You must write your full name and all answers (but nothing else) on the final page. Only the final sheet will be used for grading.

There are 22 total answers. A perfect score on this exam is 20 (of 22). Each answer is worth the same number of points. There is no partial credit, and there is no negative-point penalty for a wrong answer.

No calculators, books or other aids are to be used during the exam; only pencil/pen and paper.

1. A neural net outputs the following vector: \([\frac{1}{4}, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}]\). The target output is: \([0, \frac{1}{2}, \frac{1}{4}, \frac{1}{4}]\). Compute the cross-entropy error (with all logarithms calculated using a base of 2.) (** Answer = 2)

2. The performance of your neural network has been unsatisfactory. Which of the following actions is LEAST likely to improve the results:
   (a) Increase the learning rate to 500% of the current value.
   (b) Read literature on practical tips for backpropagation written by Geoff Hinton.
   (c) Change your loss function from cross-entropy to mean-squared error.
   (d) Double the filter size and the strides for the convolution layers.
   (e) Remove one of the Max-pooling layers.
   (f) Inside the LSTM units, replace all of the sigmoid activation functions on the gates with linear activation functions. (**)
   (g) Set the dropout rate to 30%
   (h) Scale all of the values of the input vectors to non-negative fractions.

3. In a neural network, layer U (containing many neurons) sends its activation values to the output layer, V, with the weights W on the (full set of) connections between U and V. Layer U employs a linear activation function, while layer V uses a sigmoid. In one minibatch, 5 cases pass forward through the net, producing five activation values for \(u_i\): (+0.3, + 0.4, +0.1, +4.1, +3.0). For each of the 5 cases, the output values for all neurons in V are below the corresponding values in each of the 5 target vectors, and none of the sums-of-weighted-inputs to V have an absolute value greater than 10. The loss function is mean squared error. The weight on the connection from \(u_i\) to \(v_k\) is \(w_{ik}\) and the learning rate is positive. What is the sign of \(\Delta w_{ik}\) for the ensuing round of backpropagation.
   (a) Positive (**)
   (b) Negative
   (c) Zero
   (d) Not enough information is given to make the correct determination.

4. Consider the same problem as above, but now V uses the unconventional activation function \(f(x) = x^2\), and the 5 values of \(u_i\) are: (-0.1, -0.3, -0.1, -2.3, -1.6). Also, \(w_{ik} = +0.5\). What is the sign of \(\Delta w_{ik}\) for the ensuing round of backpropagation.
   (a) Positive
   (b) Negative
   (c) Zero
   (d) Not enough information is given to make the correct determination. (**)

IT-3030 Deep Learning Midterm Exam (Spring 2020)
5. Input layer U consists of 3 feature maps (i.e. channels), each of size 15 x 15. Layer U feeds directly into convolution layer V, which consists of 5 feature maps, each of which uses a stack (of depth 3) of 3 x 3 kernels, horizontal and vertical strides of size 2, and no zero padding. How many total neurons are in layer V? (** 5 x 7 x 7 = 245)

6. In the problem above, how do the number of neurons in V change if the kernel size changes from 3 x 3 to 5 x 5 (but the strides remain the same)?

(a) Increase by 66%
(b) Decrease by 66%
(c) Decrease by 33%
(d) 0%: the number does not change.
(e) Increase by 18%
(f) Decrease by 27% (** 5 x 6 x 6 = 180 neurons)
(g) Since this is a dynamic system, the number cannot be determined without actually simulating the network.

7. Which of the following is true about the softmax activation function?

(a) It is generally more useful for classification problems than for regression problems.
(b) Its Jacobian matrix consists of one type of term along the main diagonal, and another type of term everywhere else.
(c) It can scale a vector of activations.
(d) It is technically possible to use softmax in a hidden layer of a network without violating the basic computational rules of backpropagation.
(e) It can be used to simulate competition among same-layer neurons by magnifying the strengths of the more active neurons and reducing those of the less active neurons.
(f) All of the above (a-e) are True (**)
(g) Several, but not all, of the above (a-e) are True
(h) None of the above (a-e) are True.

8. The Word2Vec system has produced the following binary embeddings for four words, denoted as W, X, Y and Z:

- W: 1 0 1 1 1 1
- X: 0 1 1 0 1 0
- Y: 0 0 1 1 1 0
- Z: 1 1 1 0 1 1

Word2Vec enables the representation of analogies: A:B = C:D, whose connotation is: A is to B as C is to D. Based on the above embeddings, which of the following is the most accurate analogy:

(a) W:X = Y:Z
(b) Y:Z = W:X
(c) W:Y = Z:X (**)
(d) W:Z = X:Y

9. For which of the following problems would it be reasonable to attempt a solution using a network with recurrent connections?

(a) Given the notes of a song, determine the composer.
(b) Generating natural-language captions for photographs.
(c) Using time series of sensor and motor readings to control the actions of a robot.
(d) Translating text from German to Italian.
(e) Analyzing video footage to detect suspicious behavior in public places.
(f) All of the above (a-e) are reasonable. (**)
(g) Several, but not all, of the above (a-e) are reasonable.
(h) None of the above (a-e) are reasonable.

10. A standard feed-forward neural network is trained by Stochastic Gradient Descent (SGD). The net has 9 input neurons, 5 output neurons and a hidden layer, H, using tanh activations. The network is fully connected, but only the hidden layer uses biases. The network has a total of 300 trainable parameters, $p_i \in P$. During a particular session, it performs the forward and backward processes of SGD on a minibatch consisting of 40 cases.

(a) How many tanh nodes are in H? ___ (** $9H + H + 5H = 300$, so $H = 20$)
(b) How many total internal calculations of gradients of the form $\frac{\partial L}{\partial p_i}$ are performed as a result of this minibatch, where $p_i$ is any trainable parameter? (** $300 \times 40 = 12,000$)
(c) How many total parameter updates (of the $p_i$) are performed as a result of this minibatch? ___ (** $300$, each $p_i$ one time)

11. Consider a deep, feed-forward neural network trained with backpropagation where layer X feeds into layer Y, which feeds into layer Z. Matrix $V$ houses the weights on the connections from X to Y, while matrix $W$ houses those on the Y-to-Z connections; $v_{ij}$ is the weight on the connection from $x_i$ to $y_j$, and $w_{jk}$ is the weight on the connection from $y_j$ to $z_k$. Layer’s X and Y use the hyperbolic tangent (tanh) activation function, while layer Z uses the sigmoid activation function. Which of the following is the proper expression for $\frac{\partial z_k}{\partial x_i}$?

(a) $\frac{\partial z_k}{\partial x_i} = \sum_j v_{ij}(1 - y_j^2)w_{jk}z_k(1 - z_k)$ (**)
(b) $\frac{\partial z_k}{\partial x_i} = \sum_j v_{ij}(1 - y_j^2)\sum_m w_{jm}z_m$ (c) $\frac{\partial z_k}{\partial x_i} = (1 - x_i^2)v_{ij}(1 - y_j^2)w_{jk}z_k(1 - z_k)$
(d) $\frac{\partial z_k}{\partial x_i} = (1 - x_i^2)v_{ij}(1 - y_j^2)w_{jk}z_k(1 - z_k)$
(e) $\frac{\partial z_k}{\partial x_i} = (1 - x_i^2)\sum_j v_{ij}(1 - y_j^2)w_{jk}z_k$
(f) $\frac{\partial z_k}{\partial x_i} = \sum_j x_i v_{ij}(1 - y_j^2)w_{jk}z_k(1 - z_k)$

12. Which of the following are true statements about the three optimizers: Adagrad, RMSProp and Adam?

(a) They base their weight updates on a history of gradients recorded over many minibatches.
(b) They include an individual learning rate for every trainable weight and bias in the neural network.
(c) When the magnitudes of gradients are high, these optimizers tend to decrease the learning rates.
(d) Adam is popular, but there is no mathematical guarantee of its superior performance.
(e) All of the above (a-d) are True (**)
(f) Several, but not all, of the above (a-d) are True
(g) None of the above (a-d) are True

13. An Elman network has a single hidden layer, $H = (h_1, h_2, h_3)$, which uses an ReLU activation function. At time $t$, the activation levels of these three neurons are $(2, 8, 3)$, respectively. At time $t+1$, their activations are $(4, 1, 6)$. Weights $w_{1,3} = 0.4$ and $w_{2,3} = 0.5$ are on the recurrent connections from $h_1$ to $h_3$, and $h_2$ to $h_3$, respectively. Compute the following values:
\[ \frac{\partial h_{t+1}}{\partial w_{2,3}} = (** \ 8) \]
\[ \frac{\partial h_{t+1}}{\partial h_{1,t}} = (** \ 0.4) \]

14. Through their writings and research, which ONE of the following people has had the LEAST direct impact on the field of neural networks.

(a) Yoshua Bengio
(b) Marvin Minsky
(c) Ian Goodfellow
(d) Frank Rosenblatt
(e) Jurgen Schmidhuber
(f) Herbert Simon (**)

15. The business leader for a large Norwegian corporation comes to your AI consulting firm, OuiNoAI, and tries to impress you with buzzwords supplied by his tech staff. All you can really understand is: vanishing gradients. To be able to charge a handsome fee, and still have a good conscience, which of the following responses could you give?

(a) Have you tried reducing the length of the sequences that you feed into your LSTM?
(b) Maybe try L2 regularization instead of L1?
(c) Teacher Forcing normally helps me.
(d) Have you replaced your sigmoids and hyperbolic tangents with RELUs, ELUs or SELUs?
(e) What if you replace those 6 hidden layers with one or two larger layers?
(f) All of the above (a-e) are legitimate suggestions. (**)
(g) Several, but not all, of the above (a-e) are legitimate suggestions.
(h) None of the above (a-e) are legitimate suggestions.

16. For each of the 4 concepts below (I - IV), choose ONE of the expressions (A-I) that is most appropriate for, or characteristic of, that concept. The expressions A-I may be used more than once, but only one can be used per concept.

(I) Contractive Autoencoder (** D)
(II) Dropout (** A)
(III) Adversarial Training (** B)
(IV) Hessian Matrix (** C)

(A) Motivated by Ensemble Learning
(B) \( x_i \leftarrow x_i + \frac{\partial L}{\partial x_i} \) where \( L \) = loss, \( x_i \) = ith input feature.
(C) Assessing a gradient’s reliability
(D) \( \Omega(H,X) = \lambda \sum_i \| \nabla_X h_i \|^2 \) where \( h_i \) = ith hidden neuron; \( X \) = input feature vector; \( \lambda \) = small weighting factor.
(E) \( \frac{\partial L}{\partial w_j} = \sum_{c \in M} \sum_{s \in S} \frac{\partial L}{\partial w_{j,c,s}} \) where \( c \) = cases in minibatch \( M \), \( s \) = locations on feature plane \( S \), \( L \) = loss, and \( w_j \) = a connection weight.
(F) \( J_L^Z = \left(J_L^S\right)^T \cdot J_Z^S \) where \( J \) = a Jacobian, \( L \) = loss, \( Z \) = neural layer, \( S \) = softmax, and \( T \) denotes matrix transpose.
(G) The Jacobian squared
(H) A variant of Teacher Forcing
(I) Invented by Herbert Simon
Print your FULL NAME below:

Print your answers below:

1.
2.
3.
4.
5.
6.
7.
8.
9.
10. (a)
    (b)
    (c)
11.
12.
13. (a)
    (b)
14.
15.
16. (I)
    (II)
    (III)
    (IV)