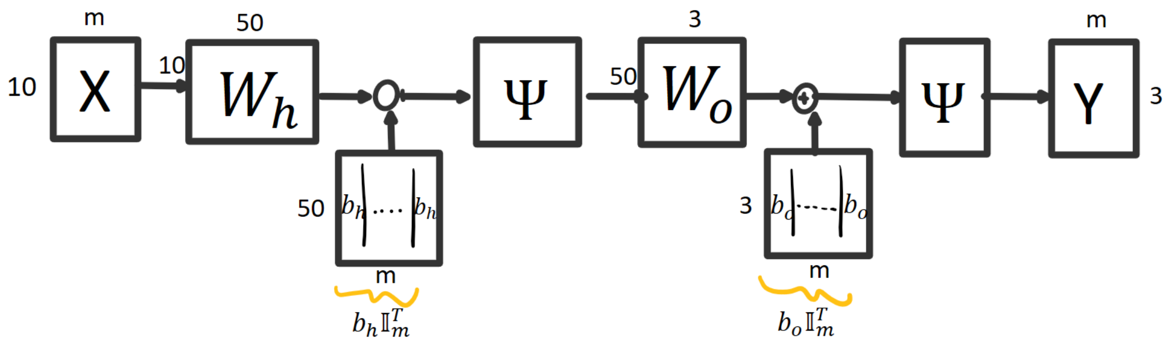
<u>תרגיל בית 1</u>

מגישים:

שם	ת.ז.
גבריאל חביב	
איטן חצרוני	302383856

<u>חלק תיאורטי</u>

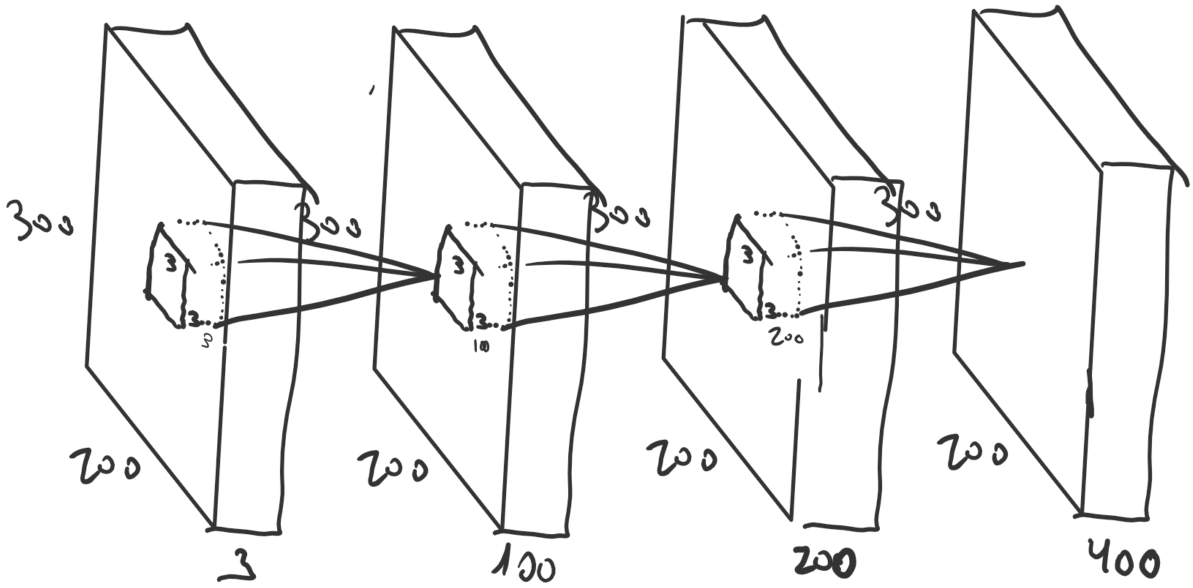
1. מבנה הרשת:



a. המימדים של X הם 10xm. 10 זהו גודל וקטור הכניסה, ו-m גודל ה-batch.
b. המימדים של W_h הם 10x5 . המימד של וקטור ה-bias, b_h הוא 50x1. יש להוסיף את אותו ה-bias לכל התמונות, לכן למעשה צריך להוסיף מטריצה בגודל 50xm, כאשר היא מורכבת מעמודות של הוקטור b_h . ניתן לכתוב זאת בתור $b_o\mathbb{I}_m^T$, כאשר $\mathbb{I}_m$ הוא וקטור אחדות בגודל m.
c. המימדים של W_o הם 50x3 ושל וקטור ה-b_o הם 3x1. גם כאן יש להוסיף מטריצה בגודל 3xm שעמודותיה בנויות מהוקטור b_o. ניתן לכתוב זאת בתור $b_o\mathbb{I}_m^T$.
d. הצורה של Y היא מטריצה בגודל $3xm$.
e. $Y = \Psi(W_0^T \Psi(W_h^T X + b_h \mathbb{I}_m^T) + b_o \mathbb{I}_m^T)$

כאשר Ψ היא פעולת ReLU איבר איבר על המטריצה

2. מבנה הרשת:



$(3 \times 3 \times 3 + 1) \times 100 \quad (3 \times 3 \times 100 + 1) \times 200 \quad (3 \times 3 \times 200 + 1) \times 400$

$2800 + 180200 + 720400 = 903400$

סה"כ ישנם 903400 פרמטרים.

3\. a. $\frac{\partial f}{\partial \gamma} = \sum_{i=1}^{m} \frac{\partial f}{\partial y_i} \cdot \frac{\partial y_i}{\partial \gamma} = \sum_{i=1}^{m} \frac{\partial f}{\partial y_i} \cdot \frac{\partial}{\partial \gamma}\left(\gamma \hat{x}_i + \beta\right) = \sum_{i=1}^{m} \frac{\partial f}{\partial y_i} \cdot \hat{x}_i$

b. $\frac{\partial f}{\partial \beta} = \sum_{i=1}^{m} \frac{\partial f}{\partial y_i} \cdot \frac{\partial y_i}{\partial \beta} = \sum_{i=1}^{m} \frac{\partial f}{\partial y_i} \cdot \frac{\partial}{\partial \beta}\left(\gamma \hat{x}_i + \beta\right) = \sum_{i=1}^{m} \frac{\partial f}{\partial y_i}$

c. $\frac{\partial f}{\partial \hat{x}_i} = \frac{\partial f}{\partial y_i} \cdot \frac{\partial y_i}{\partial \hat{x}_i} = \frac{\partial f}{\partial y_i} \cdot \frac{\partial}{\partial \hat{x}_i}(\gamma \hat{x}_i + \beta) = \frac{\partial f}{\partial y_i} \cdot \gamma$

d. $\frac{\partial f}{\partial \sigma^2} = \sum_{i=1}^{m} \frac{\partial f}{\partial y_i} \cdot \frac{\partial y_i}{\partial \sigma^2} = \sum_{i=1}^{m} \frac{\partial f}{\partial y_i} \cdot \frac{\partial y_i}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial \sigma^2} = \sum_{i=1}^{m} \frac{\partial f}{\partial y_i} \cdot \frac{\partial}{\partial \hat{x}_i}(\gamma \hat{x}_i + \beta) \cdot \frac{\partial}{\partial \sigma^2}\left(\frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}\right)$

$= -\frac{1}{2}\gamma \sum_{i=1}^{m} \frac{\partial f}{\partial y_i} \frac{x_i - \mu_B}{(\sigma_B^2 + \epsilon)^{3/2}} = \frac{-\gamma}{2(\sigma_B^2 + \epsilon)^{3/2}} \sum_{i=1}^{m} \frac{\partial f}{\partial y_i}(x_i - \mu_B)$

e. $\frac{\partial f}{\partial \mu_B} = \sum_{i=1}^{m} \frac{\partial f}{\partial y_i} \cdot \frac{\partial y_i}{\partial \mu_B} = \sum_{i=1}^{m} \frac{\partial f}{\partial y_i} \cdot \frac{\partial y_i}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial \mu_B} = \sum_{i=1}^{m} \frac{\partial f}{\partial y_i} \gamma \cdot \frac{\partial}{\partial \mu_B}\left(\frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}\right) =$

$\gamma \sum_{i=1}^{m} \frac{\partial f}{\partial y_i}\left(\frac{-1}{\sqrt{\sigma^2+\epsilon}} - \frac{x_i - \mu_B}{2(\sigma_B^2+\epsilon)^{3/2}} \cdot \frac{\partial \sigma_B^2}{\partial \mu_B}\right) = \gamma \sum \frac{\partial f}{\partial y_i}\left(-\frac{1}{\sqrt{\sigma^2+\epsilon}} + \cancel{\frac{x_i - \mu_B}{2(\sigma_B - \epsilon)^{3/2}} \cdot \frac{-2}{m} \sum (x_i - \mu_B)}\right)$

$m \cdot \underbrace{\frac{1}{m}\sum x_i}_{\mu_B} - m\mu_B = 0$

$= -\frac{\gamma}{\sqrt{\sigma_B^2 + \epsilon}} \sum_{i=1}^{m} \frac{\partial f}{\partial y_i}$

f. $\frac{\partial f}{\partial x_i} = \frac{\partial f}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial x_i} + \frac{\partial f}{\partial \mu_B} \cdot \frac{\partial \mu_B}{\partial x_i} + \frac{\partial f}{\partial \sigma_B^2} \cdot \frac{\partial \sigma_B^2}{\partial x_i} = \gamma \frac{\partial f}{\partial y_i} \cdot \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} - \frac{\gamma}{\sqrt{\sigma_B^2 + \epsilon}} \sum_{j=1}^{m} \frac{\partial f}{\partial y_j} \cdot \frac{1}{m}$

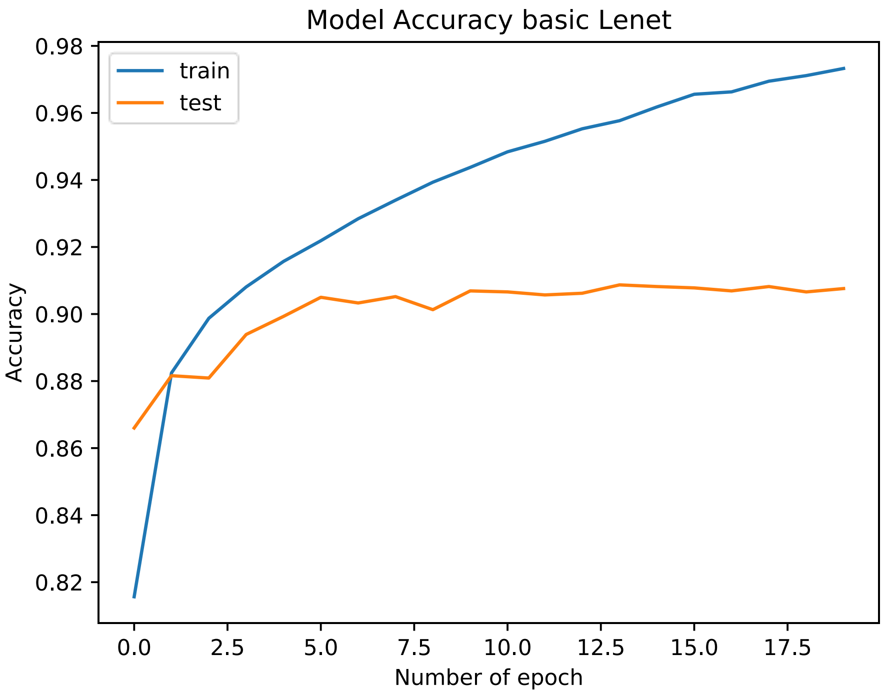
$- \frac{\gamma}{2(\sigma_B^2 + \epsilon)^{3/2}} \sum \frac{\partial f}{\partial y_j}(x_j - \mu_B) \frac{2}{m}(x_i - \mu_B)$

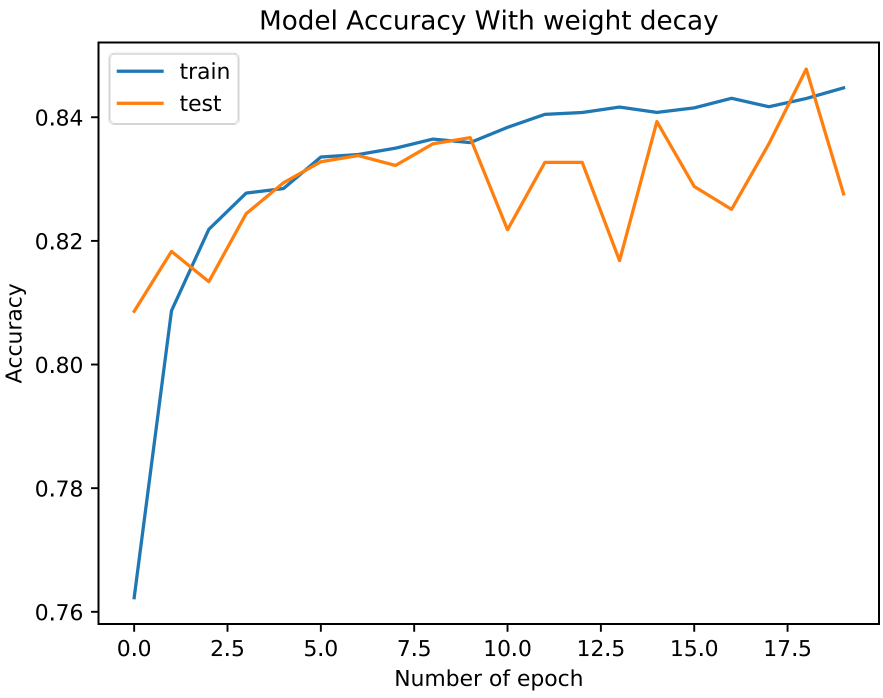
$= \frac{\gamma}{m\sqrt{\sigma_B^2 + \epsilon}}\left(m \frac{\partial f}{\partial y_i} - \sum_{j=1}^{m} \frac{\partial f}{\partial y_j} - \underbrace{\frac{1}{\sqrt{\sigma_B^2+\epsilon}}(x_i - \mu_B)}_{\hat{x}_i} \sum \frac{\partial f}{\partial y_j} \underbrace{\frac{(x_j - \mu_B)}{\sqrt{\sigma_B^2 + \epsilon}}}_{\hat{x}_j}\right)$

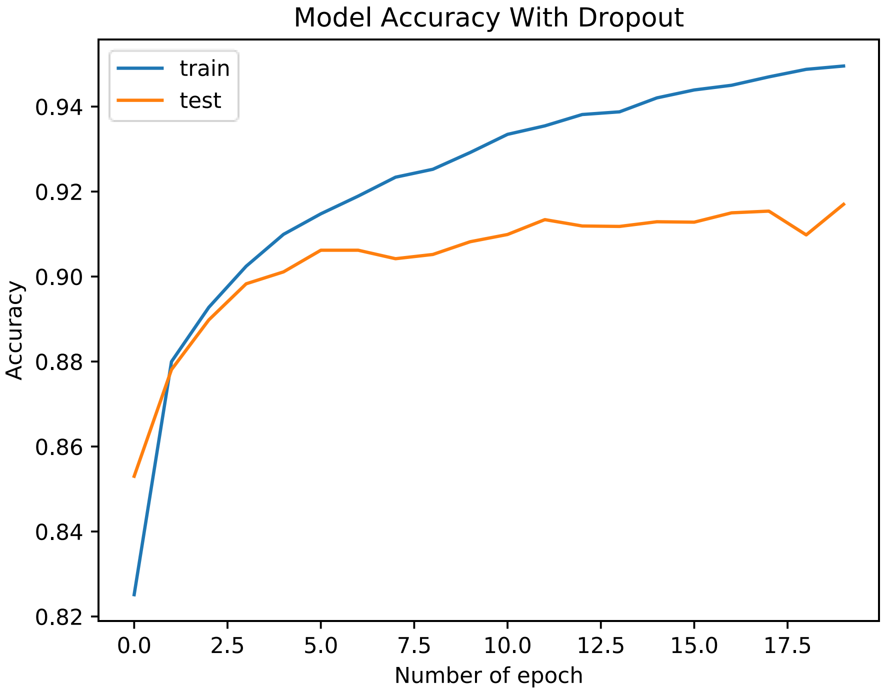
$= \frac{\gamma}{m\sqrt{\sigma_B^2 + \epsilon}}\left(m \frac{\partial f}{\partial y_i} - \sum_{j=1}^{m} \frac{\partial f}{\partial y_j} - \hat{x}_i \sum \frac{\partial f}{\partial y_j} \hat{x}_j\right)$

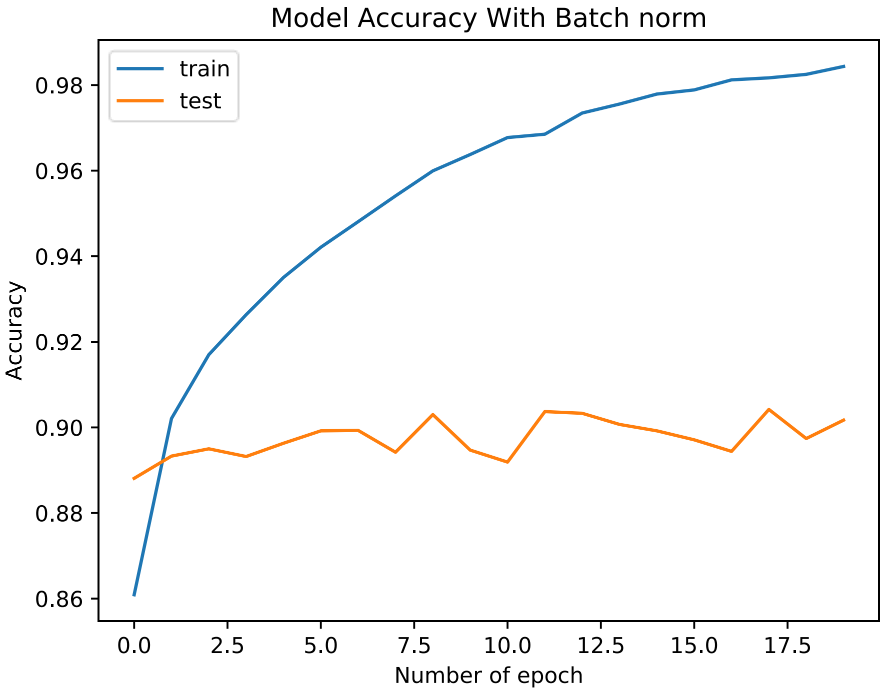
חלק מעשי

הגרפים המבוקשים:









סיכום ביצועים:

טכניקה	Accuracy
Normal	0.91
Dropout	0.92
Weight Decay	0.83
Batch Normalization	0.90

הערות:

1. ניתן לראות הביצועים של רוב הטכניקות דומים.
2. הביצועים של טכניקת weight decay נמוכים באופן יחסי, אך עם שגיאת ההכללה הנמוכה ביותר. הסיבה לכך היא שאמנם הרשת לא מגיעה למצב של over-fitting, אך הרגולריזציה מונעת ממנה להיות מורכבת מספיק בכדי לסווג את המידע. ניתן ככל הנראה בעזרת אופטימיזציה של איבר הרגולריזציה להגיע ל-trade-off מוצלח יותר בין ביצועים לשגיאת הכללה.
3. ניתן לראות שבעזרת batch normalization הרשת התכנסה לערך סופי בצורה מהירה, ומשם גדלה שגיאת ההכללה (וה-accuracy כמעט ולא השתנה).