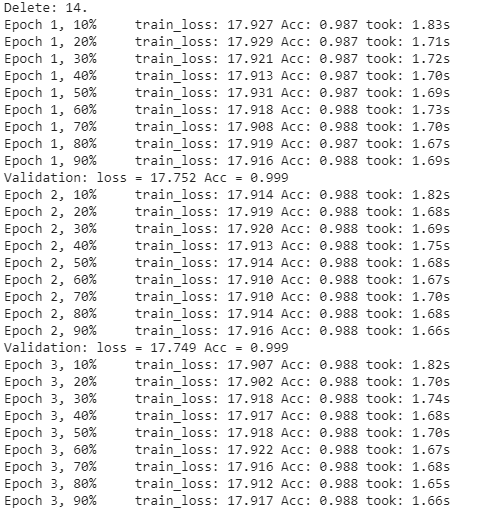
פרויקט סודוקו – תמצית

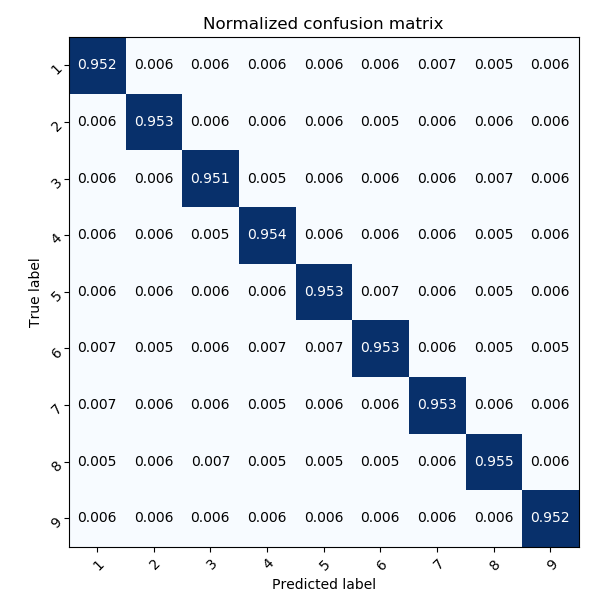
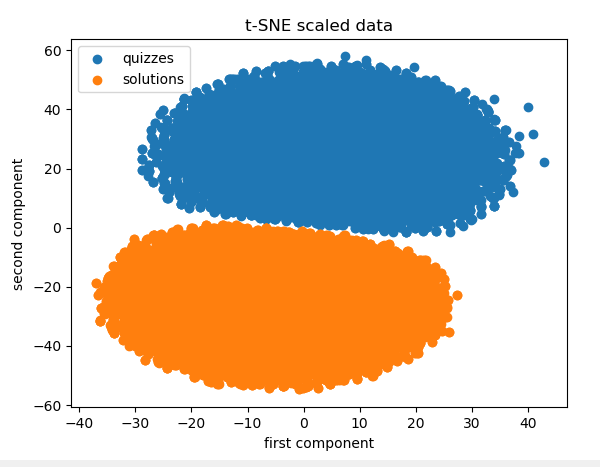
פתרונות תיאורטיים על סמך הכרת הבעיה:

1. יצירת גרעינים מותאמים לתבניות הסודוקו – שורה, עמודה, קוביה
2. עבודה עם וואן הוט וקטור מכיוון שאין ערך להפרש בין שני מספרים.
3. מילוי משבצות באופן איטרטיבי – בפתרון סודוקו בזמן אמת, נמלא כל פעם את המשבצת עם ההסתברות הגבוהה ביותר שהתקבלה מהרשת.

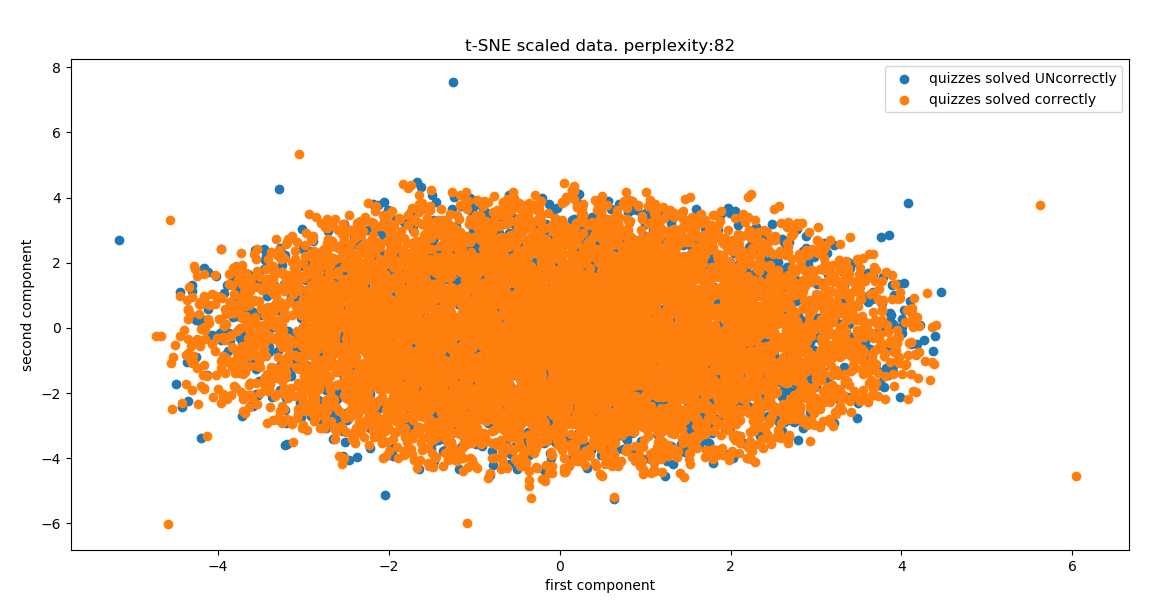
לאורך הדרך:

1. תחילה עבודה עם רשת FC פשוטה לדיבוגים כלליים ולנסות להצליח טוב על סט האימון.
2. עבודה על פונקצית לוס לטובת חישוב הלוס רק על האיברים שאותם אנחנו ממלאים. לצורך כך יש להכפיל את מטריצת הלוסים המתקבלת במסכה של האיברים שמילאנו ולסכום.
3. כדי לבדוק את הדיוק של הרשת שלנו, אנחנו צריכים לדמות מצב של טסט – כלומר מילוי איבר איבר. לכן בכל פעם שאנחנו בודקים את סט הוולידציה אנחנו למעשה לוקחים את הדיוק עליו לפי מילוי איבר איבר. לא עשינו זאת באימון מכיוון שזה לוקח הרבה זמן וזה חשוב בעיקר בשביל לראות את הדיוק במהלך האימון. מספיק בינתיים לראות את הדיוק רק בסוף כל איפוק על סט הוולידציה.
4. אחד הדברים שהיו בעייתיים זה הזמן ריצה הארוך של האימון והבדיקה את הדיוק האמיתי על הוולידציה. לכן השקענו זמן בלשפר את הזמנים בחישוב הדיוק לוולידציה. עשינו זאת על ידי זה שנפתרנו מלולאה (שהייתה בתוך לולאה אחרת) שרצה כל פעם על כל הבאטצ'. כך קיצרנו ביותר מפי 9 את הזמן של חישוב הדיוק האמיתי.
5. הגענו ככה למצב שהרשת לומדת יחסית לאט כאשר אנחנו מוחקים 14 מספרים ומעלה. כלומר הלוס משתנה ממש ממש לאט וזה מצריך הרבה איפוקים. אחת התיאוריות אומרת שזה קורה עקב כניסה של הרשת לאזור מינימום רחב כלשהו ששם הנגזרות קטנות יחסית. לכן ניסינו מדי פעם להגדיל באופן מלאכותי את קצב הלמידה. בנוסף, נשים לב שהבעיה שלנו משתנה כאשר אנחנו מוחקים מספר נוסף. כלומר, ייתכן שפונקצית המטרה משתנה ולכן מינימום שנקלענו אליו עד כה הוא פחות טוב ממינימום אחר שקיים אך לא נגיע אליו עקב קצב למידה קטן יחסית ומכך שאנחנו כלואים בתוך מינימום רחב. 

זאת אחת הבעיות העיקריות שלנו כרגע, אם כי עדיין הדיוק על סט הוולידציה משביע רצון יחסית (כשלא מוחקים יותר מדי מספרים). יוצא שבכל מחיקה של מספר נוסף אחרי ה 14 מחוקים, הלוס גדל בערך ב 5 ולא מצליח לקטון מספיק עד שמוחקים עוד מספר (וזה קורה אחרי 50 איפוקים עבור אותה מחיקה). להגדיל את מספר האיפוקים ייתכן שיעזור אך חשבנו שנוכל להגדיל את הקצב לימוד ולחסוך זמן ריצה.

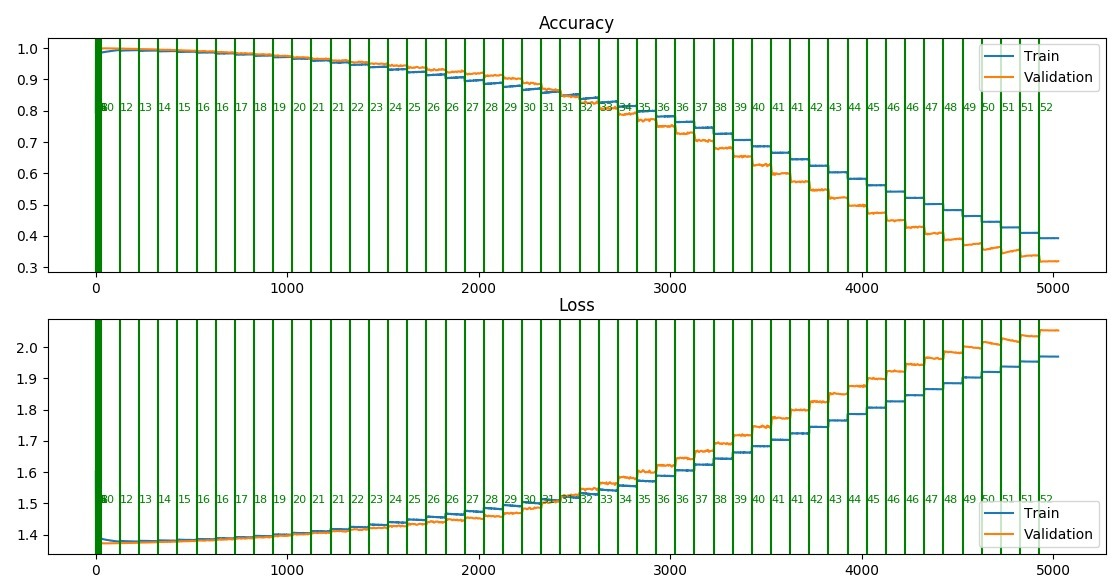
1. התחלנו להסתכל על ביצועי הרשת. ניתן לראות למשל שעבור 50 מחוקים (טסט דאטא) רשת FC שלנו הייתה עם 95% דיוק והניבה:  
   שמראה שהפילוג יחסית אחיד – כפי שציפינו. הרי אין הבדל משמעותי בין ספרה כזו או אחרת ולמעשה יש סימטריה. לכן גם ה confusion matrix יצאה סימטרית יחסית.
2. הסתכלנו על ההטלה למרחב דו מימדי של לוחות חסרים ושל לוחות פתורים מתאימים, על ידי TSNE וקיבלנו:  
   

כצפוי, ניתן לראות הבדלה מלאה בין המחלקות השונות.

זאת אבן ראשונה בהסתכלות על הטלת הלוחות. נמשיך וננסה הטלות מתוך מחלקות שונות.  
  
כעת ניסינו משהו אחר: לקחנו לוחות חסרים, והרצנו אותם ברשת. חלקם נפתרו נכון וחלקם לא. אז אנחנו מטילים את הלוחות החסרים לדו מימד בעזרת TSNE ומנסים לראות האם יש מאפיינים שדרכם ניתן לראות הבחנה בין הלוחות שהצלחנו לפתור לבין אלו שלא. התוצאות שקיבלנו לא מראות לנו הבדלים לצערנו:  


דבר נוסף שנוכל לבדוק הוא האם רשתות שונות שיצרנו טועות באותם לוחות.

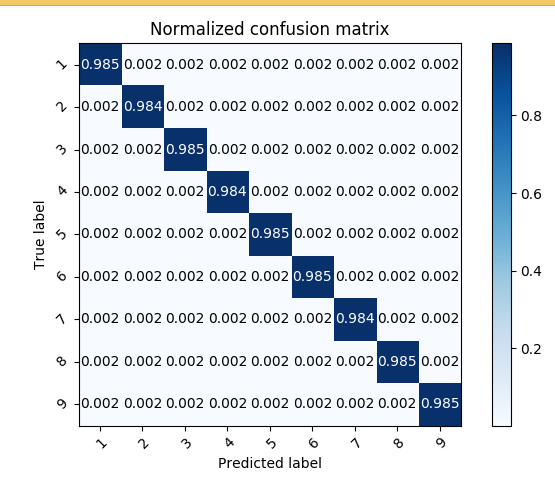
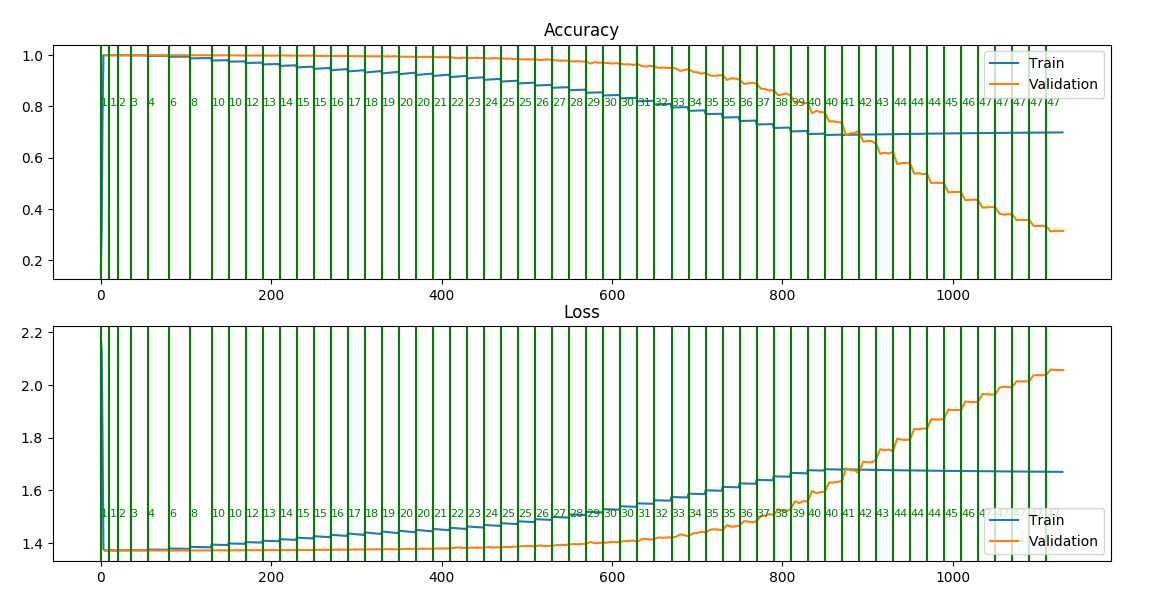
1. מסתבר שאנחנו מוחקים איברים, אז למשל כשמוחקים 60 איברים זה למעשה מוחק פחות כי זה מוחק את אותו איבר כמה פעמים.. אז לנסות לתקן את זה.. יוצא שאנחנו לא מוחקים 60 איברים אלא 38 בסוף.  
   לאחר שתיקנו את זה ופשוט מוחקים את אותם מקומות בכל הלוחות באותו הבאטץ' (תוצא לוואי בינתיים שאפשר לתקן), התוצאות נראות הרבה פחות טוב. למשל עבור רשת הקונבולוציה:



ננסה לעבור עם FC כי המידע אינו בהכרח לוקלי ואין לנו בעיה יותר מדי ממספר הפרמטרים של הרשת (האינפוט מאד קטן יחסית).

מה שמעניין הוא שרשת ה FC הביאה עדיין תוצאות טובות! כשאנחנו בודקים על טסט הדאטא כמו שהוא (ללא מחיקה מהפתרונות) אנחנו מקבלים דיוק **של 99.9% פר איבר ודיוק של 99.2% פר** **לוח**.

1. כשאנחנו מוחקים איברים, אנחנו למעשה עלולים ליצור לוחות שאינם לוחות סודוקו. מה שעלול להטעות את הרשת שלנו שלא לצורך (המטרה היא פתירת סודוקו). לכן, מה שננסה לעשות זה מתוך הלוחות המלאים, נמחק את האיברים שהיו מחוקים בלוחות החלקיים אחד אחרי השני. כלומר נניח שיש לנו לוח מלא ולוח חלקי. אז נמחק איבר אחד מהלוח המלא בהתחלה על ידי בחירת מיקום שבוא יש 0 בלוח החלקי (כלומר איבר מחוק) ואת המקום המתאים נמחק בלוח המלא. בשיטה זו בוודאות נשאר עם לוח שמתאים למשחק הסודוקו.
2. במצב הנוכחי, הרשת לומדת למלא את כל הלוח בבת אחת. בזמן מבחן אנחנו ממלאים את האיבר עם ההסברות הגבוהה ביותר. אולי נוכל לשנות את למידת הרשת כך שהיא תלמד למלא איבר אחד בלבד. אולי על ידי softmax על פני כל האיברים.
3. מה לגבי אוגמנטציות? אפשר לנסות למשל להחליף את כל המקומות בהם רשום 1 לספרה אחרת ולהפך. כך חוקיות המשחק נשמרת. או לסובב את הלוח סביב העמודה/השורה/האיבר האמצעי.
4. עבור רשת שטוחה יחסית (רק שתי שכבות FC יחד עם DROPOUT ) ועבור אימון בו בסט האימון אנחנו מוחקים באופן הולם ובולידציה לא, קיבלנו את התוצאות הבאות:

ניתן לראות שהמחיקה הלא הולמת משפיעה מאד על הולידציה. אנחנו נשווה את התוצאה הזאת מול מחיקה הולמת לבחינת כמות ההשפעה של אופן המחיקה על למידת הרשת.

1. בבואנו לשחזר את הרשת שעשינו בהתחלה ונתנה את התוצאות הטובות עד כה, רשת FC\_1 , עם אימון המכיל מחיקה לא לגיטימית עם חזרות, הפעם עשינו עד מחיקה של 55 איברים (ולא 62 כמו שהיה בפעם הקודמת). התוצאות שקיבלנו על הטסט המקורי החלקי, הוא **99.5% פר לוח ו 99.9% פר תא.** אלו התוצאות הטובות ביותר שקיבלנו עד כה. מה שמעניין הוא שהתוצאות אף השתפרו מפעם קודמת. האם ייתכן שהמחיקה של 55 רק ולא 62 זה מה שתרם לתוצאות? כלומר האם זה שאנחנו מקשים את הבעיה החל מסף מסוים פוגע בביצועי הרשת? יהיה מעניין לבדוק עם מחיקה מקסימאלית של 47 למשל ולבחון את התוצאות על הטסט. בנוסף, בדקנו את התוצאות על הטסט באופנים שונים: עבור מחיקה לגיטימית התוצאות היו טובות כמו הטסט המקורי. עבור מחיקה לא לגיטימית ללא חזרת התוצאות היו הרבה פחות טובות. עבור המחיקה לא לגיטימית עם חזרות, התוצאות היו גם פחות טובות מאשר מחיקה לגיטימית. כלומר אנחנו מבינים מכאן שהרשת למדה חלק ממאפייני לוחות הסודוקו ומחיקה לא לגיטימית יכול לפגוע בחוקיות לוח בסודוקו.

**הערות להמשך:**

**עשינו:**

* החלטנו להשאיר את המחיקה עם חזרות כדי להשתמש בזה כפיצ'ר שזוכר עם מחיקות עם מעט מספרים- להראות פילוג של כמה לוחות מכל סוג יש.
* הגיוני להשאיר לוחות מחוקים ברמות שונות. לעשות את זה.
* להשוות זמני ריצה של backtracking לעומת הרשת
* לבדוק על test עם יותר מספרים מחוקים.
* זמן ממוצע כפונקציה של מספר מחוקים (רשת ובקטרקינג)

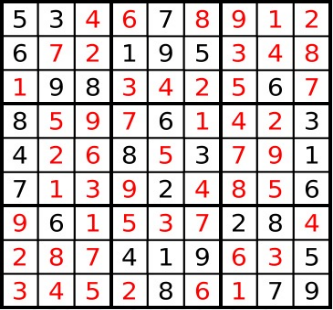
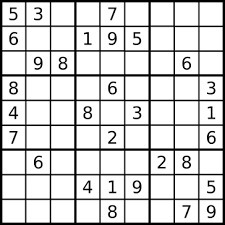
**יש לעשות:**

* הסתברויות על תא בודד של כל הספרות לאורך הריצה
* להריץ את הרשת הטובה על יותר מחוקים (על הדאטא הקשה).
* הרצה בלי חזרות לגיטימי.
* פילוג המחוקים בדאטא.

**קל לעשות בסוף:**

* לבדוק את התוצאות כאשר מוחקים בצורה לא לגיטימית בtest כאשר מאמנים בצורה לגיטימית.
* לעשות השוואה לעומת התוצאות בtest בצורה לגיטימית.
* להתמקד בתוצאות עם לוחות לגיטימיים, ולעשות השוואה בין FC ל CNN .
* להסתכל על ההסתברות של המספרים שאנחנו לא משלימים. להסתכל מה קורה להסתברויות לאורך ההשלמה של הלוחות.
* להסתכל האם ברגע שאני משלים אחד מהמספרים, אז במקום אחר באותה שורה ההסתברות של אותו מספר יורדת. כלומר האם הרשת מסיקה מסקנות באופן דומה למה שאדם עושה.
* אפשר אולי לחפור על curicculum בדו"ח כי זה נשמע מפוצץ.

לדוח:

1. האם יש משמעות לערכי המספרים בלוח סודוקו? גם אם כל מספר היה מוחלף בשם של חיה, המשחק היה נשאר זהה. לכן עבדנו עם וקטור 1hot לכל ספרה – כדי לא לתת אפשרויות להתייחסות שונה של הרשת עבור ספרות שונות.
2. **הפתרון בפרויקט שלנו – תיאור כללי**
   1. נתחיל מהדאטא. כאמור, הדאטא שלנו הם לוחות סודוקו חסרים ולוחות מלאים מתאימים. בלוחות אלו המספרים הם 1-9 כידוע, אך נשאלת השאלה האם יש משמעות לערכי המספרים בלוח סודוקו? הרי גם אם כל מספר היה מוחלף בשם של חיה, המשחק היה נשאר זהה. לכן עבדנו עם וקטור 1hot לכל ספרה – כדי לא לתת אפשרויות להתייחסות שונה של הרשת עבור ספרות שונות. כך, כל לוח סודוקו חסר, הפך למטריצה בגודל 10x9x9 (10 במימד העומק מכיוון שיש גם לספק את הספרה 0 המבטאת תא חסר). והפתרונות המלאים הפכו למטריצה בגודל 9x9x9. עבור מיקום שורה ועמודה, אנחנו מקבלים וקטור 1hot המתאים לספרה שנמצאת שם.את הדאטא פיצלנו לסט אימון, ולידציה ומבחן לפי היחס 10%, 20%, 70% בהתאמה. להוסיף תמונות (אולי גם של 1hot)
   2. פונקציית ההפסד: מכיוון שחלק מהתאים כבר ידועים מראש, מה שהיה חשוב לנו הוא שהרשת תלמד את למלא התאים שאינם ידועים, ואז את מה שידוע פשוט נדביק לבסוף על התוצאה שהתקבלה מהרשת. לכן, חישוב הלוס התבצע על כל תא חסר בלבד, אל מול התוצאה הידועה של אותו תא בלוח המלא. מכאן למעשה אפשר לדמות את הבעיה כבעיית סיווג עבור כל תא בנפרד, שמנסה לחזות את המחלקה (המספר) המתאים עבור אותו תא. ולכן השתמשנו בפונקציית הפסד שהיא Cross Entropy על כל תא חסר.
   3. ארכיטקטורה: מכיוון שמימד הבעיה יחסית נמוך (לוחות קטנים יחסית), ומתוך רצון להימנע מהתאמת יתר, החלטנו לעבוד עם רשתות פשוטות. הכניסה לרשת היא מטריצה 10x9x9 עבור לוח חסר והמוצא הוא מטריצה 9x9x9 שבמקום ה (i, j, k) אנחנו מקבלים את ההסתברות שהמספר המתאים למיקום ה (j, k) הוא i. באופן פשוט יותר, עבור שורה j , עמודה k ועומק i אנחנו מקבלים את ההסברות לאיבר i+1 באותה שורה ועמודה.
      1. רשת 1 – FC:

רשת המורכבת משלוש שכבות FC עם batch-Normalization ופונקצית אקטיבציה Relu. לאחר מכן, מסדרים את הנוירונים בצורת מטריצה 9x9x9 ולבסוף שכבת softmax על מימד העומק במטריצה הסופית לקבלת ההסתברויות לכל ספרה.

* + 1. רשת 2 – CNN:  
       במקרה זה, המסננים שלקחו אינם טריוויאלים, אלא מותאמים לבעיה הספציפית שלנו. בלוח סודוקו יש משמעות גדולה למבנה של שורה, עמודה ותיבה של 3x3. לכן בשכבת הקונבולוציה שלנו, השתמשנו במסננים בגודל 1x9, 9x1, 3x3 כאשר המסננים של השורה ועמודה בעלי צעד בגודל 1 ואילו המסנן בגודל 3x3 עם גודל צעד 3 (כיאה ללוח סודוקו). מתוך מעבר של כל מסנן מקבלים וקטור מאפיינים בגודל 9. וקטורים אלו שורשרו לוקטור אחד ארוך שנכנס לשתי שכבות FC. בסופן, סידרנו את הנוירונים לצורת מטריצה 9x9x9 וביצענו softmax על מימד העומק כמקודם. גם כאן, פונקציית האקטיבציה היא Relu.
  1. שיטת פתרון לוח סודוקו בודד: בבואנו לפתור לוח סודוקו בעזרת הרשת, אנחנו קודם הופכים אותו למטריצת 1hot ואז אנחנו מכניסים לרשת את המטריצה. תהליך מילוי האיברים הינו איטרטיבי כלומר, לאחר שהמטריצה עוברת ברשת, אנחנו ממלאים רק את המקום בעל ההסתברות הגבוהה ביותר (עם האיבר המתאים). ואז מכניסים את התוצאה שוב לרשת וממלאים עוד איבר וכך הלאה. תהליך זה, מראה הפקת מסקנות כפי שנעשות בידי אדם הממלא לוח סודוקו באופן איטרטיבי ומשנה את רמות הביטחון שלו עבור המיקומים השונים ככל שהמשחק מתקדם.

Offline:

Online:

Classifier

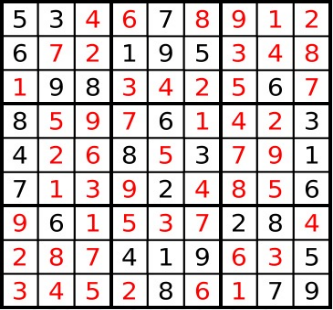
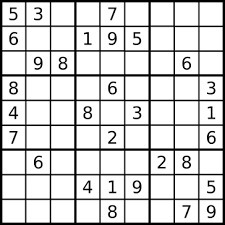
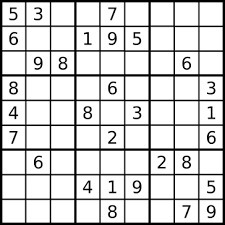
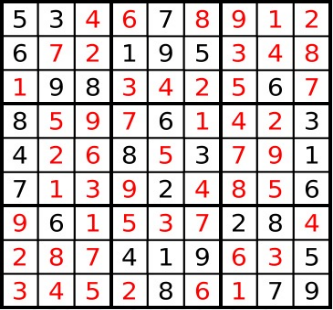
CNN training

Dataset + solution

Partial board

Trained CNN

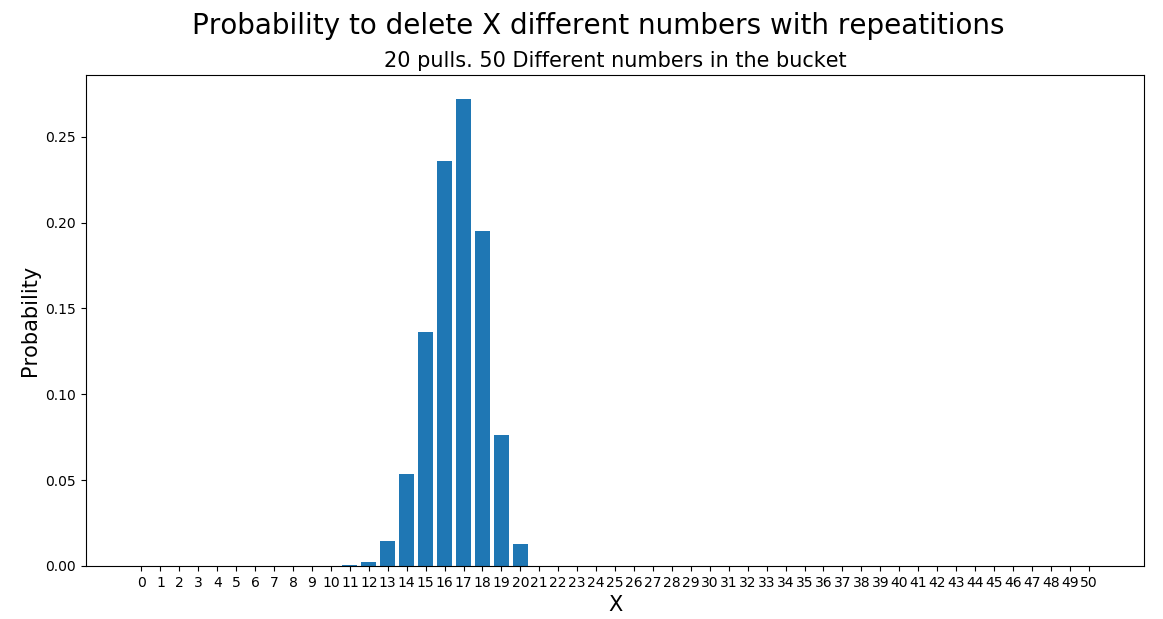
Solution



Trained CNN

predicted material

For example: grass

* 1. תהליך האימון: לצורך אימון הרשת, השתמשנו בשיטת curriculum learning כך שלמעשה עבדנו רק עם הלוחות המלאים, וכל פעם מחקנו איבר נוסף ולימדנו את הרשת להתרגל למצב החדש. עשינו זאת בצורה הבאה: תחילה מחקנו איבר בודד ולימדנו את הרשת להשלים אותו לאורך מספר איפוקים. לאחר מכן מחקנו שני איברים ולימדנו את הרשת להשלים אותם לאורך כמה איפוקים. וכך הלאה. האיברים שאותם אנחנו מוחקים, אלו איברים שהיו מחוקים בלוחות החסרים המתאימים לאלו שמהם מחקנו. כלומר, עבור מחיקה של 20 איברים מלוח מלא, אנחנו מסתכלים אילו איברים מחוקים באותו לוח חסר המתאים ללוח המלא הזה, ומהם מוחקים את האיברים. לסוג מחיקה זו אנחנו קוראים מחיקה לגיטימית. עבור מחיקה רנדומלית מתוך הלוח המלא, אנחנו עלולים למחוק איברים ככה שיהיו שני פתרונות ללוח וזו נקראת מחיקה לא לגיטימית. בהמשך נראה את המשמעות של סוג המחיקה.  
     כאשר יש לנו batch שממנו אנחנו מוחקים איברים בכל לוח, אנחנו לא מוחקים את אותו מספר איברים בכל לוח מכיוון שאנחנו לא רוצים שהרשת תשכח את מה שהיא למדה מקודם (כי בזמן מבחן המילוי איטרטיבי). לכן כשאנחנו באים למחוק 20 מספרים בכל לוח ב batch, אנחנו למעשה מוחקים בחלק מהלוחות 20 ובחלק 19 ובחלק 18 וכך הלאה לפי פילוג גאוסי סביב 17 במקרה הזה.   
     באיור זה ניתן לראות שמתוך 50 מספרים חסרים בלוח, כשנרצה למחוק 20 איברים אנחנו נמחק בלוחות שונים מספר שונה של איברים לפי האיור.  
     כך אנחנו דואגים לכך שבתהליך האיטרטיבי בזמן מבחן, הרשת עדיין תדע להתמודד עם לוח שחסרים בו פחות איברים.

1. **תוצאות:**

רעיונות להמשך:

1. שימוש באנסמבל.