# LSTM Ausarbeitung

## 1. Describe an application of **convolutional LSTMs**

One application of conv LSTM is video sequence extrapolation, so next frame prediction.
A video is a sequence of images over time.

## 2. Describe **LSTM dropout**. Where is the **dropout mask**? How are the elements of the dropout mask computed?

LSTMs need a specialized version of dropout, because the probability for a unit to survive is (1 - p)^T (decays exponentially with sequence length T). The network cannot learn because the internal memory gets cleared.

**Special dropout techniques for LSTM**

- Apply dropout mask to input connections ( x(t) ) only
- *Zoneout* - drop the updates to the memory cell

**Elements of dropout mask**

- d(t) is a random vector of length I (hidden) whose entries are all (1 - p)-Bernoulli distributed.

## 3. Describe **Schmidhuber's approach** for learning fully recurrent neural networks.

1. The idea is to divide the sequence into blocks of length I (hidden).
2. Within each block BPTT is performed.
3. Between the blocks RTRL is performed to collect all gradient infromation

Complexity of O(I^3), but only activations of sequences length I have to be stored.

## 4. Describe the Bahdanau (2014) **Additive Attention Model**.

## 5. Describe the different variants of **sequence-to-sequence learning**?

**2 different variants**

One-to-one correspondence between input and output sequence.
Input sequence and output sequence have the same length and are semantically aligned

Utilize a second LSTM network and split the translation task into two phases: *encoding* and *decoding.*

1. The encoder reads the input and learns a vector representation of the sequence.

2. This vector representation is copied to the decoder which generates a sequence in the target space.

---

## 6. Describe when **tanh** and when **sigmoid** should be used for g.

Sigmoid: For detecting patterns in sequences
Tanh: for detecting hints for or against a situation (language tasks)

---

## 7. Do RNNs with **exploding gradients** recognize the end of a sequence or the beginning of a sequence?

RNNs with exploding gradient only recognize the beginning of a sequence

---

## 8. Do RNNs with **vanishing gradients** recognize the end of a sequence or the beginning of a sequence?

RNNs with vanishing gradient only recognize the end of a sequence

---

## 9. Explain the two self-supervised learning tasks of BERT for generating good representations.

1. Mask LM task: goal is to insert missing words into a sequence
2. Next sentence prediction (NSP): goal is to understand the relation between two sentences. Task is to decide whether a sentence is successor of another sentence or not.

---

## 10. Give an application of the Luong (2015) **Multiplicative Attention Model**.

Multiplicative attention is used in Transformers in the form of dot-product attention.
Language translation

---

## 11. How is the **attention score** of the transformer computed?

The attention score of a transformer is the product of the Key and query vector, normalized with the square root of the key-dimension dk.

---

## 12. Name 3 main application domains of LSTM.

Protein classification
Language recognition
Image caption generation

---

## 13. Name 4 characteristics where RNNs differ from feedforward neural networks.

---

## 14. Name 5 companies who are using or used LSTM.

Apple
Google

Facebook
Amazon

## 15. What are the three term/values the input to **transformer attention module** is mapped to?

Query, Key and Value

## 16. What does the **Multi-Head** with the context values of the transformer

The Multi-Head concatenates the context matrices of multiple layers (heads) and multiplies them with a matrix W

## 17. What is a **PyraMiD LSTM**? Why has it been introduced compared to **multidimensional LSTM**? Describe how an input is processed.

## 18. What is **BERT**? How does it improve the original transformer?

BERT is a pretrained language representation model based on the Transformer architecture. Pre-training is done in a self-supervised fashion and only has to be done once.

## 19. What is Multi-Head-Attention?

Multi-head attention is a number of parallel scaled dot-product attention units, which calculate the attention that a key / value pair gets, given a query (output)

## 20. What is „**neural word embedding**“?

Neural word embedding is based on a next-word-perdiciton task.

## 21. What means that RNNs are **Turing complete**?

## 22. Where in the transformer is the **encoder-decoder-attention module** and what does it do and provide?

The encoder-decoder-attention module is in the decoder part of the transformer. It helps the decoder focus on the appropriate places in the input sequence.

## 23. Why should LSTM be trained with **online learning**?

## 24. What are **ticker steps** and why are they useful?

## Focused LSTM

## Lightweight LSTM