Examen – Advanced Machine Learning (M1)
2024

Calculator is allowed but lectures documents are not authorized.

Part I – General questions: true or false (/7 points)
The following statements are either true or false, for each one specify its truth value and argument your answer.

1. The *sigmoid* activation function is used in the final layer of a neural network in the cases of multiclass classification.

2. *Overfitting* is a phenomenon that arises when the training loss start to increase after some training epochs.

3. A *generative adversarial network* is trained to generate data from unlabeled examples.

4. For tasks dealing with sequences, the *convolutional neural network* is the most suited architecture.

5. Backpropagation is the method by which the loss is calculated with respect to the model's parameters.

6. The transformer architecture was initially introduced for natural language processing tasks, and it is not suitable for computer vision applications.

7. Both vanilla RNN and Transformer architectures are neural networks that iterate over a sequence (sentence) while keeping an internal memory (hidden state) after each iteration (word).

Part II – Transformers (/5 points)

1. Suppose we have the following sentences, and we would like to use them to train a Transformer Encoder to classify the emotional context.

- "The cat is sleeping."
- "A dog wagged its tail."
- "The cat plays in front of the dog."

Given the following indices:

1: "the", 2: "cat", 3: "is", 4: "sleeping",

5: "a", 6: "dog", 7: "wagged", 8: "its",

9: "tail", 10: "plays", 11: "in",

12: "front", 13: "of"

- Prepare these sentences to be suitable for training the transformer. (1.5)

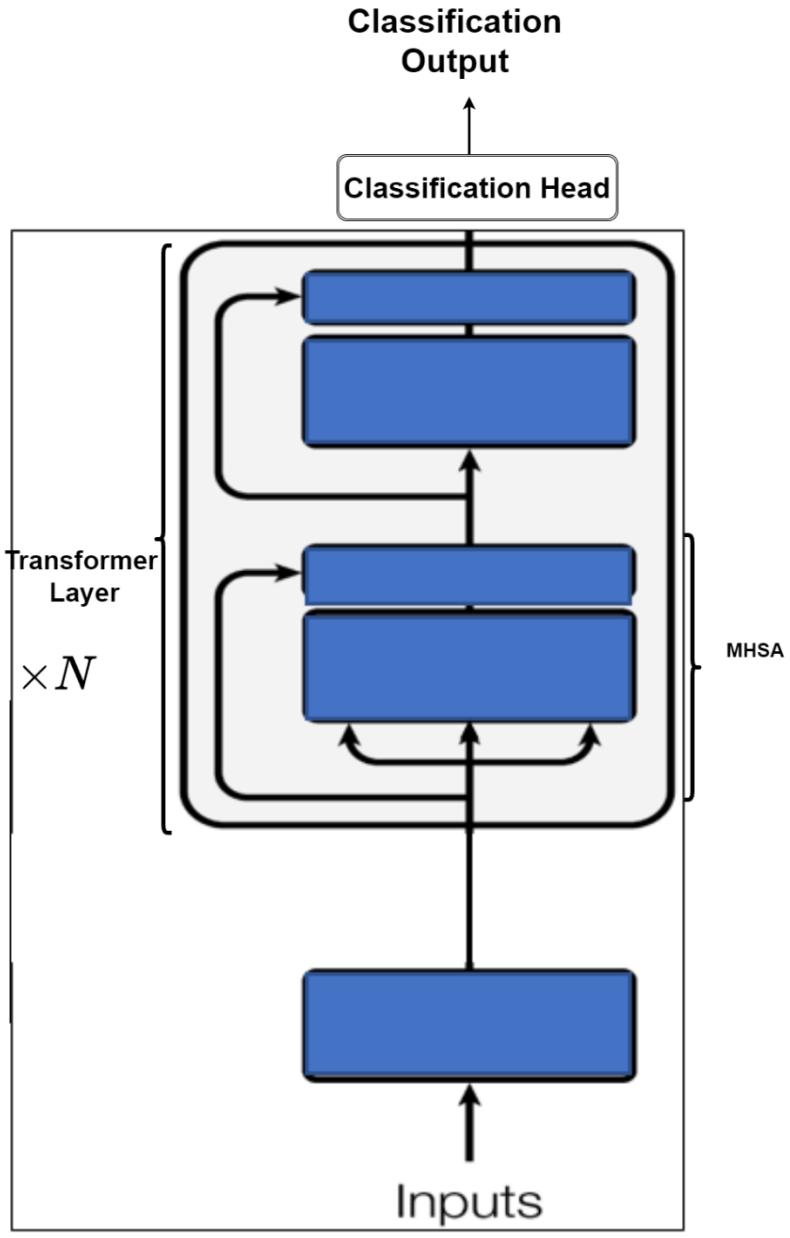
2. In order to use the above dataset to train a model, the following transformer is used with the following parameters (it should be noted that the bias is not used in this transformer):

- The number of heads of multi-head self-attention, h = 8.
- The number of layers of multi-head self-attention, l = 12.
- The embedding dimension, d = 768.
- The dimensionality of the feed-forwards network is d*4.
- The number of classes = 3.

a. Give the elements of the Multi-head Self-Attention Block with the needed number of parameters for each, then inference the total number of parameters of MHSA block. (1.5)

b. Calculate the total number of parameters for one single transformer layer. (1)

c. Calculate the total number of parameters for the full transformer encoder from inputting the data to the decision (to not be surprised or confused, the total number of parameters is more than 283 million). (1)

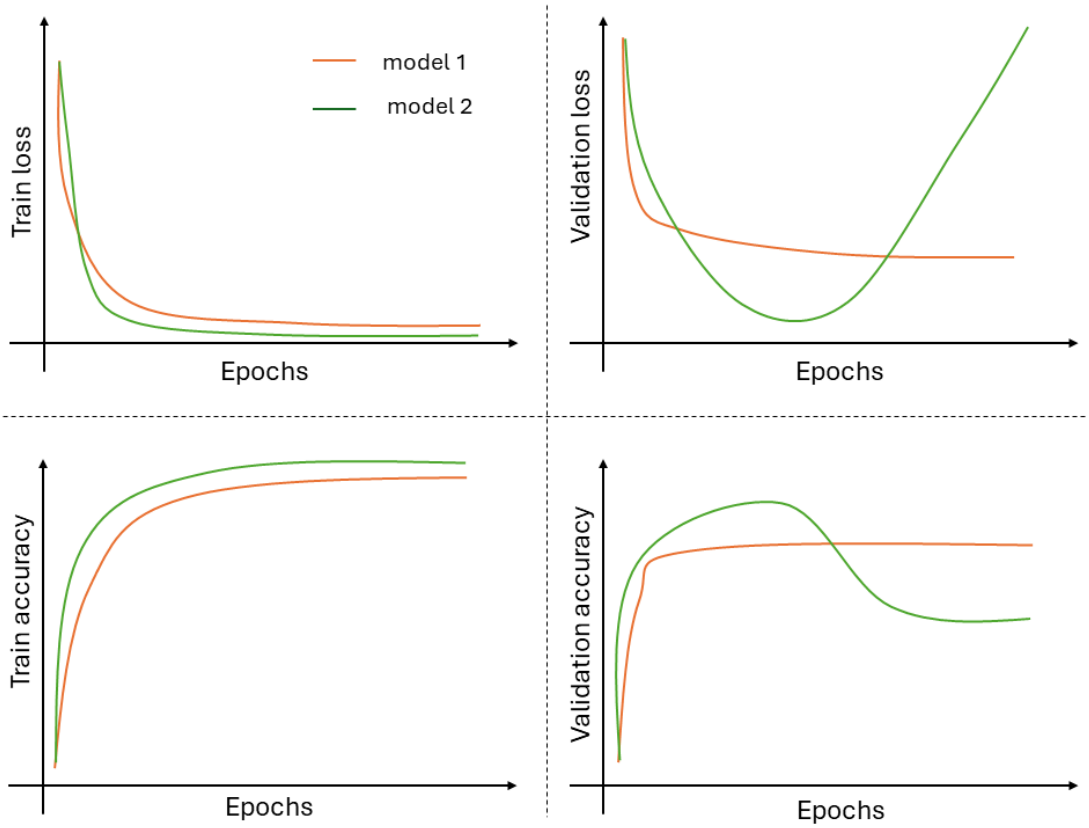

Classification Output
Classification Head
Transformer Layer
$\times N$
MHSA
Inputs

Part III – Analysis (/8 points)

1. The following code creates a CNN for a multiclass classification problem on images. However, it contains two errors. Gives the line number of each error and explain why it is an error.

```
1  import tensorflow as tf
2  from tensorflow.keras import layers
3
4  model = tf.keras.Sequential()
5
6  model.add(layers.Conv2D(64, 3, input_shape=(256, 256, 3)))
7  model.add(layers.BatchNormalization())
8  model.add(layers.Activation("relu"))
9
10 model.add(layers.MaxPool2D())
11 model.add(layers.Dropout(0.5))
12
13 model.add(layers.Conv2D(64, 3))
14 model.add(layers.BatchNormalization())
15 model.add(layers.Activation("relu"))
16
17 model.add(layers.MaxPool2D())
18 model.add(layers.Dropout(0.5))
19
20 model.add(layers.Dense(128))
21 model.add(layers.Flatten())
22 model.add(layers.BatchNormalization())
23
24 model.add(layers.Dense(3, activation="sigmoid"))
```

2. The image below depicts the learning curves of two models with loss and accuracy on the training and validation splits during the training. A model checkpoint callback is added to the training, so that the models parameters are saved each time the validation loss reach a new minimum value. Based on the curves below, which model would you select, argue your answer.



3. Since the apparition of the CNN architecture, it has been widely used as the default architecture for neural networks applied on images. Provide a well-argumented explanation of why CNNs are better suited for images compared to simple MLPs.