# Artificial Neural Networks 2020 – 2021

# Theory Exam - May 25th 2021

**Name: ……………………………………………………………………**

## GUIDELINES

- Start by writing your name on each page and the additional pages
- This is a closed book exam.
- Make sure your mobile phone is switched off and place it at the front of the room.
- Be as complete as possible in your answers. Motivate clearly why you provide this answer. **Usually answers in 1 sentence are not enough!**
- All forms of communication are prohibited.

## QUESTION 1: PERCEPTRON LEARNING

A) Parameters of artificial neurons include weights applied to the input and a bias term. 1) Indicate what is the function of the bias term in an artificial neuron, and 2) Indicate what would be the effect of removing it from an artificial neuron trained with it.

## QUESTION 2: DEEP NEURAL NETWORKS

A) One of the defining characteristics of models following the “Deep Learning” paradigm is the relatively high number of internal layers that define the “deep” architectures that they posses. Indicate the advantages and disadvantages that having a deep architecture may bring.

B) Describe the training procedure of a deep neural network ***f*** trained for a classification task of ***C*** classes from a dataset **X = [ $\mathbf{x}_\mathbf{i}$, $\mathbf{y}_\mathbf{i}$ ]** where each example $\mathbf{x}_\mathrm{i}$ is paired with an annotation $\mathbf{y}_{\mathrm{i.}}$ Describe relevant factors related to the forward-pass, backward-pass, optimization and stopping criteria involved in the training procedure.

## QUESTION 3: LEARNING AND OPTIMIZATION

A) Drop-out is a technique that is commonly used for regularizing the weights of a neural network during training. 1) Indicate how this technique works and 2) indicate what are the advantages of applying it during training.

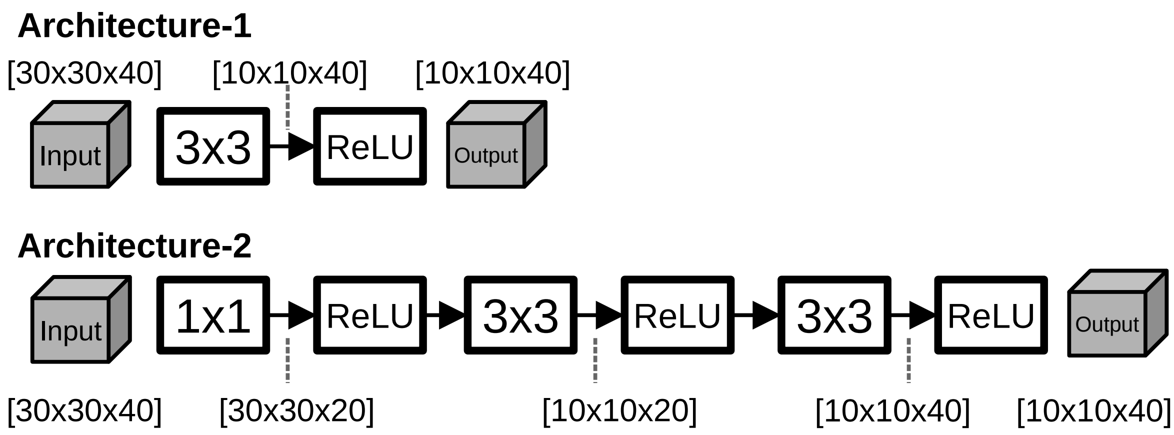
B) Following the stocastic gradient descent algorithms, parameters of a deep neural network are updated based on the equation $\theta_{t+1} = \theta_t - \alpha_t \nabla_\theta L(\theta_t)$ where $\alpha_t$ is the learning rate set during training. Early works applied a fixed learning rate while follow-up work adopted dynamic learning rates. 1) Indicate how learning rates are applied in a dynamic manner in practice. 2 Indicate what are the advantages, and potential disadvantages, of using a dynamic learning rate instead of a fixed.

## QUESTION 4: CONVOLUTIONAL NEURAL NETWORKS

A) Explain how the Squeeze and Excitation (SE) Module propagates information computed by a given filter to different spatial locations in other channels.

B) A company is developing an image recognition algorithm based on convolutional neural networks. They have produced two different architectures that are capable of producing the same internal representation from a given input, thus achieving the same classification performance. The diagram below illustrates the two models in question, each block represents a layer. Each convolutional layer indicates its kernel size within it. The size of the intermediate representation after different operations is indicated by the dashed lines.

- ○ Indicate what are the main differences between these two architectures, potential strengths and weaknesses that they may have and scenarios when it would be better using one over the other. Justify your answer appropriately.



**Architecture-1**

**Architecture-2**

## QUESTION 5: MODELING SEQUENCES WITH NEURAL NETWORKS

A) Indicate the main difference between Recurrent Neural Networks and Long Short - Term Memory Networks.

B) Data related to sequence modelling problems is provided as a set of valid sequences $\{x\}_i$. Describe the training procedure of Recurrent architectures for sequence modelling problems and indicate how they are optimized without the need of annotations.

## QUESTION 6: TRANSFER LEARNING

A) Describe the procedure of response-based knowledge distillation.

## QUESTION 7: DEEP GENERATIVE NETWORKS

A) Describe the training procedure of denoising autoencoders and indicate what is the added value of this procedure when compared to that of standard autoencoders.

B) Describe two differences between autoencoders and variational autoencoders.

## QUESTION 8: INTERPRETATION AND EXPLANATION

A) Describe two differences between input modification methods and deconvolution-based methods for model explanation.

# QUESTION1:

A)

1) Function of the Bias Term in an Artificial Neuron:
The bias term in an artificial neuron allows the model to adjust the threshold at which the neuron activates independently of the input values.It shifts the activation function giving the model more flexibility.

2) Effect of Removing the Bias Term from an Artificial Neuron:
Without the bias term, the threshold needs to be explicitly set and cannot be adjusted during the training process. This limits the neuron's ability to learn the optimal decision boundary. The removal of the bias termconstrains the types of functions the neuron can model.

# QUESTION2:

A)

+ deep neural networks can learn hierarchical representations. The first layer captures general features and by combining multiple layers the net can capture more specific features In the data.

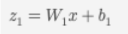
- A problem here can be that the network is too deep so that it overfits on the training data. This is because it captures features that are too specific.

- Deep neural networks are also susceptible for the vanishing gradient problem.

B)

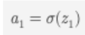
Given the network f and input xi. We first talk about the forward pass. We put the input xi through the layers in the netword and predict an output. This goes like this:

1) input x in the first linear layer:



$z_1 = W_1 x + b_1$

2) this output passed to an activation function lets say sigmoid:



$a_1 = \sigma(z_1)$

=> this repeats untill we have an output. With this output we can compute the loss L(y, ŷ). This is the loss between the preditcted lable and the ground truth. Given the loss we are going to back propagate to adjust the weights to make future predictions better.

back propagation with gradient decent:
-> We calculate the gradient of the loss given the prediction
-> Use the chain rule to propagate the gradient backwards through the each layer. Calculating the gradient off the loss with respect to the weights and biases in each layer.

stopping criteria:
-> if the Loss is under a certain threshould than we can stop and say that the model is trianed well.
-> We can aslo look at the last x epochs and if the standart deviation of the loss os low that means that the model is not really improving and we can also stop the training.

# QUESTION3:

A)

1) Dropout is a technique where each iteration we are going to "drop out" neurons with a certain probability. This dropping out is essentially setting the activation to 0 creating a "death node".

2) The advantage of doing this is overfitting. Let's look at the example in the lecture with the beaches. Some neuron might be very focussed on sand and the network might predict an image of the desert as "beach". When we dropout this neuron the network is forces to focus on different features. It also promotes enseble lreaning.

B)

1) if you have a fixed lr, then it might overshoot, which means that because of the large steps it has it will cross the optimal point, continue and lose it forever or just take wayy longer to get there. In a dynamic manner in practice the learning rate starts high and is dicreased as training progresses.

2) Dynamic learning rates can help the model converge faster by adapting the step size during training, but Implementing dynamic learning rate schedules or adaptive methods adds complexity to the training process.

# QUESTION4:

A)

fitst we have the squize function that is going to reduce the space of the entire “tensor” form H x W x C, here (C is the amoun of channels) to a 1 x 1 x C value reducing or squizin the whole channel into 1 value. Then the excitation function is going to “rank” the ouput vector of the squize function. This tells us which channels are more informative than others.

B)

Architecture 1:

Strengths:
- Simplicity: Fewer layers mean a simpler and more straightforward implementation.
- Faster Computation: With fewer layers, the forward pass and backpropagation are faster.
- Lower Memory Consumption: Fewer parameters and intermediate feature maps reduce memory usage.

Weaknesses:
- Flexibility: Limited ability to learn hierarchical features due to having only one convolutional layer.
- Depth: Lack of depth might limit the ability to capture more complex patterns in the data.

Architecture 2:

Strengths:
- Hierarchical Feature Learning: Multiple layers allow for learning more complex and hierarchical features.
- Dimensionality Reduction: The 1 x 1 convolution reduces the number of channels, reducing computational load while retaining important information.
- Rich Representations: More layers can extract more abstract features, potentially improving performance on complex tasks.

Weaknesses:
- Complexity: More layers mean a more complex implementation and longer training times.
- Computational Cost: More layers increase the computational requirements and memory usage.
- Overfitting Risk: More parameters can lead to overfitting, especially with limited data.

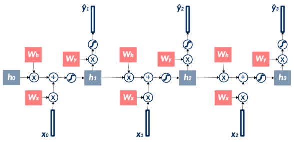
Architecture 1 is suitable for scenarios where simplicity, speed, and resource efficiency are crucial. Its straightforward design makes it ideal for simpler tasks or environments with limited computational resources.

Architecture 2, with its deeper structure and ability to learn more complex features, is better suited for challenging tasks where the richness of feature representation significantly impacts performance. The inclusion of the 1 x 1 convolution helps manage computational load while allowing the network to learn more abstract and hierarchical features. This makes it suitable for scenarios where higher accuracy and more detailed feature extraction are required, and computational resources are less of a concern.

# QUESTION5:

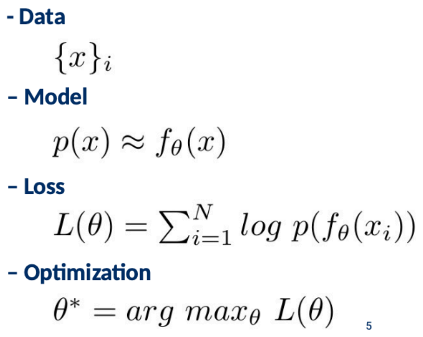
A)

in reccurent neural newerks we have the presistent state h that captures the information about the context. This persistent state h , also known as the hidden state, is designed to retain information across different time steps in a sequence, enabling the model to learn and remember temporal dependencies. the model looks like this:



The problem here is that it suffers from the vanishing gradient problem. It has only short term memory. The Long Short - Term Memory Networks fixes this with an other state. the cell state c. it has 3 gates. the forget gate, the input gate and the output gate. The cell state is the memory of the model and the gates tell the memory what to remember and forget. This is more robust for longer sequences.

B)



**- Data**

$$\{x\}_i$$

**– Model**

$$p(x) \approx f_\theta(x)$$

**– Loss**

$$L(\theta) = \sum_{i=1}^{N} log\ p(f_\theta(x_i))$$

**– Optimization**

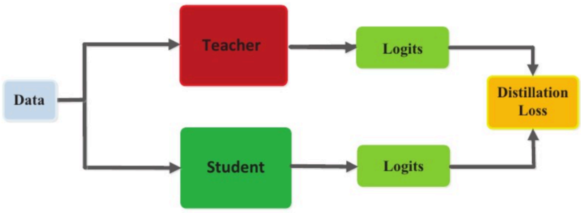
$$\theta^* = arg\ max_\theta\ L(\theta)$$

5

# QUESTION6:

A)

for response-based knowledge distillation we essentially want to train a smaller model from the most important parts of a bigger already trained model. We do this with the student- teacher architecture. For the respose based model we look at the logits. This is the layer before the fully connected one because it encodes more info than just a prediction. We are going to train the model based on the distillation loss. here is a diagram of how this works.



# QUESTION7:

A)

By training denoising auto-encoders we first add noice tot a given data sample x and we will call this sample x'. We train this by calculating the loss based on the prediction (generated image) made from the noizy input data and compare it to the ground truth. L(x, g(f(x'))). The interesting part here is that the model learns to denoice the image and by learning what noice is it also has to learn what is not noice.

B)

Normal autoencoders try to capture all the important characteristics in the laten space h. If we project this latent space we can see that there are "holes" in this space because there are images that are not defined. When we let's say want to generate images that are like the input but not an exact copy we can use variational autoencoders. This architectrures latens space consists of 2 vectors σ am μ. An other difference is the training of the models. The Normal autoencoders try to optimise the features in the latent space h. by calculating the loss $L(y, \hat{y}) + \Omega(h)$ = which is the sparecty penalty. this enshures that only a few neurons form h are active so we don't overfit the model. to train variational autoencoders we are optimising σ am μ. (We need to add a ε because the sampeler is not differentiable. )

# QUESTION8:

A)

the input modification methode masks a part of the input and than fives it to the model. Based in the prediction we know that that part is valuable to the prediction or not. by doing this for multiple parts of the image we can make a heat map of the most important features this model is independent of the model used. This is different form deconvolution- based where we give an image to the model and than reverse the operation leading to that prediction to get a heatmap of the features. So the deconvolution- based methode requires 2 passes while the input modification methode only requires one pass.

(Depending on how meany passed you do in the input modification methode) you get a more detailed feature map than in the input modification methode. The input modification methode is also more computationaly expensive.