DL concepts:

Concept	**Description**	**Effect**	**Notes**
Covariance Shift	General: Change in the distribution of data. In DL: Change in the distribution of activations of layer *n* due to change in weights of previous layers.	Makes learning slower, requiring smaller α Can cause saturation in some units, which can stop learning completely	Apply BN
Co-adaptation	Some neurons become very dependent on others. Deeper layers learn to correct errors from earlier layers.	Over-fitting	Apply Dropout

Concept	**Description**	**Effect**	**Notes**
Early stopping	Stop training if **validation** loss does not improve after *n* iterations	Prevent overfitting	
Batch Norm	Without BN, layers are constantly learning normalization, and changes in early layers screw deeper layers expectations Idea: Normalize data between network layers Idea: Have specific parameters for normalization mean and std, so weights don't have to perform the normalization	More stable learning, even with larger α Small regularization effect (adds noise to activations). Less dropout needed	Can be applied to all types of layers Learnable shift and scale are per feature
Adagrad	Per-weight adaptive LR, with larger LR for sparser	Works well on sparse data when sparse	

	parameters.	features are informative	
RMSprop	Per-weight adaptive learning rate, divided by a moving average of gradient. Extends Rprop to mini-batch training scenarios.	Faster learning	
Adam	Per-weight adaptive learning rate, divided by a moving average of gradient and gradient second moment (variance). Extends RMSprop by also using second moment.	Faster learning	
Leaky ReLU PReLU	Leaky ReLU produces non-zero gradient in the negative domain PReLU makes the slope in the negative domain a learnable parameter.	Avoid zero gradient when x < 0.	
Dropout	During training, randomly drop some neuron activations. Similar to training multiple networks in parallel and averaging their outputs, which reduces variance (over-fitting).	Regularizes, preventing over fitting.	Mostly for dense layers

Model overfit	Get more data	Get more data for real Image data augmentation Smote -like Data augmentation with generative models	
	Reduce model capacity / complexity (reduce variance)	Reduce network’s layers and nodes	
		Early stop	
		Regularize	L1 & L2 regularization Momentum, Smaller batch size Dropout Batch Normalization Use larger learning rate
		Average several models	
Model underfit	Regularize less		
	Increase model capacity	More layers, more nodes	