Final Review

CS 4277: Deep Learning

8 Measuring Performance

1. * Describe the three principle sources of errors that lead to poor generalization in machine learning and how they can be reduced. (8.2-8.3)

2. * Describe the bias-variance tradeoff. (8.3.3)

3. * Describe the double-descent phenomenon in deep neural networks. (8.4)

4. What is the typical approach to choosing hyperparameters? (8.5)

9 Regularization

5. * What is the goal of regularization?

6. * What is the standard approach to explicit regularization? (9.1)

7. Describe L2 regularization. (9.1.2)

8. How is implicit regularization accomplished by SGD? (9.2.2)

9. List 3 heuristic methods of implicit regularization. (9.3)

10. How is implicit regularization accomplished by SGD? (9.2.2)

11. * Consider a model where the prior distribution over the parameters is a normal distribution with mean zero and variance σ_ϕ^2 so that

$$Pr(\phi) = \prod_{i=1}^{J} \text{Norm}_{\phi_j}(0, \sigma_\phi^2)$$

where j indexes the model parameters. When we apply a prior, we maximize $\prod_{i=1}^{I} Pr(\boldsymbol{y}_i|\boldsymbol{x}_i, \phi)Pr(\phi)$. The associated loss function of this model is equivalent to which regularization technique?

10 Convolutional Networks

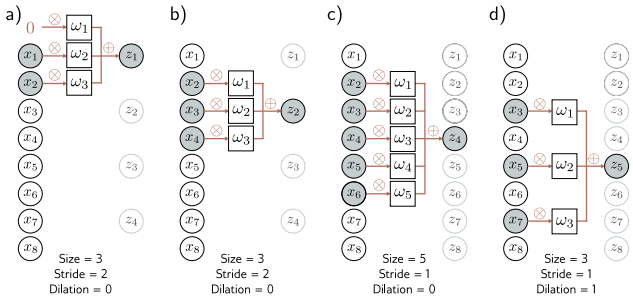
12. Invariance (10.1)

13. Equivariance (10.1)

14. * What properties of images make convolutional neural networks well-suited to them?

15. * What is the motivation for convolutional layers in a neural network?

16. * Write out the equation for the 1D dilated convolution with a kernel size of three and a dilation rate of two, as pictured in Figure 10.3d (reproduced below).



17. * T/F The convolution operation is equivariant to translation.

18. T/F The convolution operation is invariant to translation.

19. * Consider a 1D convolutional layer computed using a kernel size of three and has four channels. How many weights and biases are needed for this convolutional layer?

20. Describe three methods of downsampling. (10.4.1)

21. Describe four methods of upsampling. (10.4.2)

11 Residual Networks

22. * Describe the shattered gradients problem in deep networks. (11.1.1)

23. * What is a residual, a.k.a., skip, connection? (11.2)

24. What is the typical order of operations in a residual block? (11.2 - 11.2.1)

25. * Describe the problem of exploding gradients in residual networks. (11.3)

26. * What is batch normalization and why is it used? (11.4)

27. What is the chief drawback of batch normalization? (11.4.1)

28. What are the advantages of batch normalization? (11.4.1)

12 Transformers

29. * What are the two primary design goals acheived by dot-product self-attention in a language model? (12.2)

30. * Why is positional encoding used in language models? (12.3.1)

31. What are the typical internal sub-layers of a transformer layer? (12.4)

32. * Tokenization (12.5.1)

33. * Embeddings (12.5.2)

34. * Encoders, decoders. (12.6)

35. * Pre-training and fine-tuning (12.6.1 - 12.6.2)

36. * Auto-regressive language modeling (12.7.1)

37. Few-shot learning (12.7.4)

13 Graph Neural Networks

38. Graph-level tasks (13.3.1)

39. Node-level tasks (13.3.1)

40. Edge-prediction tasks (13.3.1)

41. * What is the defining feature of graph convolutional neural networks? (13.4)

42. * What is meant by *relational inductive bias* in graph convolutional netwroks? (13.4)

43. How is parameter sharing accomplished in graph convolutional networks? (13.4.2)

19 Deep Reinforcement Learning

44. What is meant by *temporal credit assignment*?

45. What is the *Markov property* with respect to states $s_1, s_2, \ldots, s_T$ where $t \in T$ are time steps?

46. * What is the primary advantage of deep reinforcement learning over tabular reinforcement learning?