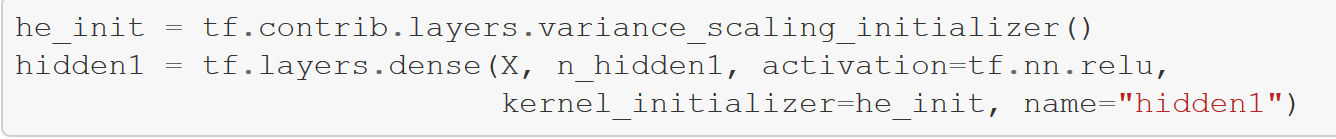
Chapter 11 Training Deep Neural Nets

Three main problems:

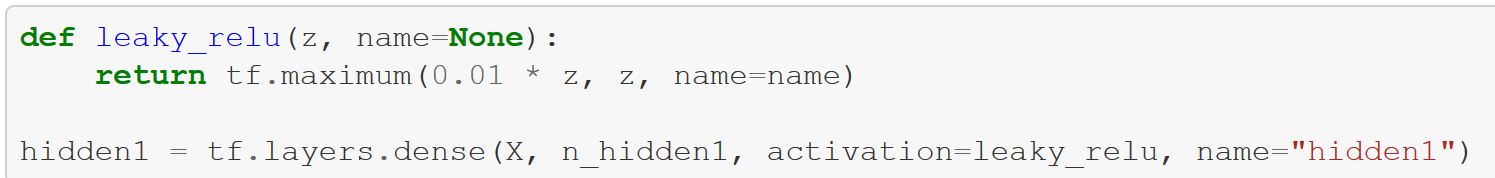
1. Vanishing gradients or exploding gradients problem, makes lower layers very head to train;
2. Training would be extremely slow;
3. A model with millions of parameters would severely risk overfitting the training set.
4. Vanishing/Exploding Gradients Problems

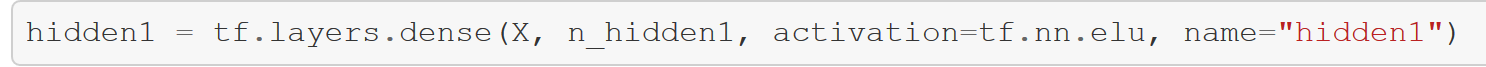
In backpropagation algorithm, it calculates the gradient of the cost function with regards to each parameter in the network, and it uses these gradients to update each parameter with a Gradient Descent Step.

1. Xavier and He Initialization



1. Non-saturating Activation Functions

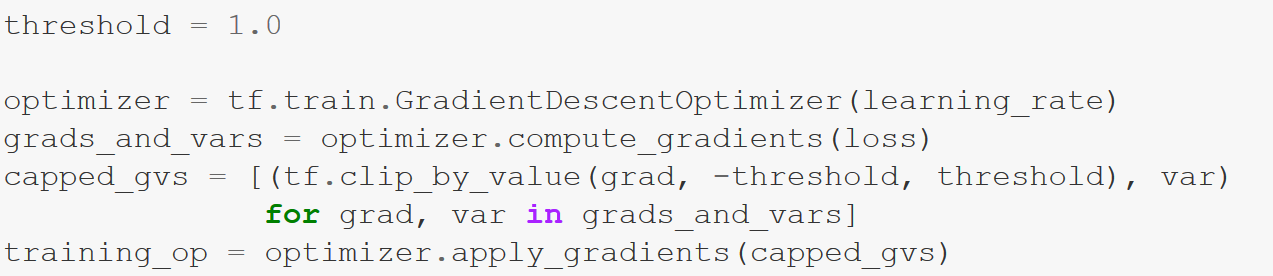




1. Batch Normalization

The technology consists of adding an operation in the model just before the activation function of each layer, simply zero-centering and normalizing the inputs, then scaling and shifting the result using two new parameters per layer, one for scaling, the other for shifting.

1. Gradient Clipping



1. Reusing Pretrained Layers
2. Reusing a TensorFlow Model
3. Reusing Models from Other Frameworks
4. Freezing the Lower Layers
5. Catching the Freezing Layers
6. Tweaking, Dropping, or Replacing the Upper Layers
7. Model Zoos
8. Unsupervised Pretraining
9. Pretraining on an Auxiliary Task
10. Faster Optimizers
11. Momentum optimization
12. Nesterov Accalerated Gradient
13. AdaGrad
14. RMSProp
15. Adam Optimization
16. Learning Rate Scheduling
17. Avoiding Overfitting Though Regularization
18. Early Stopping
19. L1 and L2 Regularization
20. Dropout
21. Max-Norm Regularization
22. Data Augmentation