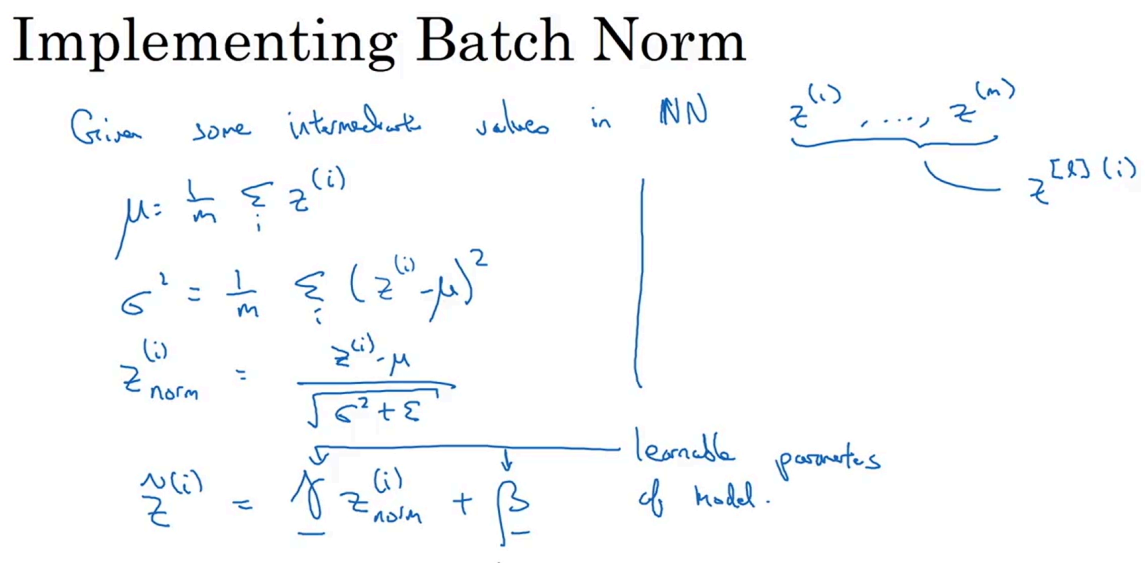
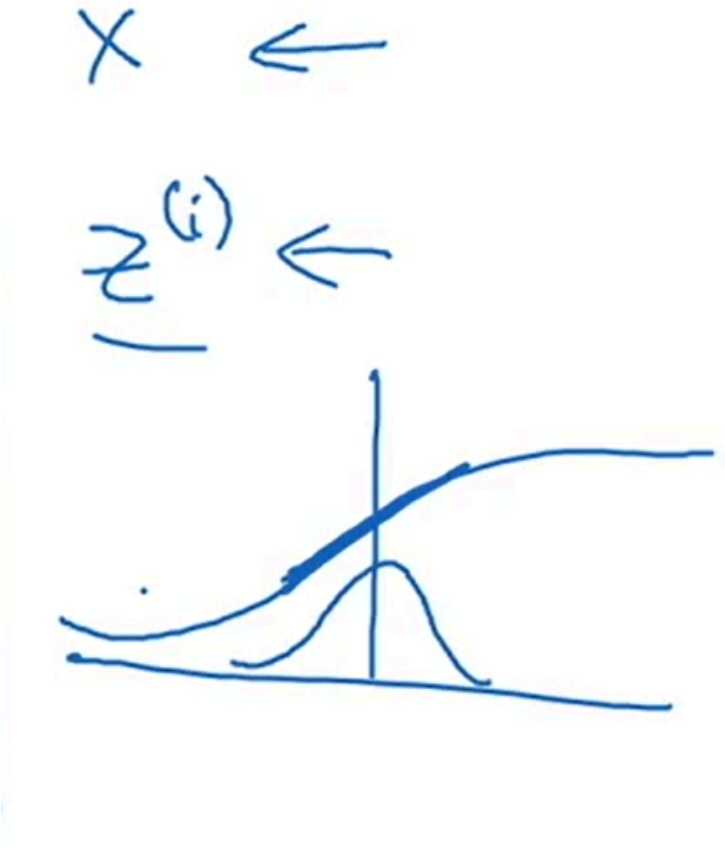
Batch Norm (1)

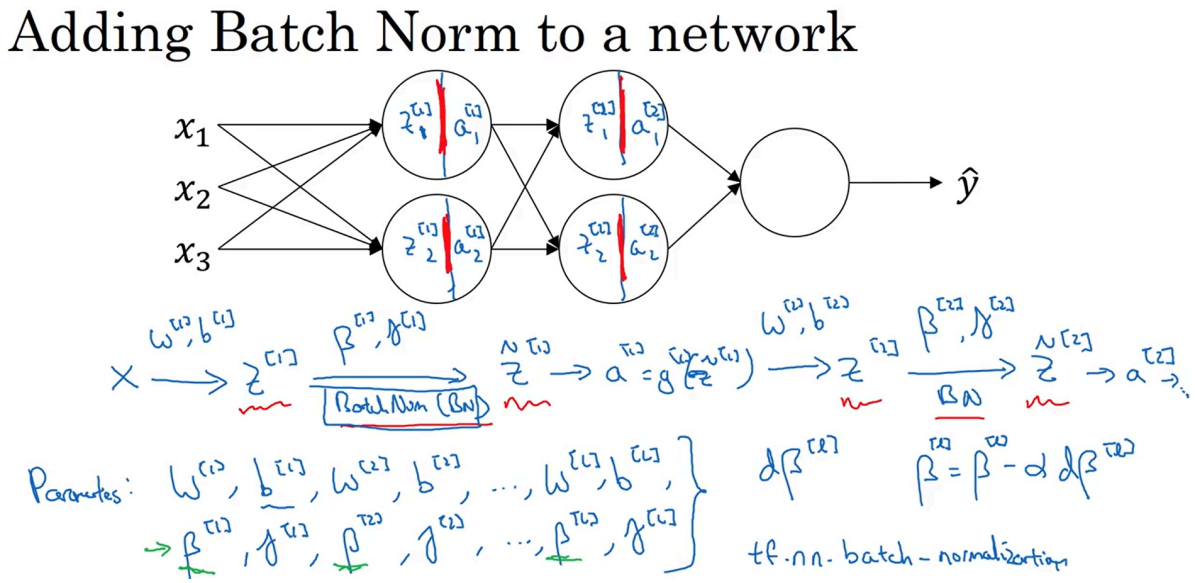


If gamma were sqrt(sigmasqrd + epsilon) and beta were mu, then no effect.

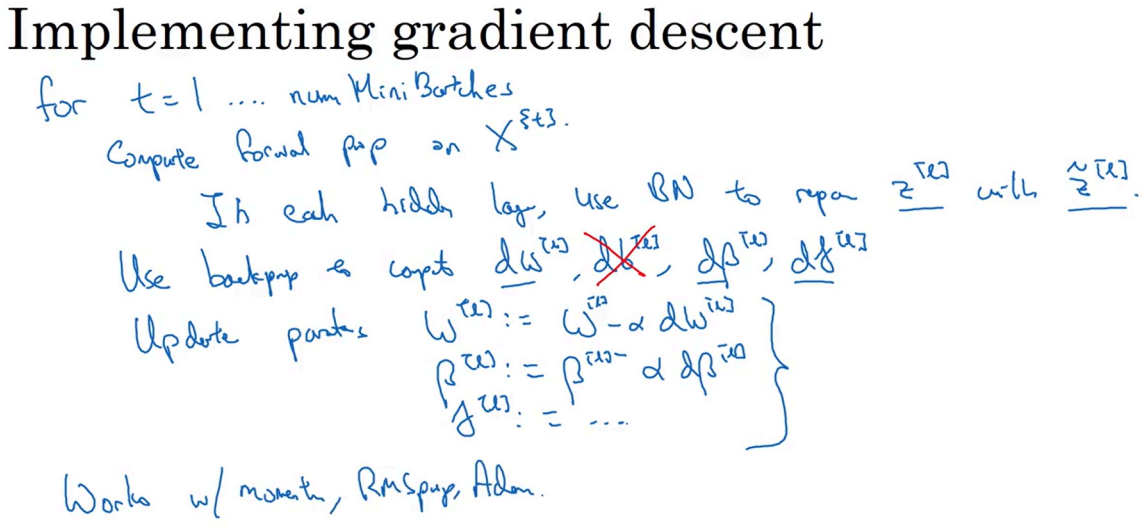


But we need to be careful to avoid the above situation (where the values are clustered around 0) and we are unable to take advantage of the non-linearity of sigmoid. That's where gamma and beta come into play

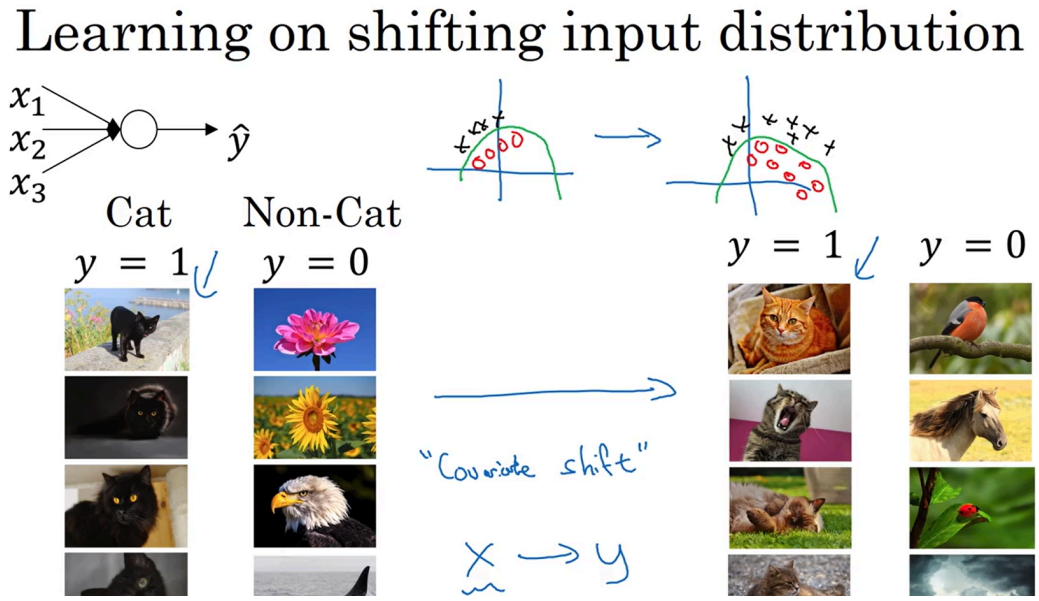
Adding BatchNorm to a network



If you use batch norm, just can ignore b

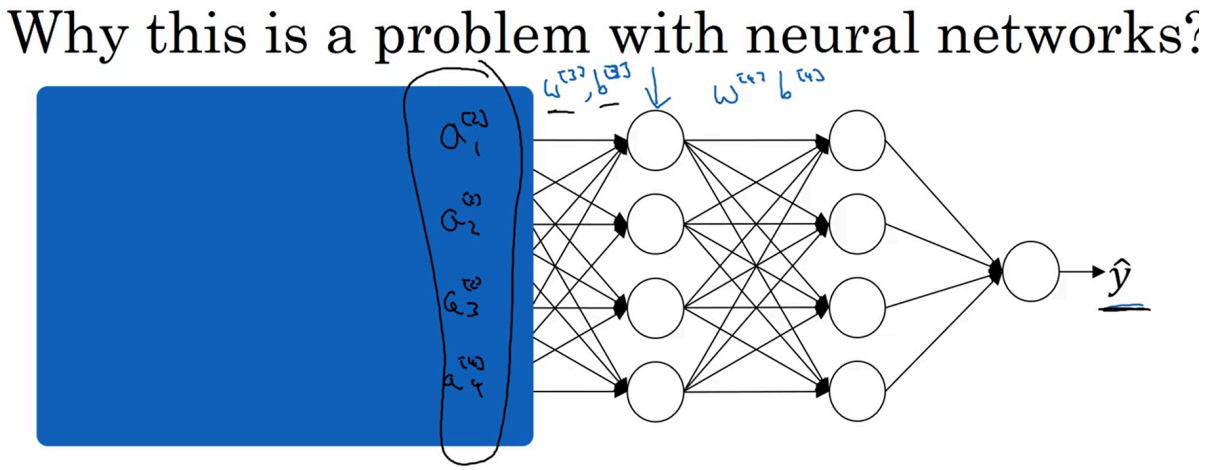


Why does BatchNorm work?



If you train to identify black cats and you test on coloured cats (:D), then your classifier may not work properly mah!

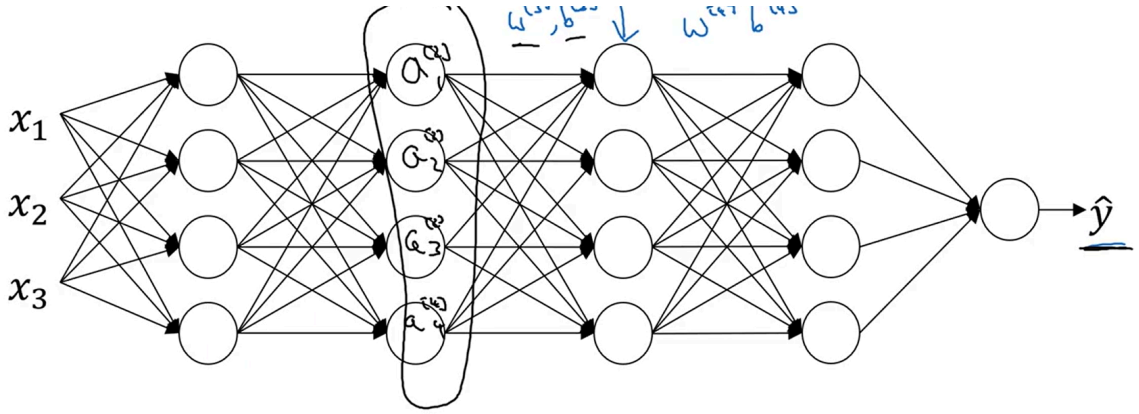
The change of the distribution from X (containing black cats) to X′ (containing coloured cats is known as covariate shift)



Ok bah?

The NN needs to try to work with the activations (can be interpreted as features). Job of the third layer needs to take these values and try to map to y

Let's uncover the blue mah?



From the view of the third layer, the inputs to it always keep changing! Come on lor! This is like that covariate shift bah!

So what batch norm does it that it reduces the amount that the distribution of hidden unit values shift around.

This is done by ensuring that the mean and variance doesn't change lor! Mean depends on gamma and beta for that BN layer

It limits the amount of how changing parameters in an earlier layer will affect the inputs to the current layer

i.e weakens coupling between weights parameters of learning almost making learning independent b/w layers.

Ok bah, got it? Sike! Here is another usage of BN! (You have more to think about lol)

Batch Norm as regularization

- Each mini-batch is scaled by the mean/variance computed on just that mini-batch.
- This adds some noise to the values $z^{[l]}$ within that minibatch. So similar to dropout, it adds some noise to each hidden layer's activations.
- This has a slight regularization effect.

Noise is added because the normalization is done w.r.t to the current minibatch and not w.r.t to the entire training set. (i.e estimate of the mean and st.dev is noisy lor!) This noise forces downstream hidden units not to be dependent on one hidden unit!

(So the larger mini-batch size you use, the less regularization you do! Ok mah? You seem confused lor!)

BN → Multiplicative noise + Additive noise

Dropout → Additive noise