UNIVERSITÄT DES SAARLANDES

Saarland University, Department of Computer Science
Neural Network Assignment 10

Deborah Dormah Kanubala (7025906) , Irem Begüm Gündüz (7026821), Anh Tuan Tran (7015463)

July 19, 2023

Exercise 10.1 Recurrent Neural Network

Exercise 10.1.a

Both are specialized kinds of NN for processing different types of data. CNNs are useful for processing data with grid-like topology. Whereas RNN is useful for processing sequential values and can scale to longer sequences. CNN's operations allow the network to share parameters across time but are shallow. The same parameter can be used for more than one function in a model. On the other hand, RNNs share parameters differently. Each output is produced using the same update rule applied to the previous output, hence predicted outputs depend not only on the current input sequence but also on past information. [1]

Exercise 10.1.b

The mathematical difficulty in training RNN is the issue of Vanishing Gradient and Exploding gradients problems, which hinders the learning of long data sequences. This occurs when the gradients calculated become very small or very large when back propagated through time. LSTM solves this problem by introducing three gates, thus: the forget gate, the input gate, and the output gate. To forget, gates handle the information that needs to be forgotten, given the new information that enters the network.

Exercise 10.2 Transformers

Exercise 10.2.a

They use self-attention to weigh the importance of different tokens in the input sequence, and a technique called positional encoding to provide the relative positional information of the tokens. This allows the model to focus on different parts of the input sequence depending on the task and take into account the order of the tokens. The attention mechanism is deterministic as it is based on a fixed set of parameters. [2]

Exercise 10.2.b

It is not necessary to use trainable positional encodings, the encoding used in transformers is typically fixed. Some variants of the transformer architecture use trainable positional encodings, however, the fixed positional encoding is commonly used and performs well in practice as it provides more computational efficiency than using trainable encodings.

Exercise 10.2.c

Single-head attention is a mechanism where the model focuses on a single representation of the input tokens when computing the attention scores, while a model with multi-head attention focuses on multiple representations of the input tokens by applying multiple linear transformations. Multi-head attention performs better because it allows the model to attend to different parts of the input sequence in parallel.

The authors of the paper propose a "selective attention" mechanism, which allows the model to selectively focus on different parts of the input sequence depending on the task. Selective attention allows the model to focus on the most important parts of the input sequence for a specific task, while multi-head attention attends to all parts of the input sequence regardless of their importance. Therefore, it outperforms multi-head attention-based models.

References

[1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. book in preparation for mit press. *URL¡ http://www. deeplearningbook. org*, 1, 2016.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.