1. Explain all the details of your algorithm and the steps of the code. Be very clear of what each algorithm does.

Self-Attention Layer: The input feature map would go through three different 1x1 convolution layers which can be regarded as linear transforms, and get hidden values f(x), g(x), and h(x). The output of a matrix multiple of f(x) and g(x) will take a softmax activation, and the result will take another matrix multiple with h(x). Finally, we can get the self-attention feature map.

Adam Optimizer: Algorithm for gradient-based optimization. It replaces the classical stochastic gradient for deep learning. It focuses on a single learning rate. It combines two other stochastic gradient descent (SGD). It uses Adaptive Gradient Algorithm and Root Mean Square Propagation. Instead of using the average first moment, Adam focuses on the average of the second moment of gradients.It calculates the exponential average of the gradient and the squared gradient. It is useful due to the fact that good results are achieved fast. Adam has a few parameters that need to be adjusted around. There is alpha which is the learning rate. There is beta1 which is the exponential decay for the first moment estimates. Beta2 is the exponential decay rate for second moment estimates. Finally epsilon is a small number to prevent any division by zero.

Binary Cross Entropy: Algorithm that compares the predicted probabilities to the output which is either 0 or 1. The score is then calculated by penalizing the probability based on the distance from the expected value. It totals the scores and takes the negative average of the logs of predicted probabilities.