دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

گزارش تمرین شماره 4

درس سیستم های هوشمند

پاییز 1401

امیرحسین بیرژندی

...

810198367

...

سوال 1:

Feedforward for X_1

$z = \tanh(W_1^T X_1 + B_1)$

$A_1 = W_1^T X_1 + B_1 = \begin{bmatrix} 42.71 \\ 40.22 \end{bmatrix} \Rightarrow z = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$k = \text{Sigmoid}(W_2^T z + B_2)$

$A_2 = W_2^T z + B_2 = \begin{bmatrix} 26.75 \\ 27.05 \\ 27.35 \end{bmatrix} \Rightarrow k = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

$p = \tanh(W_3^T k + B_3)$

$A_3 = W_3^T k + B_3 = \begin{bmatrix} 128.08 \\ 128.48 \end{bmatrix} \Rightarrow p = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$\hat{y} = \text{ReLU}(W_4^T p + B_4) = [1.78]$

$A_4 = W_4^T p + B_4 = 1.78$

Back propragation

$*$ → element wise

$$\frac{\partial E}{\partial W_4^T} = \frac{\partial E}{\partial A_4} p^T = \left(\frac{\partial E}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial A_4}\right) p^T = ((\hat{y} - y) * 1) p^T$$

$$= -5.22 \times 1 \times [1 \quad 1] = [-5.22 \quad -5.22] \longrightarrow \frac{\partial E}{\partial W_4^T} = [-5.22 \quad -5.22]$$

$$\frac{\partial E}{\partial W_3^T} = \frac{\partial E}{\partial A_3}(k)^T = \left(\frac{\partial E}{\partial p} * \frac{\partial p}{\partial A_3}\right) k^T$$

$$\frac{\partial E}{\partial p} = W_4 \cdot \frac{\partial E}{A_4} = W_4 \cdot \left(\frac{\partial E}{\partial \hat{y}} * \frac{\partial \hat{y}}{A_4}\right) = W_4 \cdot (\hat{y} - y) * \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\frac{\partial E}{\partial W_3^T} = \left(\frac{\partial E}{\partial p} * (1 - \tanh^2(A_3))\right) k^T = \left(\left[W_4 (\hat{y} - y) \begin{bmatrix} 0 \\ 0 \end{bmatrix}\right] * (1 - \tanh^2(A_3))\right) k^T$$

$$= \left(\begin{bmatrix} 1.16 \\ 1.36 \end{bmatrix} \cdot -5.22 \begin{bmatrix} 0 \\ 0 \end{bmatrix}\right)[1 \quad 1 \quad 1] = \begin{bmatrix} 0 \\ 0 \end{bmatrix}[1 \quad 1 \quad 1] \rightarrow \frac{\partial E}{\partial W_3^T} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\frac{\partial E}{\partial W_2^T} = \frac{\partial E}{\partial A_2}(z^T) = \left[\frac{\partial E}{\partial k} * \frac{\partial k}{\partial A_2}\right] z^T = \left[\left(W_3 \frac{\partial E}{\partial A_3}\right) * \frac{\partial k}{\partial A_2}\right] z^T$$

$$\frac{\partial E}{\partial A_3} = \frac{\partial E}{\partial p} * \frac{\partial p}{\partial A_3} = W_4 \cdot (\hat{y} - y) \begin{bmatrix} 0 \\ 0 \end{bmatrix} * (1 - \tanh^2(A_3))$$

$$\frac{\partial E}{\partial W_2^T} = \left[\left(W_3 ((W_4 \cdot \left(\frac{\partial E}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial A_4}\right)) * \frac{\partial p}{\partial A_3}\right)\right) * \frac{\partial k}{\partial A_2}\right] z^T$$

$$\frac{\partial E}{\partial W_2^T} = \left[\left(W_3\left(\left(W_4 \cdot (\hat{y}-y) * \begin{bmatrix}0\\0\end{bmatrix}\right) * \begin{bmatrix}1\\1\\1\end{bmatrix}\right)\begin{bmatrix}0\\0\\0\end{bmatrix}\right]\begin{bmatrix}1 & 1\end{bmatrix} = \begin{bmatrix}0 & 0\\0 & 0\\0 & 0\end{bmatrix}$$

$$\frac{\partial E}{\partial W_2^T} = \begin{bmatrix}0 & 0\\0 & 0\\0 & 0\end{bmatrix}$$

$$\frac{\partial E}{\partial W_1^T} = \frac{\partial E}{\partial A_1}(X_1^T) = \left[\frac{\partial E}{\partial Z} * \frac{\partial Z}{\partial A_1}\right]X_1^T = \left[\left(W_2 \frac{\partial E}{\partial A_2}\right) * \frac{\partial Z}{\partial A_1}\right]X_1^T$$

$$\frac{\partial E}{\partial A_2} = \frac{\partial E}{\partial K} * \frac{\partial K}{\partial A_2} = \left(W_3 \cdot \frac{\partial E}{\partial A_3}\right) * \frac{\partial K}{\partial A_2} = \left(W_3 \cdot \left[\frac{\partial E}{\partial P} * \frac{\partial P}{\partial A_3}\right]\right) * \frac{\partial K}{\partial A_2}$$

$$= \left(W_3 \cdot \left[\left(W_4 \cdot \left(\frac{\partial E}{\partial \hat{y}} \odot \frac{\partial \hat{y}}{A_4}\right)\right) * \frac{\partial P}{\partial A_3}\right]\right) \odot \frac{\partial K}{\partial A_2}$$

$$\frac{\partial E}{\partial W_1^T} = \left[\left(W_2\left[\left(W_3 \cdot \left[W_4 \cdot \left(\frac{\partial E}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial A_4}\right)\right) * \frac{\partial P}{\partial A_3}\right]\right) * \frac{\partial K}{\partial A_2}\right]\right) * \frac{\partial Z}{\partial A_1}\right]X_1^T$$

$$\frac{\partial E}{\partial W_1^T} = \left[\left(W_2 \cdot \left[W_3 \cdot \begin{bmatrix}0\\0\end{bmatrix} * \begin{bmatrix}1\\1\\1\end{bmatrix}\right]\begin{bmatrix}0\\0\\0\end{bmatrix}\right) * \frac{\partial Z}{\partial A_1}\right]X_1^T = \begin{bmatrix}0 & 0 & 0\\0 & 0 & 0\end{bmatrix}$$

$$\frac{\partial E}{\partial B_4} = \frac{\partial E}{\partial A_4} = \frac{\partial E}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial A_4} = (\hat{y}-y)[1] = -5.22 \rightarrow \frac{\partial E}{\partial B_4} = -5.22$$

$$\frac{\partial E}{\partial B_3} = \frac{\partial E}{\partial A_3} = \frac{\partial E}{\partial P} * \frac{\partial P}{\partial A_3} = \left(W_4 \cdot \frac{\partial E}{\partial A_4}\right) * \frac{\partial P}{\partial A_3} = \left(W_4 \cdot \left[\frac{\partial E}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial A_4}\right]\right) * \frac{\partial P}{\partial A_3}$$

$$\frac{\partial E}{\partial B_3} = \left(\begin{bmatrix}1.16\\1.96\end{bmatrix} \cdot -5.22\right) * (1-\tanh^2(A_3)) = \begin{bmatrix}0\\0\end{bmatrix} \rightarrow \frac{\partial E}{\partial B_3} = \begin{bmatrix}0\\0\end{bmatrix}$$

$$\frac{\partial E}{\partial B_2} = \frac{\partial E}{\partial A_2} = \frac{\partial E}{\partial K} * \frac{\partial K}{\partial A_2} = \left(W_3 \cdot \frac{\partial E}{\partial A_3}\right) * \frac{\partial K}{\partial A_2}$$

$$= \left(W_3 \cdot \left[\left(W_4 \cdot \left[\frac{\partial E}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial A_4}\right]\right) * \frac{\partial P}{\partial A_3}\right]\right) * \frac{\partial K}{\partial A_2}$$

$$\frac{\partial E}{\partial B_2} = \left(\begin{bmatrix}42.12 & 42.22\\42.32 & 42.42\\42.52 & 42.62\end{bmatrix} \cdot \begin{bmatrix}0\\0\end{bmatrix}\right) * \begin{bmatrix}1\\1\\1\end{bmatrix}\begin{bmatrix}0\\0\\0\end{bmatrix} = \begin{bmatrix}0\\0\\0\end{bmatrix} \rightarrow \frac{\partial E}{\partial B_2} = \begin{bmatrix}0\\0\\0\end{bmatrix}$$

$$\frac{\partial E}{\partial B_1} = \frac{\partial E}{\partial A_1} = \frac{\partial E}{\partial Z} * \frac{\partial Z}{\partial A_1} = \left(W_2 \cdot \frac{\partial E}{\partial A_2}\right) * \frac{\partial Z}{\partial A_1}$$

$$= \left(\begin{bmatrix}7.15 & 7.25 & 7.35\\6.45 & 6.65 & 6.65\end{bmatrix} \cdot \begin{bmatrix}0\\0\\0\end{bmatrix}\right) * \begin{bmatrix}0\\0\end{bmatrix} \rightarrow \frac{\partial E}{\partial B_1} = \begin{bmatrix}0\\0\end{bmatrix}$$

• Gradient update

$$W_4^T = W_4^T - \eta \frac{\partial E}{\partial W_4^T} = [1.16 \quad 1.36] - 0.1\,[-5.22 \quad -5.22] = [1.682 \quad 1.882]$$

$$W_3^T = W_3^T - \eta \frac{\partial E}{\partial W_3^T} = \begin{bmatrix} 42.12 & 42.32 & 42.52 \\ 42.22 & 42.42 & 42.62 \end{bmatrix} - 0.1 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 42.12 & 42.32 & 42.52 \\ 42.22 & 42.42 & 42.62 \end{bmatrix}$$

$$W_2^T = W_2^T - \eta \frac{\partial E}{\partial W_2^T} = \begin{bmatrix} 7.15 & 6.45 \\ 7.25 & 6.55 \\ 7.35 & 6.65 \end{bmatrix} - 0.1 \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 7.15 & 6.45 \\ 7.25 & 6.55 \\ 7.35 & 6.65 \end{bmatrix}$$

$$W_1^T = W_1^T - \eta \frac{\partial E}{\partial W_1^T} = \begin{bmatrix} 0.17 & 0.37 & 0.57 \\ 0.26 & 0.46 & 0.46 \end{bmatrix} - 0.1 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0.17 & 0.37 & 0.57 \\ 0.26 & 0.46 & 0.46 \end{bmatrix}$$

$$B_4 = B_4 - \eta \frac{\partial E}{\partial B_4} = [-0.74] - 0.1(-5.22) = -0.218$$

$$B_3 = B_3 - \eta \frac{\partial E}{\partial B_3} = \begin{bmatrix} 1.12 \\ 1.22 \end{bmatrix} - 0.1 \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1.12 \\ 1.22 \end{bmatrix}$$

$$B_2 = B_2 - \eta \frac{\partial E}{\partial B_2} = \begin{bmatrix} 13.15 \\ 13.25 \\ 13.35 \end{bmatrix} - 0.1 \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 13.15 \\ 13.25 \\ 13.35 \end{bmatrix}$$

$$B_1 = B_1 - \eta \frac{\partial E}{\partial B_1} = \begin{bmatrix} 0.71 \\ 0.62 \end{bmatrix} - 0.1 \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.71 \\ 0.62 \end{bmatrix}$$

Feed Forward for X_2

$$z = \tanh(W_1^T X_2 + B_1)$$

$$A_1 = W_1^T X_2 + B_1 = \begin{bmatrix} 7.74 \\ 8.16 \end{bmatrix} \Rightarrow z = \tanh(A_1) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$K = \text{Sigmoid}(W_2^T z + B_2)$$

$$A_2 = W_2^T z + B_2 = \begin{bmatrix} 26.75 \\ 27.05 \\ 27.35 \end{bmatrix} \Rightarrow K = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$P = \tanh(W_3^T K + B_3)$$

$$A_3 = W_3^T K + B_3 = \begin{bmatrix} 128.08 \\ 128.48 \end{bmatrix} \Rightarrow P = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\hat{y} = \text{ReLU}(W_4^T P + B_4) = [3.346]$$

Back propagation

$$\frac{\partial E}{\partial w_4^T} = \frac{\partial E}{\partial A_4} P^T = \left(\frac{\partial E}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial A_4}\right) P^T = (\hat{y}-y)\begin{bmatrix}1\end{bmatrix} \times [1 \quad 1] = [-2.654 \quad -2.654]$$

$$\frac{\partial E}{\partial w_3^T} = \frac{\partial E}{\partial A_3}(K)^T = \left(\frac{\partial E}{\partial P} * \frac{\partial P}{\partial A_3}\right) K^T$$

$$\frac{\partial E}{\partial P} = w_4 \cdot \frac{\partial E}{\partial A_4} = w_4\left(\frac{\partial E}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial A_4}\right) = w_4 \cdot (\hat{y}-y)\begin{bmatrix}0\\0\end{bmatrix}$$

$$\frac{\partial E}{\partial w_3^T} = \left(\frac{\partial E}{\partial P} * (1-\tanh^2(A_3))\right) K^T = \left(\left[w_4 (\hat{y}-y)\begin{bmatrix}0\\0\end{bmatrix} * (1-\tanh^2(A_3))\right) K^T$$

$$= \left(\begin{bmatrix}1.682\\1.882\end{bmatrix} \cdot -2.654 * \begin{bmatrix}0\\0\end{bmatrix}\right)[1 \quad 1 \quad 1] \Rightarrow \frac{\partial E}{\partial w_3^T} = \begin{bmatrix}0 & 0 & 0\\0 & 0 & 0\end{bmatrix}$$

$$\frac{\partial E}{\partial w_2^T} = \frac{\partial E}{\partial A_2} Z^T = \left[\frac{\partial E}{\partial K} * \frac{\partial K}{\partial A_2}\right] Z^T = \left[\left(w_3 \frac{\partial E}{\partial A_3}\right) * \frac{\partial K}{\partial A_2}\right] Z^T$$

$$\frac{\partial E}{\partial A_3} = \frac{\partial E}{\partial P} * \frac{\partial P}{\partial A_3} = w_4 \cdot (\hat{y}-y)\begin{bmatrix}0\\0\end{bmatrix} * (1-\tanh^2(A_3))$$

$$\frac{\partial E}{\partial w_2^T} = \left[\left(w_3\left(\left(w_4 \cdot \left(\frac{\partial E}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial A_4}\right)\right) * \frac{\partial P}{\partial A_3}\right)\right) * \frac{\partial K}{\partial A_2}\right] Z^T$$

$$\frac{\partial E}{\partial w_1^T} = \left[\left(w_3\left((w_4 \cdot (\hat{y}-y) * \begin{bmatrix}0\\0\end{bmatrix}\right)\right) * (1-\tanh^2(A_3))\right) * \text{Sigmoid}(A_2)(1-\text{Sigmoid}(A_2))\right] Z^T$$

$$\frac{\partial E}{\partial w_2^T} = \left[\left(\begin{bmatrix}42.12 & 42.22\\42.32 & 42.42\\42.52 & 42.62\end{bmatrix}\left(\left(\begin{bmatrix}1.682\\1.882\end{bmatrix} \cdot (-2.654)\right) * \begin{bmatrix}0\\0\end{bmatrix}\right) * \begin{bmatrix}1\\1\\1\end{bmatrix}\right)\begin{bmatrix}0\\0\\0\end{bmatrix}\right][1 \quad 1] = \begin{bmatrix}0 & 0\\0 & 0\\0 & 0\end{bmatrix}$$

$$\frac{\partial E}{\partial w_2^T} = \begin{bmatrix}0 & 0\\0 & 0\\0 & 0\end{bmatrix}$$

$$\frac{\partial E}{\partial w_1^T} = \frac{\partial E}{\partial A_1}(X_1^T) = \left[\frac{\partial E}{\partial Z} * \frac{\partial Z}{\partial A_1}\right] X_1^T = \left[\left(w_2 \frac{\partial E}{\partial A_2}\right) * \frac{\partial Z}{\partial A_1}\right] X_1^T$$

$$\frac{\partial E}{\partial A_2} = \frac{\partial E}{\partial K} * \frac{\partial K}{\partial A_2} = \left(w_3 \cdot \frac{\partial E}{\partial A_3}\right) * \frac{\partial K}{\partial A_2} = \left(w_3 \cdot \left[\frac{\partial E}{\partial P} * \frac{\partial P}{\partial A_3}\right]\right) * \frac{\partial K}{\partial A_2}$$

$$= \left(w_3 \cdot \left[\left(w_4 \cdot \left(\frac{\partial E}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial A_4}\right)\right) * \frac{\partial P}{\partial A_3}\right]\right) * \frac{\partial M}{\partial A_2}$$

$$\frac{\partial E}{\partial w_1^T} = \left[\left(w_2 \cdot \left[(w_3 \cdot [\begin{smallmatrix}0\\0\end{smallmatrix}]) * \begin{bmatrix}1\\1\\1\end{bmatrix}\begin{bmatrix}0\\0\\0\end{bmatrix}\right]\right) * \frac{\partial Z}{\partial A_1}\right] X_2^T$$

$$= \left[\left(\begin{bmatrix}7.45 & 7.25 & 7.35\\6.45 & 6.25 & 6.65\end{bmatrix} \cdot \begin{bmatrix}0\\0\\0\end{bmatrix}\right) * \frac{\partial Z}{\partial A_1}\right] X_2^T = \left(\begin{bmatrix}0\\0\end{bmatrix} * (1-\tanh^2(A_1))\right)[6 \quad 7 \quad 6]$$

$$\frac{\partial E}{\partial W_1^T} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\frac{\partial E}{\partial B_4} = \frac{\partial E}{\partial A_4} = \frac{\partial E}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial A_4} = (\hat{y} - y)\begin{bmatrix} 1 \end{bmatrix} = -2.654 \rightarrow \frac{\partial E}{\partial B_4} = -2.654$$

$$\frac{\partial E}{\partial B_3} = \frac{\partial E}{\partial A_3} = \frac{\partial E}{\partial P} * \frac{\partial P}{\partial A_3} = \left(W_4 \cdot \frac{\partial E}{\partial A_4}\right) * \frac{\partial P}{\partial A_3} = \left(W_4 \cdot \left[\frac{\partial E}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial A_4}\right]\right) * \frac{\partial P}{\partial A_3}$$

$$\frac{\partial E}{\partial B_3} = \left(\begin{bmatrix} 1.682 \\ 1.882 \end{bmatrix} \cdot -2654\right) * (1 - \tanh^2(A_3)) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \frac{\partial E}{\partial B_3} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\frac{\partial E}{\partial B_2} = \frac{\partial E}{\partial A_2} = \frac{\partial E}{\partial K} * \frac{\partial K}{\partial A_2} = \left(W_3 \cdot \frac{\partial E}{\partial A_3}\right) * \frac{\partial K}{\partial A_2}$$

$$\frac{\partial E}{\partial B_2} = \left(W_3 \cdot \left[\left(W_4 \cdot \left[\frac{\partial E}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial A_4}\right] * \frac{\partial P}{\partial A_3}\right]\right) * \frac{\partial K}{\partial A_2}\right.$$

$$\frac{\partial E}{\partial B_2} = \left(\begin{bmatrix} 42.12 & 42.22 \\ 42.32 & 42.42 \\ 42.52 & 42.62 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 0 \end{bmatrix}\right) * \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \rightarrow \frac{\partial E}{\partial B_2} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\frac{\partial E}{\partial B_1} = \frac{\partial E}{\partial A_1} = \frac{\partial E}{\partial Z} * \frac{\partial Z}{\partial A_1} = \left(W_2 \cdot \frac{\partial E}{\partial A_2}\right) * \frac{\partial Z}{\partial A_1}$$

$$\frac{\partial E}{\partial B_1} = \left(\begin{bmatrix} 7.15 & 7.25 & 7.35 \\ 6.45 & 6.55 & 6.65 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}\right) * \begin{bmatrix} 0 \\ 0 \end{bmatrix} \rightarrow \frac{\partial E}{\partial B_1} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Gradient Descent Update

$$W_4^T = W_4^T - \eta \frac{\partial E}{\partial W_4^T} = \begin{bmatrix} 1.682 & 1.882 \end{bmatrix} - 0.1 \begin{bmatrix} -2.654 & -2.654 \end{bmatrix} = \begin{bmatrix} 1.9474 & 21474 \end{bmatrix}$$

$$W_3^T = W_3^T - \eta \frac{\partial E}{\partial W_3^T} = \begin{bmatrix} 42.12 & 42.32 & 42.52 \\ 42.22 & 42.42 & 42.62 \end{bmatrix} - 0.1 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 42.12 & 42.32 & 42.52 \\ 42.22 & 42.42 & 42.62 \end{bmatrix}$$

$$W_2^T = W_2^T - \eta \frac{\partial E}{\partial W_2^T} = \begin{bmatrix} 7.15 & 6.45 \\ 7.25 & 6.55 \\ 7.35 & 6.65 \end{bmatrix} - 0.1 \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 7.15 & 6.45 \\ 7.25 & 6.55 \\ 7.35 & 6.65 \end{bmatrix}$$

$$W_1^T = W_1^T - \eta \frac{\partial E}{\partial W_1^T} = \begin{bmatrix} 0.17 & 0.37 & 0.57 \\ 0.26 & 0.46 & 0.46 \end{bmatrix} - 0.1 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0.17 & 0.37 & 0.57 \\ 0.26 & 0.46 & 0.46 \end{bmatrix}$$

$$B_4 = B_4 - \eta \frac{\partial E}{\partial B_4} = [-0.218] - 0.1(-2.654) = 0.474$$

$$B_3 = B_3 - \eta \frac{\partial E}{\partial B_3} = \begin{bmatrix} 1.12 \\ 1.22 \end{bmatrix} - 0.1 \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1.12 \\ 1.22 \end{bmatrix}$$

$$B_2 = B_2 - \eta \frac{\partial E}{\partial B_2} = \begin{bmatrix} 13.15 \\ 13.25 \\ 13.35 \end{bmatrix} - 0.1 \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 13.15 \\ 13.25 \\ 13.35 \end{bmatrix}$$

$$B_1 = B_1 - \eta \frac{\partial E}{\partial B_1} = \begin{bmatrix} 0.71 \\ 0.62 \end{bmatrix} - 0.1 \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.71 \\ 0.62 \end{bmatrix}$$

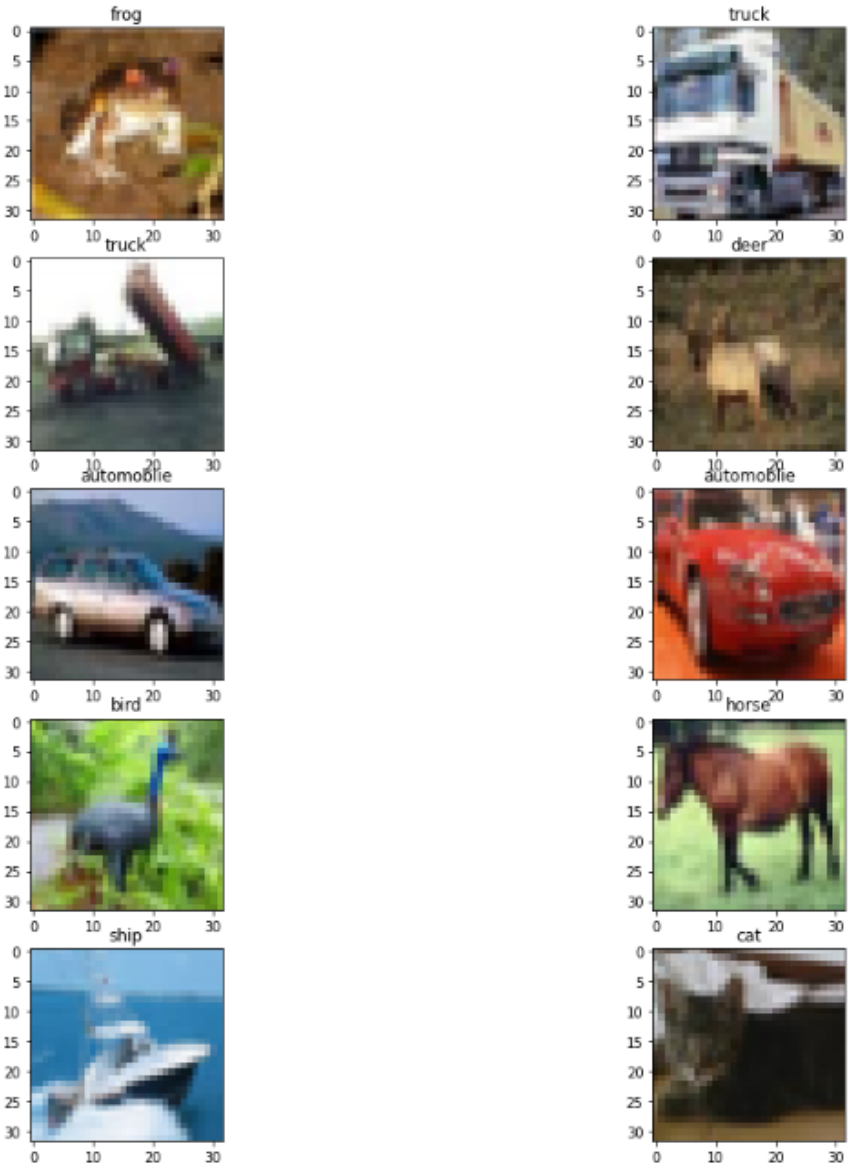
حال نتایج را با محاسبات در کد پایتون مقایسه می‌کنیم.

```
initial values:
W1:  [[0.17 0.37 0.57]
 [0.26 0.46 0.46]]
B1:  [[0.71]
 [0.62]]
W2:  [[7.15 6.45]
 [7.25 6.55]
 [7.35 6.65]]
B2:  [[13.15]
 [13.25]
 [13.35]]
W3:  [[42.12 42.32 42.52]
 [42.22 42.42 42.62]]
B3:  [[1.12]
 [1.22]]
W4:  [[1.16 1.36]]
B4:  [[-0.74]]

 first iteration:
 W1:  [[0.17 0.37 0.57]
  [0.26 0.46 0.46]]
 B1:  [[0.71]
  [0.62]]
 W2:  [[7.15 6.45]
  [7.25 6.55]
  [7.35 6.65]]
 B2:  [[13.15]
  [13.25]
  [13.35]]
 W3:  [[42.12 42.32 42.52]
  [42.22 42.42 42.62]]
 B3:  [[1.12]
  [1.22]]
 W4:  [[1.682 1.882]]
 B4:  [[-0.218]]

 second iteration:
 W1:  [[0.17 0.37 0.57]
  [0.26 0.46 0.46]]
 B1:  [[0.71]
  [0.62]]
 W2:  [[7.15 6.45]
  [7.25 6.55]
  [7.35 6.65]]
 B2:  [[13.15]
  [13.25]
  [13.35]]
 W3:  [[42.12 42.32 42.52]
  [42.22 42.42 42.62]]
 B3:  [[1.12]
  [1.22]]
 W4:  [[1.9474 2.1474]]
 B4:  [[0.0474]]
```

سوال 2 :


frog
truck
truck
deer
automobile
automobile
bird
horse
ship
cat

الف) استفاده از شبکه MLP

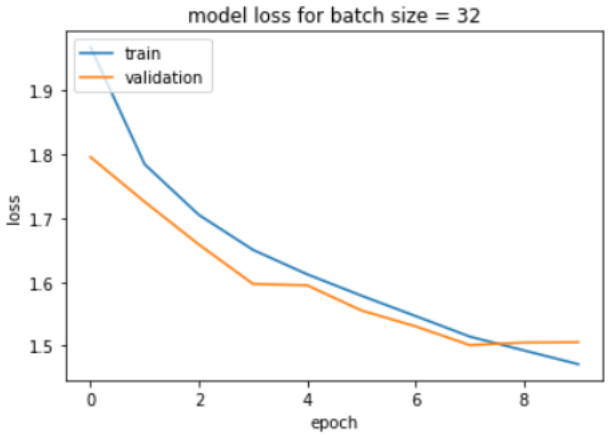
1- استفاده از Batch size های متفاوت:

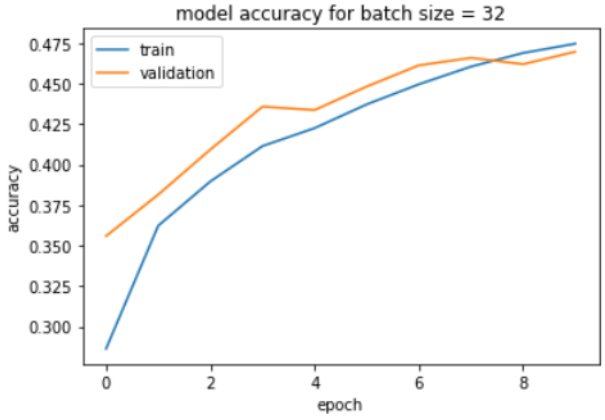
Batch size: 32

Accuracy & Loss:

```
Test loss: 1.4839056730270386
Test acc: 0.46630001068115234
```

Graphs:



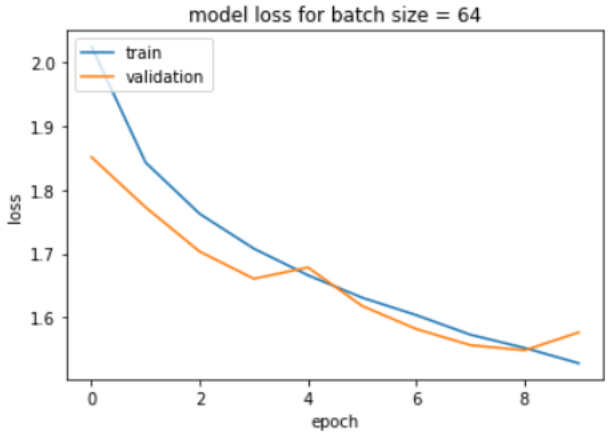


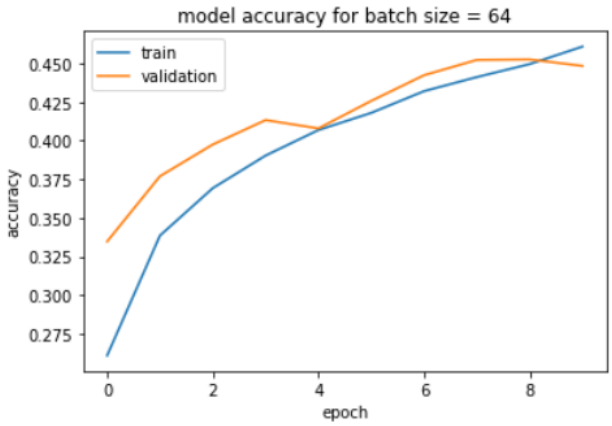
Batch size: 64

Accuracy & Loss:

```
Test loss: 1.5459495782852173
Test acc: 0.45410001277923584
```

Graphs:



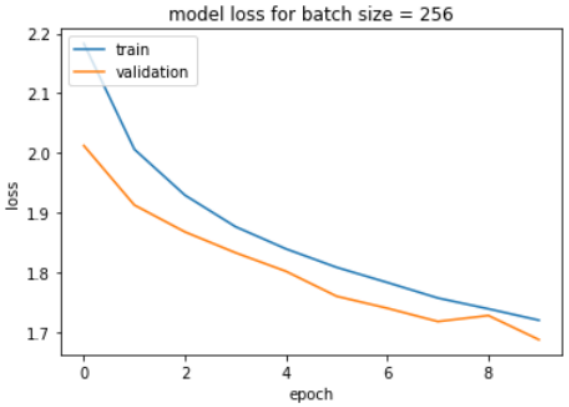


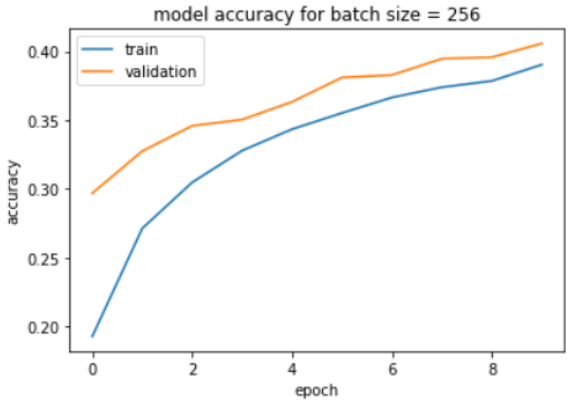
Batch size: 256

Accuracy & Loss:

```
Test loss: 1.6612892150878906
Test acc: 0.41190001368522644
```

Graphs:





تحلیل:

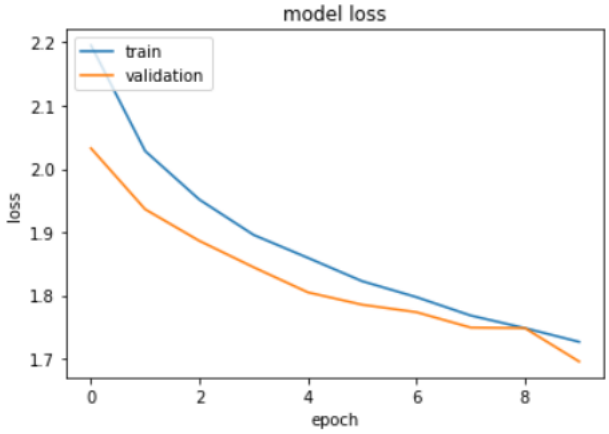
همانطور که مشاهده می‌شود با افزایش Batch Size دقت شبکه بر روی داده های تست تغییر مشهودی نکرده است(کمی کمتر شده است). نکته ای که مشهودتر است سرعت آموزش شبکه است که با افزایش Batch Size بسیار افزایش پیدا کرده است.

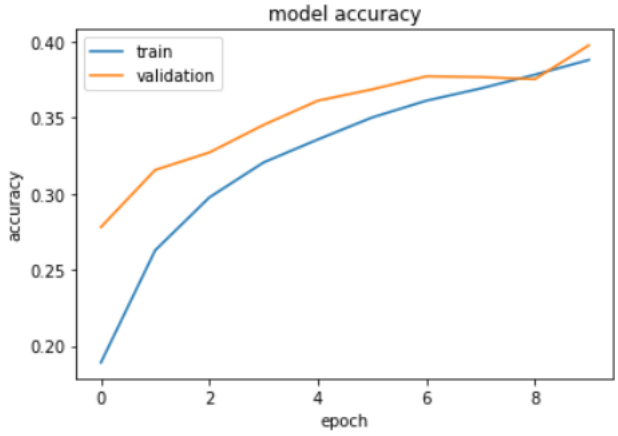
دلیل این موضوع هم این است که با افزایش Batch داده ها به قسمت های کمتری تبدیل می‌شوند و حجم محاسبات کمتر می‌شود. البته دقت شود که این موضوع یک trade off است زیرا افزایش بیش از حد این مقدار باعث زیاد شدن داده های هر بخش می‌شود که دوباره حجم محاسبات زیاد می‌شود. پس باید انتخاب بهینه Batch Size دقت کنیم که نه زیاد باشد و نه کم تا بهترین سرعت را داشته باشیم.

سوال 2 - اثر توابع فعالساز

Relu & 256

```
Test loss: 1.6691114902496338
Test acc: 0.4065000116825104
```

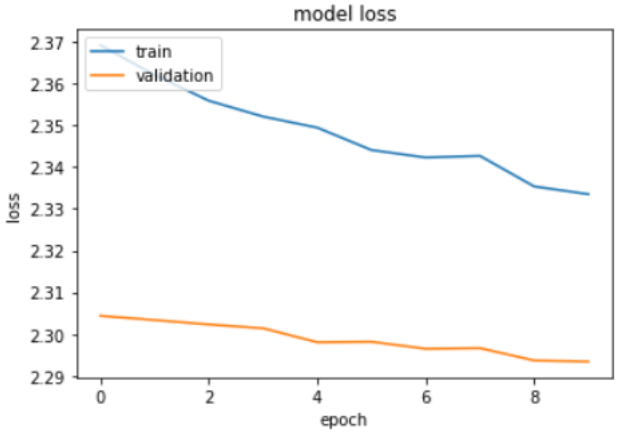


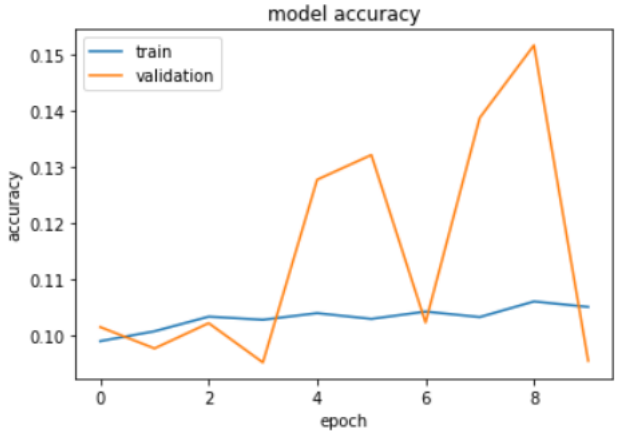



```
[[475  38  37  13  19  20  23  67 211  97]
 [ 40 430   5  13   8  50  33  52 102 267]
 [143  35 194  27 161 122 137 104  37  40]
 [ 44  65  50 134  39 291 110 134  39  94]
 [ 79  29 109  18 321  93 141 141  34  35]
 [ 33  29  71  75  53 398 105 145  48  43]
 [ 13  39  67  32 122 132 455  65  14  61]
 [ 55  45  32  23  78  80  41 516  24 106]
 [129  75   9  10   8  45   6  26 559 133]
 [ 45 145   5  14   6  25  25  56  96 583]]
```

Sigmoid & 256

```
Test loss: 2.2928385734558105
Test acc: 0.09939999878406525
```

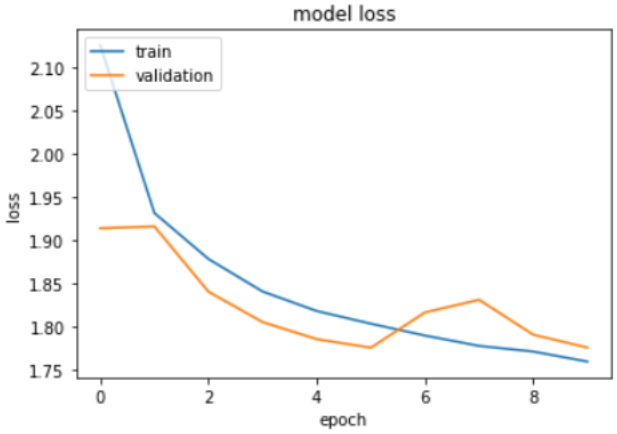


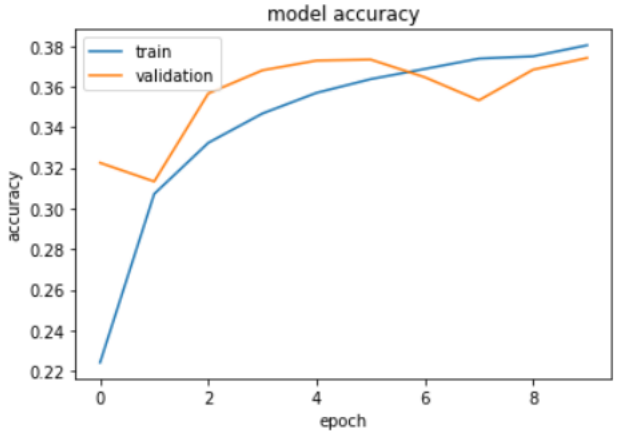



```
[[  3   0 956   0  40   0   1   0   0   0]
 [  0   0 885   0 112   0   3   0   0   0]
 [  0   0 932   0  65   0   3   0   0   0]
 [  0   0 915   0  84   0   1   0   0   0]
 [  0   0 943   0  53   0   4   0   0   0]
 [  0   0 920   0  78   0   2   0   0   0]
 [  0   0 879   0 115   0   6   0   0   0]
 [  0   0 923   0  76   0   1   0   0   0]
 [  1   0 938   0  61   0   0   0   0   0]
 [  0   0 940   0  59   0   1   0   0   0]]
```

Tanh & 256

```
Test loss: 1.7545132637023926
Test acc: 0.3824999928474426
```






```
[[718  49  16  16   8   8  31  14 109  31]
 [162 515  10  25   8  32  46  21  77 104]
 [289  61 147  74  73  72 195  39  35  15]
 [171 103  39 233  29 163 148  37  35  42]
 [169  36  97  69 227  67 235  58  26  16]
 [167  53  57 148  40 292 136  43  46  18]
 [ 74  64  47  91  55  52 557  22  18  20]
 [190  65  43  63  70  56  77 346  30  60]
 [352  81   3  21   2  33  10   6 442  50]
 [197 218   4  24   8  16  45  23 117 348]]
```

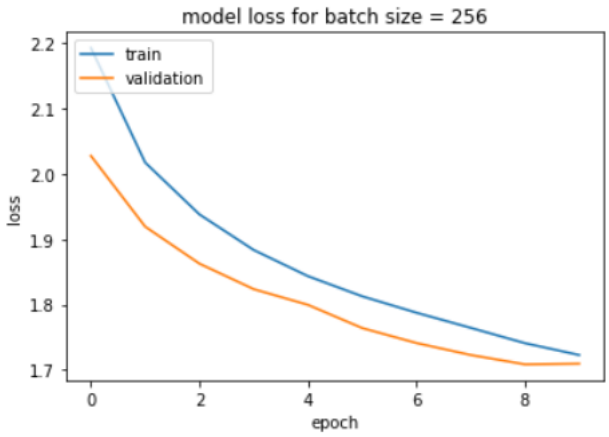
تحلیل:

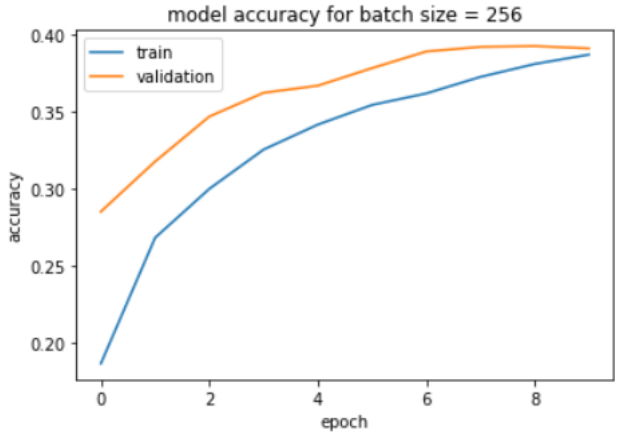
در این مرحله باید بهترین تابع فعالساز را انتخاب کنیم. همانطور که مشاهده می‌شود Relu بهترین عملکرد را دارا است. Relu بسیار بهتر روی داده ها آموزش داده می‌شود در نتیجه با تعداد ایپاک مناسب (برای جلوگیری از overfitting) Relu بهترین نتیجه را خواهد داشت. پس باید دقت کرد Relu خطر overfitting را دارد اما هم سرعت بهتری نسبت به بقیه و هم در تعداد epoch مناسب دقت بهتری خواهد داشت.

سوال 3 – اثر توابع هزینه

Categorical Cross Entropy & Relu & 256

```
Test loss: 1.685102105140686
Test acc: 0.4025999903678894
```

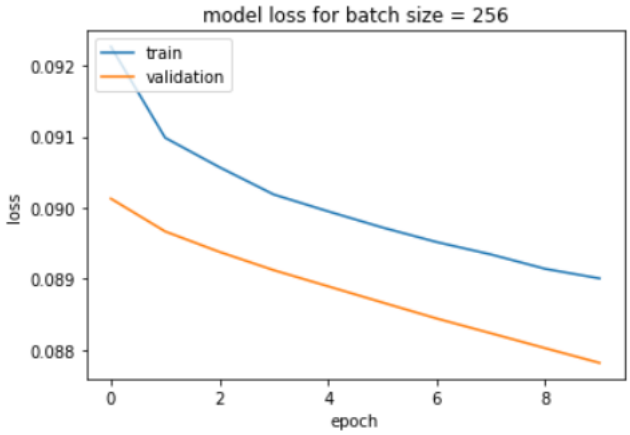


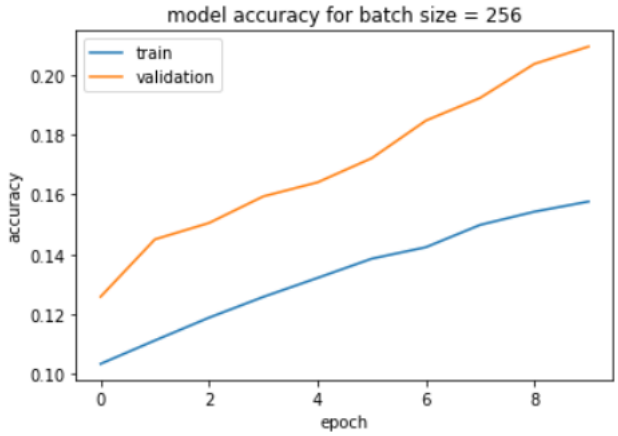



```
[[603  21 107   6  12   4  22  62 149  14]
 [160 342  25  23  24  30  42  70 151 133]
 [158  16 415  25 120  58  97  77  26   8]
 [ 87  31 192 134  73 188 114 118  35  28]
 [ 98  12 286  15 309  30 108 113  22   7]
 [ 58  13 181  81  98 297  89 131  40  12]
 [ 19  15 249  24 128  65 398  72  20  10]
 [ 78  32 106  30 103  43  32 526  25  25]
 [291  42  34   5   8  26   6  34 514  40]
 [166  96  23  17  17  20  45 117 164 335]]
```

MeanSquareError & Relu & 256

```
Test loss: 0.08791651576757431
Test acc: 0.20900000631809235
```

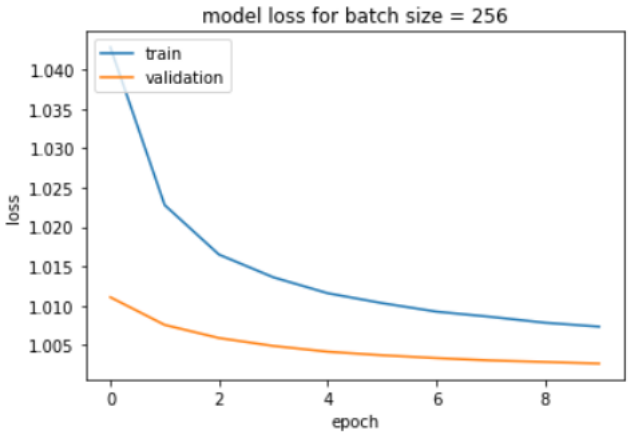


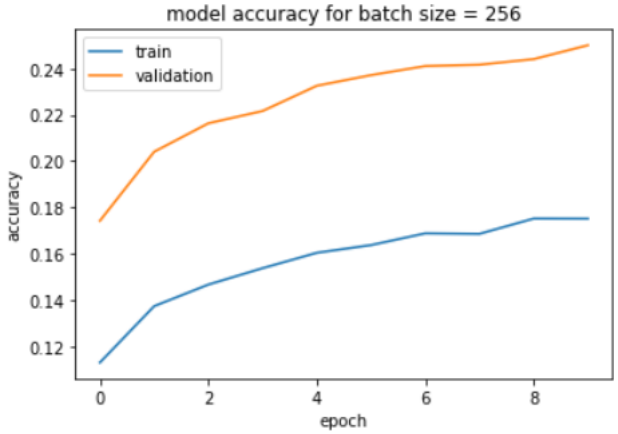



```
[[624   6  11  10  20  15   1   2 277  34]
 [288  15  31  32  55  72  16  15 371 105]
 [514   5  54  17 161  65  20   8 124  32]
 [350  11  81  55 117 162  13  11 133  67]
 [451   5  64  25 231  76  21   6  94  27]
 [367   3  76  47 127 200  18   4 119  39]
 [301   5  53  45 256 120  45  17 108  50]
 [340  10  79  49 137  80  14   4 186 101]
 [258   3   7  15   9  48   3   2 619  36]
 [258  14  25   9  41  15   5   6 424 203]]
```

Categorical Hinge Loss & Relu & 256

```
Test loss: 1.00264573097229
Test acc: 0.25290000438690186
```






```
[[369  28  90  25  19  18  31  42 224 154]
 [ 69 147  67  58  41  36  73  65 157 287]
 [142  38 243  61  79  53  90 118  80  96]
 [ 75  65 151 144  53 109 101 119  64 119]
 [ 61  31 223  91 135  45 103 146  63 102]
 [ 46  31 147 126  62 154 116 126  77 115]
 [ 28  48 165 124  78  62 209 112  49 125]
 [ 53  54 148  74  51  70  73 213  66 198]
 [127  36  35  26  19  25  39  33 453 207]
 [ 52  98  49  31  21  24  53  56 154 462]]
```

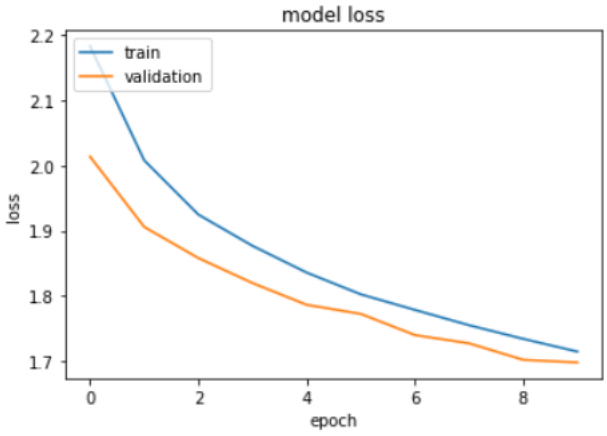
تحلیل:

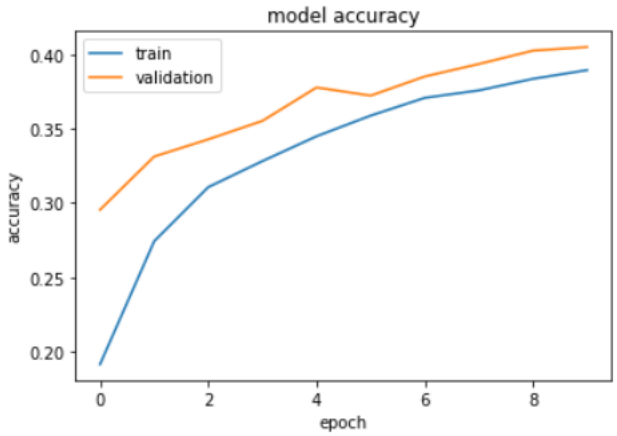
دلیل اینکه cross entropy بهترین نتیجه را منجر می‌شود این است که این تابع هزینه روی شبکه هایی که خروجی بین 0 و 1 دارند نتیجه خیلی خوبی می‌دهد و از آنجایی که لایه خروجی ما نیز با softmax است این تابع هزینه بهترین نتیجه خواهد داشت. حال MSE نتیجه مطلوبی روی این شبکه نمی‌دهد زیرا مناسب regression می باشد

سوال 4 – اثر بهینه‌ساز

SGD & Categorical Cross Entropy & Relu & 256

```
Test loss: 1.6746406555175781
Test acc: 0.4106000065803528
```

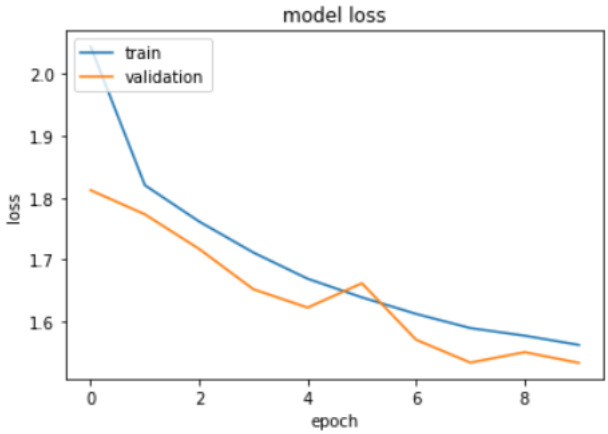


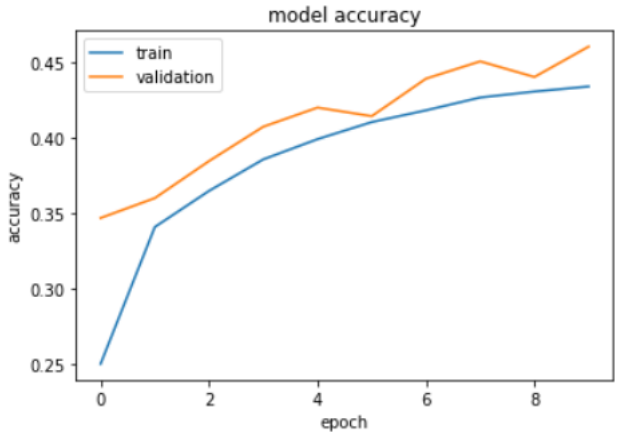



```
[[358  66  64  66  19  15  48  98 187  79]
 [ 18 546   5  63   6  29  55  52  49 177]
 [ 83  42 167 128 124  71 237 104  27  17]
 [  9  67  41 311  25 193 183  99  23  49]
 [ 30  32  82  69 274  53 268 144  26  22]
 [ 10  31  59 197  38 330 170 114  28  23]
 [  5  35  44  96  64  60 639  29   8  20]
 [  9  54  25  88  70  68 105 498  18  65]
 [ 76 122  28  40   5  40  21  55 486 127]
 [ 16 229   5  53   5  19  48  73  55 497]]
```

Adam & Categorical Cross Entropy & Relu & 256

```
Test loss: 1.5165727138519287
Test acc: 0.4681999981403351
```

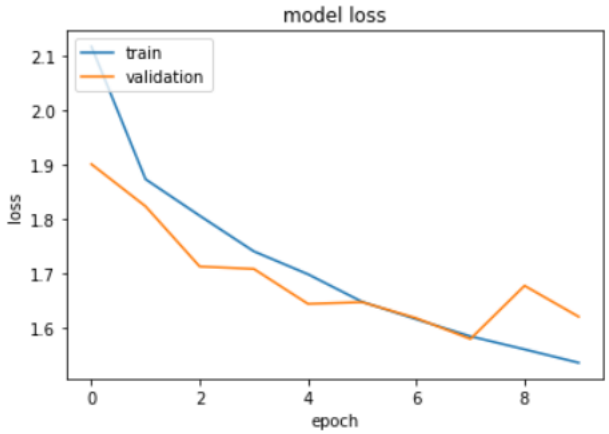


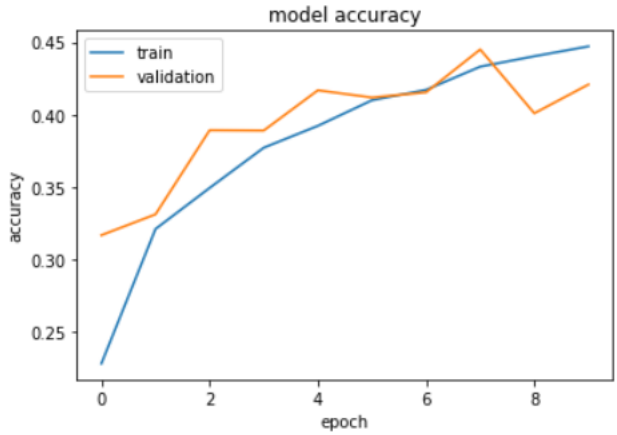



```
[[590  40 111  27  32   6  25  40  98  31]
 [ 65 574  10  26  18  15  19  34  71 168]
 [ 84  25 379  81 152  38 127  87  15  12]
 [ 33  20 128 301  74 131 163  86  25  39]
 [ 72   9 169  46 407  21 140 115  12   9]
 [ 36  14 139 211  73 277 100 106  23  21]
 [  5  17 101  76 121  31 591  36   9  13]
 [ 65  24  66  51  97  45  55 544  13  40]
 [214  63  46  27  27  11  12  16 521  63]
 [ 86 182  21  30  17  13  47  49  57 498]]
```

Nadam & Categorical Cross Entropy & Relu & 256

```
Test loss: 1.6078981161117554
Test acc: 0.4212999939918518
```






```
[[470   4  82  34  59  17  11 111 156  56]
 [109 280  20  45  26  21   8  87  61 343]
 [ 61   4 348 107 177  72  54 123  42  12]
 [ 23   4 104 277  87 195  60 166  39  45]
 [ 39   0 208  70 420  45  53 136  22   7]
 [ 12   0 117 168  97 369  43 134  42  18]
 [  3   6 125 145 183  54 347  96  14  27]
 [ 29   3  59  50 128  59  14 626  12  20]
 [143  21  14  31  34  26   4  78 509 140]
 [ 66  51  11  33  31  21  12 169  39 567]]
```

تحلیل:

همانطور که مشاهده می‌شود adam بهترین نتیجه را دارد دلیل برتری آن نسبت به سایر روش ها توانایی آن در مقابله نویز داده ها است.

سوال 5 –

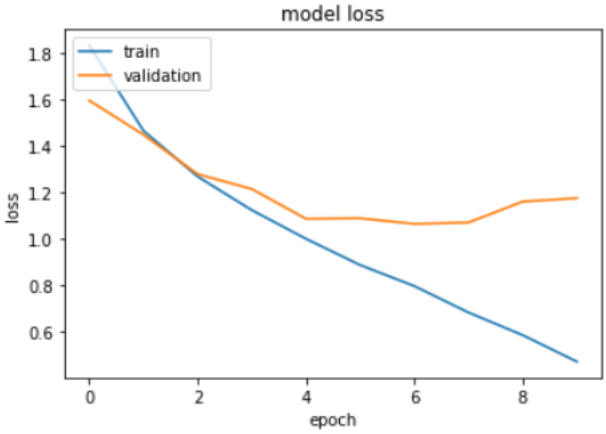
با توجه به ارزیابی های انجام شده بهترین پارامتر ها

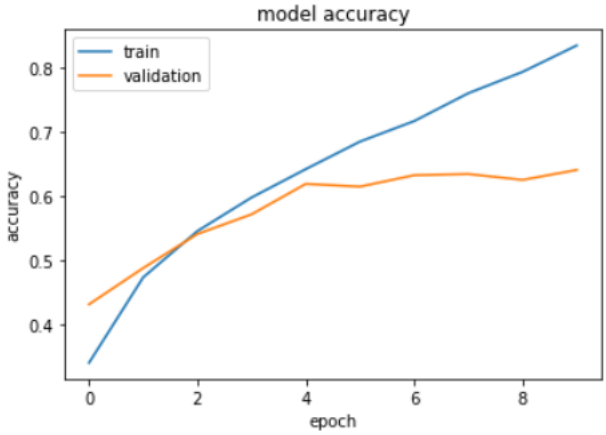
Batch size: 256, Activation function: Relu, Loss function: Categorical Cross Entropy, Optimizer: Adam

بخش ب)

MLP+CNN

```
Test loss: 1.1876391172409058
Test acc: 0.63919997215271
```



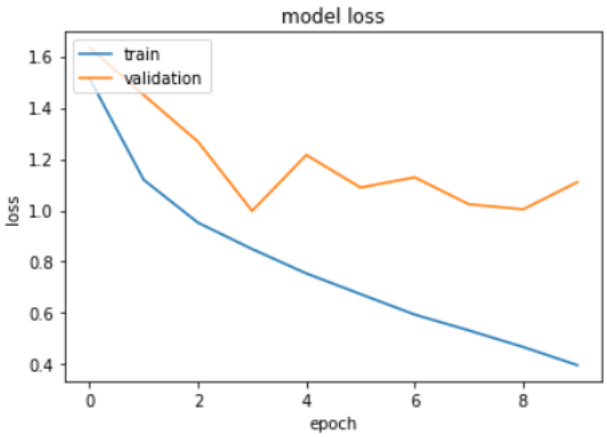


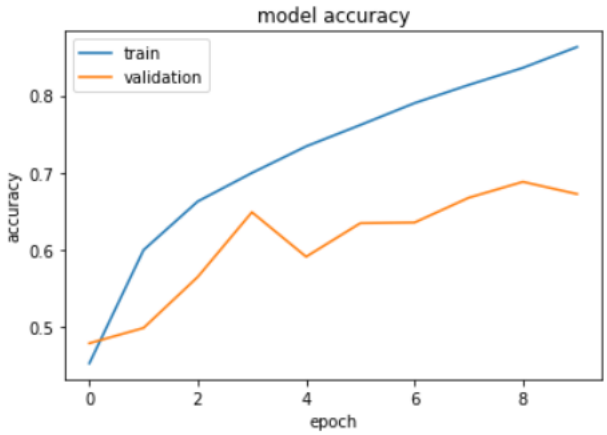

```
[[717  50  62  12  15   4  24   8  55  53]
 [ 19 845   7   7   1   5  13   4  12  87]
 [ 72  18 518  39  83  85 111  46  14  14]
 [ 31  48  76 303  59 225 183  34  14  27]
 [ 34   7 115  36 536  55 121  77  12   7]
 [ 14  16  78 110  43 576  73  66   8  16]
 [  8  17  40  18  43  20 837   5   4   8]
 [ 25  17  44  22  76  71  21 683   5  36]
 [ 95 107  13   7   7   9  12   3 697  50]
 [ 37 192  14  15   3   7  17  15  20 680]]
```

همانطور که مشاهده می‌کنیم پیشرفت مشهودی در دقت نسبت به بخش قبل می‌بینیم. همچنین خطا نیز کاهش محسوسی یافته است.

MLP+CNN+Pooling+BatchNormalization

```
Test loss: 1.12882661819458
Test acc: 0.6680999994277954
```



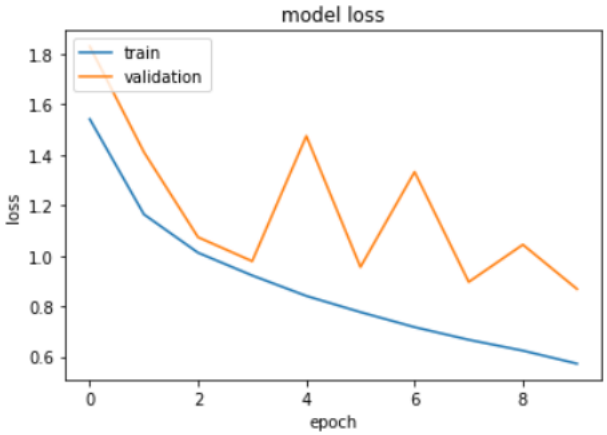


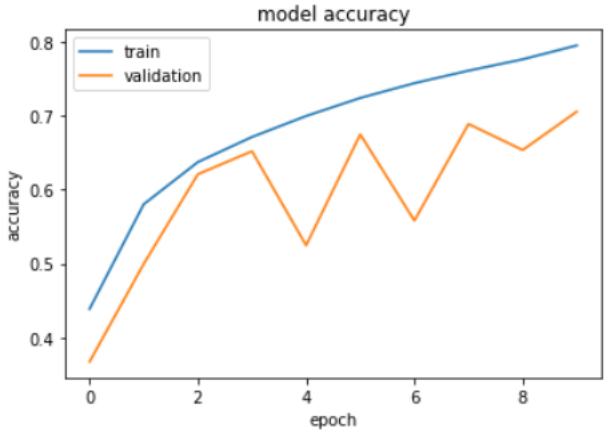

```
[[639  18  63  34  69   7   7  38  71  54]
 [ 16 772   7   9   1  10   9   9  13 154]
 [ 49   9 430  85 155  79  53 105  17  18]
 [ 13   9  47 550  89 150  35  78   9  20]
 [  6   6  26  70 655  38  30 157   6   6]
 [  9   5  28 227  66 526  15 115   4   5]
 [  9   6  22  97  72  39 725  18   1  11]
 [  8   3  17  46  38  32   5 836   4  11]
 [ 59  56  19  31  20   6   8  11 727  63]
 [ 35  54   5  14  11   7   3  36  14 821]]
```

در اینجا نیز مشاهده می‌کنیم دقت افزایش یافته و همچنین خطا نیز کمتر شده است. در نتیجه تاثیر Pooling و Batch Normalization بسیار مثبت است.

MLP+CNN+Pooling+BatchNormalization+Dropout

```
Test loss: 0.9008815288543701
Test acc: 0.7001000046730042
```






```
[[849  26  32  13  11   9   8   7  31  14]
 [ 39 870   4   4   3   4   9   3  19  45]
 [ 95   8 564  59  64  95  73  28   6   8]
 [ 45  23  48 431  50 254  89  45   5  10]
 [ 45   6  64  72 568  61  83  93   5   3]
 [ 20   6  38 122  33 691  30  52   3   5]
 [  5   8  35  57  21  33 827   8   1   5]
 [ 29   5  29  31  51  61   8 778   1   7]
 [165  50  10   7   6   5   4   7 728  18]
 [ 68 169  10  10   2   5   8  16  17 695]]
```

در آخر مشاهده می‌کنیم Dropout تاثیر بسیار مثبتی روی نتایج دارد. دلیل استفاده از این تکنیک این است که لزوما همه نورون ها باعث افزایش دقت ما نخواهند شد و با حذف آن ها دقت ما نیز بهتر خواهد شد.

سوال 4 – توقف زود هنگام

در شبکه های عصبی به مقطعی خواهیم رسید که با افزایش epoch خطای داده های Validation شروع به افزایش می‌کند که بدین معناست باید آموزش را بر روی داده ها متوقف کنیم وگرنه دقت ما روی داده های تست نیز کاهش خواهد یافت. برای تشخیص این موضوع کافیست نمودار دقت یا خطا را مشاهده کنیم و در نقطه‌ای که دو نمودار شروع به فاصله گرفتن می‌کنند شبکه را متوقف می‌کنیم.

سوال 3: یادگیری انتقال یافته برای شبکه EfficientNet

الف) آشنایی با شبکه EfficientNet

معماری شبکه:

افیشنت نت شبکه ای پیچشی است که با انتخاب پارامتر های مربوط آن مانند تعداد لایه های پنهان و ... البته دقت شود همه پارامتر ها با یک نسبت ϕ تعیین می‌شود. در این روش ها از MoblieNetV2 استفاده می‌شود.

توضیح نسخه های مختلف معماری و تفاوت آن‌ها:

در این نسخه ها که همگی با یک پسوند b0,b1,... مشخص می‌شوند در حقیقت دقت و خطا و متناظراً پیچیدگی شبکه ها متفاوت است.

پیش‌پردازش های اولیه برای تضویر ورودی:

مهمترین پیش پردازش لازم تغییر سایز تصاویر به ورودی مطلوب شبکه است که برای مثال برای b0 باید به سایز (224و224) تغییر دهیم.

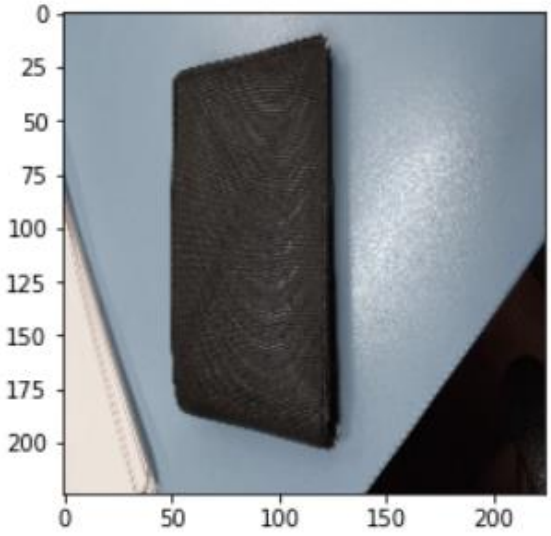
مزایا نسبت به سایر مدل‌ها:

مزیت اصلی این شبکه دقت بالای آن در آموزش داده های با کلاس های متعدد است که آن را برای استفاده در transfer learning مناسب می‌کند.

ب) پیاده سازی شبکه به کمک ایدهLearning Transfer

در این بخش قرار است با شبکه EfficientNet B0 یک عکس را تشخیص دهیم. عکس یک کیف پول را به شبکه داده و نتایج به صورت زیر می‌باشد.

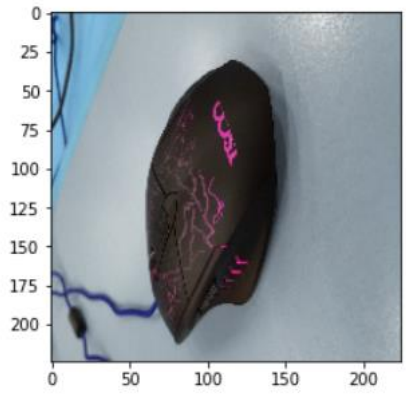
```
Label 1: wallet          Prob: 0.4664161205291748
Label 1: binder          Prob: 0.11798178404569626
Label 1: lighter         Prob: 0.06010153889656067
```



ج) رفع یک مشکل خاص در شبکه

برای جلوگیری از برچسب اشتباه زدن به اشیایی که در دیتاست وجود ندارند می‌توانیم یک حد در نظر بگیریم که اگر احتمال یک برچسب کمتر از 25 درصد بود آن را در نظر نگیریم.

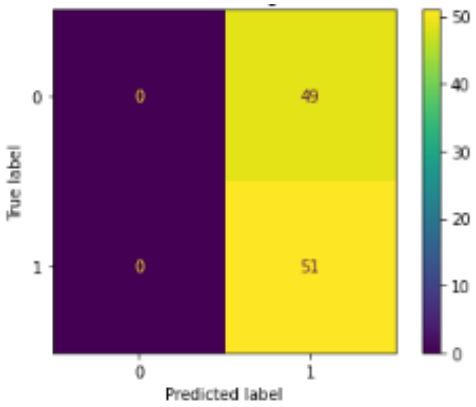
```
Not found in the dataset!
```



د) آموزش شبکه با مجموعه دادگان جدید

در اینجا با استفاده از داده های مجموعه cifar10 شبکه را آموزش می‌دهیم. نکته قابل توجه این است که به دلیل کیفیت پایین این تصاویر شبکه دقت خیلی مناسبی نداشت.

```
Accuracy = 50.999999046325684%
Test-loss = 0.693223357200622
```


0
49
0
51
0
1
True label
Predicted label
50
40
30
20
10
0