

Questions:

1. Why do we NOT use Perceptron as often as a Logistic-Regression or Neural-Network?
2. How does ReLU introduce non-linearity when it looks “linear”? It is easy to observe that Sigmoid introduces non-linearity as it is a nonlinear function.
3. Why do we prefer dropout for regularization? Why not simply use L2 or L1 reg like in LogisticRegression?
4. How does dropout work at test-time? [E]
5. Why is Max-pooling popular in CNNs? Why not any other function like mean, median, min etc? [M]
6. Why do individual learning rates per weight (like in AdaGrad) help as compared to one learning rate for all weights? [Why: M ; Proof:H]
7. Write the weight update functions for ADAM. When does ADAM behave like Adadelata? [M]
8. Number of parameters & hyper-params in a max-pooling layer? [E]
9. How do we differentiate and back-prop through a max-pooling layer? [E]
10. You are training a model and its train-loss is not changing from epoch to epoch in NN. What could be the possible reasons? [Open-ended]
11. What’s the loss function of an autoencoder? [E]
12. Why is hierarchical softmax used in Word2Vec?[M]
13. How do we update weights in a Negative sampling based training of Word2Vec model?[M]
14. What are the trainable params in a BatchNorm layer?[M]
15. Why is a ResNet able to learn models of significantly larger depth than earlier VGGNet? [E]
16. CS: How is back-prop (in NN) related to dynamic programming in Algorithms? [M]
17. You have 5000 images with 5 class-labels and you want to build a CNN model? How would you go about building a classification model? [Open-ended]

18. Why does data augmentation help in object recognition tasks? [E]
19. Why does a GPU help in deep-learning much more than a multi-core CPU with say 8 or 16 cores? [M]
20. How is convolution different from simple weights as both involve element wise multiplication followed by addition?[Tricky]
21. Derive the derivative of a sigmoid function?[M]
22. Why do we need Leaky ReLu? [E-M]
23. Why is tanh (sometimes) better than sigmoid for training a NN? [M]
24. How to fix exploding gradients in a MLP? [E]
25. You want to detect outliers using Neural Nets? How would you go about doing it? [M]

Answers:

1. Why do we NOT use Perceptron as often as a Logistic-Regression or Neural-Network?
 - a. 11:20 min
<https://www.appliedaicourse.com/lecture/11/applied-machine-learning-online-course/3364/diagrammatic-representation-logistic-regression-and-perceptron/8/module-8-neural-networks-computer-vision-and-deep-learning>
2. How does ReLU introduce non-linearity when it looks “linear”? It is easy to observe that Sigmoid introduces non-linearity as it is a nonlinear function.
 - a. 7:02 min
<https://www.appliedaicourse.com/lecture/11/applied-machine-learning-online-course/3377/rectified-linear-units-relu/8/module-8-neural-networks-computer-vision-and-deep-learning>
3. Why use dropout for regularization? Why not simply use L2 or L1 reg like in LogisticRegression?

- a. Top comments @
<https://www.appliedaicourse.com/lecture/11/applied-machine-learning-online-course/3376/dropout-layers-regularization/8/module-8-neural-networks-computer-vision-and-deep-learning>
- 4. How does dropout work at test-time?
 - a. 16:45 @
<https://www.appliedaicourse.com/lecture/11/applied-machine-learning-online-course/3376/dropout-layers-regularization/8/module-8-neural-networks-computer-vision-and-deep-learning>
- 5. Why is Max-pooling popular in CNNs? Why not any other function like mean, median, min etc?
 - a. <https://soundcloud.com/applied-ai-course/ashwani-email>
- 6. Why do individual learning rates per weight (like in AdaGrad) help as compared to one learning rate for all weights?
 - a. 1:40 @
<https://www.appliedaicourse.com/lecture/11/applied-machine-learning-online-course/3382/optimizersadagrad/8/module-8-neural-networks-computer-vision-and-deep-learning>
 - b. Sparse vs Dense features: <https://youtu.be/c86mqhdmfL0>
- 7. Write the weight update functions for Adam. When does ADAM behave like Adadelta?
 - a. 6:38 @
<https://www.appliedaicourse.com/lecture/11/applied-machine-learning-online-course/3390/adam/8/module-8-neural-networks-computer-vision-and-deep-learning>
- 8. Number of parameters & hyper-params in a max-pooling layer?
 - a. 4:40 @
<https://www.appliedaicourse.com/lecture/11/applied-machine-learning-online-course/3418/max-pooling/8/module-8-neural-networks-computer-vision-and-deep-learning>
- 9. How do we differentiate and back-prop through a max-pooling layer?
 - a. 5:17 @
<https://www.appliedaicourse.com/lecture/11/applied-machine-learning-online-course/3418/max-pooling/8/module-8-neural-networks-computer-vision-and-deep-learning>

[arning-online-course/3419/cnn-training-optimization/8/module-8-neural-networks-computer-vision-and-deep-learning](https://www.appliedaigcourse.com/lecture/11/applied-machine-learning-online-course/3419/cnn-training-optimization/8/module-8-neural-networks-computer-vision-and-deep-learning)

10. You are training a model and its train loss is not changing from epoch to epoch. What could be the possible reasons?
11. What's the loss function of an autoencoder?
 - a. $\|X_i - \hat{X}_i\|$
 - b. 12:48 @
<https://www.appliedaigcourse.com/lecture/11/applied-machine-learning-online-course/3385/auto-encoders/8/module-8-neural-networks-computer-vision-and-deep-learning>
12. Why is hierarchical softmax used in Word2Vec?
 - a. 2:45 @
<https://www.appliedaigcourse.com/lecture/11/applied-machine-learning-online-course/3396/word2vec-algorithmic-optimizations/8/module-8-neural-networks-computer-vision-and-deep-learning>
13. How do we update weights in a Negative sampling based training of Word2Vec model?
 - a. 9:45 @
<https://www.appliedaigcourse.com/lecture/11/applied-machine-learning-online-course/3396/word2vec-algorithmic-optimizations/8/module-8-neural-networks-computer-vision-and-deep-learning>
14. What are the trainable prams in a BatchNorm layer?
 - a. 14:13 @
<https://www.appliedaigcourse.com/lecture/11/applied-machine-learning-online-course/3378/batch-normalization/8/module-8-neural-networks-computer-vision-and-deep-learning>
15. Why is a ResNet able to learn models of significantly larger depth than earlier VGGNet?
 - a. Skip connections explained in
<https://www.appliedaigcourse.com/lecture/11/applied-machine-learning-online-course/3411/residual-network/8/module-8-neural-networks-computer-vision-and-deep-learning>

16. CS: How is back-prop related to dynamic programming in Algorithms?
 - a. Memoization is the core of DP
17. You have 5000 images with 5 class-labels and you want to build a CNN model? How would you go about building a classification model?
 - a. Transfer learning and change the last layer.
18. Why does data augmentation help in object recognition tasks?
 - a. Robust to noise, rotation, shear etc..
19. Why does a GPU help in deep-learning much more than a multi-core CPU with say 8 or 16 cores?
 - a. Vector/Matrix/Tensor operations(GPU) vs independent compute tasks (CPU).
20. How is convolution different from simple weights as both involve element wise multiplication followed by addition?
 - a. Image in and image out per conv-kernel. Same weights used across the image.
21. Derive the derivative of a sigmoid function?
 - a. <https://math.stackexchange.com/a/1225116>
22. Why do we need Leaky ReLu?
 - a. 21:05 @
<https://www.appliedaicourse.com/lecture/11/applied-machine-learning-online-course/3377/rectified-linear-units-relu/8/module-8-neural-networks-computer-vision-and-deep-learning>
23. Why is tanh (sometimes) better than sigmoid for training a NN?
 - a. 12:15 @
<https://www.appliedaicourse.com/lecture/11/applied-machine-learning-online-course/3368/activation-functions/8/module-8-neural-networks-computer-vision-and-deep-learning>
 - b. <https://www.quora.com/What-are-the-benefits-of-a-tanh-activation-function-over-a-standard-sigmoid-activation-function-for-artificial-neural-nets-and-vice-versa>
24. How to fix exploding gradients in a MLP?

- a. L2 norm clipping (6:47):
<https://www.appliedaicourse.com/lecture/11/applied-machine-learning-online-course/3383/gradient-checking-and-clipping/8/module-8-neural-networks-computer-vision-and-deep-learning>
25. You want to detect outliers using Neural Nets? How would you go about doing it?
- a. Use autoencoders and find points for which $\|X_i - \hat{X}_i\|$ is large at end of training.