* Deep learning
* Use Boolean logic – can use XOR to make nonlinear boundaries
* Deep learning is using multiple levels of decisions
* Can still use nonlinear transforms, but using enough linear stages can give us any non-linear boundary
* Multiple neural networks can be put into a layer, and then it’s layered multiple times
* Extra 1 is a bias term
* To train, you’re going to do gradient loss
* The non-linearities are crucial because without them the whole thing collapses into a linear equation
* Old nonlinearities were sigmoids, but newer ones are things like rectified linear unit, leaky ReLU, softplus, and exponential linear unit
* The older ones will saturate – they have deriviatives that go to zero while the new school of non linearities are non-saturating
* The gradient goes towards zero after a while if you use a vanishing gradient. Since you are taking the product of a bunch of gradients (which are really small if you use a saturating non-linearity) it gets hard to train deep layers
* Internal covariate shift – limiting the degree the weights can change
* Normalization tries to keep the values in the dynamic range where your gradient doesn’t fall to 0
* Batch normalization normalizes each layers activations depending on the statistics of the batch
* Adding stochasticity can help regularize the data
* Dropout is a bernoullli prior on the activations of the network – you randomly drop entire units from your layer
* Another way to do regularization is to have constraints like early stopping or weight penalties
* Only regularize weights not the bias term
* We initialize weights to a normal or a Gaussian
* Not much theoretical understanding of optimizers
* Most people use adam as the gradient descender
* Can do unit or data parallelization for increasing speed – done with GPUs b/c they’re good at matrix multiplication and pointwise operations
* One technique is attention which is reweighting and making say an image more important and tell it to focus on certain details (soft attention)
* Sampling operations are often non-differentiable; sampling are things that are binary or classifiers
* Standard backpropagation won’t work