Homework 2

Group 74

79730 João Silva

95598 João Câmara

João Silva worked on Question 1, and João Câmara worked on Questions 2 and 3.

Question 1

1.1) Resulting matrix of $QK^T \in \mathbb{R}^{L\times L}$, with each element requiring D additions, therefore $O(L^2D)$. *Softmax* applied on the resulting matrix is $O(L^2)$ for calculating row sums overall, and likewise for exponentiating and dividing each element of said matrix. *Softmax(*QK^T*)*$V \in \mathbb{R}^{L\times D}$, with each element requiring L additions, therefore $O(L^2D)$. **Final time complexity is then** $O(L^2D + 2L^2 + L^2D)$**, or simply, in the context of very long sequences** *(*$L \gg D$ *),* $O(L^2)$. This quadratic growth means that it is computationally too costly to train the model given long enough sequences.

1.2) $\exp(q^Tk) \approx 1 + \sum_{i=1}^{D} q_i\, k_i \;+\; \frac{1}{2}\left(\sum_{i=1}^{D} q_i\, k_i\right)^2$, by Cauchy-Schwarz inequality, $\leq\; 1 + \sum_{i=1}^{D} q_i\, k_i \;+\; \frac{1}{2}\sum_{i=1}^{D} (q_ik_i)^2 \;=\; 1 + \sum_{i=1}^{D} q_i\, k_i \;+\; \frac{1}{2}\sum_{i=1}^{D} q_i^2\, k_i^2$.

Using the sum of squares as an approximation: $\exp(q^Tk) \;\approx\; 1 + \sum_{i=1}^{D} q_i\, k_i \;+\; \frac{1}{2}\sum_{i=1}^{D} q_i^2\, k_i^2 \;=\; \phi(q)^T\phi(k)$, where $\phi(q)^T \;=\; \left[1,\, q_1,\, \dots\,,\, q_D,\, \frac{1}{\sqrt{2}}q_1^2,\, \dots\,,\, \frac{1}{\sqrt{2}}q_D^2\right]$, $\phi(k)^T \;=\; \left[1,\, k_1,\, \dots\,,\, k_D,\, \frac{1}{\sqrt{2}}k_1^2,\, \dots\,,\, \frac{1}{\sqrt{2}}k_D^2\right]$

Therefore, $M \;=\; 2\times D + 1$. Generally (including $K \geq 3$), using the same approximation, the dimensionality is given by $M = \;(K-1)\times D + 1$.

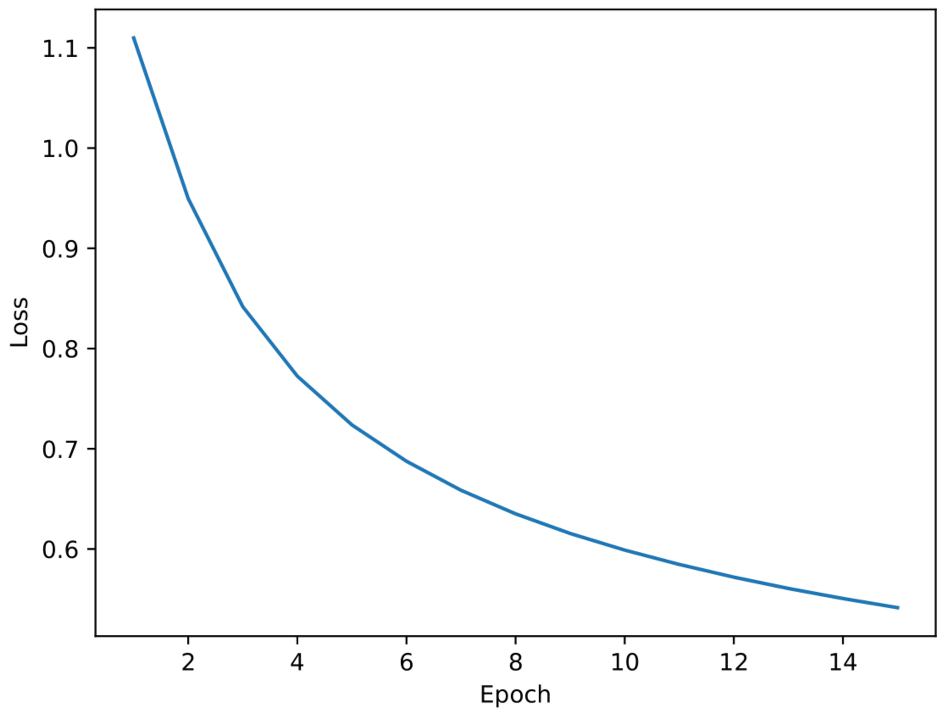
1.3) $Z \approx D^{-1}\Phi(Q)\Phi(K)^TV \approx D^{-1}\exp(QK^T)\,V \approx\; Diag(v)^{-1}\exp(QK^T)\,V$ using 1.2). Where $v \;=\; \exp(QK^T)\,1_L \;\in \mathbb{R}^L$ is the vector containing the row sums of exponentiated elements of QK^T. Assuming every sum is different than 0, $Diag(v)^{-1}$ is then the diagonal matrix whose diagonal values are $\frac{1}{v_i}$, $1 \leq i \leq L$, and $Diag(v)^{-1}\exp(QK^T)$ effectively approximates

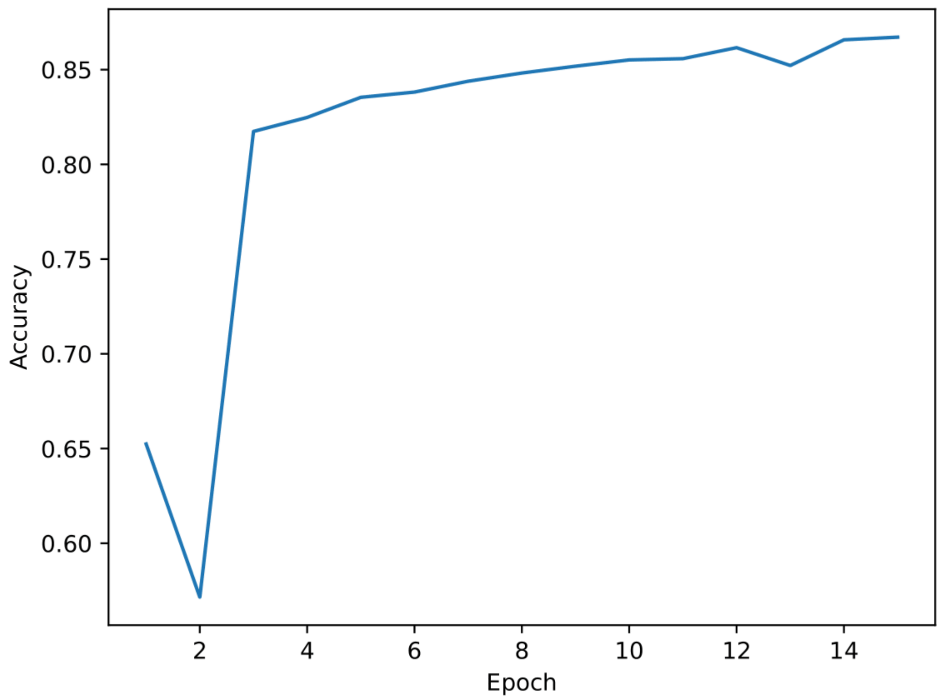
Softmax(QK^T)*, and, therefore,* $Z \;=\;$ *Softmax*$(QK^T)V \approx D^{-1}\Phi(Q)\Phi(K)^{-1}V$*.*

1.4) Worst case is still $O(L^2)$.

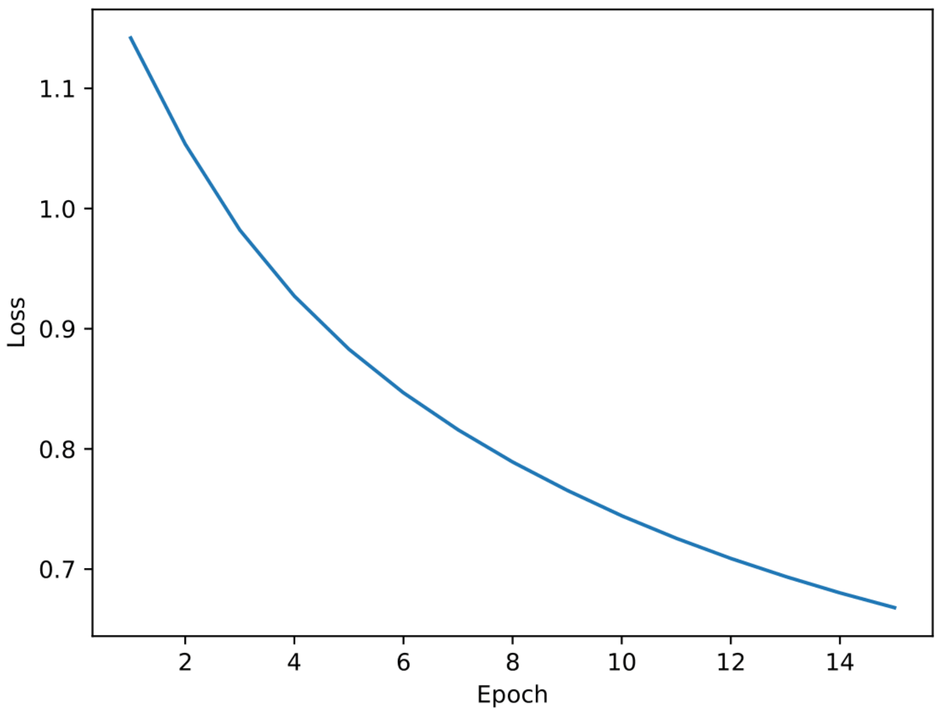
Question 2

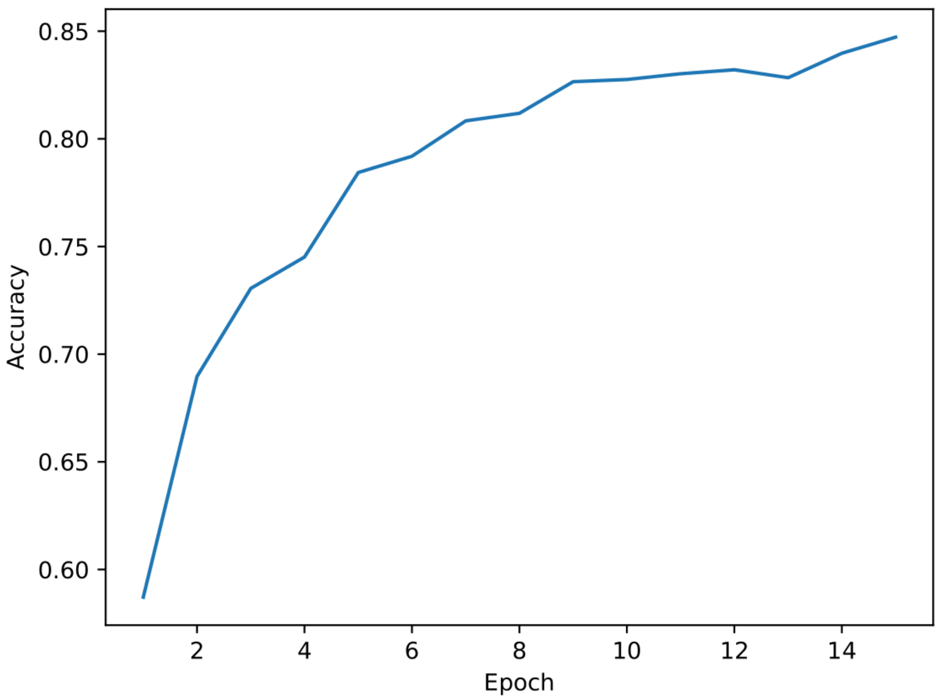
2.1) Best learning rate is 0.01.





2.2) Best learning rate is also 0.01.





2.3) (Implemented) The performance difference is due to the use of pooling layers, which effectively reduce the number of parameters of the network.

Question 3

Implemented but has bugs.