Recap: Recurrent Neural Networks

- Basic neural networks: Classify one item at a time
- Sequential labeling: Class of one item has impact on class of next item
- Recurrent neural networks
  - Additional connection along the sequence
  - Information can be passed from one sequence element to the next
- One dimension more, because training instance is a sequence

Machine Learning 6: Attention & Transformer Models

VL Sprachliche Informationsverarbeitung

Nils Reiter
nils.reiter@uni-koeln.de

January 12, 2023
Winter term 2022/23

INSTITUT FÜR
DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

Section 1

Encoder-Decoder-Networks

Introduction

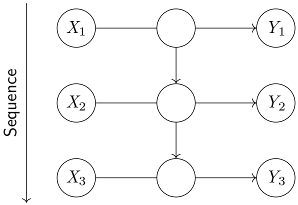


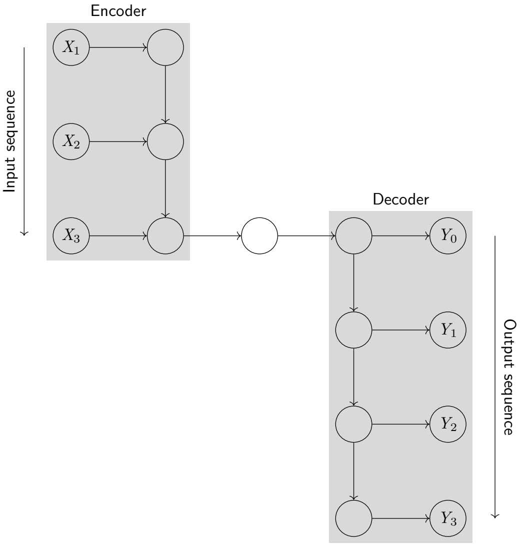
Figure: Neural network with a recurrent layer

- Each X value leads to a Y value
- Network has no way to skip a sequence element
- Many real world sequence labeling tasks are n-to-m-tasks
  - n elements in one sequence are associated with m element in the other

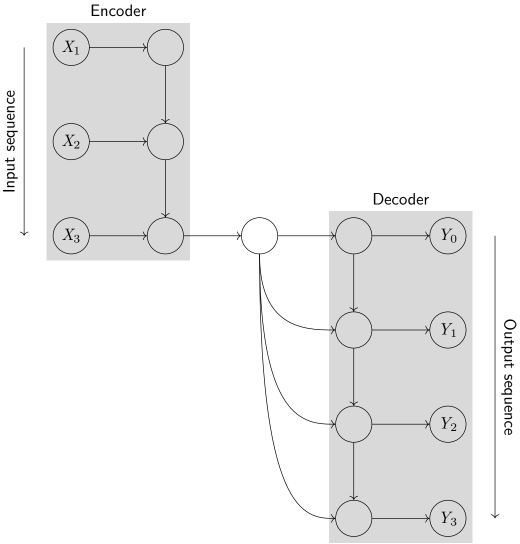
Encoder-Decoder-Architecture

- Network has two parts:
  - Encoder maps from input data to an internal representation
  - Internal representation optionally processed by a regular dense layer
  - Decoder maps from internal representation to the output
- Internal representation
  - Use the output of last recurrent neuron
    - Or internal state of last LSTM cell
  - Some vector, not interpretable

Encoder-Decoder-Architecture



Encoder-Decoder-Architecture



Encoder-Decoder-Architecture in Keras

- Encoder
  - Regular input layer
  - Recurrent layer with `return_sequences=False`
    - Because we don't want a sequence as output, but just the output of the last cell
- Decoder
  - Every output sequence element gets the internal representation as input
    - Thus, it needs to be repeated with the `RepeatVector()` layer
    - This is just copying the vector
  - Recurrent layer with `return_sequences=True`
    - Because now, we want the sequence
  - Output layer as before
    - With one-hot-encoding for multi-class problems

Encoder-Decoder-Architecture in Keras

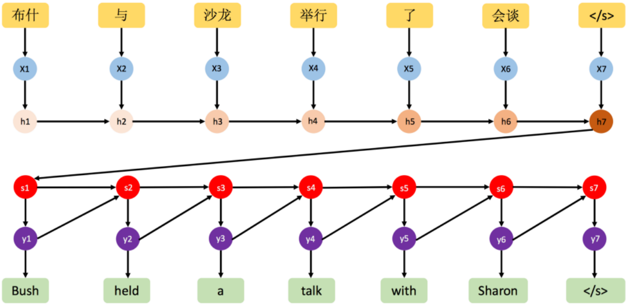
Listing 1: The Code

```
model = models.Sequential()
# Encoder
model.add(layers.Input(shape=(INPUT_LENGTH,)))
model.add(layers.Embedding(input_dim=number_of_symbols, output_dim=64,))
model.add(layers.LSTM(64, return_sequences=False))

# Copy the internal representation (optional)
model.add(layers.RepeatVector(OUTPUT_LENGTH))

# Decoder
model.add(layers.LSTM(32, return_sequences=True))
model.add(layers.Dense(number_of_symbols*2, activation='softmax'))
```

Neural Machine Translation



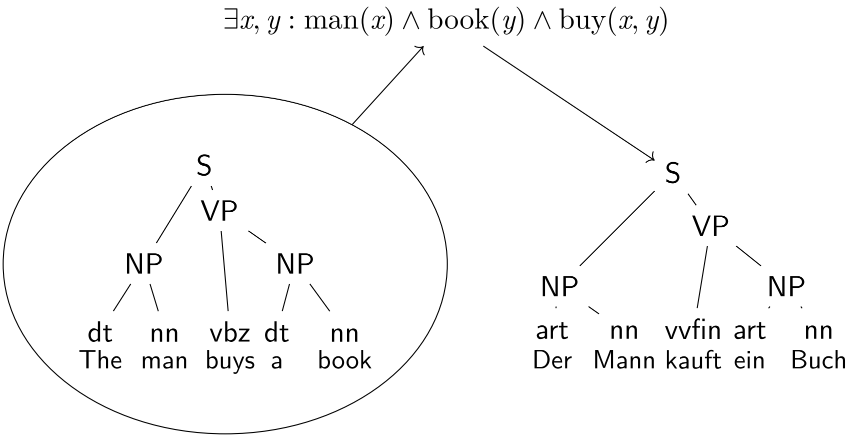
(Sutskever et al., 2014)

- This architecture was proposed in 2014 for machine translation — Sutskever et al. (2014)
- Used by Google Translate and many others

Fixed Lengths

But we still have a fixed length of output elements!
True, but: Decoupling of input and output sequences
$\Rightarrow$ Input and output length don't have to be equal, which is an improvement

Fun Fact: Symbolic Machine Translation



Section 2

Attention

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

Figure: Examples of attending to the correct object (Xu et al., 2015)

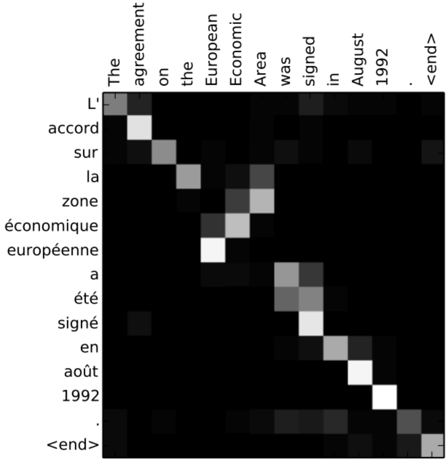


Figure: Attention paid by a neural machine translation network (Bahdanau et al., 2015)

Introduction

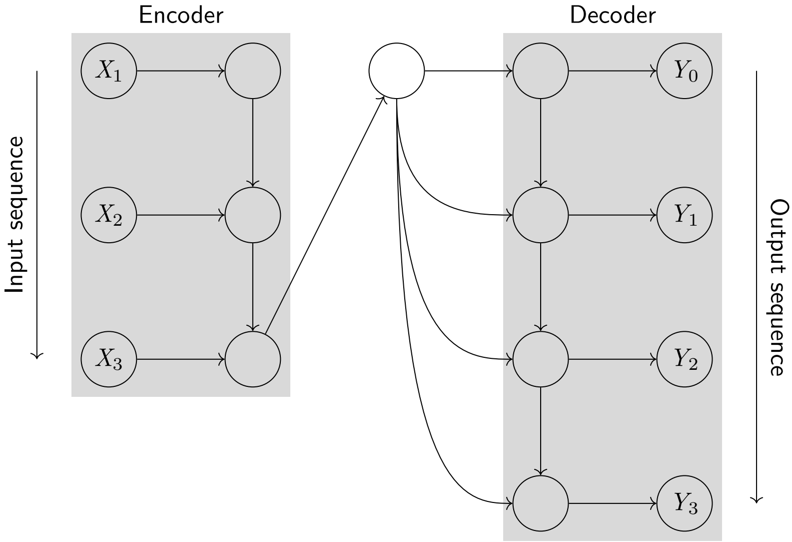
- A mechanism to allow the network to learn what to focus on
- Idea: Not all parts of the input are equally important
  - MT: “la zone économique européenne” $\rightarrow$ “the European Economic Area”, irrespective of context

Introduction

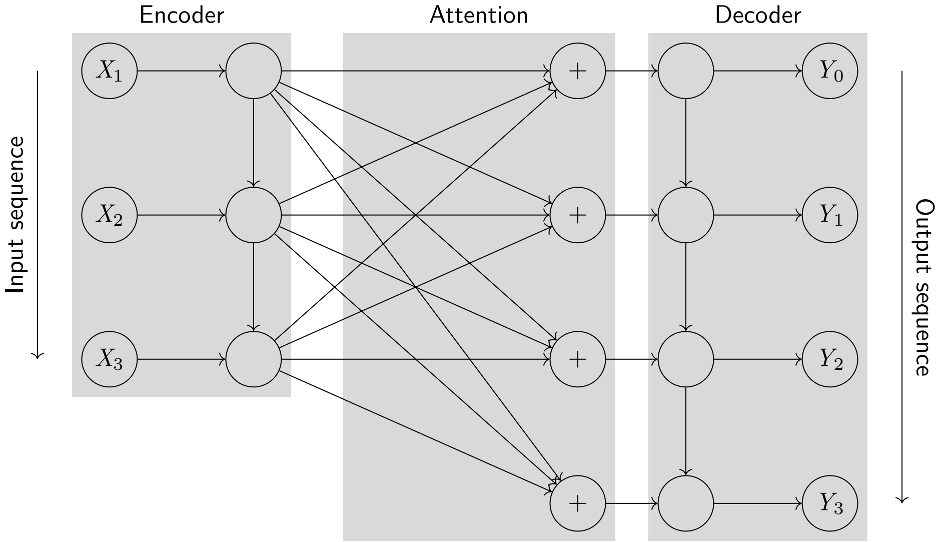
- A mechanism to allow the network to learn what to focus on
- Idea: Not all parts of the input are equally important
  - MT: "la zone économique européenne" → "the European Economic Area", irrespective of context
- Mirrows human reading/translating activities
- Developed for machine translation, then applied to other tasks

Dzmitry Bahdanau/Kyunghyun Cho/Yoshua Bengio (2015). "Neural Machine Translation by Jointly Learning to Align and Translate". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio/Yann LeCun. URL: http://arxiv.org/abs/1409.0473

From Encoder-Decoder to Attention



From Encoder-Decoder to Attention



Section 3

BERT

Introduction

- BERT has outperformed the state of the art in many NLP tasks
- Breakthrough in NLP

Introduction

- BERT has outperformed the state of the art in many NLP tasks
- Breakthrough in NLP
- General idea
  - Encoder-Attention-Decoder architecture (= transformer)
  - Process whole input at once (max. 512 tokens, = bidirectional)
  - Pre-training and fine-tuning on different tasks

Jacob Devlin/Ming-Wei Chang/Kenton Lee/Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of NAACL*. Minneapolis, Minnesota: ACL, pp. 4171–4186. DOI: 10.18653/v1/N19-1423

Pre-Training and Fine-Tuning

- BERT models are trained on huge data sets
- Training one from scratch requires significant resources (time/money)
- Pre-trained models are shared freely
- Recipe: Take a pre-trained model and fine-tune it on your task
  - Pre-trained model contains an abstract language representation

Pre-Training and Fine-Tuning

- BERT models are trained on huge data sets
- Training one from scratch requires significant resources (time/money)
- Pre-trained models are shared freely
- Recipe: Take a pre-trained model and fine-tune it on your task
  - Pre-trained model contains an abstract language representation
- Fine-tuning
  - Any language-related task!

BERT Training Tasks

Masked Language Modeling (MLM)

- Sentence-wise
- 15% of the tokens are “masked” by a special token
- Model predicts these, having access to all other tokens

BERT Training Tasks

Masked Language Modeling (MLM)

- Sentence-wise
- 15% of the tokens are "masked" by a special token
- Model predicts these, having access to all other tokens

Next sentence prediction (NSP)

- Two (masked) sentences are concatenated
- Model has to predict wether second sentence follows on the first or not

Multi-Task-Learning

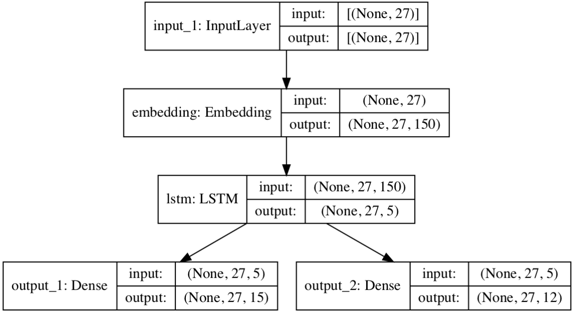
- In many situations, doing multiple tasks at the same time is beneficial
  - E.g.: Image and text recognition, part of speech and named entity tagging, …
- Solving the tasks is often also easier
- In neural networks, this is straightforward
  - Multiple output layers with the functional API
  - Multiple y data supplied to the `fit(...)`-function

Multi-Task-Learning

Example

```
l_input = layers.Input(shape = (27,))
l_emb = layers.Embedding(
  input_dim = 1000,
  output_dim = 150)(l_input)
l_lstm = layers.LSTM(
  units = 5,
  return_sequences = True)(l_emb)
l_out_1 = layers.Dense(
  15, activation = 'softmax',
  name="output_1")(l_lstm)
l_out_2 = layers.Dense(
  12, activation = 'softmax',
  name="output_2")(l_lstm)

model = models.Model(
  inputs = l_input,
  outputs=[l_out_1, l_out_2])
```



Section 4

Hugging Face

Introduction

- An AI company that provides
  - A Python library for transformer models
    - Since 2.0 compatible with tensorflow/keras and PyTorch
  - A platform to share BERT models (e.g., for different languages) and/or data sets
  - Some paid services

Introduction

- An AI company that provides
  - A Python library for transformer models
    - Since 2.0 compatible with tensorflow/keras and PyTorch
  - A platform to share BERT models (e.g., for different languages) and/or data sets
  - Some paid services

Installation

```
pip install transformers
```

Code

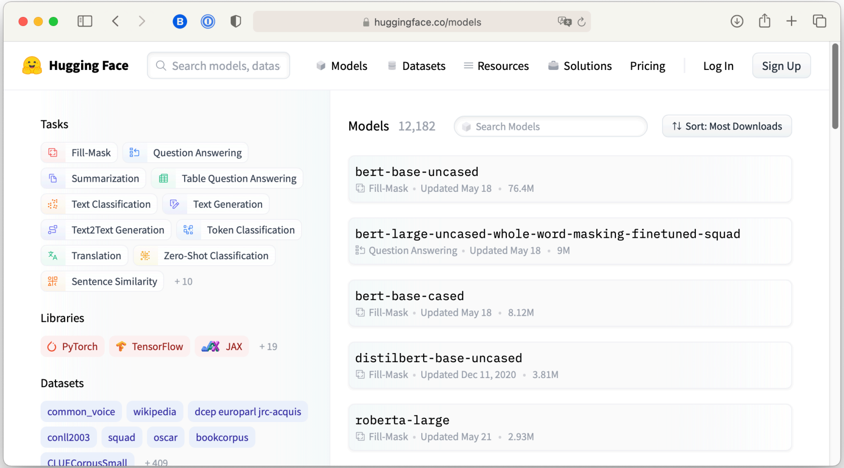
```python
import tensorflow as tf
from transformers import TFAutoModelForSequenceClassification

# Load model as keras model
model = TFAutoModelForSequenceClassification
  .from_pretrained("bert-base-cased", num_labels=2)

# do the usual keras stuff
model.compile(...)

# fine-tuning
model.fit(...)
```

https://huggingface.co/transformers/training.html


huggingface.co/models
Hugging Face
Search models, datas
Models
Datasets
Resources
Solutions
Pricing
Log In
Sign Up
Tasks
Fill-Mask
Question Answering
Summarization
Table Question Answering
Text Classification
Text Generation
Text2Text Generation
Token Classification
Translation
Zero-Shot Classification
Sentence Similarity
+ 10
Libraries
PyTorch
TensorFlow
JAX
+ 19
Datasets
common_voice
wikipedia
dcep europarl jrc-acquis
conll2003
squad
oscar
bookcorpus
CLUECorpusSmall
+ 409
Models 12,182
Search Models
Sort: Most Downloads
bert-base-uncased
Fill-Mask • Updated May 18 • 76.4M
bert-large-uncased-whole-word-masking-finetuned-squad
Question Answering • Updated May 18 • 9M
bert-base-cased
Fill-Mask • Updated May 18 • 8.12M
distilbert-base-uncased
Fill-Mask • Updated Dec 11, 2020 • 3.81M
roberta-large
Fill-Mask • Updated May 21 • 2.93M

Section 5

Summary

Summary

- Motivation: Sequence to sequence tasks (like machine translation)

Encoder-Decoder architecture

- Encoder reads in the input, generates internal representation
- Decoder produces output, consuming internal representation

Attention

- Developed for image classification, then transfered to machine translation
- Let the model learn the relevant input tokens for each output token

BERT

- Breakthrough in natural language processing
- Pre-training vs. fine-tuning
- Huggingface: Platform to make such models easy to use
  - Good documentation on transformers: huggingface.co/docs/transformers