1. (4 points) Suppose we are using the following CNN model to classify images of 3× 32×32 (channel, height, width) into 10 different categories.

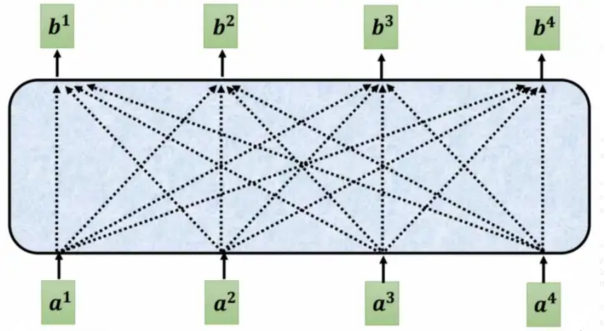
Layer	Out Channels	Kernel Size	Stride
Input	3	-	-
Convolution	8	5x5	1
Average Pooling	8	2x2	2
Convolution	16	5x5	1
Average Pooling	16	2x2	2
Fully Connected	-	-	-
Softmax	-	-	-

1) Assume that no padding is applied. Please compute the shape of the output after each layer.
2) How many parameters are there in the first convolution layer? How about the first average pooling layer?

2. (3 points) We have a sentence of four words. The corresponding word embeddings a1, a2, a3 and a4 are shown below.

$$a_1 = [1, 1, 1]^T$$
$$a_2 = [2, 2, 2]^T$$
$$a_3 = [3, 3, 3]^T$$
$$a_4 = [4, 4, 4]^T$$

Now we are using one simple self-attention layer to process these inputs.



Please compute the new hidden state b_1 for the first word. (Suppose that all the model parameters you' ll need are initialized to ones)

3. (3 points) Please briefly answer the following questions in no more than 5 sentences.
    a. (1pts) As we have learned in the lectures, gradient exploding and vanishing limits the training stage of RNN. Please explain how LSTM prevents gradient exploding and vanishing.
    b. (2pts) In CNN, we sometimes use convolution layers with 1×1 filters (e.g., in GoogLeNet). Please try to explain the utility of such 1×1 convolution.

Note: (For all the exercises above, please show brief calculation processes if there exists.)