**Assignment\_14**

1. Is it okay to initialize all the weights to the same value as long as that value is selected randomly using He initialization?

**Ans: If you initialize all the weights to be zero, then all the the neurons of all the layers performs the same calculation, giving the same output and there by making the whole deep net useless.**

1. Is it okay to initialize the bias terms to 0?

**Ans: It is possible and common to initialize the biases to be zero, since the asymmetry breaking is provided by the small random numbers in the weights.**

1. Name three advantages of the ELU activation function over ReLU.

**Ans: It avoids and rectifies vanishing gradient problem. ReLu is less computationally expensive than tanh and sigmoid because it involves simpler mathematical operations**.

1. In which cases would you want to use each of the following activation functions: ELU, leaky ReLU (and its variants), ReLU, tanh, logistic, and softmax?
2. What may happen if you set the momentum hyperparameter too close to 1 (e.g., 0.99999) when using a MomentumOptimizer?

**Ans: If you set the momentum hyperparameter too close to 1 (e.g., 0.99999) when using an SGD optimizer, then the algorithm will likely pick up a lot of speed, hopefully moving roughly toward the global minimum, but its momentum will carry it right past the minimum. Then it will slow down and come back, accelerate again** **overshoot超过 again, and so on. It may oscillate使振荡 this way many times before converging, so overall it will take much longer to converge than with a smaller momentum value.**

1. Name three ways you can produce a sparse model.

**Ans:**

* **Removing features from the model. Sparse features can introduce noise, which the model picks up and increase the memory needs of the model. ...**
* **Make the features dense.**
* **Using models that are robust to sparse features.**

1. Does dropout slow down training? Does it slow down inference (i.e., making predictions on new instances)?

**Ans: In the original implementation of dropout, dropout does work in both training time and inference time. During training time, dropout randomly sets node values to zero. In the original implementation, we have “keep probability” pkeep.**