**ניתוח ואיבוד מידע ספרתי - דוח פרויקט 20.3.2017**

מגישים: גדי גומז, מיכאל גולברייך, וזיו אורן.

הבעיה: סיווגת אותיות בעברית, מספרים וסימנים מתמטיים.

איסוף הנתונים:

מאגר התמונות שלנו כולל אוסף של כ 21000 תמונות שאספנו ושיתפנו בין הסטודנטים בקורס. התפלגות התמונות בין הספרות שווה (בקירוב). את אוסף התמונות חילקנו באופן הבא:

85% מהתמונות הלכו ללימוד, ו15% למבחן כך ש:

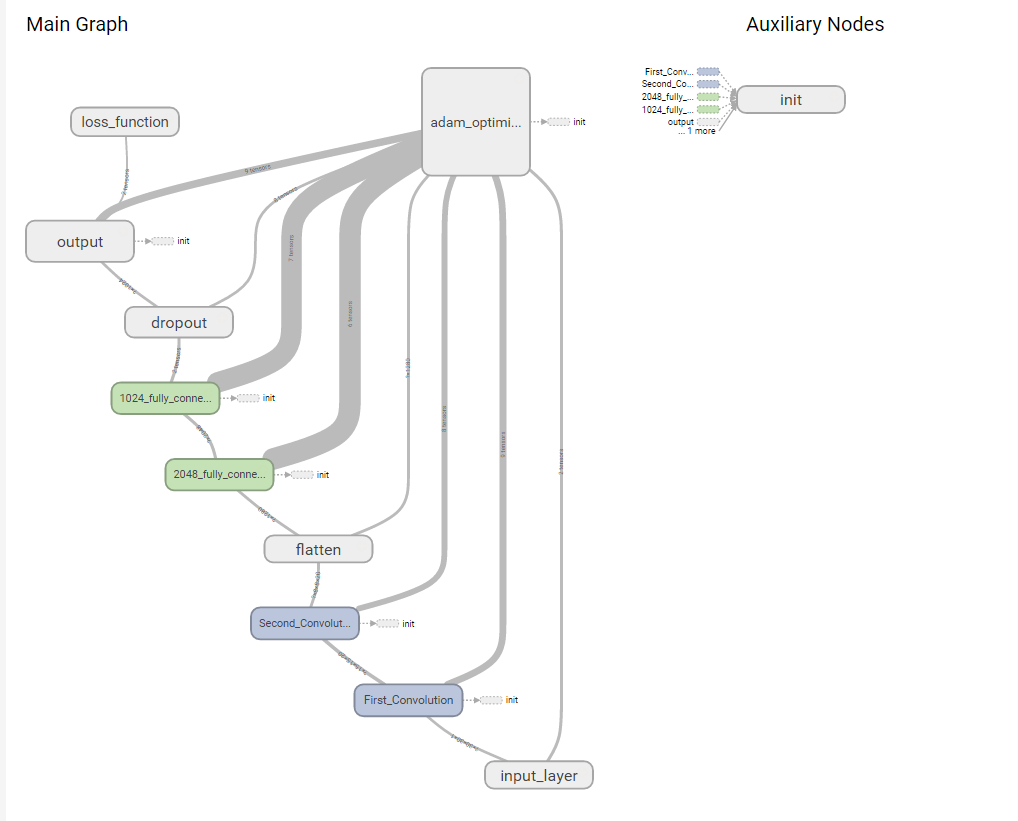
* 18000 תמונות – לימוד
* 3000 תמונות - מבחן

פתרון בעזרת CNN:

בנינו מערכת למידה המבוססת על DEEP LEARNING ברשת נוירונים מסוג CNN.

הארכיטקטורה דומה לפתרון ש-TensorFlow מציעים עבור MNIST.

* ארכיטקטורה:



* שכבת כניסה לתמונות,

1. שכבת קונבולוציה – גודל 5X5 צעד 1X1
2. MAXPOOLING
3. שכבת קונבולוציה 2 - גודל 5X5 צעד 1X1
4. MAXPOOLING נוסף
5. ושכבת FULLY CONNECTED – גודל 2048
6. שכבה FULLY CONNECTED – גודל 1024
7. ואז ביצוע DROP OUT
8. מספר הOUTPUT שיש לנו ברמה האחרונה היא 50 כמספר המחלקות שיש ועוד 1, מצופה ממחלקה 22 להיות ריקה.

המערכת קולטת נתונים לזיכרון ומשנה את גודלם להיות לגודל אחיד, ולאחר מכן מריצה את הנתונים לתוך הCNN, ולבסוף מדפיסה לפלט את התוצאה על מנת להעלות את מהירות ביצוע הריצה.

* האלגוריתם:

ממומש בPYTHON 3.5 בעזרת ספריית TENSORFLOW

1. קריאת קבצי התמונות לזיכרון
2. שינוי גודלם של התמונות לגודל אחיד
3. המרה של התמונות להיות שחור לבן (GREYSCALE)
4. אתחול של המשתנים של הרשת, learning rate, batch size, number of epochs
5. הרצה בלולאה מספר הEPOCHS
   1. הרץ את BATCH SIZE של התמונות לתוך המערכת
   2. אם זו איטרציה המתחלקת ב10 – הרץ בדיקה על הטסט דאטה.
      1. הדפס את התוצאה

* הניסויים שבוצעו במערכת:

1. שינויי LR
2. שינוי מספר הBATCH SIZE
3. שינוי גודל התמונות:

* טבלת התאמה:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| א | ב | ג | ד | ה | ו | ז | ח | ט | י | כ | ל | מ | נ | ס | ע |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| פ | צ | ק | ר | ש | ת |  | ך | ם | ן | ף | ץ | ( | ) | + | - |
| 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ∫ | <= | >= | ∞ | ∑ | ∩ | ≠ | ÷ | ͓ |
| 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 |

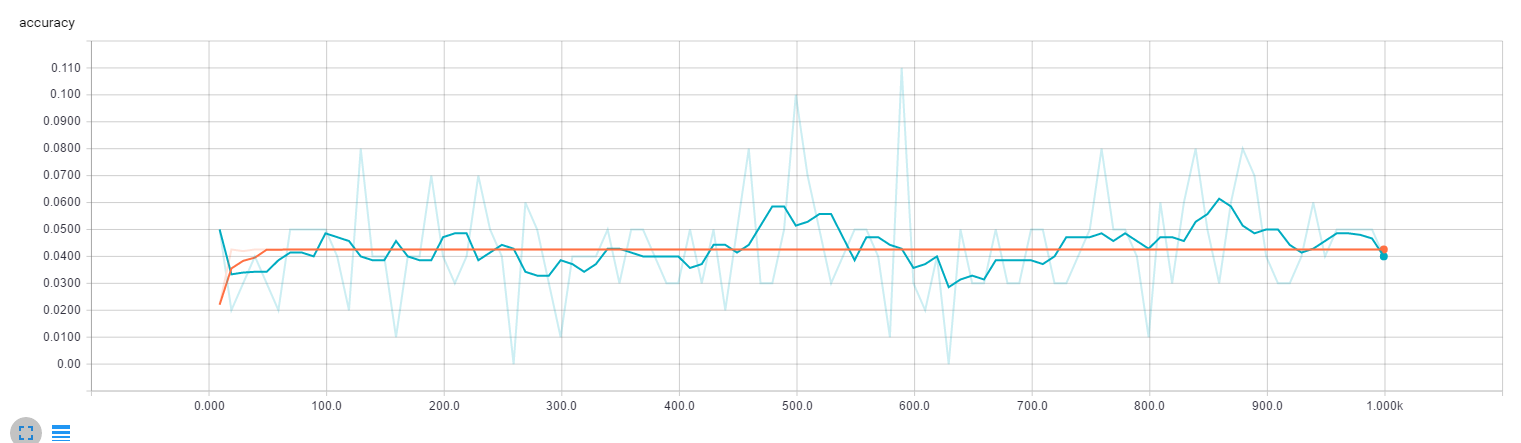
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 |

* **תוצאות הניסוי:**

**פרטים התחלתיים:**

* גודל תמונה : 30\*30
* LR = 0.005
* Batch size =100
* Epochs = 1700

היא:



ניתן לראות כי בקצב הלימדה הזה שהגענו למינימום מקומי וולא ניתן היה להמשיך, בעקבות התוצאה של הזה החלטנו להוריד בlr על מנת לנסות להיכנס לתוך המינימום הרלוונטי.

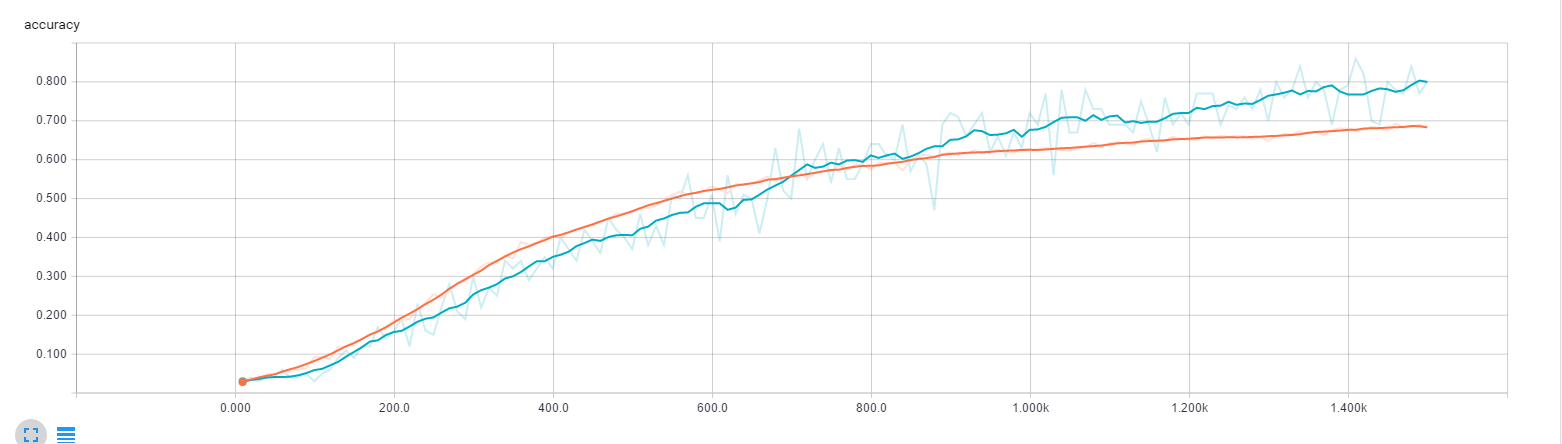
בשלב זה לא הוצאנו נתונים לאקסל.

**פרטים התחלתיים:**

* גודל תמונה : 30\*30
* LR = 0.0005
* Batch size =100
* Epochs = 1500



היא:



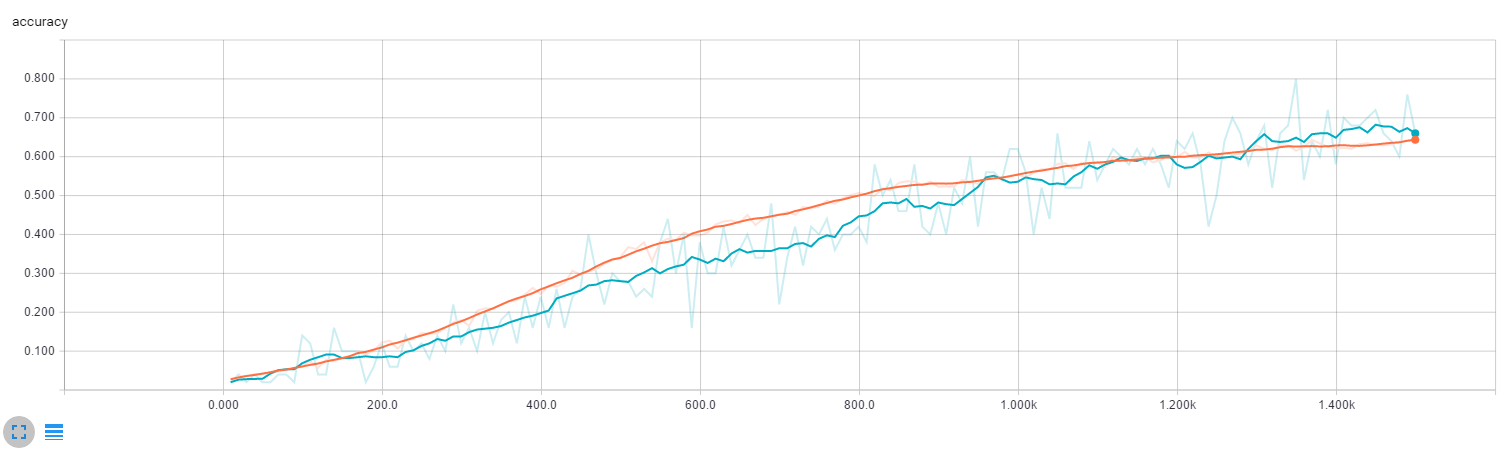
Acc: 0.675211

ניתן לראות כי בקצב הלימדה הזה שהגענו נמנענו ממינימום כלשהו וכן הייתה התקדמות אל עבר המטרה

**פרטים התחלתיים:**

* גודל תמונה : 30\*30
* LR = 0.0005
* Batch size =50
* Epochs = 1500





Acc = 0.653712

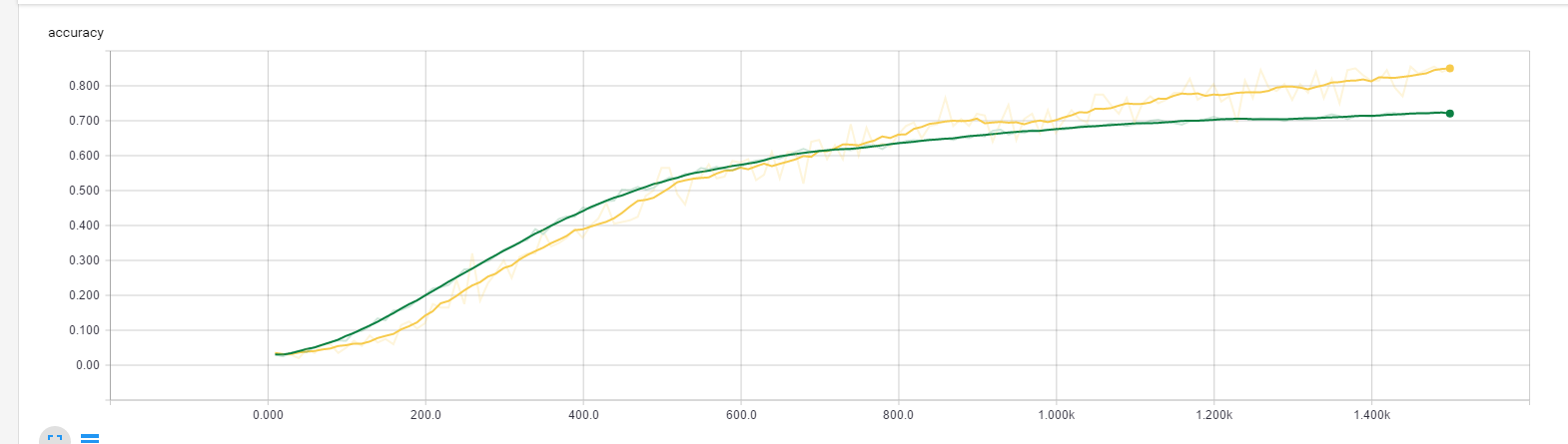
ניתן לראות כי גודל הBATCH לא היה משמעותי במקרה זה.

**פרטים התחלתיים:**

* גודל תמונה : 30\*30
* LR = 0.0005
* Batch size =200
* Epochs = 1500



תוצאה:

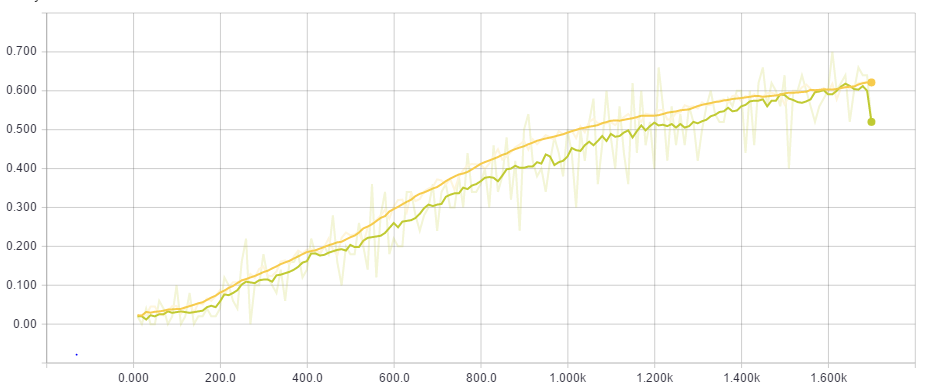


Acc = 0.723756

ניתן לראות כי גודל הBATCH בגודלו כעת כן מהווה פקטור למידה משמעותי.

**פרטים התחלתיים:**

* גודל תמונה : 20X20
* LR = 0.001
* Batch size =50
* Epochs = 1700

תוצאות:  


Confusion Matrix:

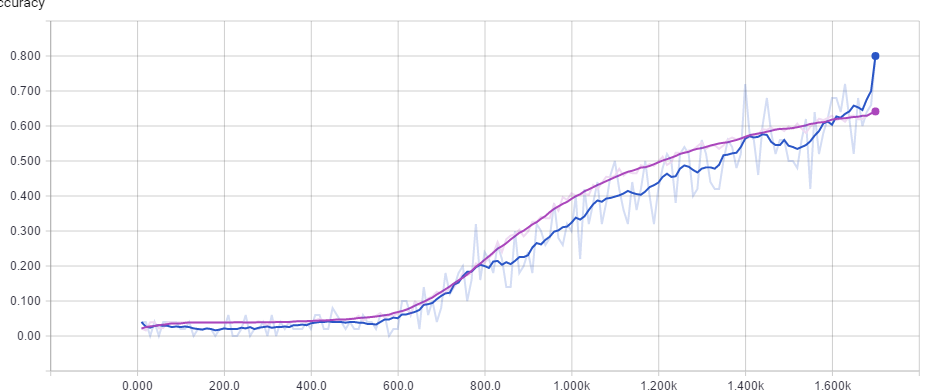
Acc: 0.5788668394088745

מסקנות:

ריצה בגודל תמונה קטן הביאה תוצאה פחות טובה ממקבילתה בגודל 30\*30. ככל הנראה נגרם מכך שבתמונה קטנה יותר מספר הפיצ'רים קטן ולכן גם יכולת הזיהוי קטנה.

**פרטים התחלתיים:**

* גודל תמונה 28\*28
* LR = 0.001
* Batch size =50
* Epochs = 1700

תוצאות:  


Confusion Matrix:

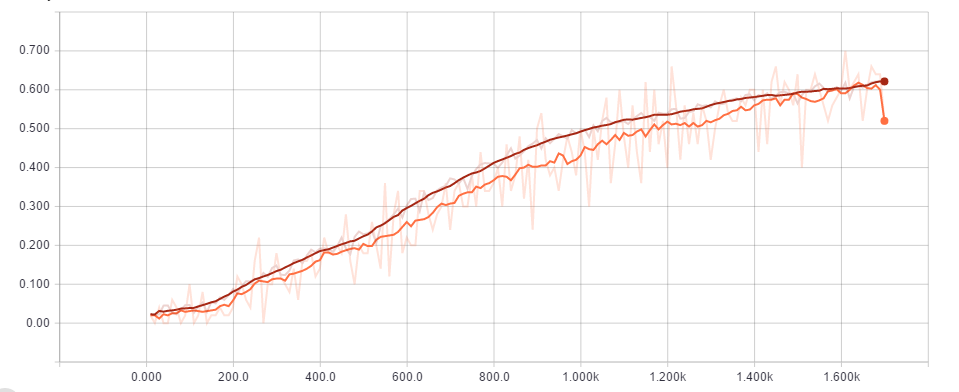
Acc: 0.6416538953781128

מסקנות:

ריצה בגודל תמונה קטן הביאה תוצאה פחות טוב. ככל הנראה נגרם מכך שבתמונה קטנה יותר מספר הפיצ'רים קטן ולכן גם יכולת הזיהוי קטנה.

**פרטים התחלתיים:**

* גודל תמונה : 28X28
* LR = 0.001
* Batch size =50
* Epochs = 1700

תוצאות:  


Confusion Matrix:

Acc: 0.621439516544342

מסקנות:

ברוב הריצות תמונה גדולה יותר נתנה תוצאה טובה יותר. ניתן לראות שבפרמטרים האלה אין הבדל בין גודל 28X28 ל-30X30

נקודות נוספות שניתן לראות מכלל הריצות:

* מחלקה מספר 26 ומחלקה מספר 27 (האותיות 'ף' ו-'ץ') דומות במציאות ונראה שגם הלמידה התבלבלה בין השתיים לא מעט.
* מחלקות 14 ו-41 ('0' ו-'ס') גם כן יצרו בלבול בשל הדמיון שלהן

**מסקנות:**

* גודל התמונה הוא מפקטור מרכזי ברשת שלנו על מנת להצליח במטרת המשימה
* גודל הLR קובע בבירור את איכות הלמידה ואת ההגעה לתוצאות הטובות ביותר
* לאחר ריצות רבות שבוצעו טרם כתיבת מסמך זה, הבנו כי קיימים גורמים מרכזיים המשפיעים על איכות הלמידה: **איכות** **המדגם**(תמונות עם פסים של מחברת לעומת לא), **כמות** **הדגימות**, **גודל** **ההתפלגות** **המחלקות** **הסופיות**(במידה והיה קטן יותר התוצאות יכלו להשתנות).