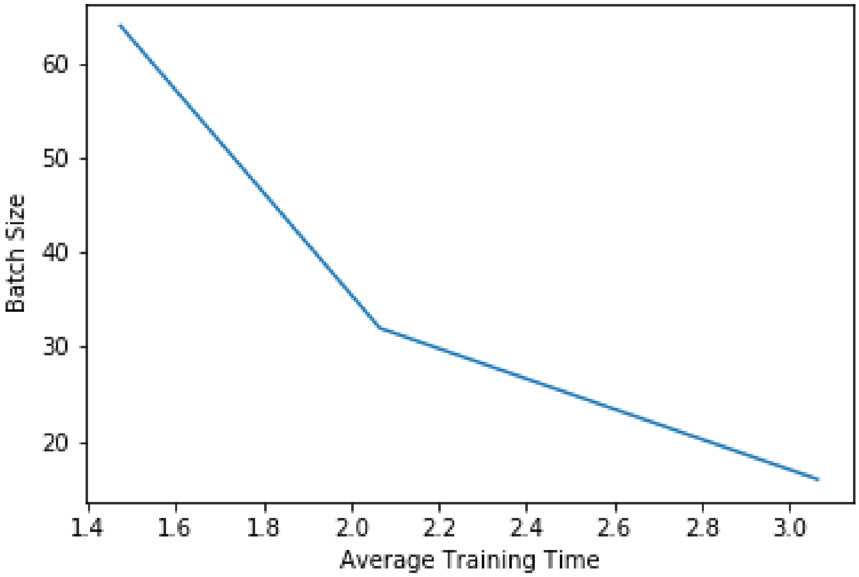
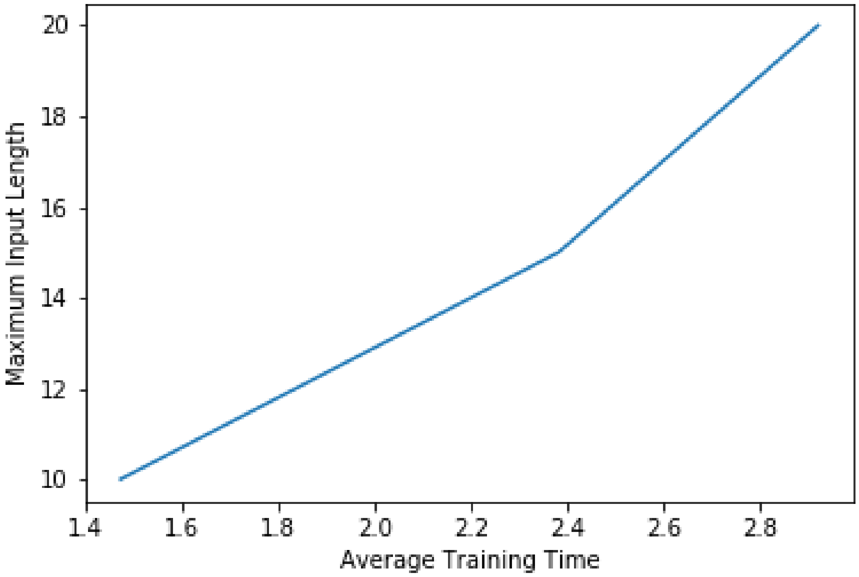
HW4 Report

1.



The average training time per epoch decreases as the batch size decreases.

2.



The average training time increases as the maximum input length increases.

3.

O1=**mul(**Q, D.T**)**

O2=**exp(**O1**)**

S=**row-sum(**O1**)**

softAttention=O2/S

4.

Our goal is to have the term in the **tanh()** small enough, so that the gradient is large. For this question, let the term in **tanh()** $\in [-1,1]$.

a) One hot

a=1

b)x$\in [0,1]$

a=1/1000

c) x$\in [1000,1001]$

a=1/10^6