1. **For which label distribution and with which loss is reasonable to adopt the softmax activation function for the output layer?**

a) Gaussian distribution / Mean Squared Error

b) Gaussian distribution / Cross-Entropy

c) Multinoulli distribution / Cross-Entropy

d) Multinoulli distribution / Mean Squared Error

2. **For which label distribution and with which loss is reasonable to adopt the sigmoid activation function for the output layer?**

a) Gaussian distribution / Mean Squared Error

b) Gaussian distribution / Cross-Entropy

c) Multinoulli distribution / Cross-Entropy

d) Bernoulli distribution / Cross-Entropy

3. **For which label distribution and with which loss is reasonable to adopt the linear activation function for the output layer?**

a) Gaussian distribution / Mean Squared Error

b) Gaussian distribution / Cross-Entropy

c) Multinoulli distribution / Cross-Entropy

d) Multinoulli distribution / Mean Squared Error

e) Bernoulli distribution / Cross-Entropy

f) Bernoulli distribution / Mean Squared Error

4. **Consider a CNN layer with 10 filters of size 4x4, a stride of 1 and input images of size 8x8. How many parameters are we required to train for such a layer? (don't consider the bias terms). Please answer with the exact number of the parameters (no formulas)**

Answer: 4*4*10=160

5. **Consider a CNN layer with 10 filters of size 4x4, a stride of 2 and input images of size 8x8 on a single channel (i.e. black and white). How many multiplications (between two numbers) are we required to compute the output (feature map) of such a layer? (don't consider bias terms or activation functions) Please answer with the exact number of the parameters (no formulas).**

Answer: 4*4*4*10=640

6. **Consider a CNN layer with 16 filters of size 3x3, a stride of 1 and input images of size 10x10. How many parameters are we required to train for such a layer? (don't consider the bias terms) Please answer with the exact number of the parameters (no formulas)**

Answer: 3*3*16=144

7. **Consider an unshared CNN layer (i.e. a Convolution layer where the weights are not shared across different positions) with just 1 filter of size 3x3, a stride of 1, input images of size 4x4 and no**

padding. How many parameters are we required to train for such a layer? (don't consider the bias terms). Please answer with the exact number of the parameters (no formulas).

Answer: 3*3*4=36

8. **Consider a CNN layer with 10 filters of size 4x4, a stride of 2 and input images of size 6x6 on a single channel (i.e. black and white). How many multiplications (between two numbers) are we required to compute the output (feature map) of such a layer? (don't consider bias terms or activation functions) Please answer with the exact number of the parameters (no formulas).**

Answer: 4*4*4*10=640

9. **Consider a CNN layer with 10 filters of size 4x4, a stride of 1 and input images of size 6x6 on a single channel (i.e. black and white). How many multiplications (between two numbers) are we required to compute the output (feature map) of such a layer? (don't consider bias terms or activation functions) Please answer with the exact number of the parameters (no formulas).**

Answer: 4*4*9*10=1440

10. **Consider a time series prediction task where, given a time sequence in input up to time t, the output should predict the value of the input at time f+1. Which architecture cannot be used to perform task?**

a) A feed-forward network.

b) A recurrent network with short-cut connections from the input to the output.

c) A recurrent network with feedback connection from output at time t to input at time t+1.

d) A bidirectional recurrent network.

e) A recurrent network with short-cut connections from the hidden state a time t to the output at time t+1.

11. **Given stochastic variables X_1, X_2, X_3, X_4 and the fact that:**
    - **X_1 is conditionally independent from X_4 given X_2 and X_3;**
    - **X_1 is conditionally independent from X_3 given X_2;**

    The joint probability $P(X_1, X_2, X_3, X_4)$ can be factorized as:

a) $P(X_1)\ P(X_2|\ X_1)\ P(X_3|\ X_2, X_1)\ P(X_4|\ X_3, X_2, X_1)$

b) $P(X_1)\ P(X_2)\ P(X_3)\ P(X_4|\ X_3, X_2)$

c) $P(X_1)\ P(X_2|\ X_1)\ P(X_3|\ X_1)\ P(X_4|X_2)$

d) $P(X_1)\ P(X_2|\ X_1)P(X_3|\ X_2)\ P(X_4|X_1)$

e) $P(X_1)\ P(X_2|\ X_1)P(X_3|\ X_1)\ P(X_4|X_1)$

12. **Given stochastic variables X_1, X_2, X_3, X_4 and the fact that:**
    - **X_4 is conditionally independent from X_3 given X_2;**
    - **X_4 is conditionally independent from X_2 given X_1;**

    the joint probability distribution $P(X_1, X_2, X_3, X_4)$ can be factorized as:

a) $P(X_1)\ P(X_4|\ X_2, X_1)\ P(X_2|\ X_1)\ P(X_3|\ X_1)$

b. $P(X_1)\ P(X_4|\ X_3, X_1)\ P(X_3|\ X_2)\ P(X_2|\ X_1)$

c. $P(X_2 \mid X_1)\ P(X_4 \mid X_3, X_2)\ P(X_3 \mid X_1)\ P(X_1)$

d. $P(X_4 \mid X_3, X_2)\ P(X_3 \mid X_2)\ P(X_2 \mid X_1)\ P(X_1)$

e. $P(X_3 \mid X_1)\ P(X_2 \mid X_1)\ P(X_1)\ P(X_4 \mid X_3, X_1)$

13. Given stochastic variables X_1, X_2, X_3, X_4 and the fact that:

- **X_2 is conditionally independent from X_3 given X_4;**
- **X_1 is conditionally independent from X_4 given X_3;**

the joint probability distribution $P(X_1, X_2, X_3, X_4)$ can be factorized as:

a) $P(X_2)\ P(X_4 \mid X_2)\ P(X_1 \mid X_2)\ P(X_3 \mid X_1, X_2, X_4)$

b. $P(X_1)\ P(X_2)\ P(X_3)\ P(X_4 \mid X_3, X_2)$

c. $P(X_3)\ P(X_2 \mid X_1)\ P(X_3 \mid X_1)\ P(X_4 \mid X_2)$

d. $P(X_1)\ P(X_2 \mid X_1)\ P(X_3 \mid X_2)\ P(X_4 \mid X_3)$

e. $P(X_3)\ P(X_2 \mid X_1)\ P(X_3 \mid X_1)\ P(X_4 \mid X_2)$

14. Given stochastic variables X_1, X_2, X_3, X_4 and the fact that:

- **X_1 is conditionally independent from X_4 given X_2 and X_3;**
- **X_1 is conditionally independent from X_3 given X_2;**

The joint probability $P(X_1, X_2, X_3, X_4)$ can be factorized as:

a) $P(X_1)\ P(X_2 \mid X_1)\ P(X_3 \mid X_2, X_1)\ P(X_4 \mid X_3, X_2)$

b) $P(X_1)\ P(X_2)\ P(X_3)\ P(X_4 \mid X_3, X_2)$

c) $P(X_1)\ P(X_2 \mid X_1)\ P(X_3 \mid X_1)\ P(X_4 \mid X_2)$

d) $P(X_1)\ P(X_2 \mid X_1) P(X_3 \mid X_2)\ P(X_4 \mid X_1)$

e) $P(X_1)\ P(X_2 \mid X_1) P(X_3 \mid X_1)\ P(X_4 \mid X_1)$

15. Given stochastic variables X_1, X_2, X_3, X_4 and the following Markov Network with factors φ_i. The joint probability distribution $P(X_1, X_2, X_3, X_4)$ can be factorized as:

a) $\frac{1}{z}\ \varphi_1(X_1, X_2)\ \varphi_2(X_2, X_3)\ \varphi_3(X_2, X_4)$

b) $\frac{1}{z}\ \varphi_1(X_1, X_2)\ \varphi_2(X_1, X_4)\ \varphi_3(X_2, X_3)$

c) $\frac{1}{z}\ \varphi_1(X_2, X_4)\ \varphi_2(X_4, X_1)\ \varphi_3(X_1, X_2)$

d) $\frac{1}{z}\ \varphi_1(X_1, X_2, X_4)\ \varphi_2(X_2, X_3)$

e) $\frac{1}{z}\ \varphi_1(X_2, X_4, X_3)\ \varphi_2(X_1, X_4)\ \varphi_3(X_2, X_4)$

16. Let P and Q be two probability distributions, where Q is parametric. Which of the following sentences about the Kullback-Leibler divergence is true?

◯ KL divergence is distance between two probability distributions.

◯ KL divergence is a measure of dissimilarity between two probability distributions.

◯ KL divergence is the same as Cross Entropy.

◯ Minimising the Cross-Entropy of P w.r.t. Q is equivalent to minimising the KL divergence between P and Q

◯ Minimising the Mean Squared Error of P w.r.t. Q is equivalent to minimising the KL divergence between P and Q.

17. **Which of the following statements about the Universal Approximation theorem for neural networks is true? Choose one or more choices:**

◯ A Neural Network with at least two hidden layers and a squashing activation function can approximate any continuous function.

◯ A Neural Network with at least one hidden layer and a squashing activation function can approximate arbitrarily well any continuous function with any number of hidden neurons.

◯ A Deep Neural Network with one hidden layer and a squashing activation function can approximate arbitrarily well any continuous function given enough hidden neurons.

◯ The number of hidden neurons required for a neural network with one hidden layer and a squashing activation function to approximate up to a given extent a continuous function may be exponential.

◯ The number of hidden neurons required for a neural network with many hidden layers and a squashing activation function to approximate up to a given extent a continuous function is linear in the depth of the network.

18. **Multi-task learning can improve the predictive performance over single-task learning. Select your choice:**

a) because there are more training data to train on.

b) because the network is more complex and thus more expressive.

c) only if the considered tasks are somehow related.

d) only if there are many tasks to use (usually at least 3).

e) None of the other answers.

19. **A sequential transduction is stationary if:**

a) It is casual.

b) The output is a linear function of its inputs.

c) It has finite memory.

d) It can be learned by a feed-forward network with a time window of size equal to the memory of the transduction.

e) None of the above answer is correct.

20. **Which one of the following properties about optimizing the weights of deep neural networks is true? Choose one or more choices:**

◯ Local minima are probably associated to high cost;

◯ Gradient clipping can alleviate the vanishing gradient problem;

◯ Gradient clipping can alleviate the exploding gradient problem;

◯ Local minima are probably associated to low cost;

◯ Saddle points are probably associated to low cost;

◯ Saddle points are probably associated to high cost.

21. Which one of the following sentences about regularization are true? Choose one or more choices:

◯ Parameter norm penalties are useless when there is dropout because it is a stronger regularizer;

◯ Dropout is useless when Parameter norm penalties are used because it is a stronger regularizer;

◯ With Parameter norm penalties, the higher we weight the norm the more the model is regularized;

◯ Weight decay is the same as L1 regularization;

◯ Weight decay is the same as L2 regularization.

22. Which one of the following sentences about regularization are true? Choose one or more choices:

◯ Regularization reduces the training error;

◯ Regularization reduces the generalization error;

◯ Regularization increases the inductive bias of the model;

◯ Regularization increases the variance of the model;

23. In order for a sequential transduction to have a recursive state representation, the following property should hold:

a) The corresponding Recursive Network should not have short-cut connections from input to output;

b) The corresponding Recursive Network should have for each hidden node a shift-time connection q^{-i} with $i>0$;

c) The transduction should have finite memory;

d) The transduction should be causal;

e) Any sequential transduction admits a recursive state representation.

24. Which one of the following sentences about the Unshared Convolution (i.e. a convolution where the weights are not shared across different locations) is true?

a) The Unshared Convolution is slower (from a complexity point of view) compared to standard Convolution;

b) The Unshared Convolution is faster (from a complexity point of view) compared to standard Convolution;

c) The Unshared Convolution maintains the translational equivariance of the computed representation;

d) The Unshared Convolution can be combined with Pooling similarity to standard Convolution. In this case the property of pooling of producing a representation approximately invariant to small translations is lost;

e) The Unshared Convolution can be combined with Pooling similarity to standard Convolutions. The property of pooling of producing a representation approximately invariant to small translations still holds;

f) None of the above.

25. Which of the following sentences about optimisation of deep neural networks are true?

a) In the optimisation of deep neural networks, the weights corresponding to low training error are the ones that generalise better;

b) Minimising the true error would be more prone to overfitting than minimising the empirical error;

c) We don't necessarily need a differentiable loss function to perform gradient descent;

d) Gradient descent many times converges to the global minima of an optimization problem even though it is not guaranteed to do so;

e) None of the above

26. Which component of an LSTM cell is very important for the cell to work in a proper way?

a) Peepholes connections.

b) Input gate.

c) Output gate.

d) Forget gate.

e) Input activation function.

27. A leaky integrator with a ReLU activation function and a = 0.1 can be implemented by a GRU with:

a) Activation function = ReLU, z=0.5, r=0.1, input and recurrent weights multiplied by 2;

b) Activation function = tanh, z=0.1, r=1;

c) Activation function = ReLU, z = 0.1, r = 0, input and recurrent weights multiplied by 0.5;

d) Activation function = ReLU, 2 = 0.5, r = 1, input and recurrent weights multiplied by 0.1;

e) None of the above.

28. A leaky integrator with a ReLU activation function and a = 0.2 can be implemented by a GRU with:

a) Activation function = ReLU, z=0.2, r=1, input and recurrent weights multiplied by 5;

b) Activation function = tanh, z=0.2, r=1;

c) Activation function = ReLU, z = 0.2, r = 0, input and recurrent weights multiplied by 0.2;

d) Activation function = ReLU, z = 0.2, r = 1, input and recurrent weights multiplied by 2;

e) A GRU cannot implement the described leaky integrator.

29. Consider a single GRU. What are the values to use for z and r to obtain the same behaviour of a standard RNN unit?

a) z = 1 and r = 1;

b) z = 1 and r = 0;

c) z=0 and r=1;

d) z=0 and r = 0;

e) It is not possible to reproduce the behaviour of an RNN unit.

30. Which one of the following is an advantage of using deep neural networks over linear models?

a) A Deep neural network has the same expressive power of linear models;

b) A Neural Network always performs better than linear models;

c) A deep neural network has better generalization than a linear model;

d) The functions we want to learn are always a composition of simpler functions, so deep neural networks are always more suited for learning problems;

e) None of the above

31. Under which condition Back-propagation Through Time and Real-Time Recurrent Learning compute exactly the same gradient?

a) If the Recurrent network contains at least one hidden layer;

b) If the Recurrent network contains high-order recurrent connections;

c) If the learning task is sequence classification;

d) If the learning task is a proper IO-transduction, i.e. if there is a target for any time step t;

e) There are no restrictions: they compute the same gradient for any Recurrent network.

32. The main feature of a Denoising Autoencoder is:

a) The use of an architecture with a hidden layer with a number of units that is much lower that the dimension of the input space;

b) The use of an architecture with a hidden layer of linear units;

c) The use of data that has been pre-processed to remove noise;

d) The use of an architecture with a first recurrent layer of sigmoidal units to reduce the noise in input;

e) The use of input data corrupted by noise.

33. Training of a Restricted Boltzmann Machine is performed thanks to:

a) The standard Back-propagation algorithm;

b) Gradient descent plus ancestral sampling;

c) Gradient ascend plus ancestral sampling;

d) A multi-phase algorithm based only on Gibbs sampling;

e) Gradient ascent plus Gibbs sampling.

34. Teacher forcing can be used for which kind of architecture?

a) Any Kind of architecture;

b) Only for Feed-forward architectures;

c) Only for Recurrent architectures;

d) Only for Recurrent architectures with short-cut connections from the input to the output;

e) Only for Recurrent architectures with feedback from output connections.

35. **Assume to have a classification task for sequences were the target value at time t only depends on inputs at time $\tau > t - 51$. Which neural network architecture would be the "best" one to use, especially from the point of view of training?**

a) A Feed-forward network with time window equal to 50;

b) A shallow Recurrent network with one hidden layer of 50 units;

c) A shallow Recurrent network with one hidden layer of 50 units, and short-cut connections from the input to the output;

d) A Recurrent network with two hidden layers of 25 units each;

e) A Recurrent network with two hidden layers of 25 units each, and short-cut connections from the units in the first hidden layer to the output.

36. **Which of the following are advantages of using deep neural networks over shallow ones? Note that when we say "required" we mean "required for achieving a pre-defined error rate". Choose one or more choices:**

◯ The number of neurons required at each layer is exponentially smaller compared to the neurons required for a shallow network;

◯ The number of neurons required at each layer in a deep network may be exponentially smaller compared to the neurons required for a shallow network;

◯ A Deep network with N neurons per layer overfit less than a shallow network with a single layer of N neurons;

◯ A Deep network tend to overfit less than a wider shallow network (i.e. a shallow network with more neurons).

37. **For which reasons using CNN on images gives, in general, better results (in terms of test error) than using dense layers? Choose one or more choices:**

◯ Because it significantly reduces the number of parameters to learn;

◯ Because it is faster;

◯ Because of the translation equivariance given by tied weights;

◯ Because of the translation invariance given by tied weights.

38. **Which one of the following sentences about Second-order optimisation methods is WRONG?**

a) They consider the Hessian of the function to optimise;

b) They are computationally slower than first-order methods;

c) If used instead of first-order gradient descent would result in faster convergence in terms of number of steps needed;

d) They empirically converge to a worse solution compared to first-order methods.

39. **When the input data are images, the learning bias induced by a CNN layer allows to improve the predictive performance over a dense layer all the times?**

a) Yes

b) No

40. The Encoding Network is:

a) Any network that maps the input of dimension n into a hidden representation of size m where n>>m;

b) It is the initial part of any Autoencoder;

c) It is the network obtained by unrolling a recursive network on a specific sequence in input.

d) It is the initial part of Autoencoders for which the size m of the hidden representation is smaller of the dimension n of input data.

e) None of the above answer is correct.

41. Why is Mini-batch gradient descent commonly used for training deep neural networks instead of gradient descent or stochastic (online) gradient descent? Choose one or more choices:

◯ Mini-batch gradient descent estimates the gradient better than gradient descent;

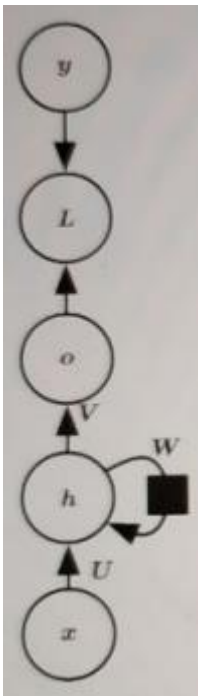
◯ Mini-batch gradient descent estimates the gradient better than online gradient descent;

◯ The gradient in Mini-batch gradient descent is faster to compute compared to gradient descent;

◯ The gradient in Mini-batch gradient descent is faster to compute compared to online gradient descent;

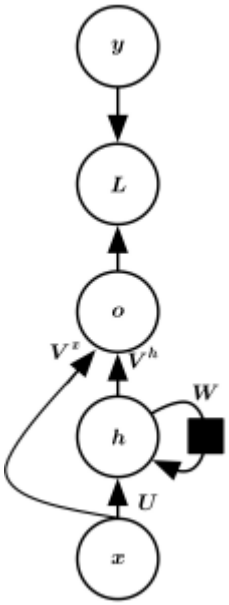
◯ The gradient in online gradient descent is too noisy, and in gradient descent it may be too slow to compute.

42. Suppose to have a IO-isomorphic prediction task and a RNN with the following Recurrent Network: where y is the target, L the loss function, o the RNN output, h the hidden state, x the input at time t and the black square represents the time-shift operator q^{-1}. U, W, V^x and V^h are weights matrices. Suppose to use back-propagation through time with mini-batch equal to 2 for training. Given an input sequence composed of 3 items, how many terms should be summed up to compute the gradient of the loss with respect to U?



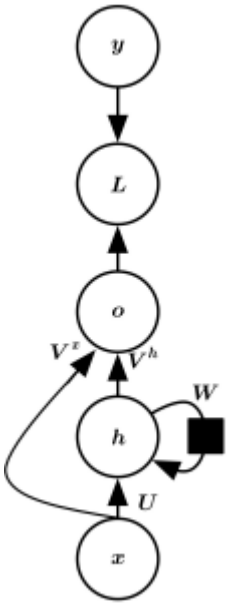
Answer:

43. Suppose to have a IO-isomorphic prediction task and a RNN with the following Recurrent Network: where y is the target, L the loss function, o the RNN output, h the hidden state, x the input at time t and the black square represents the time-shift operator q^{-1}. U, W, V^x and V^h are weights matrices. Suppose to use back-propagation through time with mini-batch equal to 2 for training. Given an input sequence composed of 4 items, how many terms should be summed up to compute the gradient of the loss with respect to U?



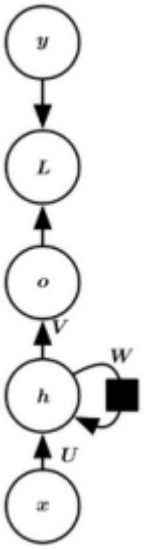
Answer:

44. Suppose to have a IO-isomorphic prediction task and a RNN with the following Recurrent Network: where y is the target, L the loss function, o the RNN output, h the hidden state, x the input at time t and the black square represents the time-shift operator q^{-1}. U, W, V^x and V^h are weights matrices. Suppose to use back-propagation through time with mini-batch equal to 2 for training. Given an input sequence composed of 4 items, how many terms should be summed up to compute the gradient of the loss with respect to W? (don't consider the contribution of $h^{(0)}$ which is the zero vector)



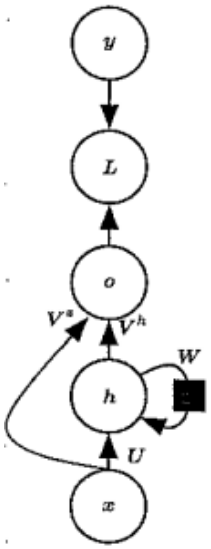
Answer:

45. Suppose to have a IO-isomorphic prediction task and a RNN with the following Recurrent Network: where y is the target, L the loss function, o the RNN output, h the hidden state, x the input at time t and the black square represents the time-shift operator q^{-1}. U, W, V^x and V^h are weights matrices. Suppose to use back-propagation through time with mini-batch equal to 1 for training. Given an input sequence composed of 4 items, how many terms should be summed up to compute the gradient of the loss with respect to U?



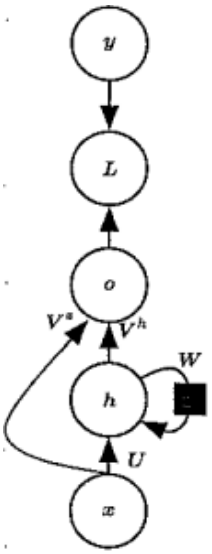
Answer:

46. Suppose to have a IO-isomorphic prediction task and a RNN with the following Recurrent Network: where y is the target, L the loss function, o the RNN output, h the hidden state, x the input at time t and the black square represents the time-shift operator q^{-1}. U, W, V^x and V^h are weights matrices. Suppose to use back-propagation through time with mini-batch equal to 3 for training. Given an input sequence composed of 3 items, how many terms should be summed up to compute the gradient of the loss with respect to U?



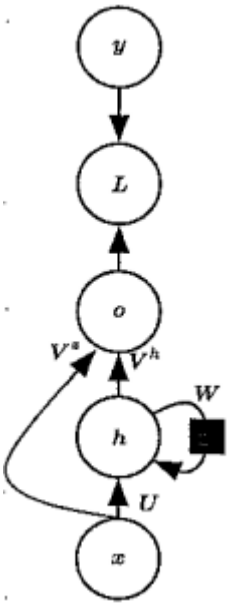
Answer:

47. Suppose to have a IO-isomorphic prediction task and a RNN with the following Recurrent Network: where y is the target, L the loss function, o the RNN output, h the hidden state, x the input at time t and the black square represents the time-shift operator q^{-1}. U, W, V^x and V^h are weights matrices. Suppose to use back-propagation through time with mini-batch equal to 2 for training. Given an input sequence composed of 2 and 4 items, respectively, how many terms should be summed up to compute the gradient of the loss with respect to U?



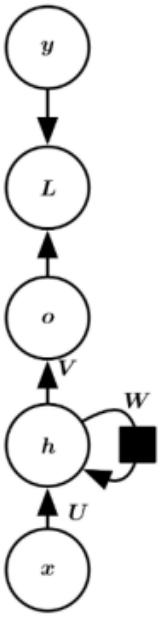
Answer:

48. Suppose to have a IO-isomorphic prediction task and a RNN with the following Recurrent Network: where y is the target, L the loss function, o the RNN output, h the hidden state, x the input at time t and the black square represents the time-shift operator q^{-1}. U, W, V^x and V^h are weights matrices. Suppose to use back-propagation through time with mini-batch equal to 3 for training. Given an input sequence composed of 3 items, how many terms should be summed up to compute the gradient of the loss with respect to U?



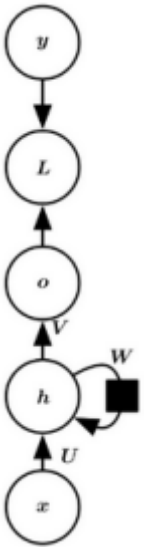
Answer:

49. Suppose to have a IO-isomorphic prediction task and a RNN with the following Recurrent Network: where y is the target, L the loss function, o the RNN output, h the hidden state, x the input at time t and the black square represents the time-shift operator q^{-1}. U, W, V^x and V^h are weights matrices. Suppose to use back-propagation through time with mini-batch equal to 1 for training. Given an input sequence composed of 4 items, how many terms should be summed up to compute the gradient of the loss with respect to W? (don't consider the contribution of $h^{(0)}$ which is the zero vector)



Answer: 1+2+3+4=10

50. Suppose to have a IO-isomorphic prediction task and a RNN with the following Recurrent Network: where y is the target, L the loss function, o the RNN output, h the hidden state, x the input at time t and the black square represents the time-shift operator q^{-1}. U, W, V^x and V^h are weights matrices. Suppose to use back-propagation through time with mini-batch equal to 1 for training. Given an input sequence composed of 3 items, how many terms should be summed up to compute the gradient of the loss with respect to U?



Answer: 3+2+1=6