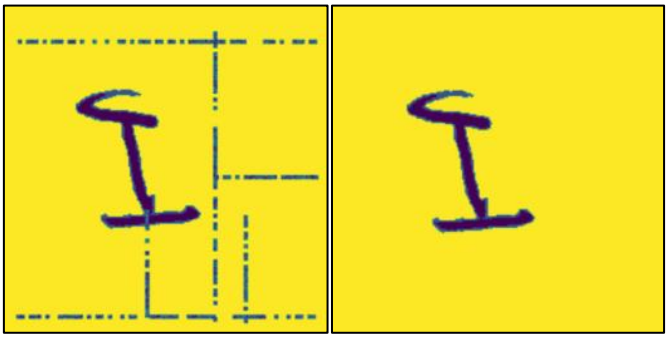
א. **הוספת נקודות:** רצינו להוסיף את הרעש בצורה אקראית ובכמויות משתנות בהתאם לדוגמאות שראינו. הגרלנו רנדומלית אחוז פיקסלים בין 0 ל 0.02% אותם נרצה להפוך לרעש. לכל פקסל הגרלנו מיקום ועל מנת שהנקודה תהיה מספיק בולטת צבענו סביב המיקום המוגרל בלוק 3 על 3 בצבע רנדומלי בין 0 (שחור) ל0.5 (אפור).

דוגמה: בצד ימין אנו רואים את התמונה המקורית מסט האימון. מספר הנקודות שהוגרלו והתווספו לתמונה הוא 28.



ב. **הוספת קווים:** רצינו להוסיף בצורה אקראית קווים אנכיים, אופקיים ולפעמים מקווקוים. ראשית, הגרלנו מספר הקווים האנכיים (בין 0 ל- 3) ומספר קווים אופקיים (בין 0 ל- 3), דאגנו שההגרלה לא תהיה (0,0). לאחר מכן לכל קו הגרלנו את מיקום התחלתו (מספר מ 1 עד רוחב או גובה התמונה בהתאם לסוג הקו). הצביעה מתחילה מהפקסל שנבחר בבלוקים של 3 על 3 ועל מנת ליצור קווים מקווקוים, כל פקסל בקו נצבע בהסתברות שליש. בדומה למקרה הקודם לכל נקודה שצבענו הגרלנו צבע בין 0 ל- 0.5.

דוגמה: בצד ימין אנו רואים את הדוגמה המקורית מסט הנתונים, אליה התווספו 3 קווים אנכיים ו 3 אופקיים כפי שרואים בצד שמאל.



2. **הגדלה והקטנה:**

על סמך דוגמאות שראינו בסט הנתונים וגם מההנחה שאנשים עלולים לכתוב בגדלים שונים ולצלם ממרחקים שונים, בחרנו להגדיל ולהקטין חלק מהתמונות בסט הנתונים. דאגנו שמידות ההגדלה וההקטנה תהיינה בטווח סביר שישמור על המידע בתמונה כך שיהיה אפשר לזהות את הספרות.

דוגמה:

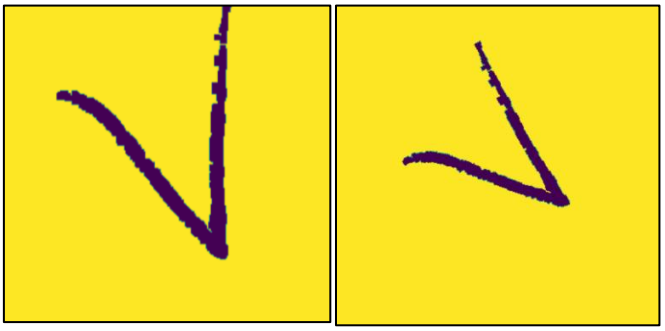
בשורה הראשונה : התמונה הימנית היא התמונה המקורית, האמצעית היא המוגדלת והשמאלית היא המוקטנת.

בשורה השנייה : הגדלה והקטנה מופרזים. אלו מקרים שמהם רצינו להימנע מפני שאי אפשר לזהות את הספרות.

3. **סיבוב:**

שמנו לב שיש תמונות בסט האימון שלא צולמו בזווית ישרה, ולכן בחרנו להוסיף תמונות מסובבות נוספות. ולכל אחת מהן בחרנו דרגת סיבוב הנעה בין -30 ל 30 דרגות. בחרנו את טווח הדרגות הנ"ל מפני שרצינו להגביל את הסיבובים לטווחים שבהם עדיין אפשר לזהות את האות בקלות וגם אלו הטווחים שהבחנו בהם בסט האימון המקורי ומפני שאנו מניחים שסט הטסט מגיע מאותה התפלגות הסתפקנו בדרגות אלו.

להלן דוגמה לסיבוב של 30 מעלות ימינה, התמונה המקורית מוצגת מימין והמסובבת משמאל:



- יצירת תמונות חדשות:

על מנת להגיע למסד נתונים בגודל 10,000, בחרנו להשתמש בשמונה אוגמנטציות המשלבות את חמשת הפונקציות שהצגנו לעיל. יש המון דרכים לשלב את הפונקציות וגם לבחור שמונה שילובים (2^5 בחר 8) לכן בחרנו להסתפק בשני ניסיונות להעשרת הדאטה ולבדוק איזה מהם טוב יותר. לכל אחת מהניסיונות נציג שמונה אוגמנטציות ועבור אוגמנטציות שמשלבות כמה פונקציות נציג דוגמה לתמונה אחרי האוגמנטציה.

<u>הצעה 1:</u>

בניסיון ראשוני הפעלנו כל אוגמנטציה לבדה והוספנו שלושה שילובים.

א. השילוב הראשון כלל הוספת רעש בצורת נקודות לאחר מכן הגדלה וסיבוב. לדוגמה:

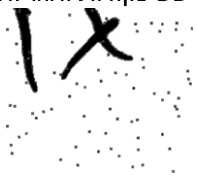
ב. השילוב השני מתחיל בסיבוב הוספת רעש בצורת קווים ולבסוף הקטנה. לדוגמה:

ג. השילוב השני משתמש בכל חמשת הפונקציות בסדר הבא: הקטנה, רעש בצורת נקודות, הקטנה, רעד בצורת קווים וסיבוב. לדוגמה:

הצעה 2:

בניסיון השני רצינו לבדוק אם אוגמנטציות שמשלבות כמה פונקציות נותנות תוצאות יותר טובות לכן בחרנו להשתמש בשמונה שילובים הבאים:

רעש נקודות והגדלה



סיבוב והגדלה

רעש משני סוגים

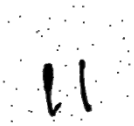
רעש נקודות וסיבוב

רעש קווים וסיבוב

סיבוב והקטנה

רעש קווים והגדלה

רעש נקודות והקטנה



כפי שרואים בתמונות דאגנו שלכל תמונה יתווסף יותר מסוג רעש אחד אך התמונות לא עמוסות מדי.

- בחירת מסד נתונים:

החלטנו לחלק את הדאטה ביחס של 80-20 בין סט האימון לוולדציה אך רצינו לדעת איזה חלוקה טובה יותר וגם איזה אוגמנטציות עדיפות לכן החלטנו לעשות 5-fold cross validation. לתהליך זה יש שני יתרונות עיקריים, ראשית זה יראה לנו אם בהינתן חלוקות שונות של הדאטה ההעשרה שבחרנו תמיד תתן תוצאות פרדקציה טובות. למשל אם נקבל תוצאות מאוד טובות על חלוקה מסויימת של הדאטה ותוצאות מאוד גרועות על שאר החלוקות נסיק שההעשרה לא מספיק טובה מפני שטיב הפרדקציה יהיה תלוי באיזה טסט נקבל. שנית, לאחר שנחליט איזה אוגמנטציה נותנת תוצאות טובות יותר על חלוקות שונות, נרצה לבחור את חלוקת הנתונים שנותנת את תוצאות הפרדקציה הטובות ביותר. כלומר, אם נבחר העשרה שנותנות תוצאות דומות עבור חלוקות שונות אבל בכל זאת חלוקה אחת תתן תוצאות קצת יותר טובות נרצה לבחור בה.

על מנת לחלק את הדאטה ל 5-folds השתמשנו בפונקציה Kfolds לכל fold העברנו 20% מהתמונות המקוריות מסט האימון לסט הוולדציה וזה כי רצינו להעריך את יכולת החיזוי של המודל על תמונות מקוריות הבאות מאותה התפלגות של הטסט ולא על תמונות שיצרנו בעצמנו. לאחר מכן, בחרנו 100 תמונות מכל לייבל בסט האימון באופן אקראי והפעלנו עליהן את אחת האוגמנטציות ושמרנו אותן תחת הלייבל המתאים. כתוצאה מתהליך זה קיבלנו בערך 10,000 תמונות.

כעת עבור כל ניסיון שמשלב 8 אוגמנטציות יש לנו 5 מסדי נתונים כל מסד מכיל fold אחר בוולדציה וסט האימון מכיל את שאר ה folds לאחר העשרה.

- הערכה:

על מנת להעריך את הביצועים של האלגורצם על כל סט נתונים אנחנו בודקים את הדיוק (accuracy) לאחר 10 epochs לכל מסד נתונים ובוחנים גם את ה confusion matrix על מנת לבדוק באיזה ספרות האלגורתם נופל הכי הרבה.

<u>תוצאות עבור הצעה 1:</u>

fold	accuracy	Confusion matrix
1	0.8258	[[49 1 1 0 0 1 0 0 0 0] [4 22 2 1 0 1 1 0 0 0] [4 2 21 0 1 0 2 6 0 1] [0 0 5 45 3 1 1 0 1 0] [0 0 1 0 44 0 0 0 1 1] [0 2 0 1 0 32 2 0 0 2] [0 0 0 0 0 2 33 0 0 1] [0 1 3 1 0 0 3 30 1 0] [0 0 2 1 0 1 0 1 35 0] [1 0 1 0 1 3 0 0 0 30]] [1]
2	0.8675	[[50 0 0 0 1 0 1 0 0 0] [0 23 7 0 1 0 0 0 0 1] [2 0 31 0 1 1 0 1 1 1] [0 1 3 49 1 1 0 0 1 0] [0 0 0 2 43 0 1 0 1 0] [2 0 1 1 0 31 1 0 0 3] [0 0 1 0 0 2 32 0 0 1] [0 0 1 0 0 0 3 33 2 0] [0 0 0 1 3 0 0 1 35 0] [0 1 0 0 2 0 0 0 0 33]]
3	0.8257	[[49 1 1 0 0 1 0 0 0 0] [4 22 2 1 0 1 1 0 0 0] [4 2 21 0 1 0 2 6 0 1] [0 0 5 45 3 1 1 0 1 0] [0 0 1 0 44 0 0 0 1 1] [0 2 0 1 0 32 2 0 0 2] [0 0 0 0 0 2 33 0 0 1] [0 1 3 1 0 0 3 30 1 0] [0 0 2 1 0 1 0 1 35 0] [1 0 1 0 1 3 0 0 0 30]]

[1] Every row corresponds to a real label by the order ['i', 'ii', 'iii', 'iv', 'ix', 'v', 'vi', 'vii', 'viii', 'x'] and each column corresponds to a predict label by the same order in the array above.

[[49 0 0 1 0 2 0 0 0 0] [3 21 4 1 0 1 1 0 0 0] [4 1 28 0 0 1 1 0 1 0] [1 0 2 43 6 1 1 0 2 0] [0 1 0 2 40 1 0 1 2 0] [1 1 0 0 0 34 0 0 0 3] [0 0 0 0 2 1 31 1 0 1] [0 0 1 0 0 1 0 31 5 0] [0 0 1 1 2 0 0 0 36 0] [1 2 0 0 0 0 0 0 1 32]]	0.8394	4
[[44 5 3 0 0 0 0 0 0 0] [2 18 5 0 1 1 0 0 1 3] [0 3 25 1 1 1 1 1 4 0] [0 0 4 46 2 2 1 0 0 1] [1 0 1 6 35 0 0 1 1 1] [0 0 1 0 0 36 0 0 0 2] [0 0 1 2 0 2 28 2 1 0] [0 0 0 0 0 0 0 35 3 0] [0 0 1 0 1 0 1 0 36 0] [1 0 2 0 2 1 0 0 0 29]]	0.8117	5

באופן ככלי קיבלנו תוצאות דיוק קרובות על סט הוולדציה בנוסף לא היה פער גדול בין הדיוק על האימון ועל הוולדציה (קיבלנו בממוצע 0.9 דיוק על האימון).

<u>תוצאות עבור הצעה 2:</u>

Confusion matrix	accuracy	fold
[[42 4 0 0 0 0 0 0 0 0] [1 24 3 1 0 0 1 0 0 0] [2 1 24 2 0 0 1 2 1 0] [0 1 0 54 0 0 0 0 0 0] [0 0 0 1 41 0 0 0 0 1] [0 0 0 1 0 34 0 0 0 2] [0 1 0 0 1 0 35 0 0 0] [0 0 3 0 0 1 1 32 1 0] [0 0 0 1 0 0 0 0 35 0] [3 1 0 0 0 1 0 0 0 32]]	0.9005	1
[[41 2 2 0 0 0 1 0 0 0] [1 25 2 0 0 0 1 0 0 0] [0 1 22 1 0 0 2 2 5 0] [0 0 1 51 2 0 0 0 0 0] [0 0 0 0 40 1 0 0 0 2] [1 0 0 0 0 33 1 1 0 1] [0 0 1 0 0 0 35 0 1 0] [0 0 1 0 0 0 0 32 4 0] [0 0 0 0 0 0 0 0 36 0] [0 0 0 0 0 2 2 0 1 32]]	0.8920	2
[[45 1 0 0 0 0 0 0 0 0] [0 26 1 0 1 1 0 0 0 0] [0 1 27 0 1 0 1 1 1 0] [0 1 3 46 3 1 0 0 0 0] [0 0 1 0 42 0 0 0 0 0] [0 1 0 2 1 33 0 0 0 0] [0 2 0 0 0 0 35 0 0 0] [0 0 0 0 0 0 0 3 34 0] [0 0 0 0 0 0 0 3 33 0] [0 0 0 0 1 0 0 0 0 36]]	0.9201	3
[[45 1 0 0 0 0 0 0 0 0] [3 25 0 0 0 0 1 0 0 0] [0 0 30 0 0 1 0 0 0 0] [0 0 4 48 1 0 0 0 0 1] [0 0 0 2 39 0 0 1 0 1] [0 0 0 1 0 34 0 0 0 2] [0 0 1 0 0 0 32 2 0 2] [0 0 4 0 0 0 1 29 3 0] [0 0 0 0 0 0 0 3 33 0] [0 0 0 2 0 1 1 0 0 33]]	0.8992	4
[[44 0 1 0 0 0 0 0 0 0] [3 22 1 0 0 1 1 0 0 1] [0 1 28 1 1 0 0 1 0 0] [0 0 1 50 1 1 0 1 0 0] [0 0 0 1 40 0 0 0 0 1] [0 0 0 0 0 35 2 0 0 0] [0 4 1 0 0 2 28 1 0 0] [0 0 1 0 0 0 0 35 1 0] [0 0 0 0 0 0 0 1 34 0] [1 0 0 0 1 0 0 1 0 34]]	0.9115	5

ניתן לראות שבניסיון זה קיבלנו דיוק יותר טוב בכל ה-folds ושוב התוצאות קרובות ב folds השונים. כמו כן, קיבלנו בממוצע דיוק יותר גבוה על סט האימון מהניסיון הראשון, 0.97.

מצד אחד, היינו רוצים לבחור את הניסיון השני בגלל הדיוק היותר גבוה אך מצד שני הדיוק הגבוה מאוד על סט האימון מעלה את החשש ל over-fitting בשלבים יותר מתקדמים של ההרצה. בנוסף מהסתכלות על ה confusion matrix של הסטים שבהם קיבלנו דיוק גבוהה ביותר בשני הניסיונות (2 בראשון ו-3 בשני) אנו רואים הרבה פחות טעויות בניסיון השני כך שייתכן ששילוב זוגות של אוגמנטציות עזר למודל להבחין יותר טוב בין המספרים השונים. ייתכן שבניסיון הראשון בחרנו שילובים שהעמיסו על התמונות ובכך לא לימדו את המודל מספיק טוב ובפונקציות הראשונות שבהם לא היו שילובים לא היה למודל מספיק אלמנטים ללמידה.

לבסוף החלטנו להגיש את האוגמנטציה בניסיון השני עבור הדאטא סט השלישי. מסד הנתונים מכיך סה"ב 9940 תמונות. מסקנה אחת מה confusion matrix של סט זה היא שאפשר להעשיר את סט האימון עבור הספרה iii, בה המודל טעה הכי הרבה אך בשל ההגבלה בגודל המסד לא התאפשר לנו לעשות זאת.

לסיום, אחרי הרצה של 100 epochs קיבלנו דיוק הכי טוב של 0.92 לוודלדציה.