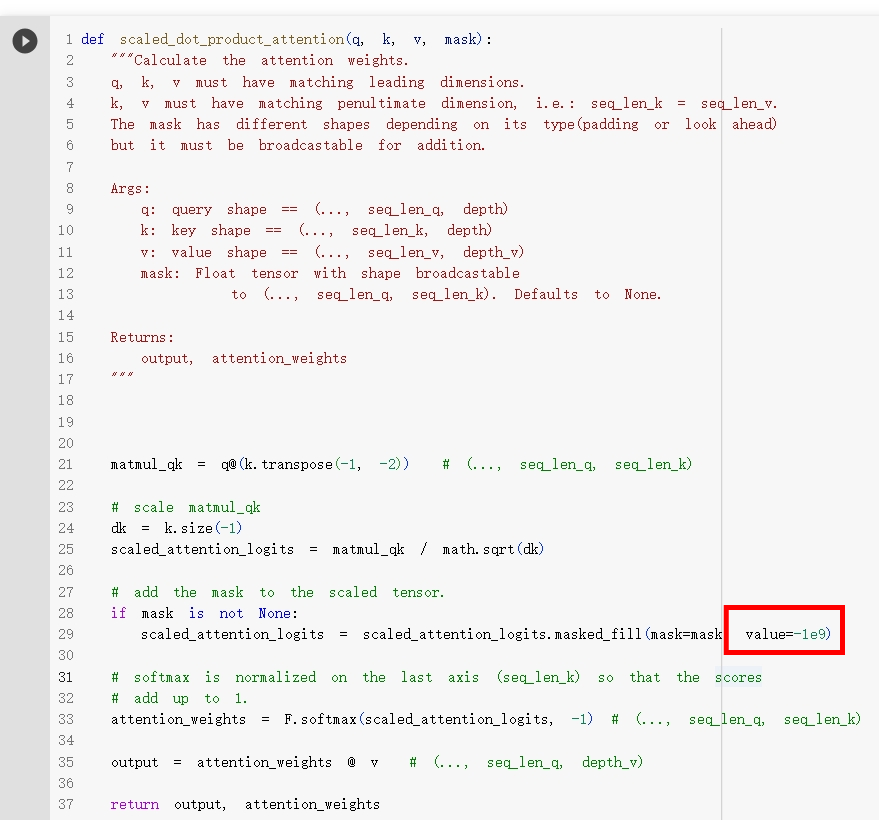
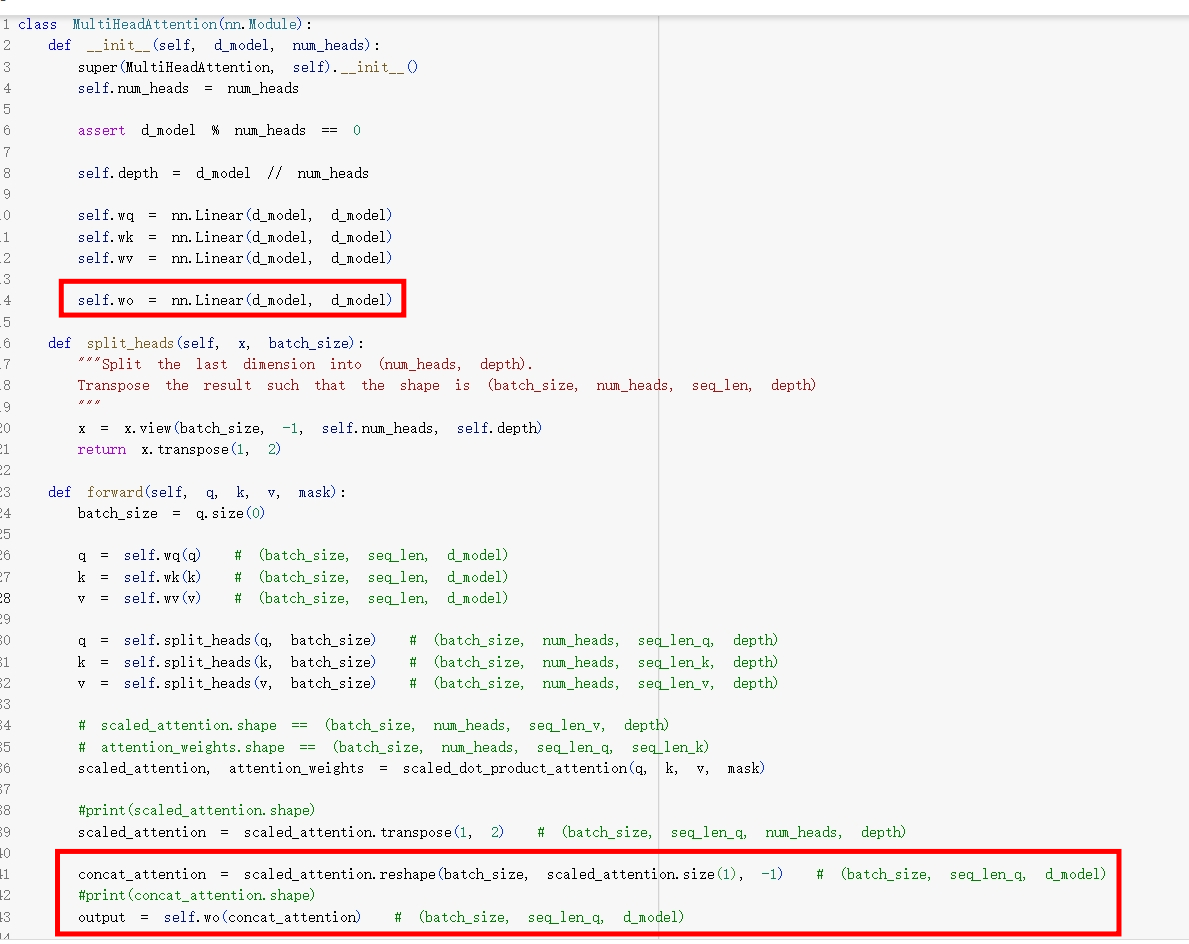
上述程式中，為何attention\_vector[2]與attention\_vector[3]有一樣的值? (hint: scale dot product)

In the above program, why attention\_vector[2] and attention\_vector[3] have the same value? (hint: scale dot product)



我們都知道masked\_fill可以把mask tensor裡等於1(True)的位置，對應到attention tensor的位置設為value，那為何value要設為一個絕對值很大的負值? (hint: softmax)

We all know that masked\_fill can fill elements of attention tensor with the value where the mask is True, so why should the value be set to a negative value with a large absolute value? (hint: softmax)



從上述code可知道Mutihead attention在經過 Wo linear layer時維度並沒有任何改變，那.....Wo linear layer是用來做什麼的? (hint: Linear layer)

From the above code, we can see that Mutihead attention does not have any change in dimension when it passes through the Wo linear layer, so ... What is the Wo linear layer used for? (hint: Linear layer)