Natural Language Modelling

PA154 Jazykové modelování (12)

Pavel Rychlý

pary@fi.muni.cz

May 18, 2021

Deep Learning

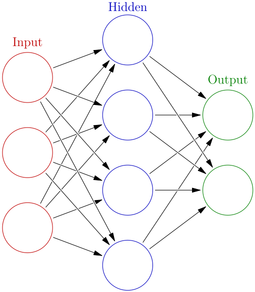
- deep neural networks
- many layers
- trained on big data
- using advanced hardware: GPU, TPU
- supervised, semi-supervised or unsupervised

Neural Networks

- Neuron: many inputs, weights, transfer function (threshold), one output:

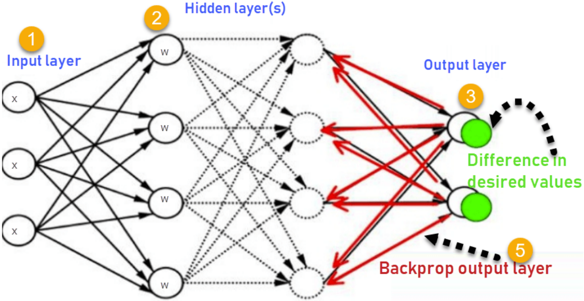
$$y_k = \phi(\sum_{j=0}^{m} w_{kj} x_j)$$

- Input/Hidden/Output layer
- One-hot representation of words/classes: `[ 0 0 0 1 0 0 0 0 ]`



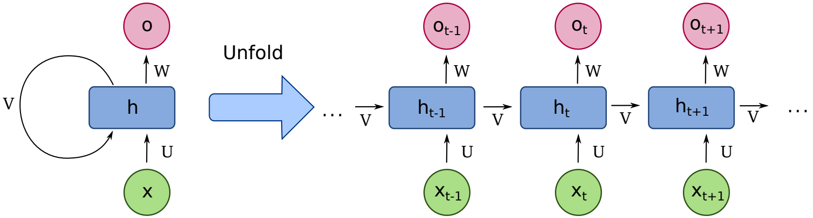
Training Neural Networks

- supervised training
- example: input + result
- difference between output and expected result
- adjusts weights according to a learning rule
- backpropagation (feedforward neural networks)
- gradient of the loss function, stochastic gradient descent (SGD)



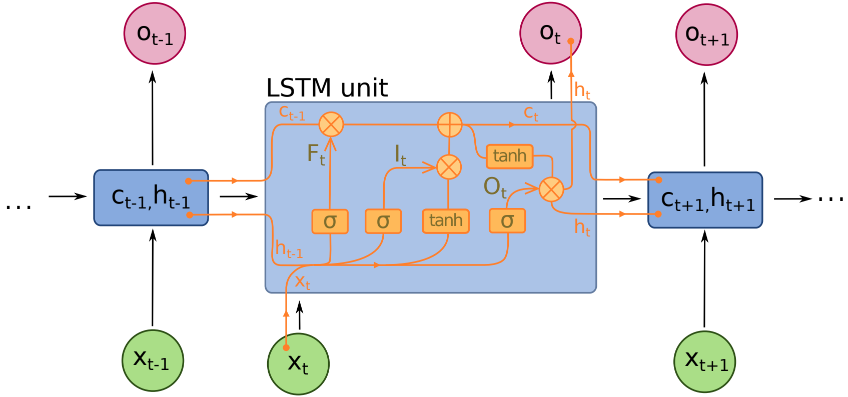
Recurrent Neural Network (RNN)

- dealing with long inputs
- feedforward NN + internal state (memory)
- finite impulse RNN: unroll to strictly feedforward NN
- infinite impulse RNN: directed cyclic graph
- additional storage managed by NN: gated state/memory
- backpropagation through time



Long short-term memory (LSTM)

- LSTM unit: cell, input gate, output gate and forget gate
- cell = memeory
- gates regulate the flow of information into and out of the cell

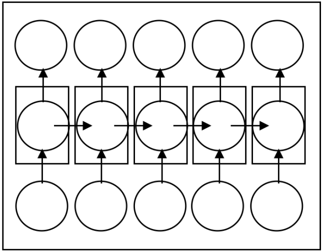


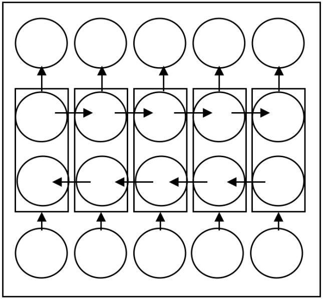
GRU, BRNN, ...

- Gated recurrent unit (GRU)
- fewer parameters than LSTM
- memory = output

- Bi-directional RNN
- two hidden layers of opposite directions to the same output

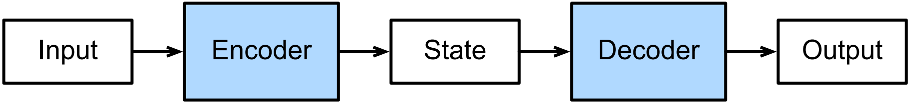
- hierarchical, multilayer





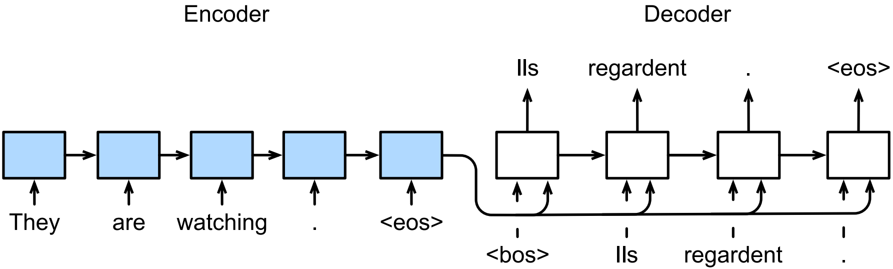
Encoder-Decoder

- variable input/output size, not 1-1 mapping
- two components
- Encoder: variable-length sequence → fixed size state
- Decoder: fixed size state → variable-length sequence



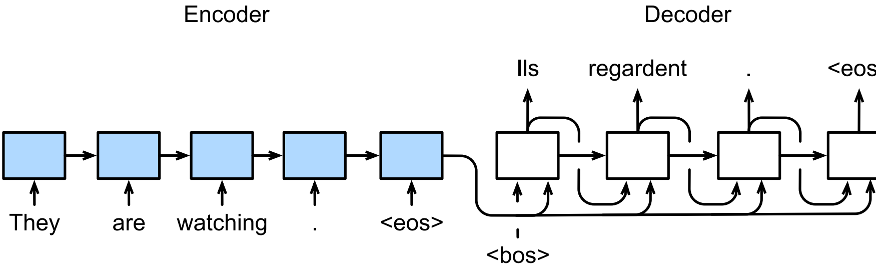
Sequence to Sequence

- Learning
  - Encoder: Input sequence → state
  - Decoder: state + output sequence → output sequence



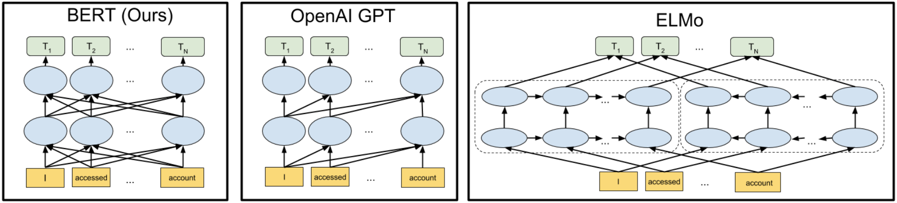
Sequence to Sequence

- Using
  - Encoder: Input sequence → state
  - Decoder: state + sentence delimiter → output



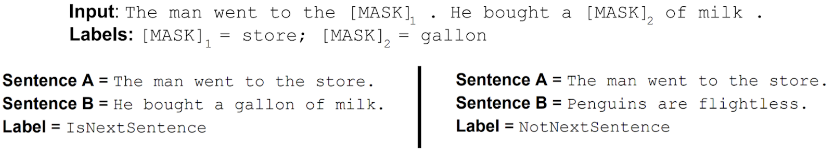
Transformers

- using context to compute token/sentence/document embedding
- BERT = Bidirectional Encoder Representations from Transformers
- GPT = Generative Pre-trained Transformer
- many varians: tokenization, attention, encoder/decoder connections



BERT

- Google
- pre-training on raw text
- masking tokens, is-next-sentence
- big pre-trained models available
- domain (task) adaptation



Input: The man went to the [MASK]$_1$. He bought a [MASK]$_2$ of milk .
Labels: [MASK]$_1$ = store; [MASK]$_2$ = gallon

Sentence A = The man went to the store.
Sentence B = He bought a gallon of milk.
Label = IsNextSentence

Sentence A = The man went to the store.
Sentence B = Penguins are flightless.
Label = NotNextSentence