Supplementary Materials for Solving Machine Learning Problems

Sunny Tran SUNNYT@MIT.EDU
MIT EECS

Pranav Krishna PKRISHNA@MIT.EDU
MIT EECS

Ishan Pakuwal IPAKUWAL@MIT.EDU
MIT EECS

Prabhakar Kafle PKAFLE@MIT.EDU
MIT EECS

Nikhil Singh NSINGH1@MIT.EDU
MIT Media Lab

Jayson Lynch JAYSONL@MIT.EDU
University of Waterloo

Iddo Drori IDRORI@MIT.EDU
MIT EECS

Editors: Vineeth N Balasubramanian and Ivor Tsang

Appendix A. Open Response Questions and Answers by our Model

A.1. Sample correct questions and model answers

Basic Questions	**Answer**
1. Compute the magnitude of $[3, 12]$.	12.37
2. If $x = [16, 4, 9]$, what is $\|\|x\|\|$?	18.79
3. Find the Euclidean length of $[7, 0, 1]$.	7.07
4. Compute the magnitude of $[10, 10, 1]$.	14.18
5. What is the magnitude of the vector $[0, 7]$?	7

Table 1: Example basic questions and answers generated by our model.

Perceptrons Questions	**Answer**
1. If the decision boundary of a classifier is θ, where $\theta = (4, 1)$, how does it classify point p, where $p = (2, -4)$?	4
2. How does a classifier with decision boundary θ classify a point p if $\theta = (2, 4)$ and $p = (0, -4)$?	-16
3. What is the most number of mistakes made by the perceptron algorithm if 13 is the maximum magnitude of a point in the dataset and the dataset has a margin of 4 to the separator.	2704
4. Determine if the following two classifiers represent the same hyperplane, $[0, 1, 1]$ and $[0, 1, 1]$. If so, return 1, and return anything else otherwise.	1
5. A classifier has a decision boundary where $\theta = (1, 0)$. What value does it classify p, where $p = (3, 0)$?	3

Table 2: Example perceptrons questions and answers generated by our model.

Features Questions	**Answer**
1. What is the margin of a classifier with $\theta = -1$ and $\theta_0 = -6$ on a point 0 with label 1?	-6
2. What is the loss for the data point $(0, -1)$ if we use NLL. Let $\theta = 2$ and $\theta_0 = 0$. Also use natural log where the base is 2.71828.	-0.69
3. A point 1 has label 1. Compute the margin of a classifier on this point. Let the θ of the classifier be -1 and the θ_0 of the classifier be -1.	2
4. Given the values for $\theta = 2$ and $\theta_0 = 1$, compute the NLL loss on the data point $(-1, 0)$. Use log base e of 2.71828 for the log.	-0.31
5. What does the sigmoid function return when you pass into it 1? Hint: have $e = 2.71828$.	0.73

Table 3: Example feature questions and answers generated by our model.

Logistic Regression Questions	**Answer**
1. What is the result of $\theta x + \theta_0$ if $x = (-1, 0)$, $\theta = (0, -1)$, and $\theta_0 = 3$?	3
2. Let $\theta = (1, -1)$, $\theta_0 = 2$, and $x = (-1, 0)$. Compute $\theta x + \theta_0$.	1
3. If you let $\theta = 1$ and $\eta = 0.05$, what is the updated θ value after one gradient descent step if the loss function is given by $(0 \times \theta + 3)^2$?	1
4. If we have $x = (0, -1)$, $\theta = (1, -2)$, and $\theta_0 = -3$, then what is the result of $\theta x + \theta_0$?	-1
5. What is the value of $\theta x + \theta_0$ if $x = (-1, 0)$, $\theta = (-1, 0)$, and $\theta_0 = -3$?	-2

Table 4: Example logistic regression questions and answers generated by our model.

Regression Questions	Answer
1. If $f(\theta) = (3 \times \theta + 3)^2$ and $\theta = 4$ what is $f(\theta)$?	225
2. Given $\theta = 3$ and $\lambda = 1$, compute the mean squared error with the data points $[(2, 0), (2, 5)]$.	19.5
3. With $\lambda = 1$, the optimal $\theta = 1$. If the data points are $[(0, 1), (1, 2), (2, y)]$, what is the value of y? The optimal θ is computed by mean squared error.	0.27
4. If we let $\theta = 1$ and $\lambda = 0.5$, what is the mean squared error of the given points $[(2, 0), (1, 1)]$?	2.25
5. If $f(\theta)$ is $(7 \times \theta + 8)^2$, what is $f(\theta)$ when $\theta = 15.4$?	13409.64

Table 5: Example regression questions and answers generated by our model.

Neural Networks I Questions	Answer
1. Neurons A and B take inputs 4 and 2 with weights 2 and 1, respectively. Neuron A has offset 0.5 and neuron B has offset 1. Neuron C takes in the output of A and B with weights 1 and 3, respectively, and with offset 3. What is the output?	20.5
2. Neuron C is the output neuron and neuron A takes the input. Compute the output with the given architecture and inputs. Neuron C takes in the offset value $oC = 4$ with weight $wOC = 2$. Neuron C also takes in the neuron A with weight $wAC = 1$. Neuron A takes in the input value $x1 = 3$ output of with weight $w1 = 1$ and offset value $oA = 0.5$ and offset weight $wOC = 2$.	11
3. In a fully-connected feed-forward network, how many weights (including biases) are there for one layer with 40 inputs and 40 outputs?	3200
4. A fully-connected neural network has 90 outputs and 40 inputs. How many total weights are there including the biases?	7200
5. Neuron A and Neuron C are the input and output neurons of a neural network. Neuron A takes in value $x1 = 2$ with weight $w1 = 1$ and offset value $oA = 0.5$ with weight $wOA = 4$. Neuron C takes in the output of neuron A with weight $wAC = 1$ and offset value $oC = 2$ with weight $wOC = 2$. What is the output of neuron C?	8

Table 6: Example neural network questions and answers generated by our model.

Neural Networks II Questions	Answer
1. A ReLU is applied to the output of neuron C, which takes in outputs from neurons A with weight $wAC = 1$ and B with weight $wBC = 2$ and offset $oC = 1$. Neuron A takes in value $x1 = 4$ with weight $w1 = 1$ and offset value $oA = 0.5$. Neuron B takes in input $x2 = 4$ with an offset of 1.	15.5
2. A neural network has inputs $x1 = 4$ with weight 2 and $x2 = 2$ with weight 1 and offset value $oA = 0.5$. Neuron B inputs $x2$ with offset 1. Neuron C takes in the output of neurons A and B with offsets $wAC = 1$ and $wBC = 2$, respectively. Neuron C has offset value $oC = 2$ and applies a ReLU on its output. Compute the output.	16.5
3. Neuron C is the output neuron which applies a ReLU on its output and neuron A is the input neuron to a neural network. Compute the output of a neural network with the given architecture and inputs. Neuron C takes in the offset value $oC = 2$ with weight $wOC = 3$. Neuron C takes in the output of neuron A with weight $wAC = 1$. Neuron A takes in the input value $x1 = 0$with weight $w1 = 2$ and offset value $oA = 0.5$ and offset weight $wOC = 3$.	6
4. Neurons A and B take inputs 2 and 1 with weights 2 and 1, respectively. Neuron A has offset 0.5 and neuron B has offset 1 . Neuron C takes in the output of A and B with weights 1 and 2, respectively, and with offset 2. Neuron C also applies a ReLU on its output. What is the output?	10.5
5. Compute the ReLU output of neuron C which takes the output of neuron A with weight 1 and neuron B with weight 2 and offset 1. Neuron B has input 0 and offset 1. Neuron A has input -1 and offset 2 with offset 0.5.	1.5

Table 7: Example neural network questions and answers generated by our model.

Convolutional Neural Networks Questions	**Answer**
1. An image I has length 5 and filter F has length 2, what is the length of the result of applying F to I?	4
2. The row of an image $[1, 1, 1]$ has a filter $[0, 3, 0]$ applied to it. What is the resulting value if they both align?	3
3. What is the minimum number of padding needed to maintain the same output size if the input image is 50 by 50 and the filter is 17 by 17?	8
4. Given that there are 22 inputs to a zero-padded max pooling layer and a stride length of 2, compute the number of output units if we also know the pooling filter size of 3?	11
5. What is the length of the output when we use an image of length 52 and a filter of length 5 if we use a stride length of 2?	24

Table 8: Example convolutional neural network questions and answers generated by our model.

Recurrent Neural Networks Questions	**Answer**
1. Consider a very simple RNN, defined by the following equation: $s_t = w * s_{t-1} + x_t$. Given $s_0 = 0$, $w = 0.1$, and $x = [0.25, 0.5]$, what is s_2?	0.525
2. An RNN is defined as $s_t = w * s_{t-1} + x_t$. If $s_0 = 1$, $w = 1$, and $x = [2, 2, 0]$, what is s_3?	5
3. What is the RNN result s_2 if $s_0 = 2$, $w = 0$, and $x = [9, 5]$ if we let $s_t = w * s_{t-1} + x_t$?	5
4. We define an RNN as $s_t = w * s_{t-1} + x_t$. What is s_2 if $s_0 = 3$, $w = 0.5$, and $x = [2, 2]$?	3.75
5. Let $s_0 = 1.5$, $w = 1.5$, and $x = [1, 0, 2]$. Compute s_3 if $s_t = w * s_{t-1} + x_t$.	9.31

Table 9: Example recurrent neural network questions and answers generated by our model.

State Machines and MDP Questions	**Answer**
1. Let a state machine be described with the equations $s_t = f(s_{t-1}, x_t)$ and $y_t = g(s_t)$, where x_t is the input. If $s_0 = 6$, $f(s_{t-1}, x_t) = \max(s_{t-1}, x_t)$, and $g(s_t) = s_t$, what is the output y_3 after the inputs $[16, 9, 18]$?	18
2. If we have a state machine, defined as $s_t = f(s_{t-1}, x_t)$ and $y_t = g(s_t)$, where x_t is the input, what is the output y_3 if we have $s_0 = 14$, $f(s_{t-1}, x_t) = \max(s_{t-1}, x_t)$, $g(s_t) = 0s_t$, and we input $[14, 15, 9]$?	0
3. Consider the input $x_t = [7, 14, 13, 10]$ to a state machine with equations $s_t = f(s_{t-1}, x_t)$ and $y_t = g(s_t)$. Compute y_4 if our initial conditions are $s_0 = 2$, $f(s_{t-1}, x_t) = \max(s_{t-1}, x_t)$, and $g(s_t) = 4s_t$.	56
4. A state machine is defined by the equations $s_t = f(s_{t-1}, x_t)$ and $y_t = g(s_t)$. Given the conditions $s_0 = 7$, $f(s_{t-1}, x_t) = \max(s_{t-1}, x_t)$, and $g(s_t) = 3s_t$, compute y_5 if the input is $x_t = [17, 4, 14, 2, 16]$.	51
5. What is the output y_5 of a state machine with equations $s_t = f(s_{t-1}, x_t)$ and $y_t = g(s_t)$, conditions $s_0 = 10$, $f(s_{t-1}, x_t) = \max(s_{t-1}, x_t)$, and $g(s_t) = 4s_t$, and input $x_t = [19, 17, 5, 9, 18]$?	76

Table 10: Example state machines and MDP questions and answers generated by our model.

Reinforcement Learning Questions	Answer
1. What is the updated Q value of a tuple (s, a) if $q = 7$, the $a = 0.1$, and $t = 6$?	6.9
2. If $q = 8$, what is its updated value after applying Q learning if $a = 0.1$ and $t = 4$?	7.6
3. Let $q = 2$. After Q learning, what is q if $a = 0.3$ and $t = 8$?	3.8
4. If $a = 0.4$ and $t = 10$, what is the Q learning value after applying one tuple (s, a) if $q = 6$?	7.6
5. After applying Q learning to $q = 4$, what is its value? Let the $t = 10$ and $a = 0.8$.	8.8

Table 11: Example reinforcement learning questions and answers generated by our model.

Decision Trees and Nearest Neighbors Questions	Answer
1. A left region has 6 points classified as positive. There are 44 points in the plane, and 31 points on the left. Compute the entropy.	0.71
2. Consider a 1D classification line on a 2D plane. There is a total of 46 points, 30 of which are on the right and the rest on the left of the 5 boundary. points on the left are classified positive. What is the entropy of the left region?	0.90
3. If a region has 24 points on the left and 46 points total. 8 points that are on the left are positive. Compute the entropy.	0.92
4. The left side of a region has 26 points. Of the 26 points, 4 are classified as positive. What is the entropy of the left region if there are 45 points in total?	0.62
5. What is the entropy of the left side of a region containing 30 points where the plane has 47 points in total and 8 points on the left are positive?	0.83

Table 12: Example decision trees and nearest neighbor questions and answers generated by our model.

A.2. Expression Trees generated by our model for open response questions.

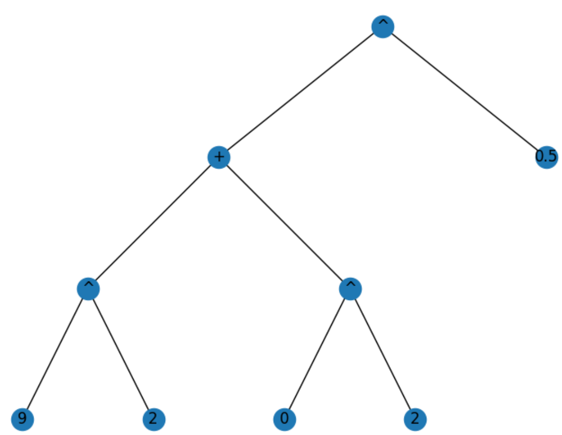

Figure 1: Basics: $(x^2 + y^2)^{0.5} = (9^2 + 0^2)^{0.5}$

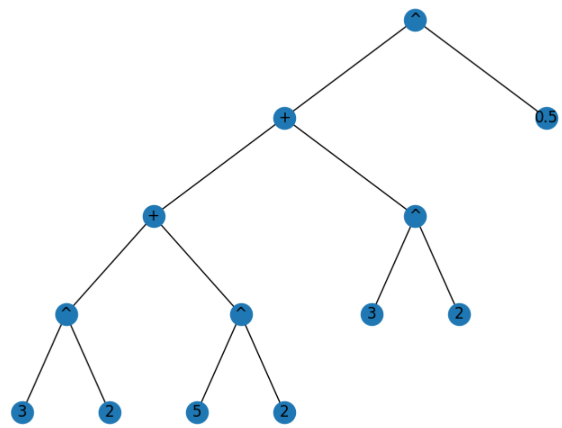

Figure 2: Basics: $(x^2 + y^2 + z^2)^{0.5} = (3^2 + 5^2 + 3^2)^{0.5}$

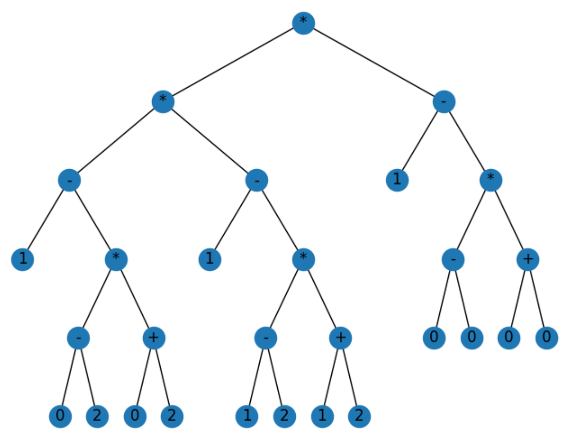

Figure 3: Perceptrons: $(a - (b - c) * (d + p)) * (f - (g - h) * (i + j)) * (k - (l - m) * (n + o)) = ((1 - (0 - 2) * (0 + 2)) * (1 - (1 - 2) * (1 + 2))) * (1 - (0 - 0) * (0 + 0))$

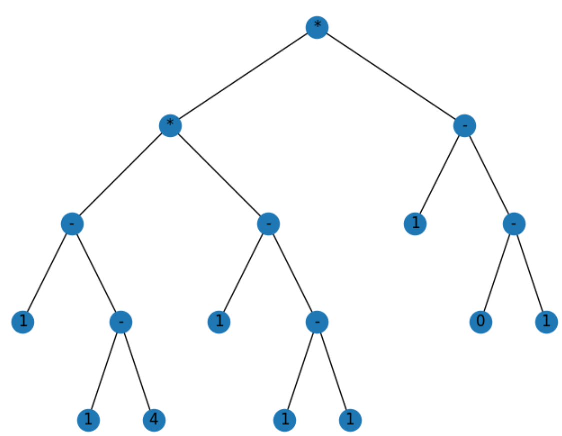

Figure 4: Perceptrons: $((a - (b - c)) * (d - (e - f))) * (g - (h - i)) = ((1 - (1 - 4)) * (1 - (1 - 1))) * (1 - (0 - 1))$

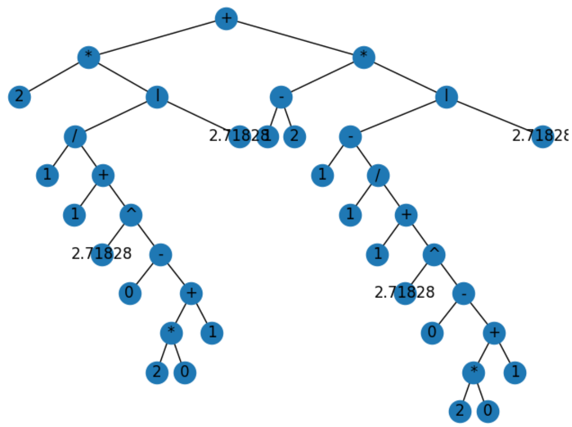

Figure 5: Features: $s * \ln \frac{1}{1+e^{(m-(n*o+p)}} + (q-r)*(1-\ln \frac{1}{1+e^{(m-(n*o+p)}}) = 2*\ln \frac{1}{1+e^{(0-(2*0+1)}} + (1-2)*(1-\ln \frac{1}{1+e^{(0-(2*0+1)}})$

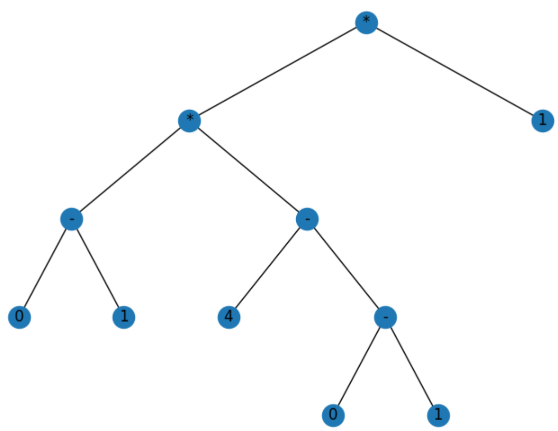

Figure 6: Features: $(u-v)*(v-(w-x))*y = ((0-1)*(4-(0-1))*1$

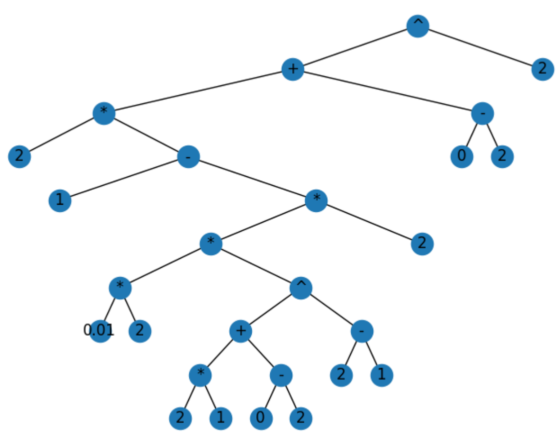

Figure 7: Logistic regression: $(f*(g-((h*i)*(j*k)+(l-m)^{n-o})*p))+(q-r)))^2 = (2*(1-((0.01*2)*(2*1)+(0-2)^{2-1})*2))+(0-2)))^2$

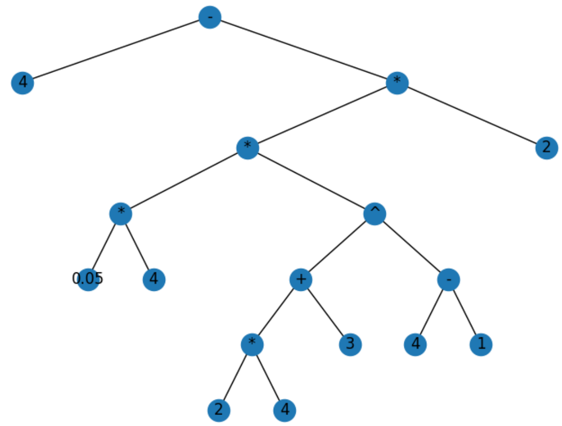

Figure 8: Logistic regression: $g-(h*i)*((k*l)+m)^{n-o})*p = 4-(0.05*4)*((2*4)+3)^{4-1})*2$

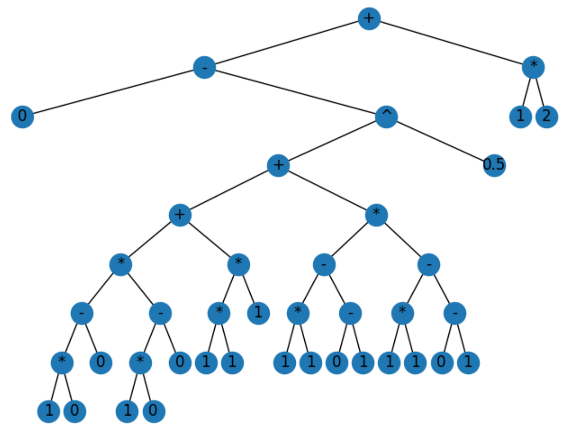

Figure 9: Regression: $(0-((a*b)-c)*((d*e)-f)+((g*h)*i)+(((j*k)-(l-m))*((n*o)-(p-q)))^{0.5})+(r*s) = (0-((1*0)-0)*((1*0)-0)+((1*1)*1)+(((1*1)-(0-1))*((1*1)-(0-1)))^{0.5})+(1*2)$

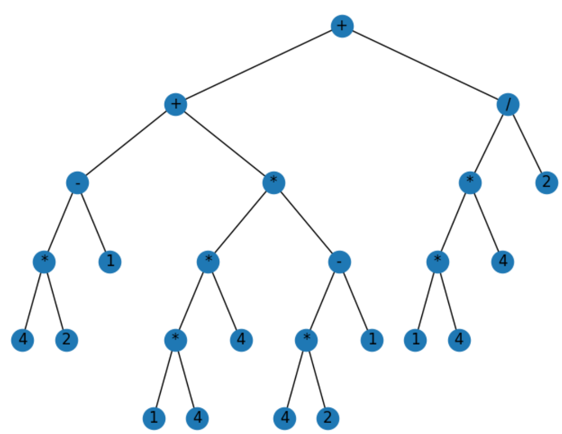

Figure 10: Regression: $((a*b)-c)+(((d*e)*f)*((g*h)-i))+(((j*k)*l)/m) = ((4*2)-1)+(((1*4)*4)*((4*2)-1))+(((1*4)*4)/2)$

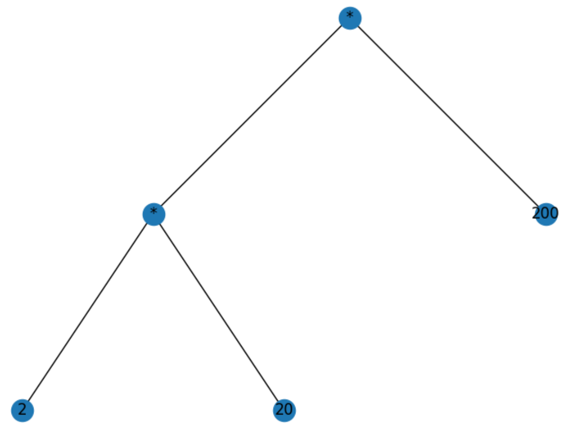

Figure 11: $(x * y) * z = (2 * 20) * 200$

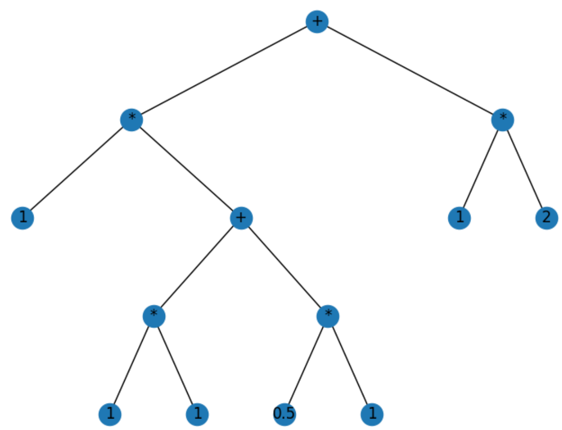

Figure 12: Neural networks I: $(u * ((v * w) + (w * x))) + (y * z) = (1 * ((1 * 1) + (0.5 * 1))) + 1 * 2$

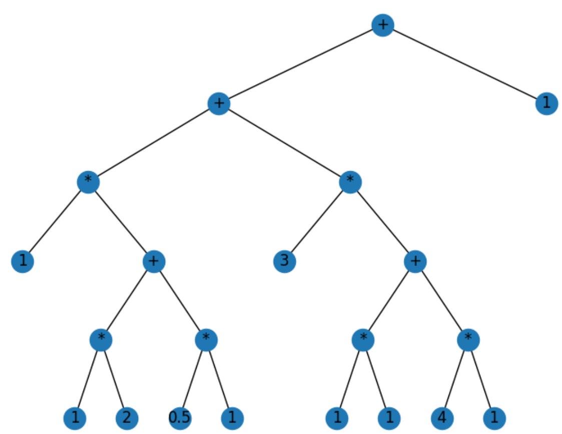

Figure 13: Neural networks I: $p * (q * r + s * t) + u * (v * w + x * y) + z = (1 * (1 * 2 + 0.5 * 1)) + 3 * (1 * 1 + 4 * 1) + 1$

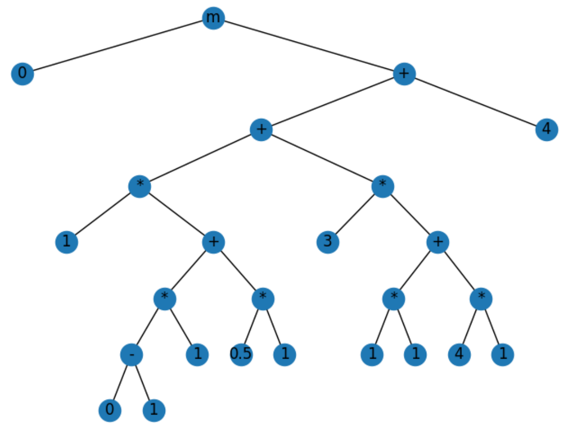

Figure 14: Neural networks II: $\max(0,(g*(h-i)*j+k*l)+m*(n*o+p*q)+r) = \max(0,(1*(0-1)*1+0.5*1)+3*(1*1+4*1)+4)$

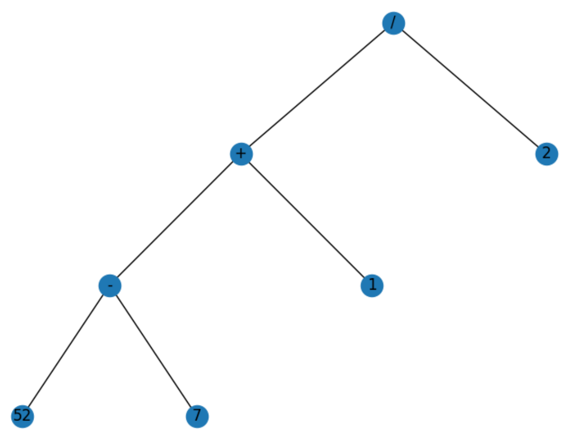

Figure 15: Convolutional neural networks: $((a-b)+c)/2 = ((52-7)+1)/2$

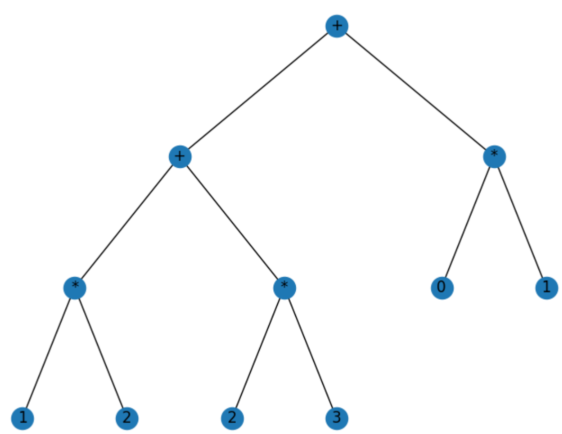

Figure 16: Convolutional neural networks: $(a*b+c*d)+(g*h) = (1*2+2*3)+(0*1)$

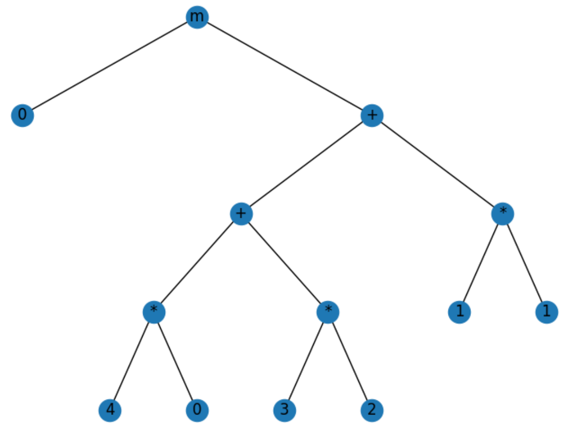

Figure 17: Convolutional neural networks: $\max(0, (p * q + r * s) + (t * u)) = \max(0, (4 * 0 + 3 * 2) + (1 * 2))$

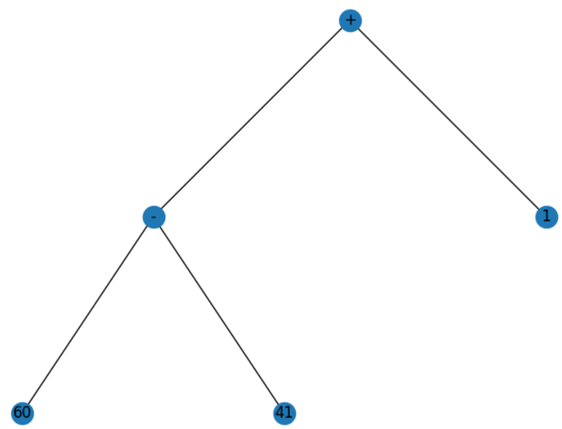

Figure 18: Convolutional neural networks: $(a - b) + c = (60 - 41) + 1$

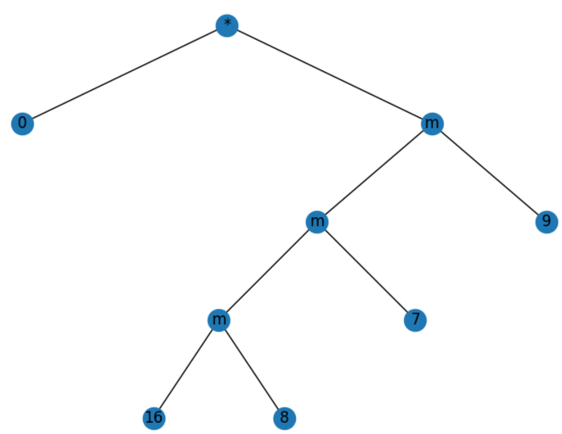

Figure 19: State machines and MDPs: $c * \max(\max(\max(m, n), o), p) = 0 * \max(\max(\max(16, 8), 7), 9)$

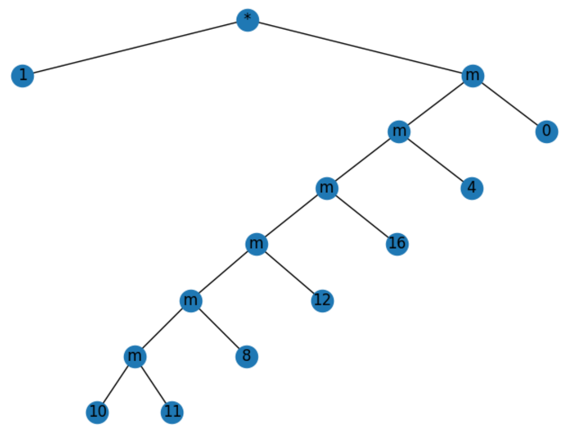

Figure 20: State machines and MDPs: $f \max(\max(\max(\max(\max(g, h), i), j), k), l), m) = 1 * \max(\max(\max(\max(\max(10, 11), 8), 12), 16), 4), 0)$

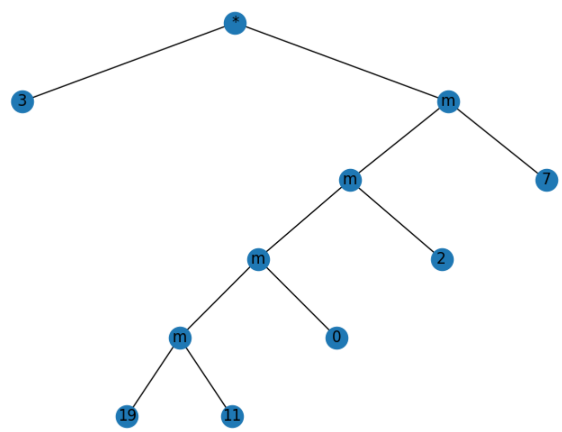

Figure 21: State machines and MDPs: $a \max(\max(\max(\max(p, q), r), s), t) = 3 * \max(\max(\max(\max(19, 11), 0), 2), 7)$

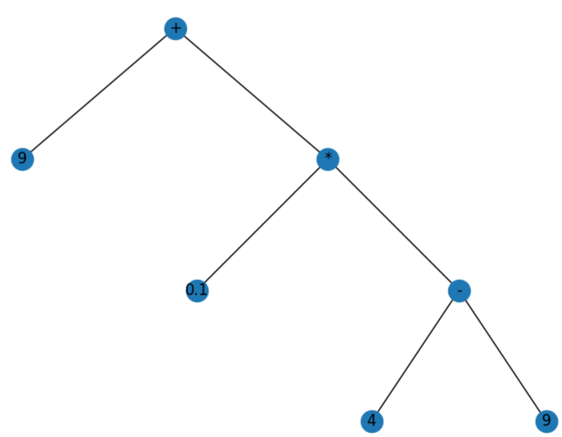

Figure 22: Reinforcement learning: $a + b * (c - d) = 9 + 0.1 * (4 - 5)$

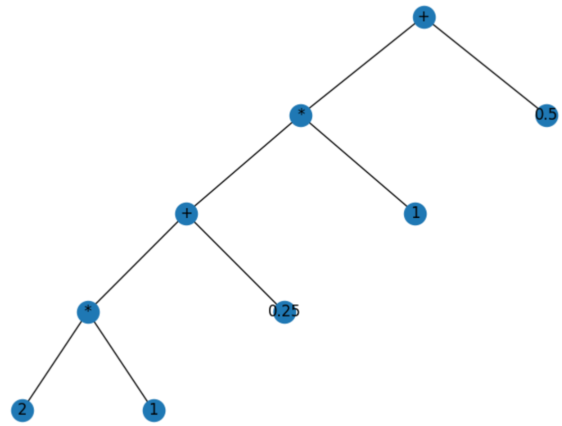

Figure 23: Recurrent neural networks: $((a*b+c)*d+e)*f = ((2*1+0.25)*1+0.5)$

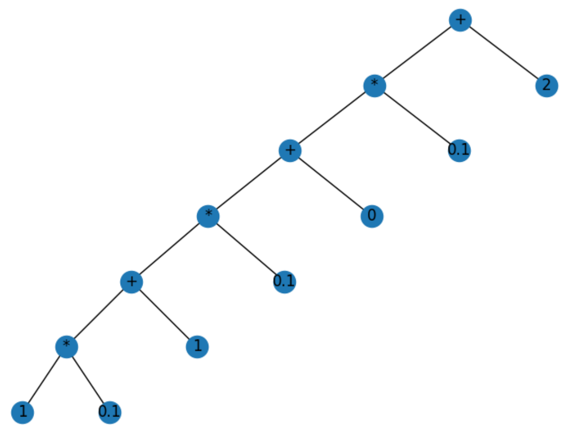

Figure 24: Recurrent neural networks: $(a*b+c)*d+e)*f+g = ((1*0.1+1)*0.1+0)*0.1+2$

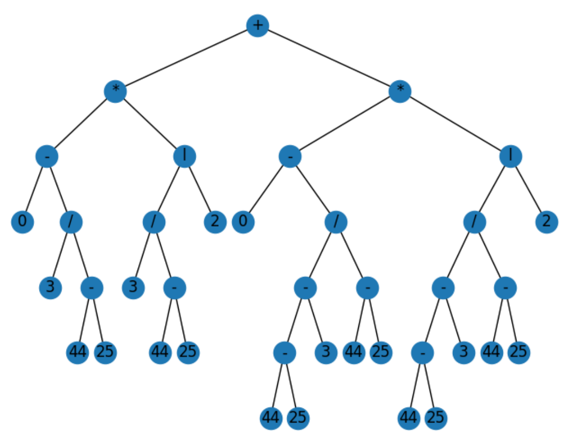

Figure 25: Decision trees: $(0-a/(b-c))*\log_2(a/(b-c))+(0-(b-c-a)/(b-c))*\log_2((b-c-a)/(b-c)) = (0-3/(44-25)*\log_2(3/(44-25))+(0-(44-25-3)/(44-25)*\log_2((44-25-3)/(44-25))$

Appendix B. Multiple Choice Questions and Answers by our Model

Topic	Question/Multiple Choice Options	Our Model's Answers
Basics	**Question:** What is the magnitude of the vector $[1, 6, 4]$? **Answer Options:** $[8, 4.12, 2.38, 1.41]$	[1.41]
Perceptrons	**Question:** Do the two classifiers $[0, 1, 0]$ and $[2, 2, 0]$ represent the same hyperplane? Return 1 if true and anything else otherwise. **Answer Options:** $[230.0, -40.0, 23.2, 20.0]$	$[-40, 23.2, 20]$
Features	**Question:** Consider the point $(0, 2)$, the $\theta = 2$ and the $\theta_0 = 1$. What is the NLL loss? Use natural log, where the base is 2.71828. **Answer Options:** $[-0.02, 0.69, 2.06, 1.06]$	[0.69]
Logistic regression	**Question:** Given a function $(2\theta - 2)^2$, calculate the value of the function after one gradient descent update if $\theta = 1$ and $\eta = 0.01$. **Answer Options:** $[\infty, 0.1, 0.41, 0]$	$[0.1, 0.41, 0]$
Regression	**Question:** Let 1 be the optimal θ by mean squared error. Given the datapoints $[(0, 0), (1, -1), (2, y)]$ and $\lambda = 1$, compute the value of y. **Answer Options:** $[-1.61, -0.28, -0.24, -0.83]$	$[-0.24]$
Neural networks I	**Question:** If we have a neural network layer with 20 inputs and 200 outputs, how many weights (including biases) are needed to describe each connection? **Answer Options:** $[32000, 80000, 64000, 8000]$	$[64000, 32000, 8000]$

Table 13: Examples of the questions given to the model, the answers generated by the evaluator, and the output answers made by the model in order of the appearance. Note that the final value in output is the correct answer of the question, and the answering stops once the models achieves the correct answer.

Topic	Question/Multiple Choice Options	Our Model's Answers
Neural networks II	**Question:** Neuron C is the output neuron which applies a ReLU on its output and neuron A is the input neuron to a neural network. Compute the output of a neural network with the given architecture and inputs. Neuron C takes in the offset value $oC = 1$ with weight $wOC = 3$. Neuron C takes in the output of neuron A with weight $wAC = 1$. Neuron A takes in the input value $x1 = -1$ with weight $w1 = 2$ and offset value $oA = 0.5$ and offset weight $wOC = 3$. **Answer Options:** $[3, 2.63, 0, 1.5]$	$[3, 2.63, 1.5]$
Convolutional neural networks	**Question:** Using a stride length of 2, what is the output from applying a filter of length 7 to an image of length 52? **Answer Options:** $[2.2, 25, 22.75, 23]$	$[23]$
Recurrent neural networks	**Question:** An RNN is defined as $s_t = w \times s_{t-1} + x_t$. If $s_0 = 2$, $w = 1$, and $x = [0.25, 0.5]$, what is s_2? **Answer Options:** $[2.75, 0.95, 1.75, 2.63]$	$[1.75, 0.95, 2.63, 2.75]$
State machines and MDPs	**Question:** A state machine is defined by the equations $s_t = f(s_{t-1}, x_t)$ and $y_t = g(s_t)$. Given the conditions $s_0 = 16$, $f(s_{t-1}, x_t) = \max(s_{t-1}, x_t)$, and $g(s_t) = 0 * s_t$, compute y_3 if the input is $x_t = [8, 7, 9]$. **Answer Options:** $[9.6, 0, 12.8, 40]$	$[40, 0]$
Reinforcement learning	**Question:** Let $q = 9$. After Q learning, what is q if $a = 0.1$ and $t = 4$? **Answer Options:** $[3.1, 8.7, 8.5, 8.95]$	$[8.7, 8.95, 3.1, 8.5]$
Decision trees	**Question:** What is the entropy of the left side of a region containing 27 points where the plane has 45 points in total and 4 points on the left are positive? **Answer Options:** $[0.61, 0.299, -0.52, 0.297]$	$[-0.52, 0.61]$

Table 14: Examples of the questions given to the model, the answers generated by the evaluator, and the output answers made by the model in order of the appearance. Note that the final value in output is the correct answer of the question, and the answering stops once the models achieves the correct answer. Thus a single numbers was correct on the first try and longer lists required more attempts.

Appendix C. Example Questions with Incorrect Answers

Incorrect Questions	Answer	Solution
1. Given the values for $\theta = 2$ and $\theta_0 = 1$, compute the NLL loss on the data point $(-2, 0)$. Use log base e of 2.71828 for the log.	-0.31	-0.05
2. Consider the point $(2, 2)$, the $\theta = 2$ and the $\theta_0 = 1$. What is the NLL loss? Use natural log, where the base is 2.71828.	3.98	4.99
3. Given a loss function, $(-2\theta + 3)^4$, for gradient descent, compute the updated θ value after one gradient descent step. Let $\theta = 2$ and $\eta = 0.05$.	2.4	0.4
4. The optimal θ value computed by mean squared error is 1 using the datapoints $[(0, -1), (1, -2), (2, y)]$. If $\lambda = 0.5$, what is y?	-0.345	2.75
5. Compute the value returned from aligning the filter $[1, 4, 1]$ to the image $[1, 4, 1]$ on top of one another.	21	18
6. What is the length of the result from applying F to I if F has length 17 and I has length 90?	34	74

Table 15: Example questions our model incorrectly answers, its answer, and the solution to the question.

Appendix D. Data Augmentation

Description	Text
Question	Let $s_0 = 0$, $w = 1.5$, and $x = (2, 2, 0)$. Compute s_3 if $s_t = w * s_{t-1} + x_t$.
Template	Let $s_0 = \{a\}$, $w = \{b\}$, and $x = \{c\}$. Compute $s_{\|\{c\}\|}$ if $s_t = w * s_{t-1} + x_t$.
Paraphrased Template	An RNN is defined as $s_t = w * s_{t-1} + x_t$. If s_0 is $\{a\}$, w is $\{b\}$, and x is $\{c\}$, what is $s_{\|\{c\}\|}$?
Paraphrased Question	An RNN is defined as $s_t = w * s_{t-1} + x_t$. If s_0 is 0, w is 1.5, and x is $(2, 2, 0)$, what is s_3?
Augmentation	An RNN is defined as $s_t = w * s_{t-1} + x_t$. If s_0 is 1, w is 0.5, and x is $(0.25, 0.25)$, what is s_2?
Expression	$s_3 = w * (w * (w * s_0 + x_1) + x_2) + x_3$
Values	$s_3 = 1.5 * (1.5 * (1.5 * 0 + 2) + 2) + 0$
Answer	7.5

Table 16: An example of converting an original question into a new one containing different values and phrasing. Note that $|\cdot|$ represents taking the length of the input.

Appendix E. Transformer Hyperparameters

Hyperparameters	T5 Transformer
Learning rate	1e-4
Batch size	32
Epochs	25
Number of embeddings	100
Number of hidden layers	512
Number of layers	3
Number of heads	8

Table 17: Hyperparameters of the T5 Transformer.

Appendix F. Reproducibility

In the spirit of reproducible research, we make our code available[1], and our models and data available[2].

1. https://github.com/stran123/solving-mlp
2. https://osf.io/eryg7/?view_only=edec7fa83ee74b1eb5cffe25555c4d89