**חלק I – קריאת נתונים**

<להוסיף חלק על קריאת הנתונים>

**חלק II- preproccesing**

בתחילה עלינו לנסות לסדר את הנתונים בצורה הטובה ביותר ע"מ לסלק הפרעות בשלב האימון, המטרה היא לסדר את הדאטה בשאיפה שיהיה מסודר כמו MNIST מהמטלה הקודמת, כלומר כלל האותיות, ישרות, בלי הפרעות מהרקע.

**שלב 1: חיתוך**

* חתכתי את האות מן התמונה ע"י תיחום הBB (Bounding Box) הנתון ע"מ ריבוע באמצעות פונקציית עזר cv2.drawContours
* ע"מ להוריד רעש מן התמונות החתוכות, שנוצר בעקבות התוספות שנוספות לתמונה בגלל תיחום המלבני, החלטתי להוסיף mask לאזורים אלו, ולמלא אותם בממוצע הצבעים בתמונה

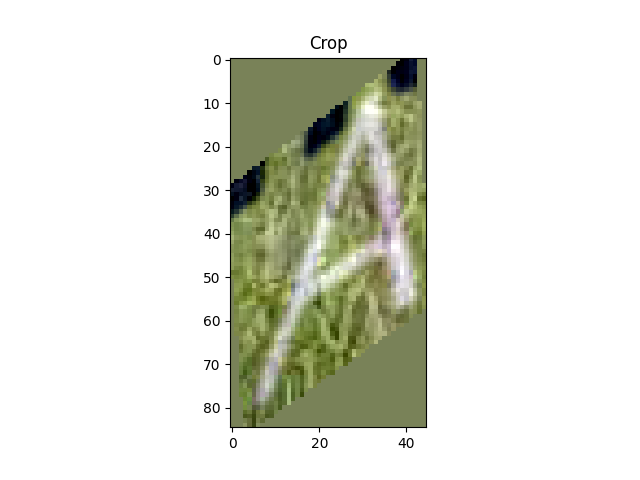
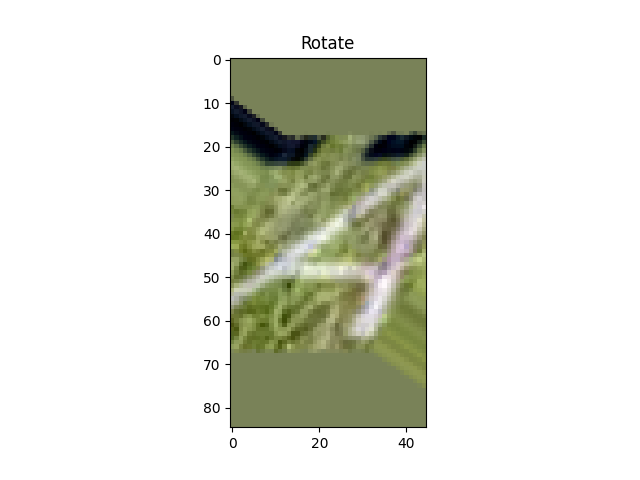
**שלב 2: סיבוב**

* הנחתי כי סיבוב הBB שיהיה מקביל לצירים יקל על הרעש בזמן האימון
* חישבתי את הזווית בין הציר האופקי של האות לבין הצירים וסיבבתי את התמונה בזווית זו, כאשר הפיקסלים שנוספו בעקבות הסיבוב מקבלים את הnearest

**שלב 3: סטנדרטיזציה**

* הפכתי את כל התמונות לגודל 32 על 32
* הפכתי את כל התמונות לשחור לבן

**דוגמאות:**



A picture containing text

Description automatically generated

תמונות מציגות את שלושת השלבים (מימין לשמאל):

התמונה לאחר החיתוך והוספת הMASK, התמונה לאחר סיבוב כך שהBB יהיה מקביל לציר הX והתמונה השלישית – הפיכה לשחור לבן וגודל 32 על 32

**שלב 4: הכנת הDATASET הסופי**

* יצרתי פונקציה אשר עוברת על כלל התמונות ועושה את פעולות הסינתזה שתוארה בשלבים 1-3, לאחר מכן היא שומרת את התמונה (כ – X) ואת הפונט (כ – Y)

**חלק III – מידול ואימון**

* יצרתי רשת קונבולוציה באמצעות keras עם backend של tensorflow
* התחלתי לחקור כל מיני מימושים של CNN, חשבתי שאופי הבעיה מזכירה קצת את MNIST ולכן חיפשתי מאמרים על CNN שהביאו מודל מוצלח לבעיה זאת
* Graphical user interface, text

  Description automatically generatedלבסוף בחרתי ברשת המורכבת משכבות הבאות:

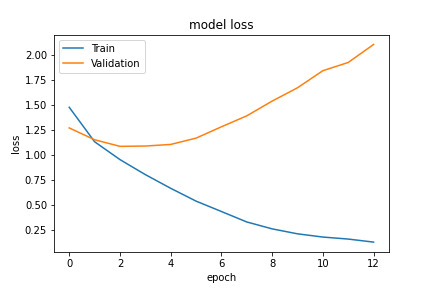
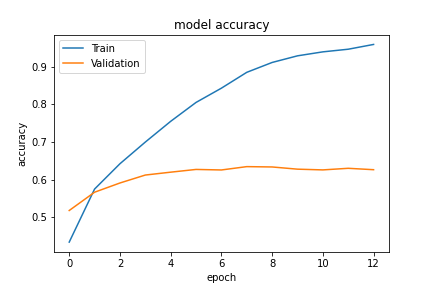
שכבת קונבולוציה עם 32 פילטרים, נורמליזציה, פולינג, לאחר מכן עוד קונבלוציה עם 32 פילטרים, ושכבה נסתרת עם 256 נוירונים

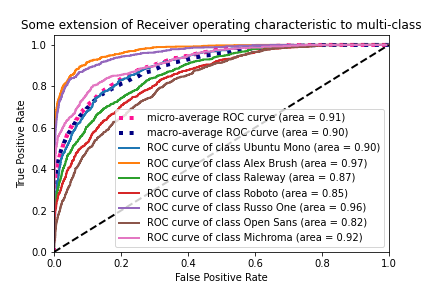
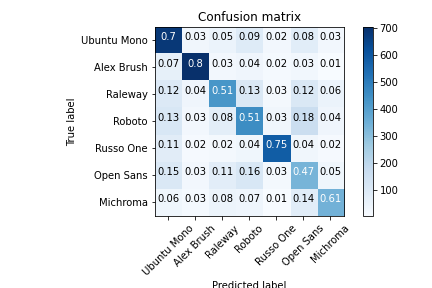
* תוצאות המודל (המשך גם בדף הבא):

Accuracy: 0.6260638236999512

Recall: 0.6166666746139526

Precision: 0.6462281942367554

AUC: 0.8733553886413574



* נראה שהמודל סובל מoverfit ומתקשה לזהות תמונות חדשות, ניסיתי לשחק עם היחס בין הtrain לtest אבל זה לא נחל הצלחה
* גם משחק עם השכבות, וקריאה של כל מיני שיטות לשכבות שונות, תקרת הדיוק הייתה באיזור ה65%

**חלק IV – ניסיון ל"הילחם" בoverfit**

<להוסיף חלק על קריאת הנתונים> והטעות