CSC 412 Machine Learning and Knowledge Discovery

Exam II

Name: _____

Total: 32 points

Part I. Training Deep Neural Networks

1. (3 points) **Fancy Optimizers**

When we are using some fancy optimizers:

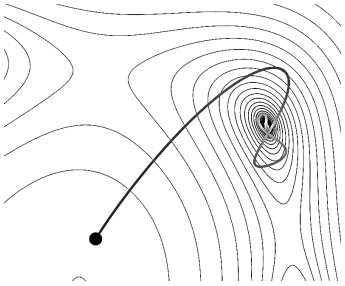
$$\boldsymbol{m}^{(\text{next})} \leftarrow \beta \cdot \boldsymbol{m} + (1 - \beta) \cdot \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$
$$\boldsymbol{\theta}^{(\text{next})} \leftarrow \boldsymbol{\theta} - \eta \cdot \boldsymbol{m}^{(\text{next})}$$

, and we see this (figure below):

(a) What is the name of this optimizer?

(b) Should you increase or decrease the β to make it descend better?

(c) If you want to use vanilla or original Gradient Descent, which value should you set your β to?



2. (2 points) **Adam**

(a) Adam = [ ] + [ ]

(b) Adam + Nesterov = [ ]

(c) RMSProp's idea? [ ]

3. (5 points) **What are they talking about?**

   (a) "... Consider a network with two hidden units, and assume we set all the biases to 0 and the weights with some constant w. If we forward propagate an input (x_1, x_2), the output of both hidden units will be $\sigma(wx_1 + wx_2)$. Thus, both hidden units will have an identical influence on the cost, leading to identical gradients. ... "

   (b) " ... PawsTech used an existing model originally trained to classify cat species and adapt it for dog breed classification. They quickly achieved impressive results. The adapted model was able to accurately identify dog breeds, from Poodles to Dalmatians, with minimal retraining effort, demonstrating the power of ..."

   (c) " ... This unfortunate behavior was empirically observed long ago, and it was one of the reasons deep neural networks were mostly abandoned in the early 2000s. It wasn't clear what caused the gradients to be so unstable when training a DNN ... "

   (d) "... Would a company perform better if its employees were told to toss a coin every morning to decide whether or not to go to work? Well, who knows; perhaps it would! The company would be forced to adapt its organization; it could not rely on any single person to work the coffee machine or perform any other critical tasks, so this expertise would have to be spread across several people. ... "

   (e) "... The technique consists of adding an operation in the model just before or after the activation function of each hidden layer. This operation simply zero-centers and normalizes each input, then scales and shifts the result using two new parameter vectors per layer ..."

Part II. Convolutional Neural Networks

4. (7 points) The table below presents a CNN. Answer the questions below:

Layer	Type	Maps	Size	Kernel size	Stride	Activation
Out	Fully connected	–	10	–	–	RBF
F6	Fully connected	–	84	–	–	tanh
C5	Convolution	120	1×1	5×5	1	tanh
S4	Avg pooling	16	5×5	2×2	2	tanh
C3	Convolution	16	10×10	5×5	1	tanh
S2	Avg pooling	6	14×14	2×2	2	tanh
C1	Convolution	6	28×28	5×5	1	tanh
In	Input	1	32×32	–	–	–

(a) What is the input size n?

(b) What is the number of parameters of Layer **C1**?

(c) How many neurons in Layer **C3**?

(d) What is the number of parameters of Layer **C5**?

(e) What is the number of parameters of Layer **F6**?

(f) Which padding is used in **C1**? (Valid or Same)

(g) In the Layer **S2**, if we set Kernel size 4×4 and Stride 4. What is the new size of this layer?

5. (3 points) What are the values? (1) — (5)

1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

(a) Input Image

-1	-1	1
-1	1	-1
1	-1	-1

(b) Filter (Kernel)

-	-	(1)	(2)
-	-	(3)	(4)
-	-	-	-
-	-	-	-

(c) Convolutional Layer

-	(5)
-	-

(d) Average Pooling

Part III. Recurrent Neural Networks

6. (5 points) Here is the equation for a layer of recurrent neurons: (Let $\boldsymbol{W}_x$ be 64×128 and $m = 1024$)

$$\boldsymbol{Y}_{(t)} = \phi\big(\boldsymbol{X}_{(t)}\boldsymbol{W}_x + \boldsymbol{Y}_{(t-1)}\boldsymbol{W}_y + \boldsymbol{b}\big)$$

   (a) What is the shape of $\boldsymbol{W}_y$?

   (b) What is the shape of $\boldsymbol{b}$?

   (c) What is the shape of $\boldsymbol{X}_{(t)}$?

   (d) What is the shape of $\boldsymbol{Y}_{(t)}$?

   (e) This equation is for a batch. What is the similar equation for a single instance?

7. (2 points) In a layer of 3 recurrent neurons, let all weights be **2**, with no bias terms and activation functions. Here are a series of inputs. What are the outputs?

$$\boldsymbol{x} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} \cdots \rightarrow \text{Time}$$

8. (1 point) List the gates of an LSTM cell.

9. (2 points) What does each term represent in this equation?

$$\boldsymbol{c}_{(t)} \leftarrow \boldsymbol{f}_{(t)} \cdot \boldsymbol{c}_{(t-1)} + \boldsymbol{i}_{(t)} \cdot \boldsymbol{g}_{(t)}$$

10. (2 points) What is the story behind the figure below? And give an application of this model.