UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Exam:	**INF 5860 / INF 9860 – Machine learning for image analysis**
Date:	**Thursday June 8, 2019**
Exam hours:	**9.00-13.00 (4 hours)**
Number of pages:	**xx pages of sketches to a solution**
Enclosures:	**None**
Allowed aid:	**Calculator**

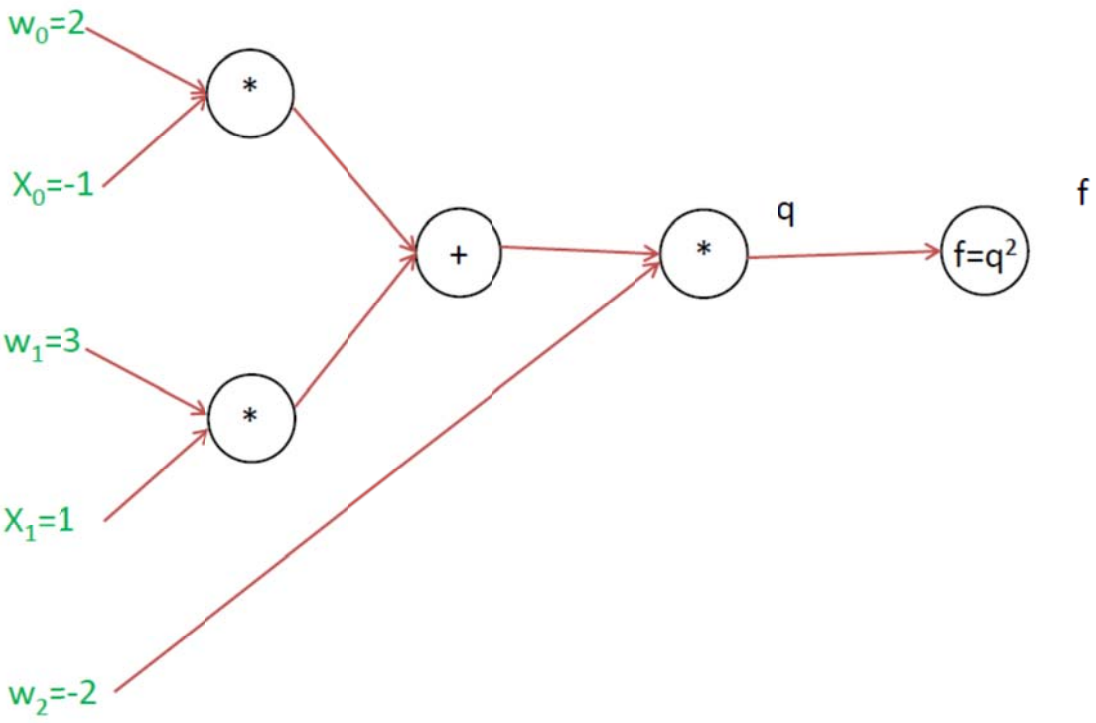
- Read the entire exercise text before you start solving the exercises. Please check that the exam paper is complete. If you lack information in the exam text or think that some information is missing, you may make your own assumptions, as long as they are not contradictory to the "spirit" of the exercise. In such a case, you should make it clear what assumptions you have made.

- You should spend your time in such a manner that you get to answer all exercises shortly. If you get stuck on one question, move on to the next question.

- Your answers should be **short**, typically a few sentences and / or a sketch should be sufficient.

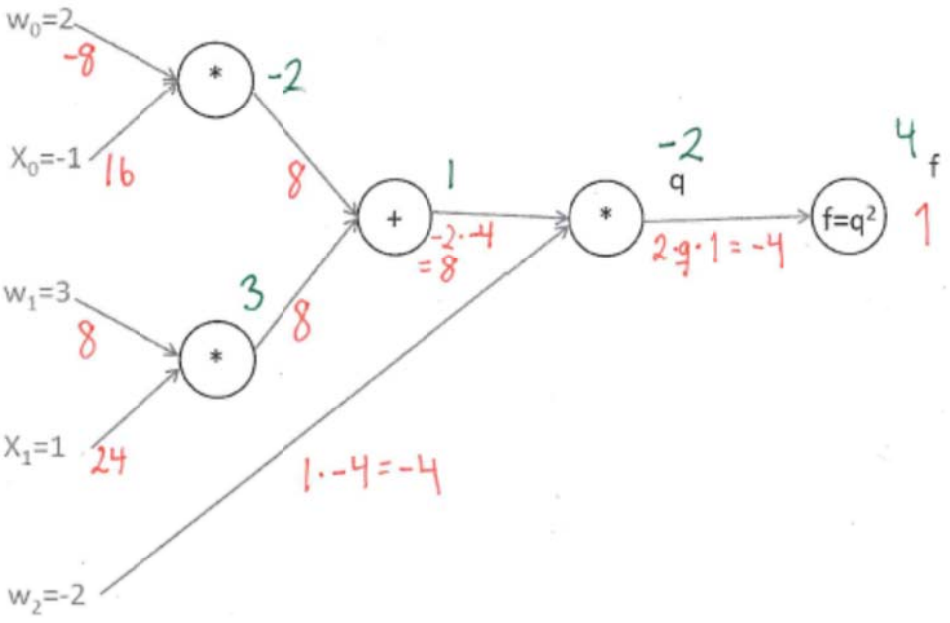
Good luck!!

Exercise 1: A simple network



a) Perform backpropagation on this example and compute one value for each arrow. You get 1/2 point per correct answer (**t**otal 4.5 points). Make a sketch of the net and draw you answers on the corresponding arrow.

Answer:



Exercise 2:

a) Assume that all weights in a net are initialized to the same random number. Discuss if this is a valid initialization

If all weights have the same value, they will learn the same thing, and this will not work.

b) Discuss briefly if the ReLU activation function has some shortcomings.

Because it is zero for all negative inputs, nodes can die. Networks should be monitored.

c) Explain briefly the main principle with dropout, and what is it used for.

Keywords: dropping random nodes in hidden layers to minimize overtraining and help generalization.

d) Given a cost function J(θ). Explain how gradient checking is performed for one parameter θ.

We use the implemented cost function to compute

$$\frac{J(\theta+\varepsilon)-J(\theta-\varepsilon)}{2\varepsilon}$$

and check that the difference between this and the dJ/dθ from backpropagation is small.

e) Explain how gradient descent with momentum updates works.

*v=mu*v - learning_rate*df # Integrate velocity*
f += v

f) Why should the range or standard deviations of weights in a neural network depend on the size of the input to a layer?
Each output is a weighted sum of the number of inputs. If the size of the weights do not depend on the size of the input, the scale of the output would depend on the number of inputs. Many inputs would mean growing output values, and few would mean vanishing outputs.

Exercise 3:

You have a convolutional neural network (CNN) with the following kernel sizes 3x3, 5x5 and 7x7. This is a 3-layer CNN with filters in that order. Consider the output activation map of the last layer. *By field-of-view we mean the number of pixels in each dimension influencing the output activation map.*

a) What will be the theoretical field-of-view for that network, if all layers use a stride of 1?

1 + (3-1) + (5-1) + (7-1) = 13

b) What will be the theoretical field-of-view be if the first layer has a stride of 2 and the following layers have a stride of 1?

1 + (3-1) + 2(5-1) + 2*(7-1) = 23*

c) The input image has 3 channels, and each layer has 2 filters. Excluding biases, how many parameters are in this model?
*3*3*3*2 + 2*5*5*2 + 2*7*7*2 = 350*

d) Dilated convolutions are a common way to increase the field-of-view of a convolutional network without reducing the spatial size. How can you use dilated convolutions to make the field of view as large as possible?

Use a growing dilation factor

e) You have raw audio data and want to detect different words. Why can a convolutional neural network be a better architecture for this application than a standard feed forward neural network?

Translation invariance/equivariance, since words don't start at same location etc.
Reuse sounds/part of words.

Exercise 4:

a) Explain the difference between residual networks and standard feed forward networks.

Residual networks add the outputs of the convolutions or multiplications to the previous input. This makes them model the difference from the identity function or "error" from the previous layer.

b) List two techniques for training deep neural networks, when you don't have much data.
Keywords: pretraining, adverserial domain adaption networks, multitask learning, regularization/augmentation (count as one)

Exercise 5:

When running *occlusion experiments* for visualization, you often get large responses for regions that does not belong to the target category. For example, the output for the category car, can also give a large output for regions of road and street signs.

a) What do we mean by occlusion experiments?
Running inference with different parts of the image occluded. Then mesuring the change in the softmax value for a given class. Then you can display the values at each occluded posistion as an image, indicating what part of the image that is important for a given class.
b) Why do we often get high responses for other regions, even though they look very different from the target object? Give some different examples.
Not all patches in the image that correlate with the object label is necessarily a part of the object.

Exercise 6:

a) You know the architecture and the weights of a given neural network. Discuss how could you construct images that you are almost certain would be classified differently by humans and the neural network. Both the humans and the neural network should also be very confident in their decision.

Describe adversarial fooling or another approach that could work.

b) t-SNE is optimized with gradient descent. What are you differentiating with respect to?

For t-SNE you are differentiating with respect to the points position in the output space, y.

Exercise for INF9860 only:

a) Give one example of data that deep learning will probably work well for, and one example of data where other machine learning techniques may work better. Briefly point out the differences between the two.
Work well for: raw image data, raw audio data, text
Work less well for: specific high-level features, that is not reasonable to combine with many other features

Deep learning typically work good for raw image data and audio data. It may on the other hand work badly on very different high level features. Deep learning relies on: Features that can be combined in simple way, into stronger features (hierarchical) Features (lower level) can be shared among different classes.
Example of bad data could be: Classification of customers, based on a set of features like: "is member of club", "male/female", "money spent", "average shopping category"

b) In reinforcement learning, what does the output of a Q-network represent? Why do you only need two steps to update a Q-network.
In Deep Q-learning we estimated the expected final reward, given an action.
We only need two steps, since we can use the mean estimated reward over all actions in the next step as ground truth (and backpropagate frome there...).

c) An example of *hard attention* could be that one part of the network select a sub-region of an image, another part of the network does classification with the sub-region as input.
Why do we often use reinforcement learning to optimize networks using hard attention?
The network cannot update weights for what it did not include in the crop.
Normal training can cause the network to converge to irrelevant regions.
Reinforcement learning sample with a probability and therefore will investigate all possible crops. So this will find a good solution/relevant crops.

Thank You for Your Attention!