1.

Momentum:

From the momentum formula in the slides we can see that, it modifies gradient descent buy adding the previous steps updates to the current one so based on the lecture one of the main benefits that this have is escaping the local minima problem however better algorithms like adagard or adam exists. And another thing that momentum helps is the problem of slow convergence in gradient descent.

Learning rate:

Learning rate tells the step sizes of gradient descent, when it is too high the model cant learn good and the accuracy decreases and when it is too low the model learning will become very slow.

Learning rate decay:

a fixed learning rate can overshoot the optimum or fail to refine the solution in detail. Learning rate decay decreases the learning rate to fine-tune convergence and with that comes more stability and better results in later training stages.

2.

Dropout is a method of generalization for preventing the model from overfitting it is being done by shotting down some random neurons (setting to zero) during each epoch this helps model not to rely on specific neurons.

3.

Formula is:

Output = (InputSize - FilterSize + 2.PaddingSize)/stride + 1.

Outputsize = (10-F+0)/2 + 1 = 5 => F = 2

Then =>

A **filter size of 2x2** is needed to reduce a 10x10 image to a 5x5 feature map using a stride of 2 with no padding.