

תרגיל 3

רשת נוירונים עם 1 hidden layer:

המודל שלי כולל את הפרמטרים הלומדים הבאים:

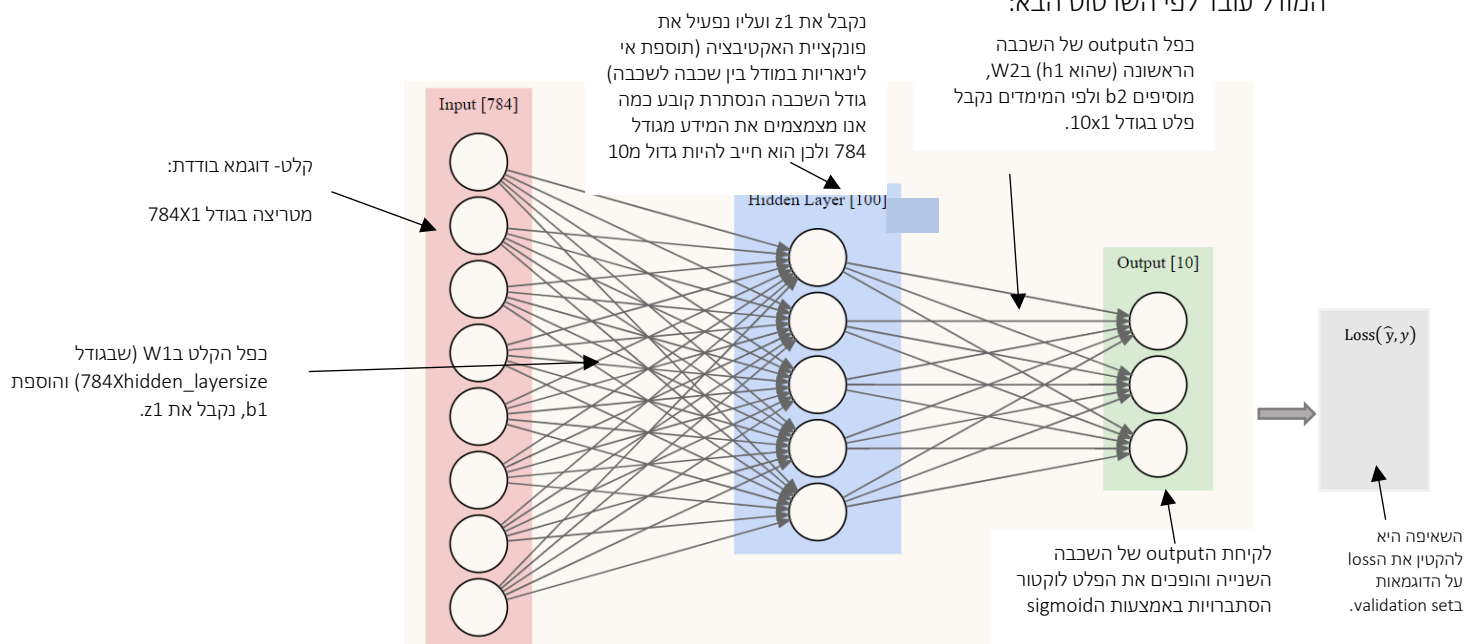
$W1, b1, W2, b2$.

פרמטרים אלו אתחלתי באופן רנדומלי (התפלגות יוניפורמית בין 0.08- ל 0.08 לפי המלצה מהאינטרנט) להיות נמוכים כיוון שפונקציית הסיגמואיד העבירה את ערכים גדולים יחסית להיות 1, ולכן העדפתי להשתמש בערכים נמוכים).

בנוסף המודל שלי כולל את ההיפר פרמטרים (לא לומדים) הבאים:

Hidden layer size, learning rate (eta), epochs, activation function and batch size.

המודל עובד לפי השרטוט הבא:



(כאשר השתמשתי ב batch הגדול מ 1 המודל למד פחות אפקטיבי, לכן נשארתי עם batch בגודל 1).

את הפרמטרים אימנתי במהלך פונקציית ה training במעבר על test set, באמצעות הגרדיאנטים שקיבלתי מה back propagation.

את המודל אימנתי ע"י כלל עדכון זהה ל SGD (אחרי כל דוגמא - מעדכנים את הפרמטרים של התרגיל).

כעת אפרט על כל היפר-פרמטר ומדוע בחרתי לאמן את המודל שלי כפי שבחרתי לאמנו:

- 1 Batch size: כאשר ניסיתי לחשב loss עבור 10 דוגמאות בבת אחת, המודל למד באופן פחות טוב (לא היה עקבי על הדוגמאות ופונק' ה loss אמנם ירדה, אך באופן לא עקבי).
- 2 Hidden layer size: כדי שהמודל יהיה אפקטיבי, נרצה לצמצם את הבעיה (מעבר מבעיה בגודל 784 לגודל קטן יותר) באופן הדרגתי, גודל ה hidden layer חייב להיות בין מספר הקלאסים לגודל הקלט. אחרי הבדיקות מצאתי שגודל ה hidden layer האפקטיבי למודל הוא: 100.
- 3 Learn rate (η): עדכון הפרמטרים $W1, b1, W2, b2$ לפי הנוסחה $W_i + 1 = W_i - \mu * \nabla W_i$ ככל ש η גדולה יותר, כך המודל יותר "מושפע" מכל דוגמא, מצאתי שה η האופטימלית היא: 0.005. (כאשר הייתי עם η גדולה מכך, ה loss על validation set היה לא יציב (קיבלנו מודל לא אמין יחסית) וכאשר הקטנתי את ה η מדי האופטימיזציה לקחה המון זמן (נדרשה כמות epochs גדולה)).
- 4 Epochs: מספר הפעמים שעוברים על training set. נשים לב שמספר epochs משפיע על כמה המודל "כפוף" ל training set, זאת אומרת, שמצד אחד המודל לומד מה training set בכל epoch ונרצה שהוא יעבור במידה מספקת על training set, אך לא יותר מדי, כדי שלא יהיה מצב Overfitting ואז התוצאות על train יהיו מדויקות, אך test לא יהיה מדויק. מצאתי שבכמות epochs האופטימלית היא: 35. את מספר ה epochs האופטימלי מצאתי בהתאם ל learn rate.
- 5 Activate function: בחירת פונקציית האקטיבציה מהאפשרויות הבאות Sigmoid, ReLU, PReLU, tanh ועוד. מטרת פונקציית האקטיבציה היא הכנסת אלמנט אי לינאריות במודל במעבר בין השכבות (מתבצע רק פעם אחת כאן סה"כ, בחרתי בפונקציית הסיגמואיד (החזירה סיכויי הצלחה (true positive) של 89% על validation set לעומת שאר הפונקציות שבדקתי, שמגיעות לסיכויים פחות טובים).