

Lecture 12

Deep Learning for Intelligent Systems

COMP 474/6741, Winter 2021

[Introduction](#)

[ImageNet](#)

[AlexNet](#)

[RNNs and LSTMs](#)

[Recurrent Neural Networks \(RNNs\)](#)

[Long Short-Term Memory Networks \(LSTMs\)](#)

[Sequence-to-Sequence Models](#)

[Transformers](#)

[BERT](#)

[GPT-2](#)

[GPT-3](#)

[Notes and Further Reading](#)

René Witte
Department of Computer Science
and Software Engineering
Concordia University

1 Introduction

Introduction

ImageNet

AlexNet

2 RNNs and LSTMs

RNNs and LSTMs

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

3 Transformers

Transformers

BERT

GPT-2

GPT-3

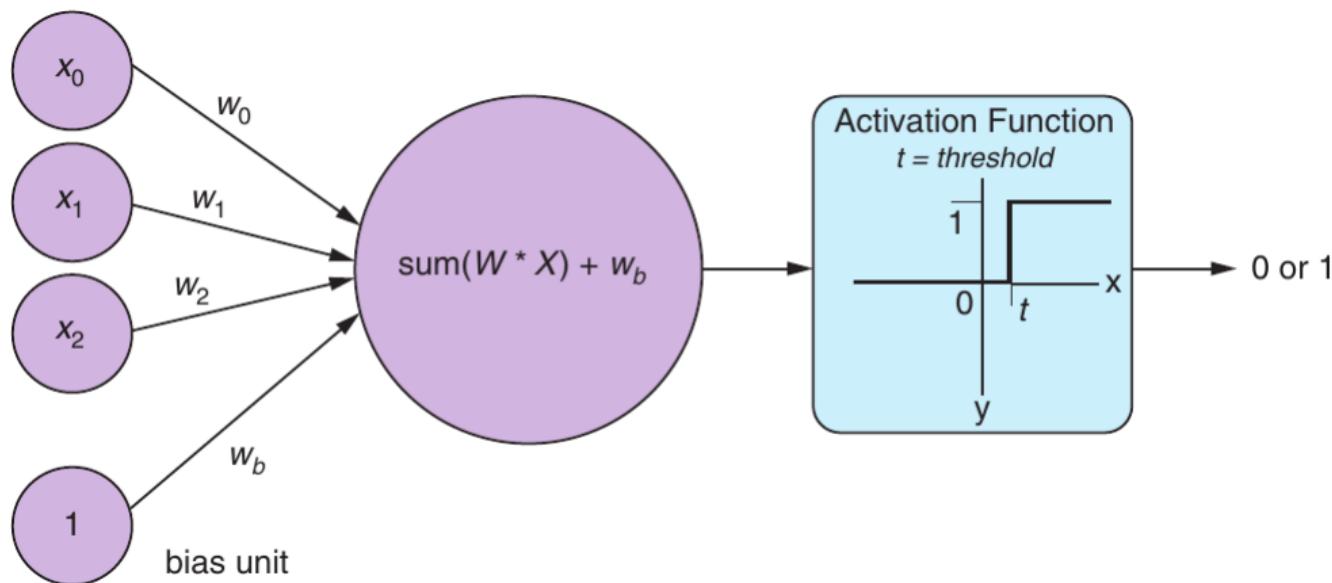
4 Notes and Further Reading

Notes and Further

Reading

Single Neuron (Perceptron)

René Witte



Copyright 2019 by Manning Publications Co., [LHHT19]

Introduction

ImageNet
AlexNet

RNNs and LSTMs

Recurrent Neural Networks (RNNs)
Long Short-Term Memory Networks (LSTMs)
Sequence-to-Sequence Models

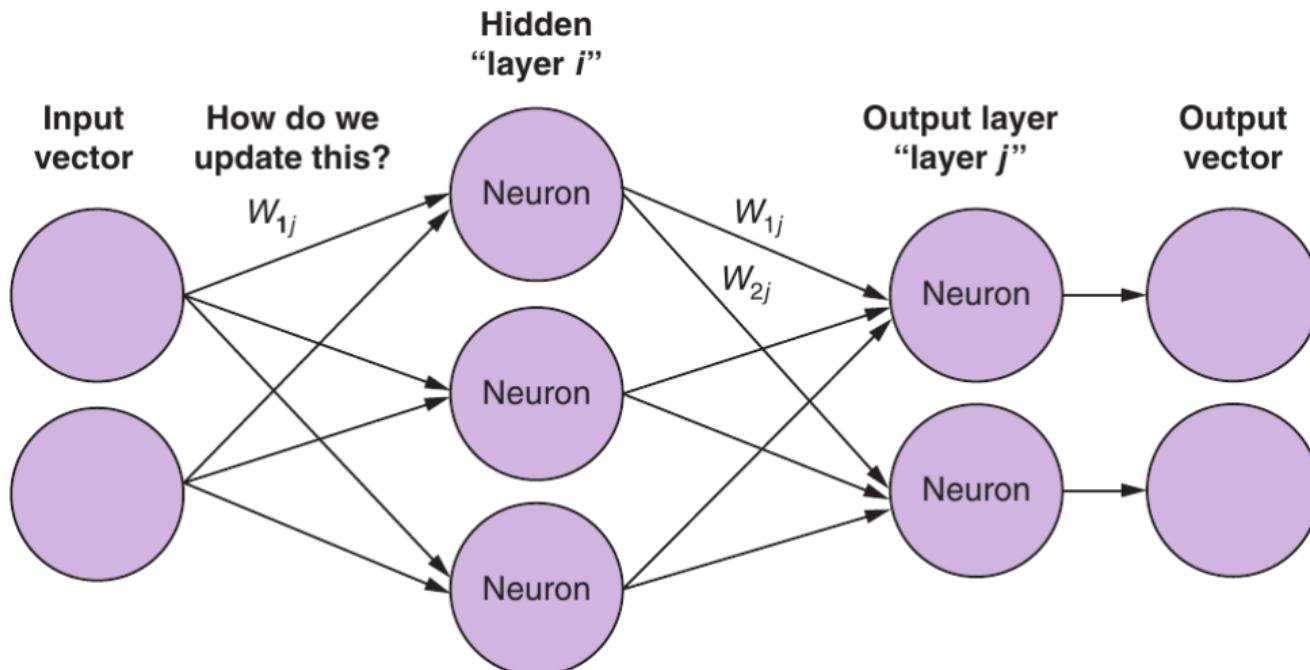
Transformers

BERT
GPT-2
GPT-3

Notes and Further Reading

Multi-layer neural networks with hidden weights

René Witte



Copyright 2019 by Manning Publications Co., [LHII19]

Introduction

[ImageNet](#)
[AlexNet](#)

RNNs and LSTMs

[Recurrent Neural Networks \(RNNs\)](#)
[Long Short-Term Memory Networks \(LSTMs\)](#)
[Sequence-to-Sequence Models](#)

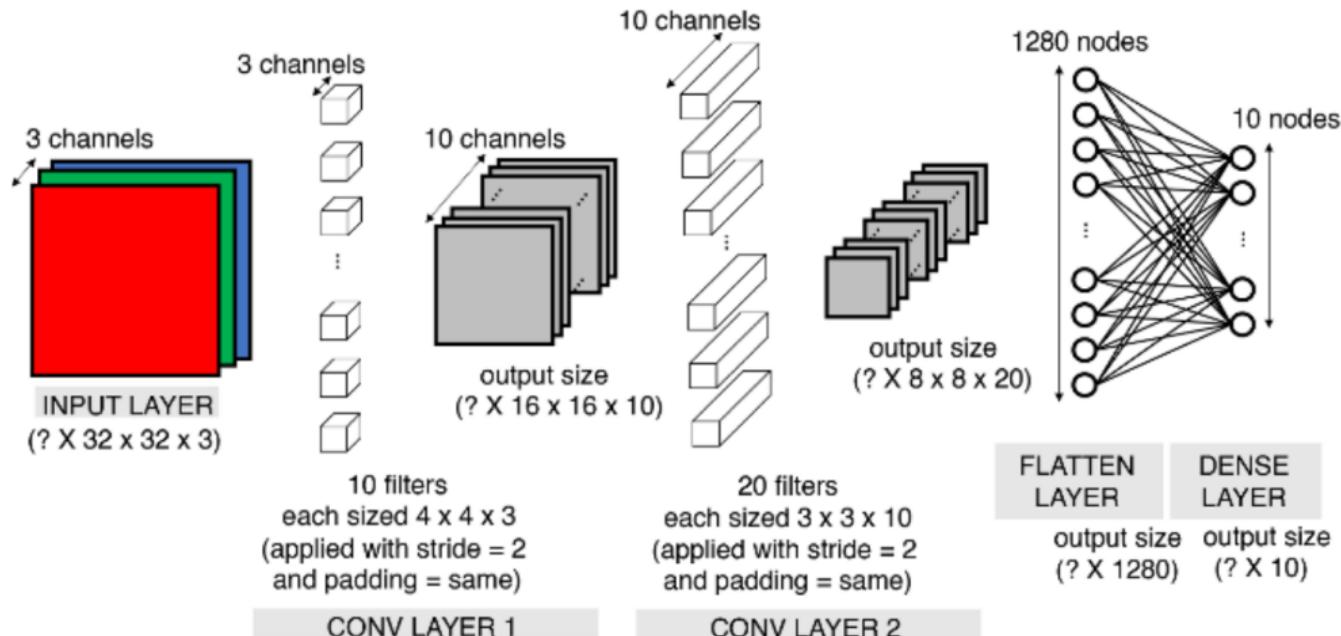
Transformers

[BERT](#)
[GPT-2](#)
[GPT-3](#)

Notes and Further Reading

Convolutional Neural Network

René Witte



Copyright 2019 Applied Data Science Partners Ltd., [Fos19]

Introduction

ImageNet
AlexNet

RNNs and LSTMs

Recurrent Neural Networks (RNNs)
Long Short-Term Memory Networks (LSTMs)
Sequence-to-Sequence Models

Transformers

BERT
GPT-2
GPT-3

Notes and Further Reading

ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

10,000,000 labeled images depicting 10,000+ object categories (2012)

René Witte



mite

container ship

motor scooter

leopard

mite	container ship	motor scooter	leopard
black widow	lifeboat	go-kart	jaguar
cockroach	amphibian	moped	cheetah
tick	fireboat	bumper car	snow leopard
starfish	drilling platform	golfcart	Egyptian cat



grille

mushroom

cherry

Madagascar cat

convertible	agaric	dalmatian	squirrel monkey
grille	mushroom	grape	spider monkey
pickup	jelly fungus	elderberry	titi
beach wagon	gill fungus	fordshire bullterrier	indri
fire engine	dead-man's-fingers	currant	howler monkey

Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks (RNNs)

Long Short-Term Memory Networks (LSTMs)

Sequence-to-Sequence Models

Transformers

BERT

GPT-2

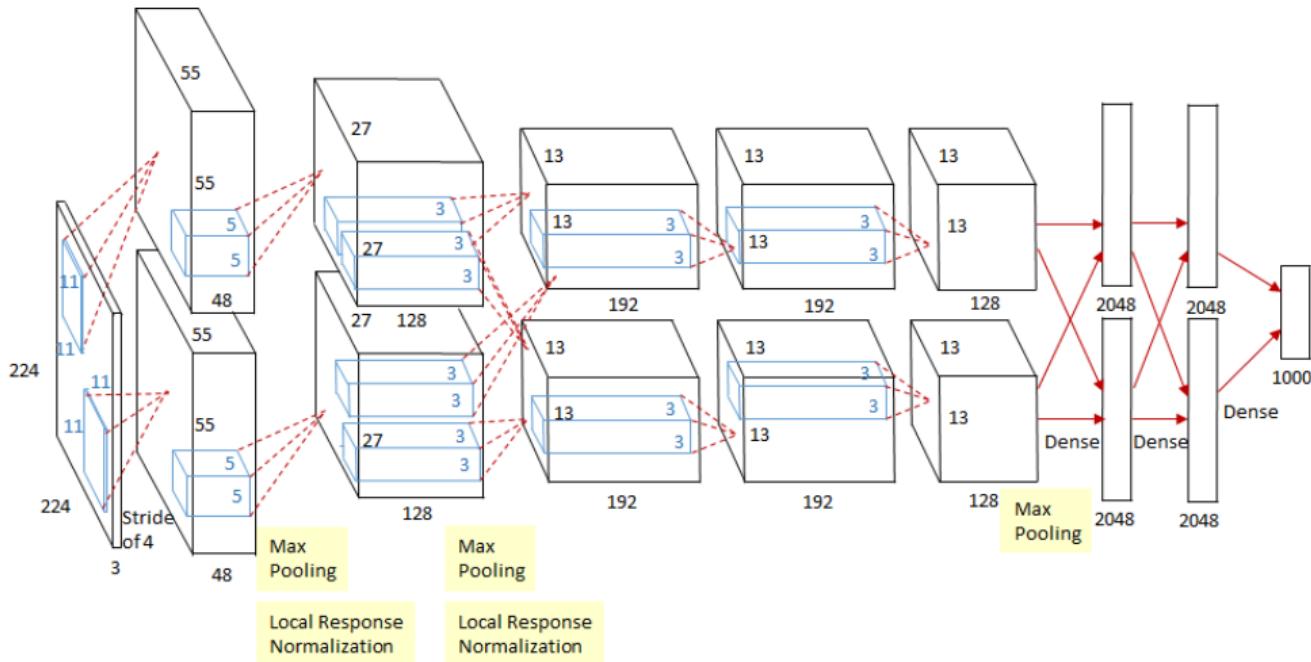
GPT-3

Notes and Further Reading

AlexNet (8-layer CNN, 2012)

60 million parameters, trained using two GPUs

René Witte



<https://medium.com/coinmonks/paper-review-of-alexnet-caffenet-winner-in-ilsvrc-2012-image-classification-b93598314160>

Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks (RNNs)

Long Short-Term Memory Networks (LSTMs)

Sequence-to-Sequence Models

Transformers

BERT

GPT-2

GPT-3

Notes and Further Reading

1 Introduction

Introduction

ImageNet

AlexNet

2 RNNs and LSTMs

Recurrent Neural Networks (RNNs)

Long Short-Term Memory Networks (LSTMs)

Sequence-to-Sequence Models

RNNs and LSTMs

Recurrent Neural Networks (RNNs)

Long Short-Term Memory Networks (LSTMs)

Sequence-to-Sequence Models

3 Transformers

Transformers

BERT

GPT-2

GPT-3

4 Notes and Further Reading

Notes and Further Reading

[Introduction](#)[ImageNet](#)[AlexNet](#)[RNNs and LSTMs](#)[Recurrent Neural Networks \(RNNs\)](#)[Long Short-Term Memory Networks \(LSTMs\)](#)[Sequence-to-Sequence Models](#)[Transformers](#)[BERT](#)[GPT-2](#)[GPT-3](#)[Notes and Further Reading](#)

Word history is important

Consider:

The stolen car sped into the arena.

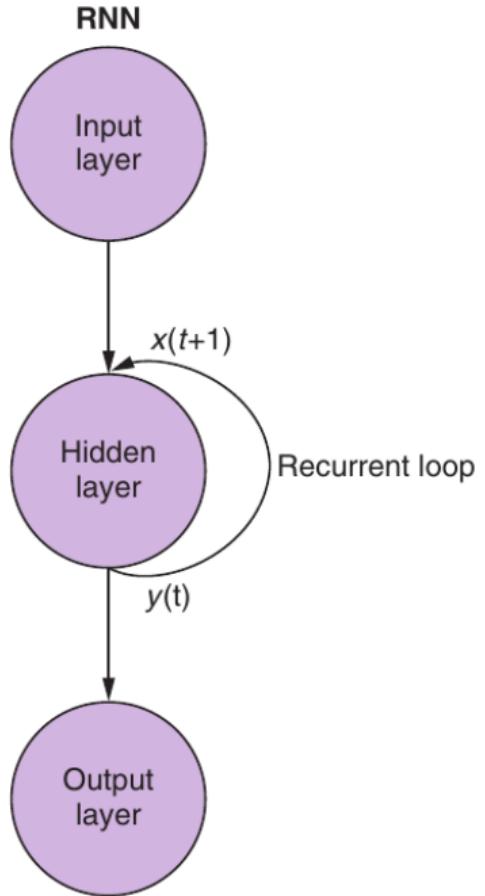
The clown car sped into the arena.

Remembering past words

- So far, we considered a sentence as whole
- Now we allow a NN to look at the sequence of words

Recurrent Neural Net (RNN)

René Witte



[Introduction](#)

[ImageNet](#)

[AlexNet](#)

[RNNs and LSTMs](#)

[Recurrent Neural Networks \(RNNs\)](#)

[Long Short-Term Memory Networks \(LSTMs\)](#)

[Sequence-to-Sequence Models](#)

[Transformers](#)

[BERT](#)

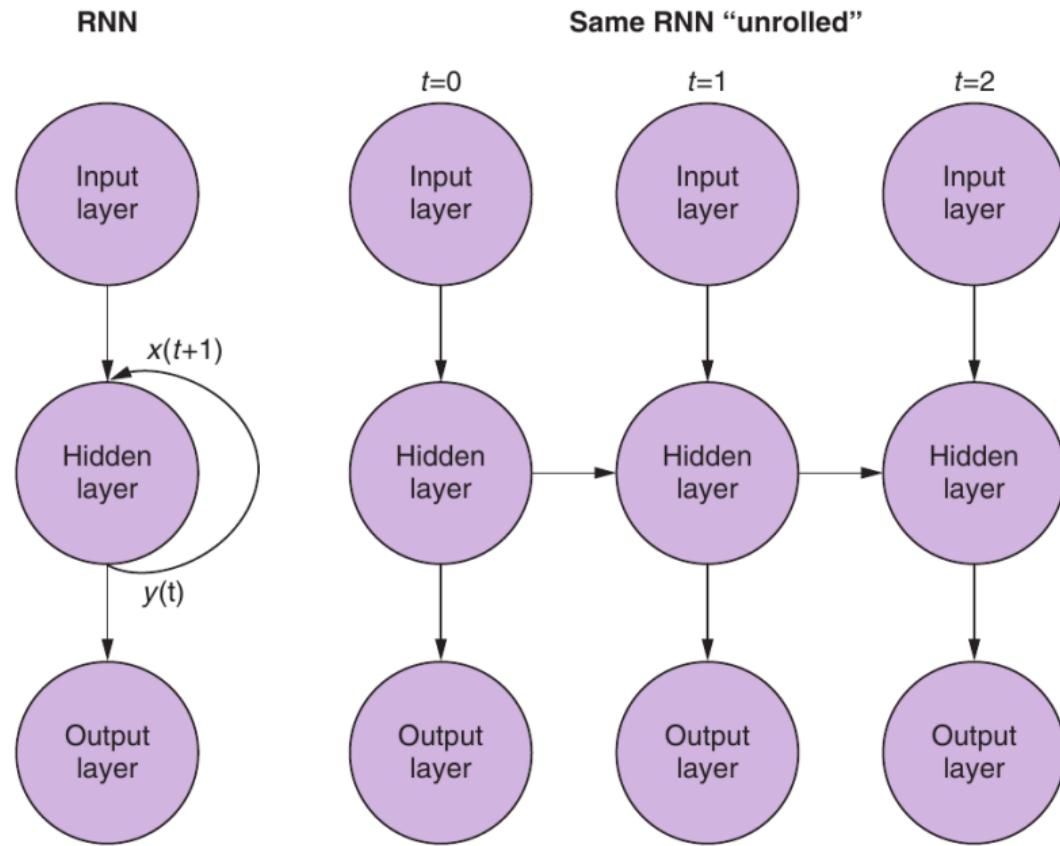
[GPT-2](#)

[GPT-3](#)

[Notes and Further Reading](#)

Unrolled Recurrent Neural Net

René Witte

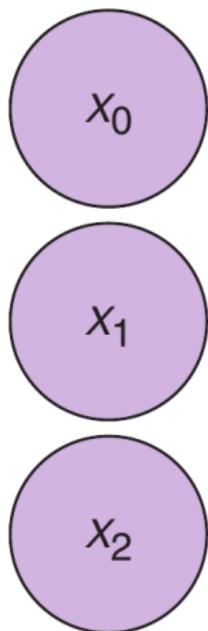


Detailed recurrent neural net at time step $t = 0$

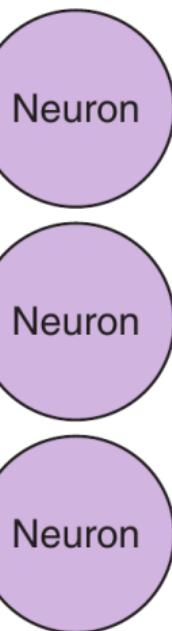
René Witte



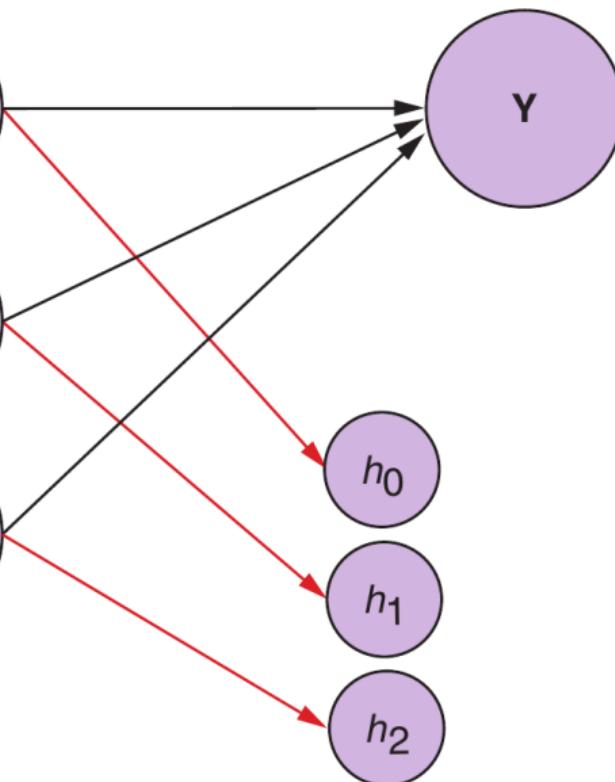
Input vector



Hidden



Output layer



Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks (RNNs)

Long Short-Term Memory Networks (LSTMs)

Sequence-to-Sequence Models

Transformers

BERT

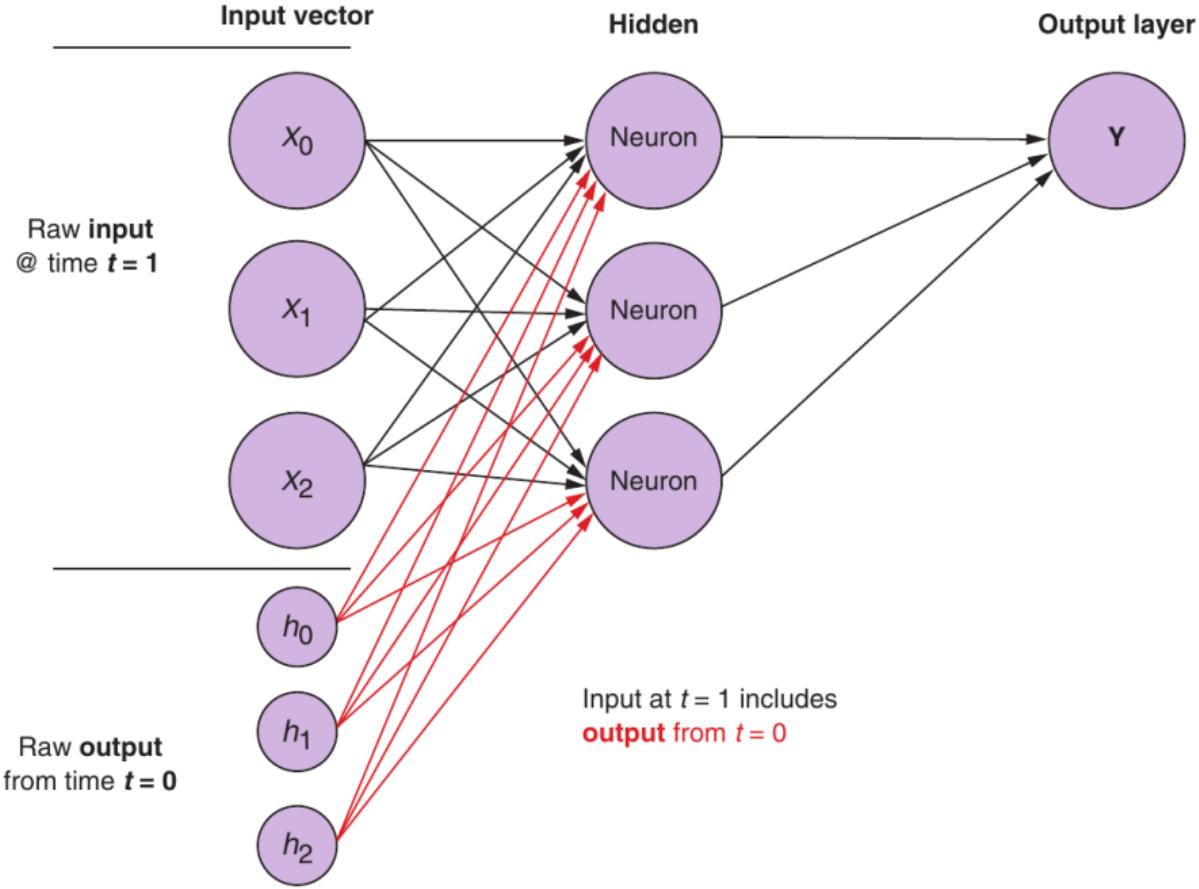
GPT-2

GPT-3

Notes and Further Reading

Detailed recurrent neural net at time step $t = 1$

René Witte



Introduction

ImageNet
AlexNet

RNNs and LSTMs

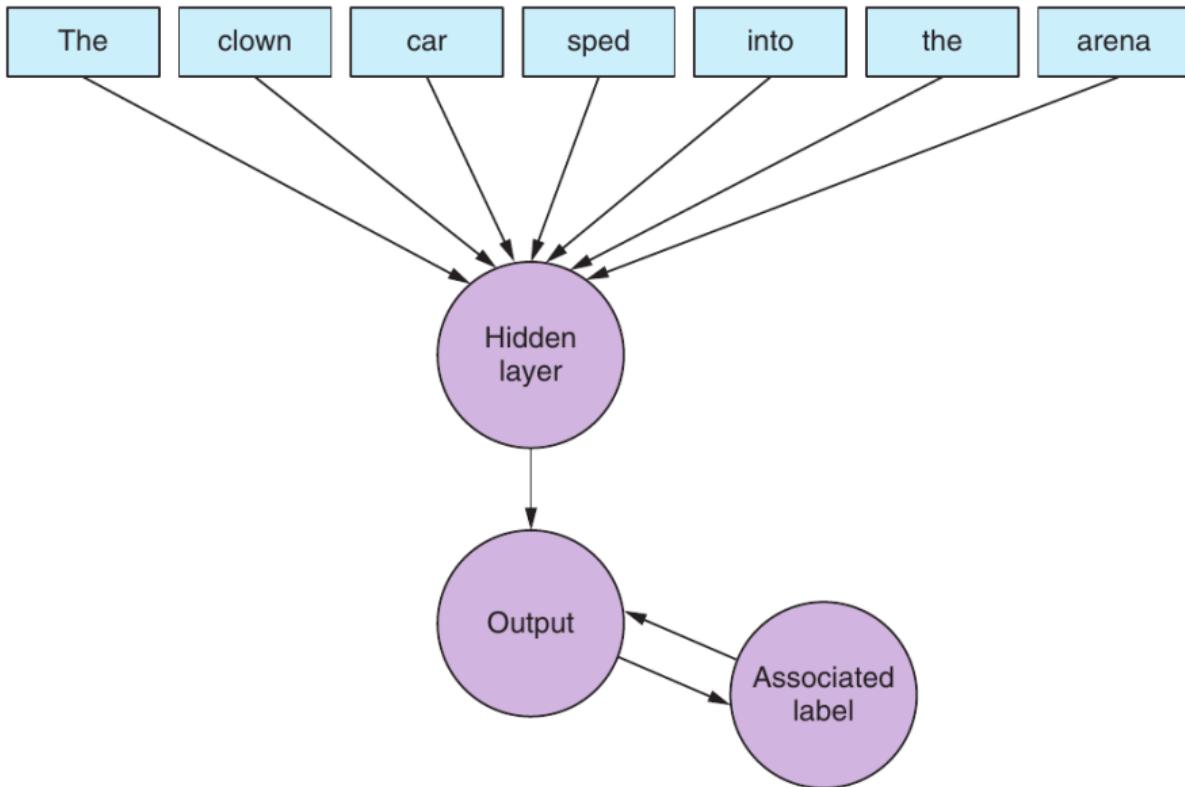
Recurrent Neural Networks (RNNs)
Long Short-Term Memory Networks (LSTMs)
Sequence-to-Sequence Models

Transformers

BERT
GPT-2
GPT-3

Notes and Further Reading

Text fed into a convolutional network



Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

Transformers

BERT

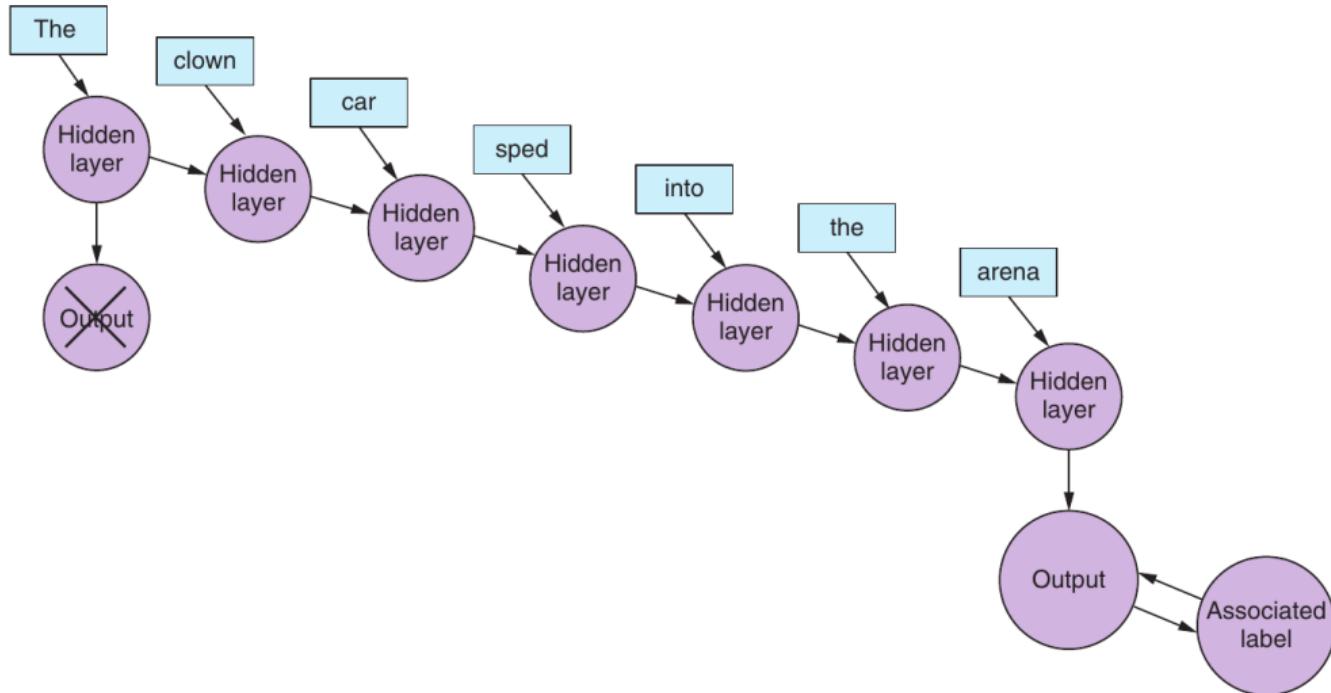
GPT-2

GPT-3

Notes and Further
Reading

Text fed into a recurrent network

René Witte



Introduction

ImageNet
AlexNet

RNNs and LSTMs

Recurrent Neural Networks (RNNs)
Long Short-Term Memory Networks (LSTMs)
Sequence-to-Sequence Models

Transformers

BERT
GPT-2
GPT-3

Notes and Further Reading

Only last output matters here

René Witte



Introduction

ImageNet
AlexNet

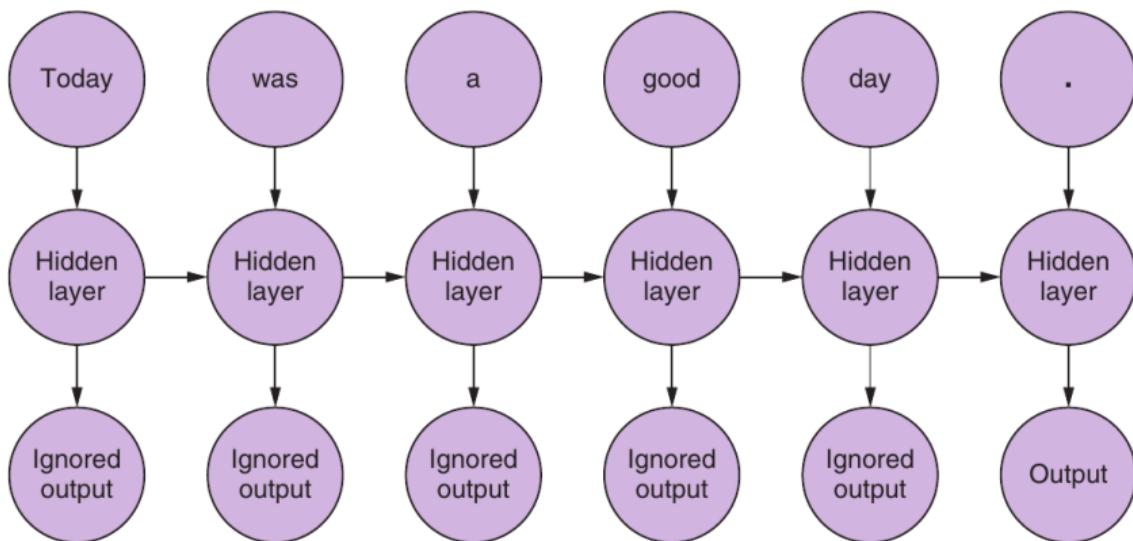
RNNs and LSTMs

Recurrent Neural Networks (RNNs)
Long Short-Term Memory Networks (LSTMs)
Sequence-to-Sequence Models

Transformers

BERT
GPT-2
GPT-3

Notes and Further Reading



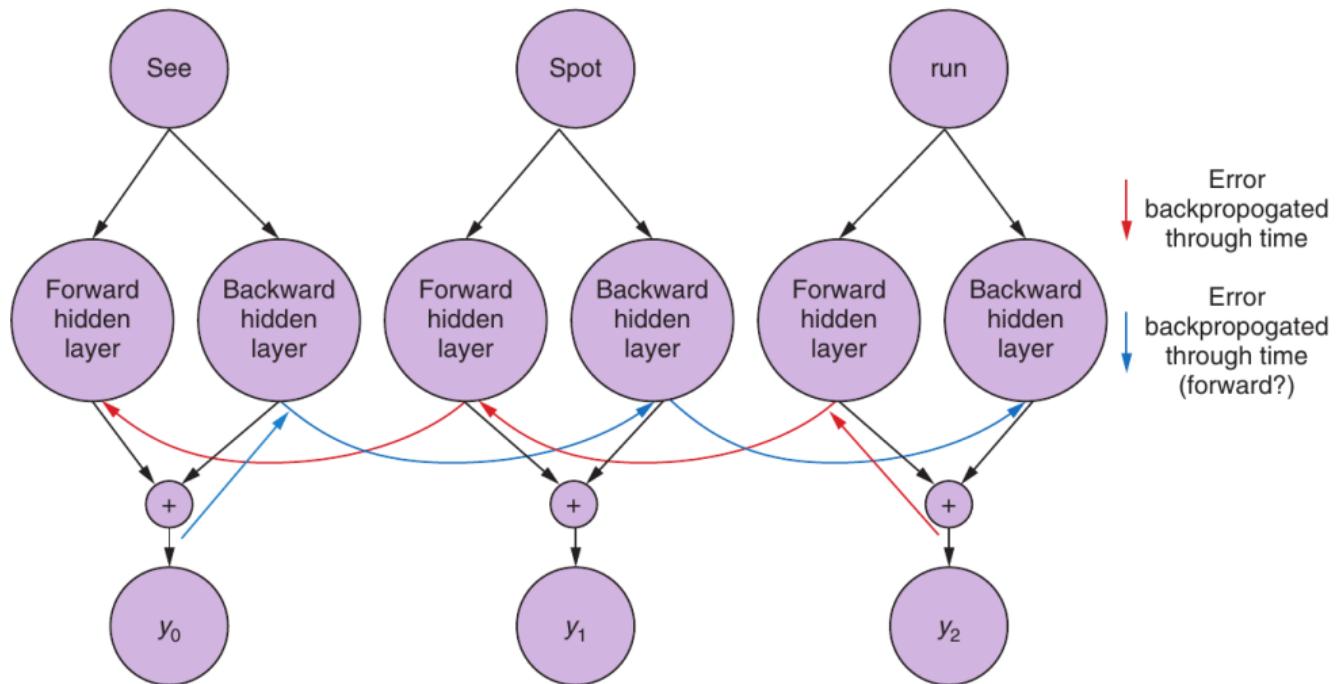
$$\text{error} = \text{y_true_label} - \text{y_output}$$

```
model.add(SimpleRNN(num_neurons,  
                    return_sequences=True,  
                    input_shape=( maxlen, embedding_dims) ))
```

[Introduction](#)[ImageNet](#)[AlexNet](#)[RNNs and LSTMs](#)[Recurrent Neural Networks \(RNNs\)](#)[Long Short-Term Memory Networks \(LSTMs\)](#)[Sequence-to-Sequence Models](#)[Transformers](#)[BERT](#)[GPT-2](#)[GPT-3](#)[Notes and Further Reading](#)

Bidirectional RNN

René Witte



Copyright 2019 by Manning Publications Co., [LHH19]

Introduction

ImageNet
AlexNet

RNNs and LSTMs

Recurrent Neural Networks (RNNs)
Long Short-Term Memory Networks (LSTMs)
Sequence-to-Sequence Models

Transformers

BERT
GPT-2
GPT-3

Notes and Further Reading

Example

Consider the sentences:

The young woman went to the movies with her friends.

The young woman, having found a free ticket on the ground, went to the movies.

Finding relationship between subject and verb

- In the second sentence, “woman” and “went” are too far apart for either a CNN or RNN to pick up
- Need some kind of “memory” to keep the core **thought** of the sentence
- Use a **Long Short-Term Memory** (LSTM) network

[Understanding LSTM Networks, <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>]

[Introduction](#)

[ImageNet](#)

[AlexNet](#)

[RNNs and LSTMs](#)

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

[Transformers](#)

BERT

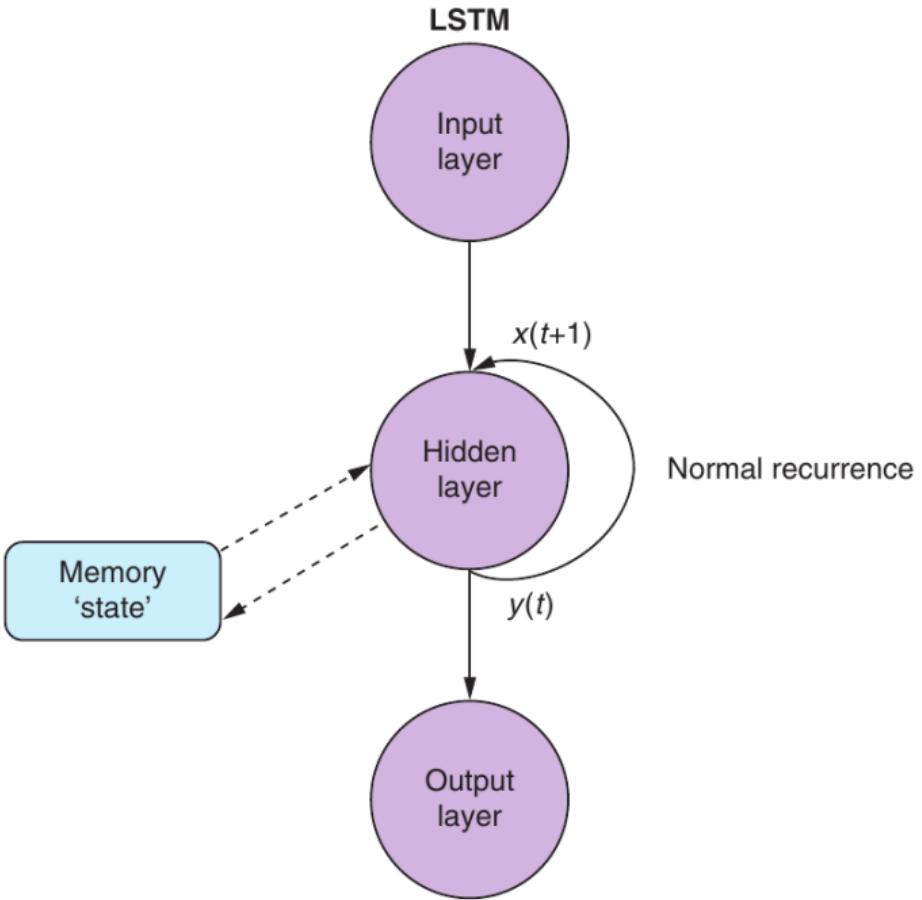
GPT-2

GPT-3

[Notes and Further
Reading](#)

A LSTM Network with Memory

René Witte



[Introduction](#)

[ImageNet](#)

[AlexNet](#)

[RNNs and LSTMs](#)

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

[Transformers](#)

BERT

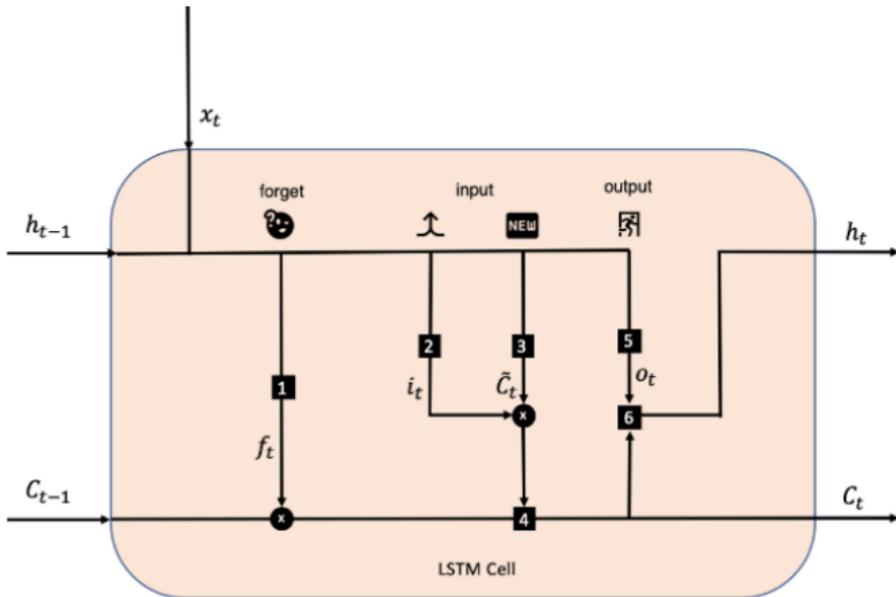
GPT-2

GPT-3

[Notes and Further
Reading](#)

An LSTM Cell

René Witte



- 1 $f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$
- 2 $i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$
- 3 $\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$
- 4 $C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$
- 5 $o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$
- 6 $h_t = o_t * \tanh(C_t)$

Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

Transformers

BERT

GPT-2

GPT-3

Notes and Further
Reading

Using an LSTM to Generate Text

René Witte



Task

- Train LSTM on a text corpus (here: *Aesop's Fables*)
- Goal: predict next word after a given sequence of words

The screenshot shows the Project Gutenberg website. At the top, it says "Project Gutenberg offers 61,713 free ebooks to download." Below the header are links for "Search", "Latest", "Terms of Use", "Donate?", "Mobile", and search and help buttons. The main content area shows the title "Aesop's Fables; a new translation by Aesop". On the left, there's a thumbnail image of the book cover, which is green with purple geometric shapes. The right side shows a table titled "Download This eBook" with various file formats and their sizes:

Format	Size	Links
Read this book online: HTML	292 kB	
EPUB (with images)	4.5 MB	
EPUB (no images)	134 kB	
Kindle (with images)	9.2 MB	
Kindle (no images)	594 kB	
Plain Text UTF-8	237 kB	
More Files...		

At the bottom left are social media sharing icons for Twitter and Print.

Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

Transformers

BERT

GPT-2

GPT-3

Notes and Further
Reading

Step 1: Preprocessing

René Witte



Tokenization

Can use basic tokenizer from Keras

`tokenizer.word_index`

```
{'|': 1,  
',': 2,  
'the': 3,  
'and': 4,  
'.': 5,  
'a': 6,  
'to': 7,  
'"': 8,  
'of': 9,  
'he': 10,  
'his': 11,  
'was': 12,
```

`token_list`

```
1, |  
3, the  
56, fox  
4, and  
3, the  
940, grapes  
5, .  
6, a  
382, hungry  
56, fox  
94, saw
```

Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

Transformers

BERT
GPT-2
GPT-3

Notes and Further
Reading

Step 2: Generate the Dataset

René Witte



Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

Transformers

BERT

GPT-2

GPT-3

Notes and Further
Reading

Goal: Sequences

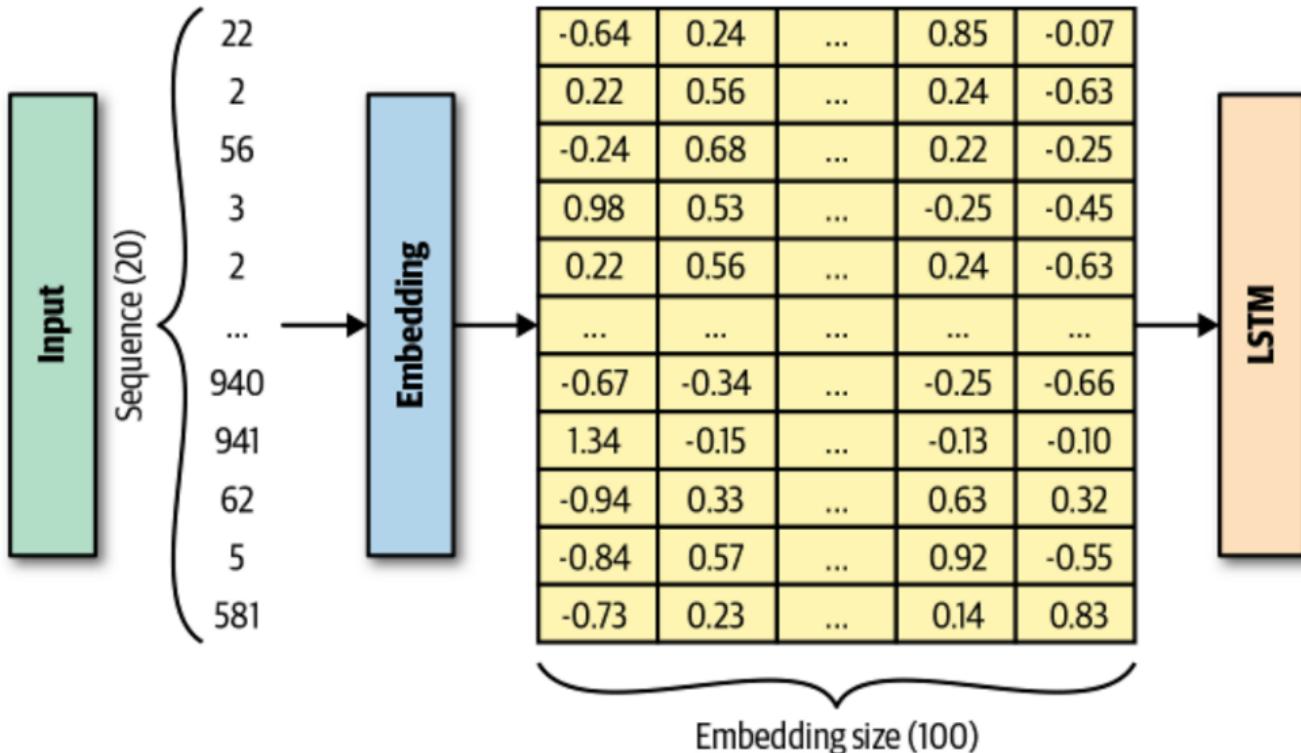
- We want the LSTM to learn the **next word in a sequence**
- E.g., *The fast and the ...* (cat? at? house? ...)
- Here, use 20-word sequences from input text
- Expected output for input sequence is a one-hot vector with the next word

See

https://github.com/davidADSP/GDL_code/blob/master/06_01_lstm_text_train.ipynb

The Embedding Layer

René Witte



[Introduction](#)

[ImageNet](#)

[AlexNet](#)

[RNNs and LSTMs](#)

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

[Transformers](#)

BERT

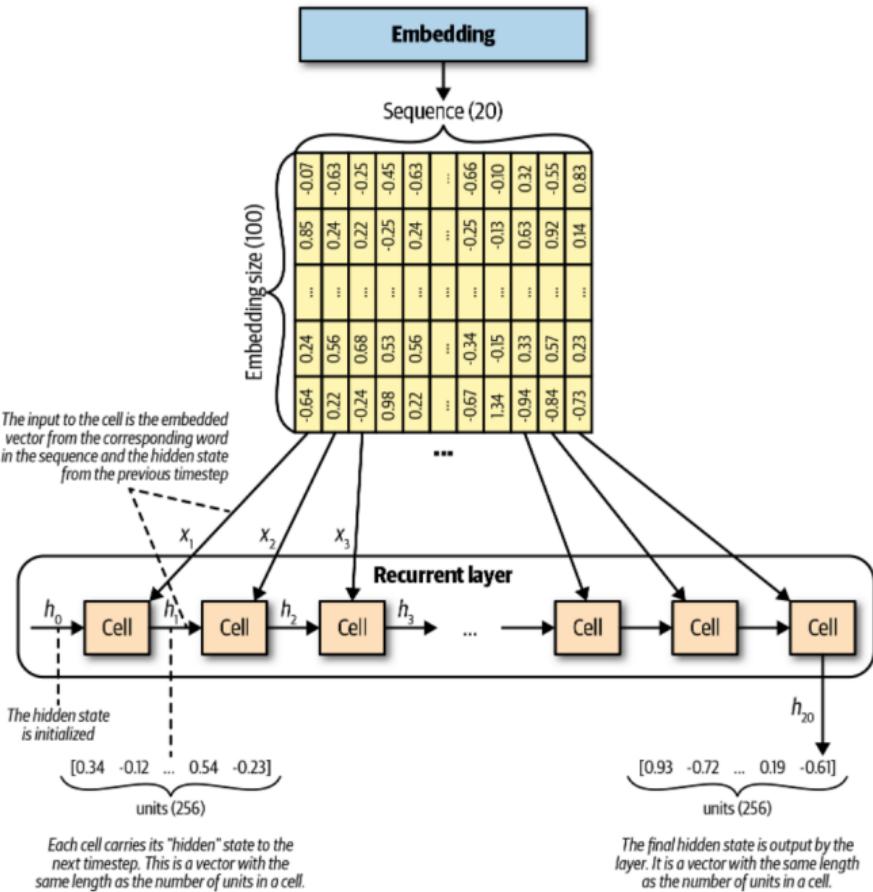
GPT-2

GPT-3

[Notes and Further
Reading](#)

The Recurrent Layer

René Witte



Introduction

ImageNet
AlexNet

RNNs and LSTMs

Recurrent Neural Networks (RNNs)
Long Short-Term Memory Networks (LSTMs)

Sequence-to-Sequence Models

Transformers

BERT
GPT-2
GPT-3

Notes and Further Reading

Implementation in Keras

René Witte



```
from keras.layers import Dense, LSTM, Input, Embedding, Dropout
from keras.models import Model
from keras.optimizers import RMSprop

n_units = 256
embedding_size = 100

text_in = Input(shape = (None,))
x = Embedding(total_words, embedding_size)(text_in)
x = LSTM(n_units)(x)
x = Dropout(0.2)(x)
text_out = Dense(total_words, activation = 'softmax')(x)

model = Model(text_in, text_out)

opti = RMSprop(lr = 0.001)
model.compile(loss='categorical_crossentropy', optimizer=opti)

epochs = 100
batch_size = 32
model.fit(X, y, epochs=epochs, batch_size=batch_size, shuffle = True)
```

Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

Transformers

BERT

GPT-2

GPT-3

Notes and Further
Reading

Our Model

René Witte



[Introduction](#)

[ImageNet](#)

[AlexNet](#)

[RNNs and LSTMs](#)

[Recurrent Neural Networks \(RNNs\)](#)

[Long Short-Term Memory Networks \(LSTMs\)](#)

[Sequence-to-Sequence Models](#)

[Transformers](#)

[BERT](#)

[GPT-2](#)

[GPT-3](#)

[Notes and Further Reading](#)

Layer (type)	Output Shape	Param #
=====		
input_20 (InputLayer)	(None, None)	0
embedding_4 (Embedding)	(None, None, 100)	443800
lstm_4 (LSTM)	(None, 256)	365568
dense_41 (Dense)	(None, 4438)	1140566
=====		

Total params: 1,949,934

Trainable params: 1,949,934

Non-trainable params: 0

Distribution of word probabilities following various sequences

René Witte



Noun

the fox and the stag . there was a
21.3% : time
19.1% : lion
15.7% : man
11.6% : ass
8.6% : good
4.3% : fox
2.0% : a
1.8% : once
1.7% : old
1.3% : dogs

Verb

the fox and the snake . one day a fox
49.9% : came
19.8% : saw
8.0% : went
6.0% : ,
5.5% : was
4.2% : were
1.2% : said
1.1% : go
0.4% : had
0.3% : put

Preposition

the dog and the hare . a dog was lying
68.6% : in
21.6% : on
6.7% : at
0.6% : by
0.5% : into
0.4% : from
0.3% : to
0.3% : with
0.2% : about
0.2% : for

Adjective / Verb

the farmer and his sheep . a farmer was
17.2% : unable
8.3% : afraid
6.1% : so
5.4% : good
5.0% : lying
4.6% : sitting
4.2% : who
2.4% : by
2.3% : going
2.3% : caught

Article

the eagle and the sea .
89.4% : an
4.1% : the
1.5% : just
1.2% : thus
0.7% : two
0.7% : presently
0.6% : a
0.5% : there
0.4% : jupiter
0.1% : at

Punctuation

the lion said ,	the lion said , and
98.0% : "	84.4% : the
1.1% : and	5.2% : you
0.4% : that	3.4% : that
0.1% : the	3.0% : if
0.1% : much	0.3% : well
0.0% : you	0.2% : "
0.0% : do	0.2% : i
0.0% : time	0.2% : my
0.0% : as	0.2% : we
0.0% : a	0.2% : very

Introduction

ImageNet
AlexNet

RNNs and LSTMs

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

Transformers

BERT
GPT-2
GPT-3

Notes and Further Reading

Algorithm

- Provide some input text to the LSTM
- Let it predict the [next word](#)
- Output vector (softmax) provides probabilities for next word
- Pick one using a stochastic approach (rather than always the maximum)
- We use a `temperature` parameter; lower temperature means more deterministic choice

[Introduction](#)[ImageNet](#)[AlexNet](#)[RNNs and LSTMs](#)[Recurrent Neural Networks \(RNNs\)](#)[Long Short-Term Memory Networks \(LSTMs\)](#)[Sequence-to-Sequence Models](#)[Transformers](#)[BERT](#)[GPT-2](#)[GPT-3](#)[Notes and Further Reading](#)

Output Examples

René Witte



Temperature 0.2

the ass and his lion . a certain man who had an ass and a stag at them in a good . " do , " said the other saying , " you but the ass to be so . to come and you are in the way of ever see how you are , in the man of a lion . " the lion replied , " you not very much to me , and i will be a eyes upon you , and , as that you were for the good of a on tha t men will be had all the dog to be a lion . as the cock came , and the ground for it in a very much to wolf , " you take my oh , if you will do not get me in any time the a very way of some day . " well again , which had very well on ly a man , he turned up and said to the , " but i see what you see how i are day : but you not one of them to be not well ? " . he saw what be a of master in the ass was said to him , " the other , you are a man of me , and i were so . " the wolf was so much to the man and said to his master , " you , in the friend of you are ; and then he should be running up they found for you said he had a way to make them . " as the very way , for it was that he could not see a t the eagle and said to him , " i will be a lion and i have was one day , and you know that you go again , but come a nd see what you are of the better . "

Temperature 1.0

the great man and the time . a sheep master fell on a head and had been in a other to get them . so the ass , caught a young horse and was by me , and so he got himself into a horse than the lion of their put out ; and why , he found a lion in a ass . you must go , he they we not only could not a bull by one day the tortoise in gratitude in a place live of the very do wish you . so , as he could not the fox no one make the ass of the day , where in the well have w ould all she had been in a good first , but the other dog and then a cock in a moment . i'm not a lion , who had " co me off , that do you see that you find that the lion , who they fell on to be very much more good good strength , and have to a mouse go . when the other who is same , who had no town mouse , and found a cat up out of a great much . " the better feed in the water , that she came to a bird . the eagle never do so good a frog or and drink ; so my maste r , saw that the master is was well and the ass at my a very good time . the ox turned too use for a way to have to h ertext : but if the will very fox came to let him one as one as it , and man as he should there well on by the place and stood there . for the time the lamb ass in the ass , for his the lost these all , should found turned down . a us e of the hard went to make it , " and the never soon can be of out of a ass . " the master , so not much be a eagle n ever should be oh ! " presently , not to be a ass , who do not how were heavy : i'm not a do you , you get off in the water , and not come to we are very take the just so ever even in my master .

Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

Transformers

BERT

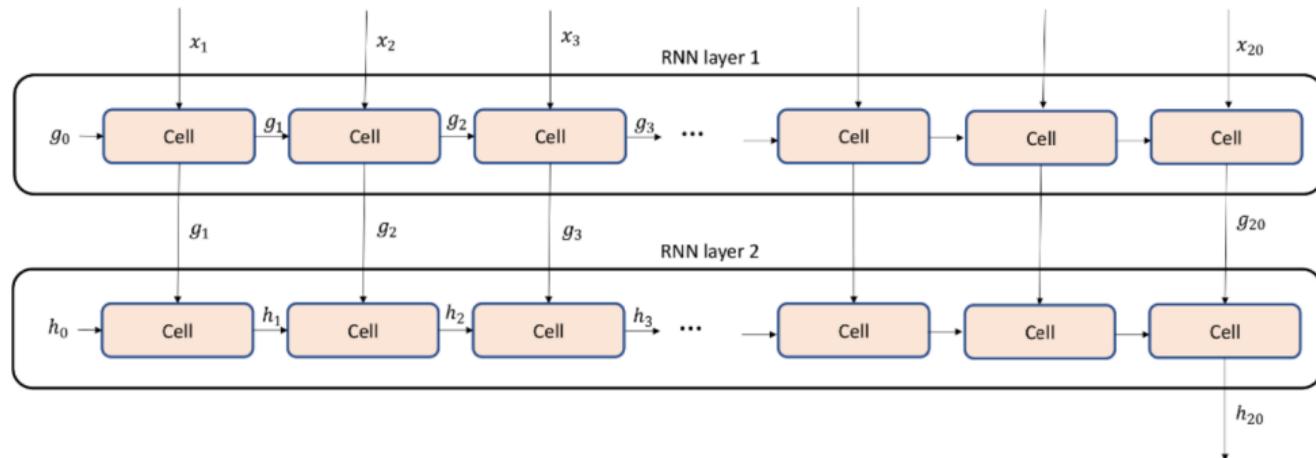
GPT-2

GPT-3

Notes and Further
Reading

More advanced networks

- Use multiple stacked LSTM layers to learn deeper features



Copyright 2019 Applied Data Science Partners Ltd., [Fos19]

[Introduction](#)

[ImageNet](#)

[AlexNet](#)

[RNNs and LSTMs](#)

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

[Transformers](#)

BERT

GPT-2

GPT-3

[Notes and Further
Reading](#)

Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

Transformers

BERT
GPT-2
GPT-3

Notes and Further
Reading

Other tasks

Not just predicting next word:

Translation: from one language into another (e.g., *English* → *German*)

Question generation: input text, generate question that could be asked about the text

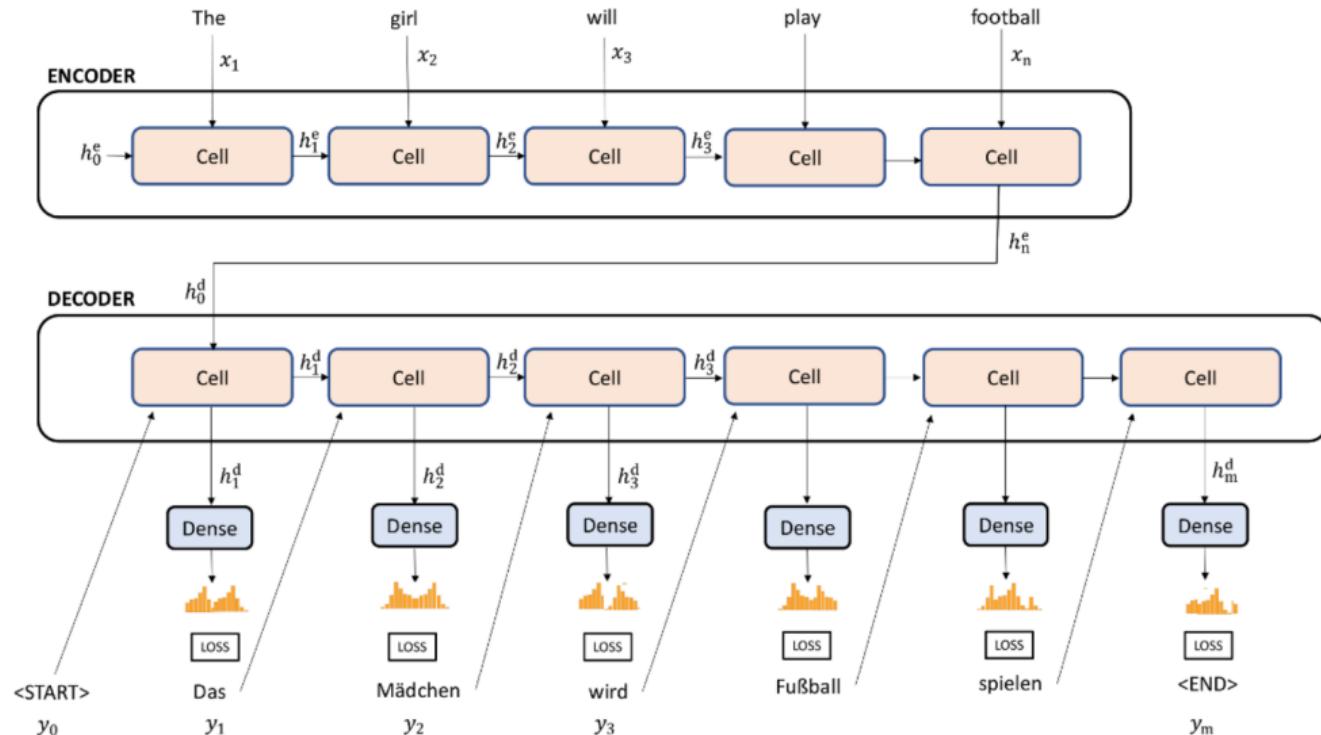
Summarization: condense a longer text into a shorter one

Encoder-Decoder

For this type of problem, we use an **encoder-decoder** network

Encoder-Decoder Network for Translation

René Witte



Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks (RNNs)

Long Short-Term Memory Networks (LSTMs)

Sequence-to-Sequence Models

Transformers

BERT

GPT-2

GPT-3

Notes and Further Reading

Given an input text

The winning goal was scored by 23-year-old striker Joe Bloggs during the match between Arsenal and Barcelona . Arsenal recently signed the striker for 50 million pounds . The next match is in two weeks time, on July 31st 2005 . "

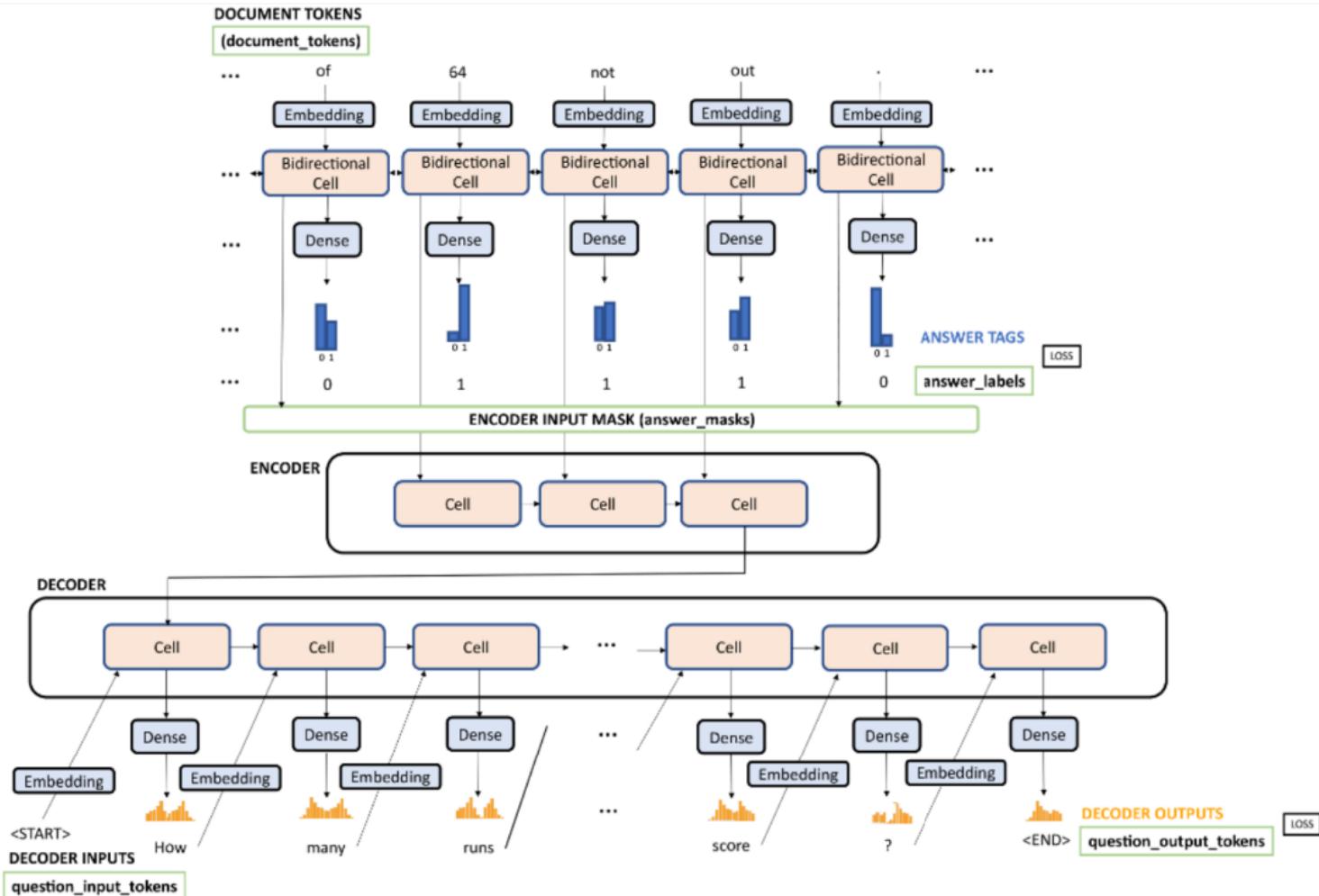
Answer Generation Network

"Joe Bloggs"
"Arsenal"
"Barcelona"
"50 million pounds"
"July 31st 2005"

Question Generation Network

"Who scored the winning goal?"
"Who won the match?"
"Who were Arsenal playing?"
"How much did the striker cost?"
"When is the next match?"

[Introduction](#)[ImageNet](#)[AlexNet](#)[RNNs and LSTMs](#)[Recurrent Neural Networks \(RNNs\)](#)[Long Short-Term Memory Networks \(LSTMs\)](#)[Sequence-to-Sequence Models](#)[Transformers](#)[BERT](#)[GPT-2](#)[GPT-3](#)[Notes and Further Reading](#)



Model in Keras for generating question-answer pairs

René Witte



```
from keras.layers import Input, Embedding, GRU, Bidirectional, Dense, Lambda
from keras.models import Model, load_model
import keras.backend as K
from qgen.embedding import glove

#### PARAMETERS ####
VOCAB_SIZE = glove.shape[0] # 9984
EMBEDDING_DIMENS = glove.shape[1] # 100
GRU_UNITS = 100
DOC_SIZE = None
ANSWER_SIZE = None
Q_SIZE = None

document_tokens = Input(shape=(DOC_SIZE,), name="document_tokens")

embedding = Embedding(input_dim = VOCAB_SIZE, output_dim = EMBEDDING_DIMENS,
                      weights=[glove], mask_zero = True, name = 'embedding')
document_emb = embedding(document_tokens)

answer_outputs = Bidirectional(GRU(GRU_UNITS, return_sequences=True),
                               name = 'answer_outputs')(document_emb)
answer_tags = Dense(2, activation = 'softmax',
                   name = 'answer_tags')(answer_outputs)
```

Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

Transformers

BERT

GPT-2

GPT-3

Notes and Further
Reading

Chosen answer range 6 : 9

['23-year-old', 'striker', 'joe', 'bloggs']

Predicted Question

<START> who scored the <UNK> ? <END>

Chosen answer range 14 : 16

['arsenal', 'and', 'barcelona']

Predicted Question

<START> who defeated <UNK> ? <END>

Chosen answer range 24 : 26

50 million pounds

Predicted Question

<START> how much money did he lose ? <END>

0 the 0.21

1 winning 0.18

2 goal 0.18

3 was 0.13

4 scored 0.16

5 by 0.13

6 23-year-old 0.72

7 striker 0.72

8 joe 0.84

9 bloggs 0.68

10 during 0.07

11 the 0.08

12 match 0.11

13 between 0.12

14 arsenal 0.25

15 and 0.25

16 barcelona 0.26

17 . 0.01

18 arsenal 0.31

19 recently 0.09

20 signed 0.06

21 the 0.08

22 striker 0.17

23 for 0.17

24 50 0.49

25 million 0.56

26 pounds 0.48

27 . 0.0

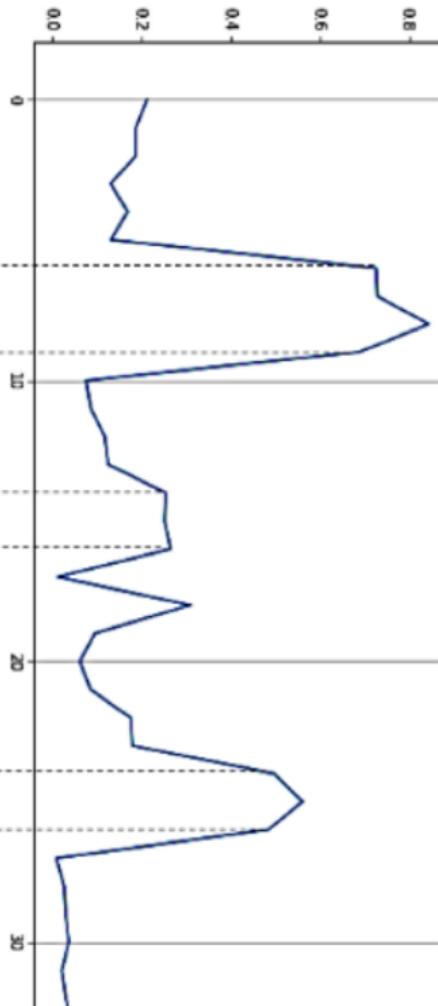
28 the 0.02

29 next 0.03

30 match 0.03

31 is 0.02

32 in 0.03



Outline

René Witte



1 Introduction

Introduction

ImageNet

AlexNet

2 RNNs and LSTMs

RNNs and LSTMs

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

3 Transformers

Transformers

BERT

GPT-2

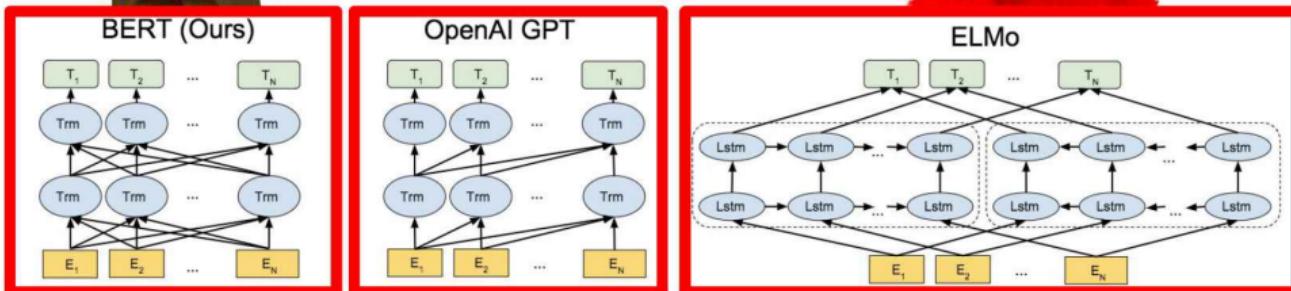
GPT-3

4 Notes and Further Reading

Notes and Further
Reading



BERT



<https://www.infoq.com/presentations/nlp-practitioners/>

Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks (RNNs)

Long Short-Term Memory Networks (LSTMs)

Sequence-to-Sequence Models

Transformers

BERT

GPT-2

GPT-3

Notes and Further Reading



The latest news from Google AI

Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing

Friday, November 2, 2018

Posted by Jacob Devlin and Ming-Wei Chang, Research Scientists, Google AI Language

<https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html>



Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

Transformers

BERT

GPT-2

GPT-3

Notes and Further
Reading

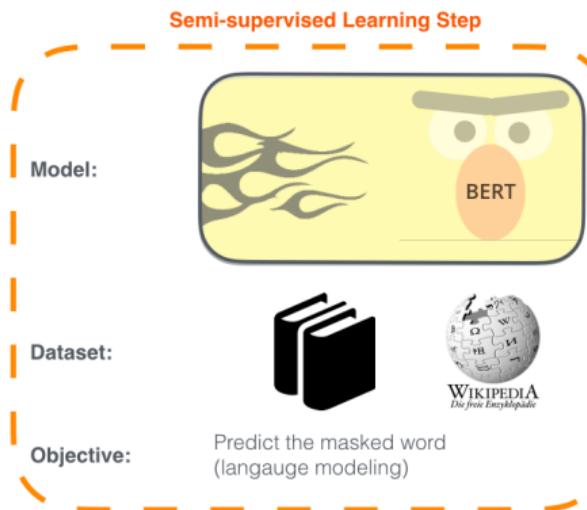
BERT Transfer Learning

René Witte

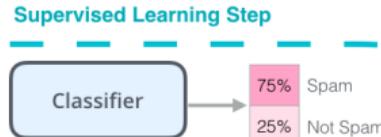


1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.



2 - Supervised training on a specific task with a labeled dataset.



Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks (RNNs)

Long Short-Term Memory Networks (LSTMs)

Sequence-to-Sequence Models

Transformers

BERT

GPT-2

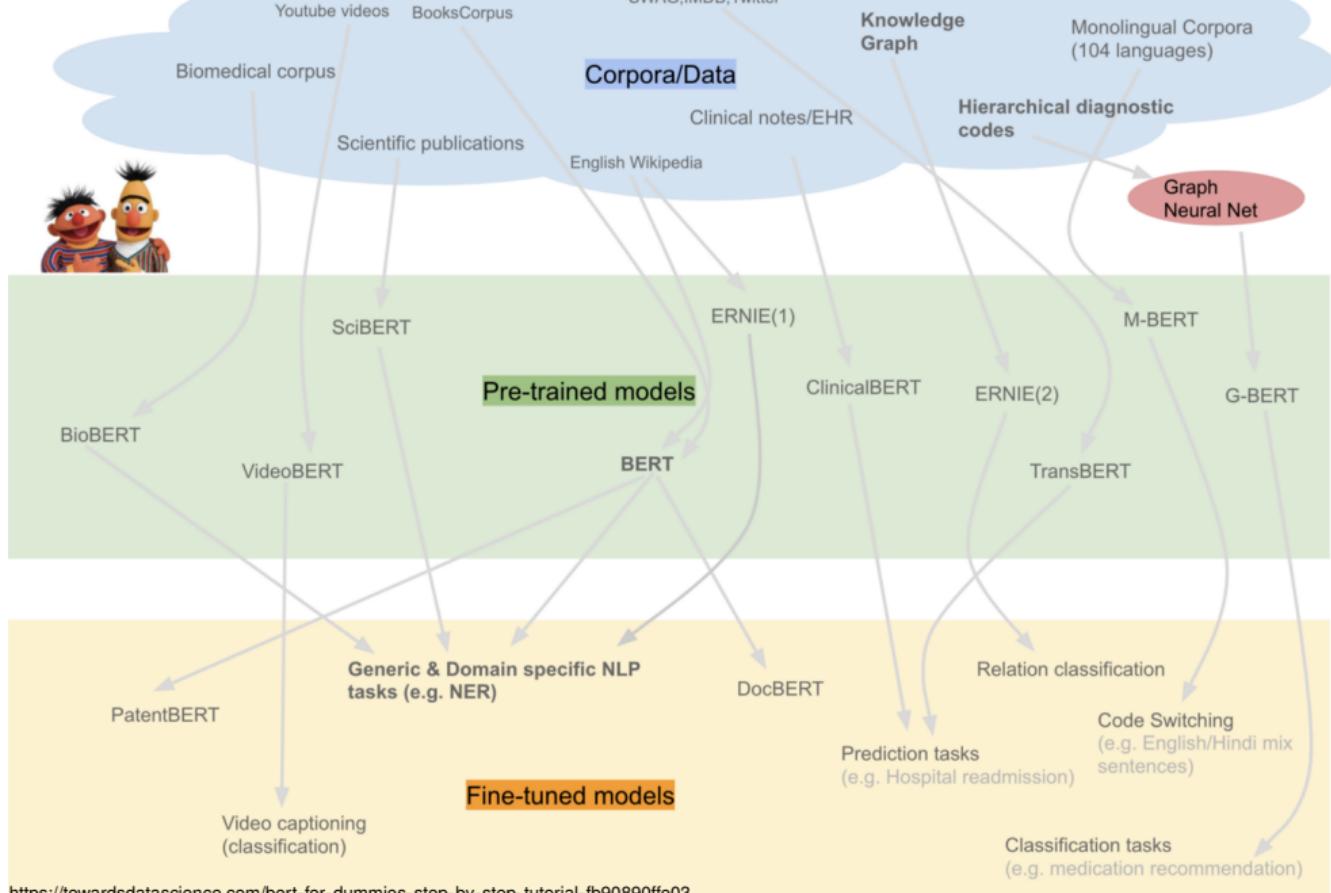
GPT-3

Notes and Further Reading

[The Illustrated BERT, ELMo, and co., <http://jalammar.github.io/illustrated-bert/>]

BERT Pre-trained Variants

René Witte



Introduction

ImageNet
AlexNet

RNNs and LSTMs

Recurrent Neural Networks (RNNs)
Long Short-Term Memory Networks (LSTMs)
Sequence-to-Sequence Models

Transformers

BERT
GPT-2
GPT-3

Notes and Further Reading

The Stanford Question Answering Dataset (SQuAD)

René Witte



SQuAD

Home Explore 2.0 Explore 1.1

Following the Cretaceous–Paleogene extinction event, the extinction of the dinosaurs and the wetter climate may have allowed the tropical rainforest to spread out across the continent. From 66–34 Mya, the rainforest extended as far south as 45°. Climate fluctuations during the last 34 million years have allowed savanna regions to expand into the tropics. During the Oligocene, for example, the rainforest spanned a relatively narrow band. It expanded again during the Middle Miocene, then retracted to a mostly inland formation at the last glacial maximum. However, the rainforest still managed to thrive during these glacial periods, allowing for the survival and evolution of a broad diversity of species.

Which type of climate may have allowed the rainforest to spread across the continent?

Ground Truth Answers: the wetter climate may have allowed the tropical rainforest to spread out across the continent. wetter wetter

Prediction: wetter

What has allowed for the Savanna region to expand into the tropics?

Ground Truth Answers: Climate fluctuations during the last 34 million years have allowed savanna regions to expand into the tropics. Climate fluctuations during the last 34 million years Climate fluctuations

Prediction: Climate fluctuations

During what time did the rainforest spanned a narrow band?

Ground Truth Answers: During the Oligocene, for example, the rainforest spanned a relatively narrow band. Oligocene Oligocene

Prediction: Oligocene

When did it retract to a inland formation?

Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks (RNNs)

Long Short-Term Memory Networks (LSTMs)

Sequence-to-Sequence Models

Transformers

BERT

GPT-2

GPT-3

Notes and Further Reading

SQuAD1.1 Leaderboard

Rank	Model	EM	F1
	Human Performance <i>Stanford University</i> (Rajpurkar et al. '16)	82.304	91.221
1	BERT (ensemble) <i>Google AI Language</i> https://arxiv.org/abs/1810.04805	87.433	93.160
2	nlnet (ensemble) <i>Microsoft Research Asia</i>	85.356	91.202
3	QANet (ensemble) <i>Google Brain & CMU</i>	84.454	90.490

[Introduction](#)[ImageNet](#)
[AlexNet](#)[RNNs and LSTMs](#)[Recurrent Neural Networks \(RNNs\)](#)
[Long Short-Term Memory Networks \(LSTMs\)](#)
[Sequence-to-Sequence Models](#)[Transformers](#)[BERT](#)
[GPT-2](#)
[GPT-3](#)[Notes and Further Reading](#)

BERT is open source, so train it yourself?

"A DGX-2 server with 16 V100s trained BERT-Large in under 3 days"

<https://devblogs.nvidia.com/training-bert-with-gpus/>



NVIDIA TESLA V100 FOR PCIe

<https://www.nvidia.com/en-gb/data-center/tesla-v100/>

Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

Transformers

BERT

GPT-2

GPT-3

Notes and Further
Reading



Search models, datasets, users...

Models

Datasets

Pricing

Resources

We're hiring!

Log In

Sign Up

Announcing AutoNLP: A new automatic way to train and deploy NLP models.



The AI community building the future.

Build, train and deploy state of the art models powered by the reference open source in natural language processing.

Star

43,786



Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks (RNNs)

Long Short-Term Memory Networks (LSTMs)

Sequence-to-Sequence Models

Transformers

BERT

GPT-2

GPT-3

Notes and Further Reading

<https://huggingface.co/>

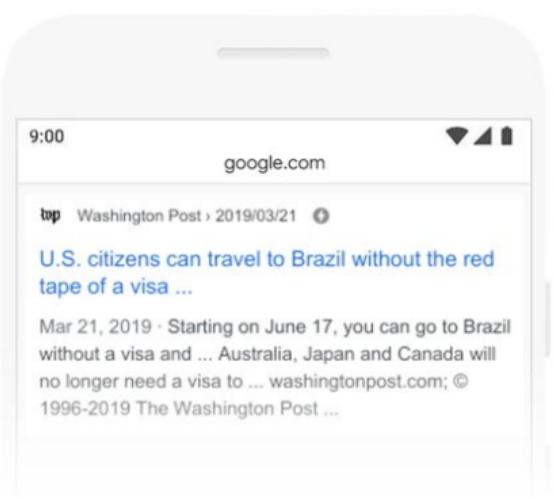
BERT in Google Search (I)

René Witte

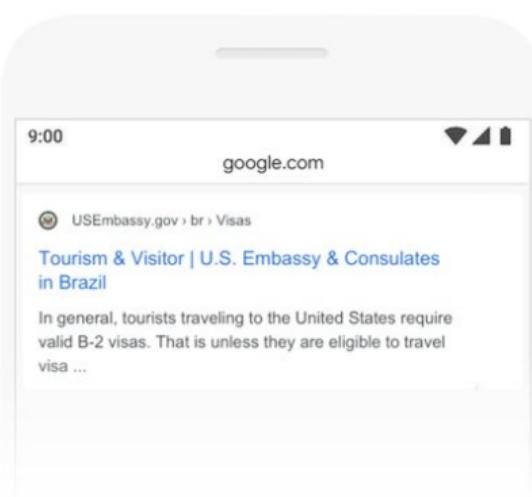


2019 brazil traveler to usa need a visa

BEFORE



AFTER



Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks (RNNs)

Long Short-Term Memory Networks (LSTMs)

Sequence-to-Sequence Models

Transformers

BERT

GPT-2

GPT-3

Notes and Further Reading

BERT in Google Search (II)

René Witte



Can you get medicine for someone pharmacy

BEFORE

A screenshot of a smartphone displaying a Google search result. The search bar at the top contains the query "Can you get medicine for someone pharmacy". Below the search bar, the word "BEFORE" is centered. The main content area shows a search result from "MedlinePlus (.gov)" titled "Getting a prescription filled: MedlinePlus Medical Encyclopedia". The snippet text reads: "Aug 26, 2017 · Your health care provider may give you a prescription in ... Writing a paper prescription that you take to a local pharmacy ... Some people and insurance companies choose to use ..." The status bar at the top of the phone screen shows the time as 9:00 and signal strength indicators.

AFTER

A screenshot of a smartphone displaying a Google search result. The search bar at the top contains the query "Can you get medicine for someone pharmacy". Below the search bar, the word "AFTER" is centered. The main content area shows a search result from "HHS.gov" titled "Can a patient have a friend or family member pick up a prescription ...". The snippet text reads: "Dec 19, 2002 · A pharmacist may use professional judgment and experience with common practice to ... the patient's best interest in allowing a person, other than the patient, to pick up a prescription." The status bar at the top of the phone screen shows the time as 9:00 and signal strength indicators.

Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks (RNNs)

Long Short-Term Memory Networks (LSTMs)

Sequence-to-Sequence Models

Transformers

BERT

GPT-2

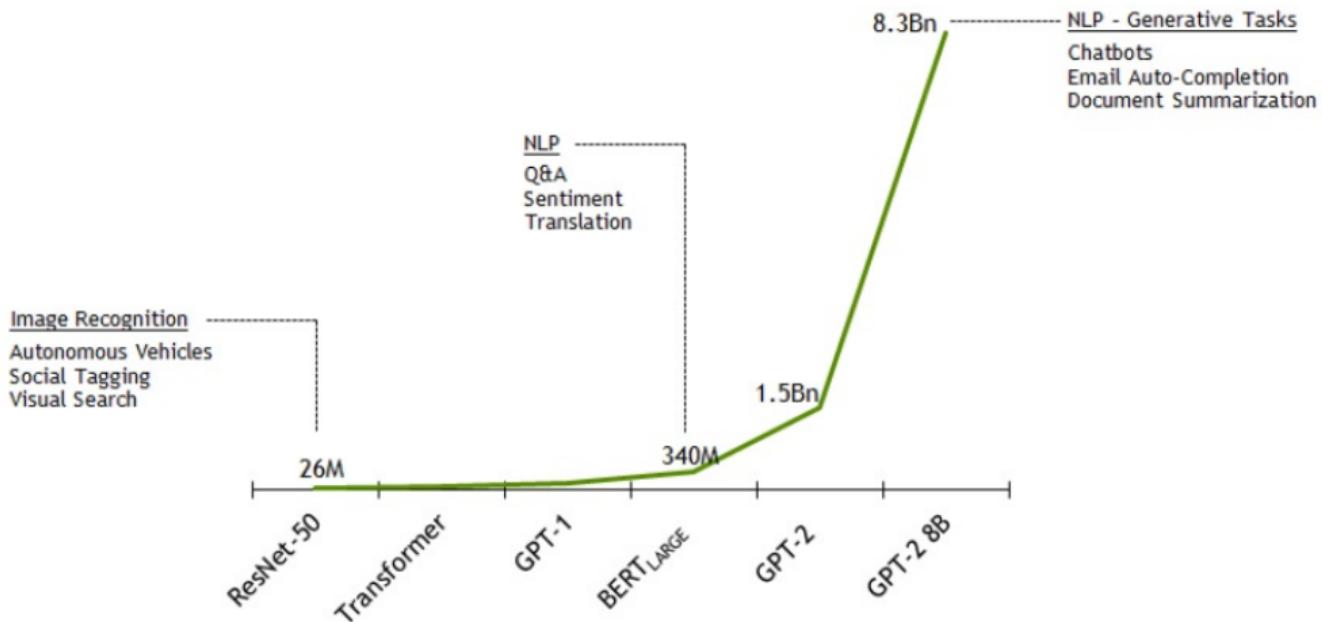
GPT-3

Notes and Further Reading

Started rollout for English Google search in October 21, 2019

Model size comparison

René Witte



<https://devblogs.nvidia.com/training-bert-with-gpus/>

Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks (RNNs)

Long Short-Term Memory Networks (LSTMs)

Sequence-to-Sequence Models

Transformers

BERT

GPT-2

GPT-3

Notes and Further Reading

Transformer Architecture

René Witte



Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

Transformers

BERT

GPT-2

GPT-3

Notes and Further
Reading



The Illustrated Transformer, <http://jalammar.github.io/illustrated-transformer/>

Transformer Architecture (II)

René Witte



Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

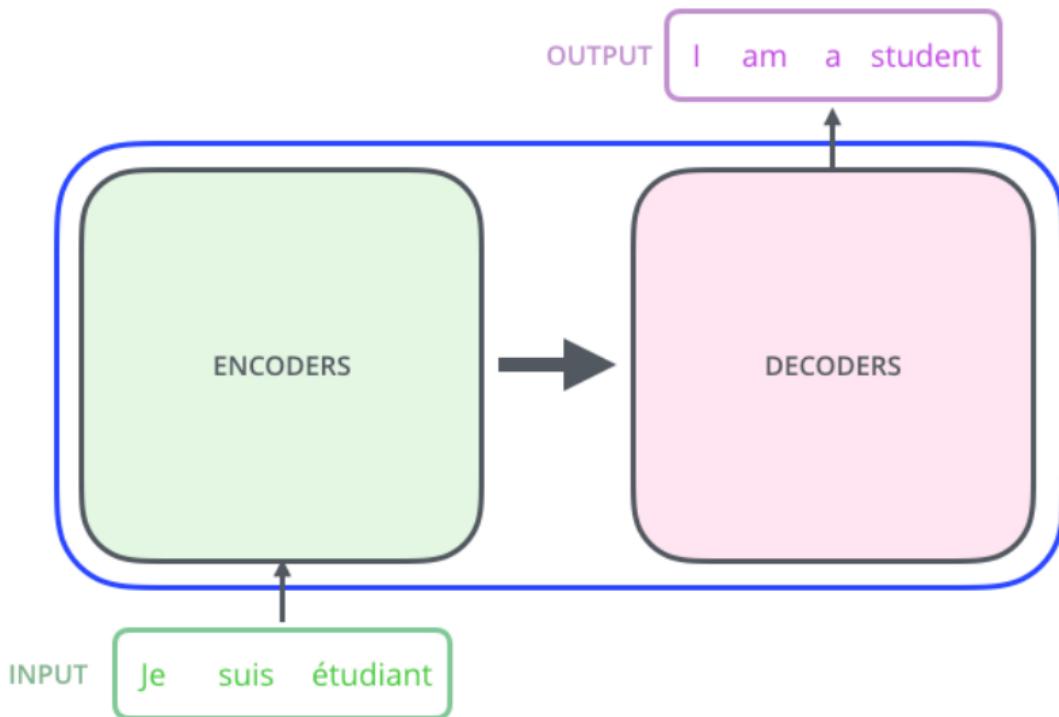
Transformers

BERT

GPT-2

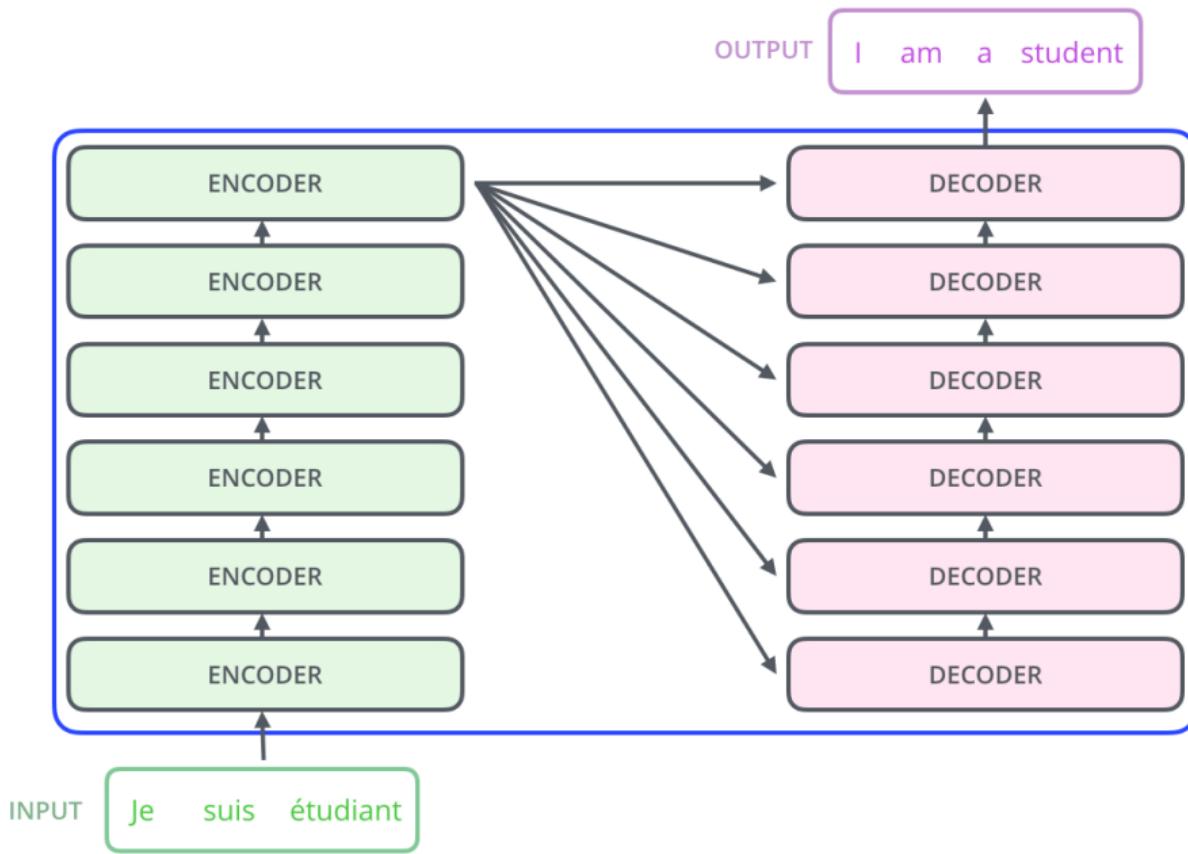
GPT-3

Notes and Further
Reading



Transformer Architecture (III)

René Witte



Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks (RNNs)

Long Short-Term Memory Networks (LSTMs)

Sequence-to-Sequence Models

Transformers

BERT

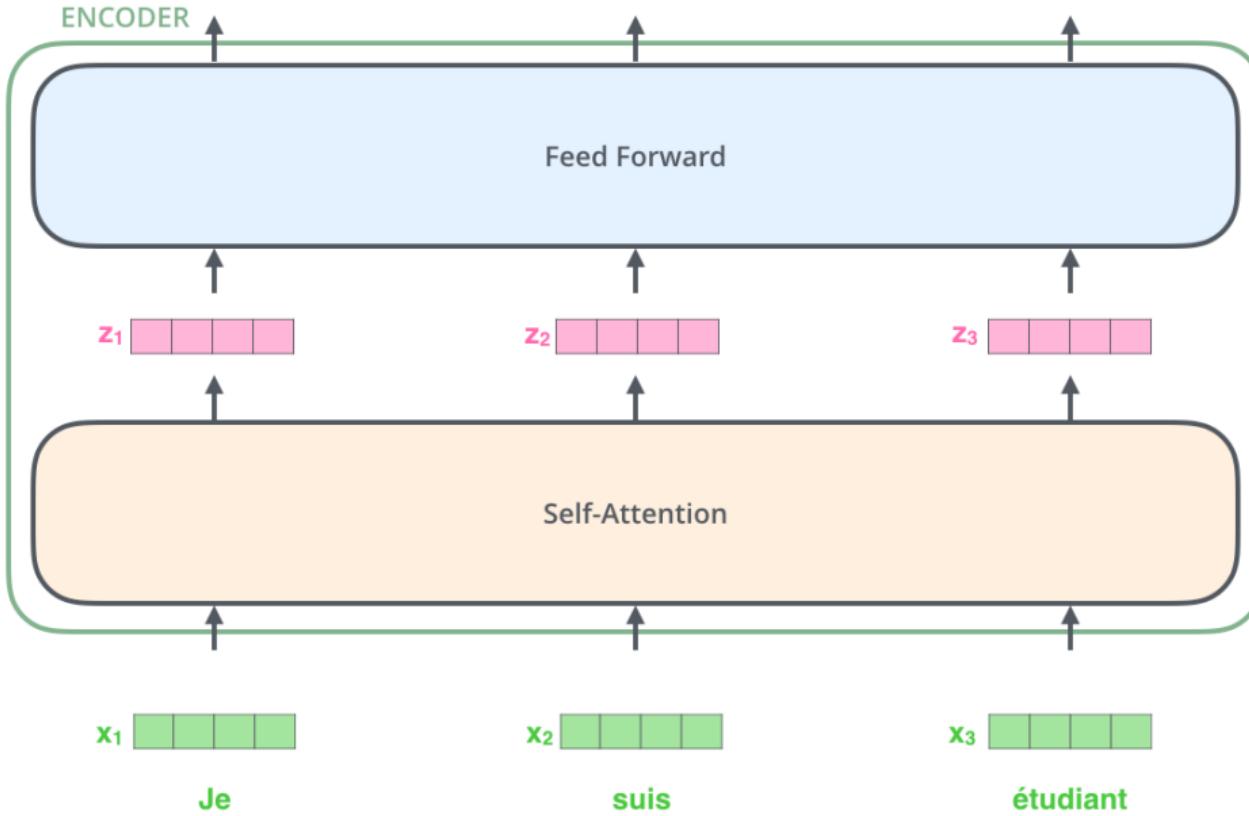
GPT-2

GPT-3

Notes and Further Reading

Transformer Architecture (IV)

René Witte



Introduction

ImageNet
AlexNet

RNNs and LSTMs

Recurrent Neural Networks (RNNs)
Long Short-Term Memory Networks (LSTMs)
Sequence-to-Sequence Models

Transformers

BERT
GPT-2
GPT-3

Notes and Further Reading

Transformer Architecture (V)

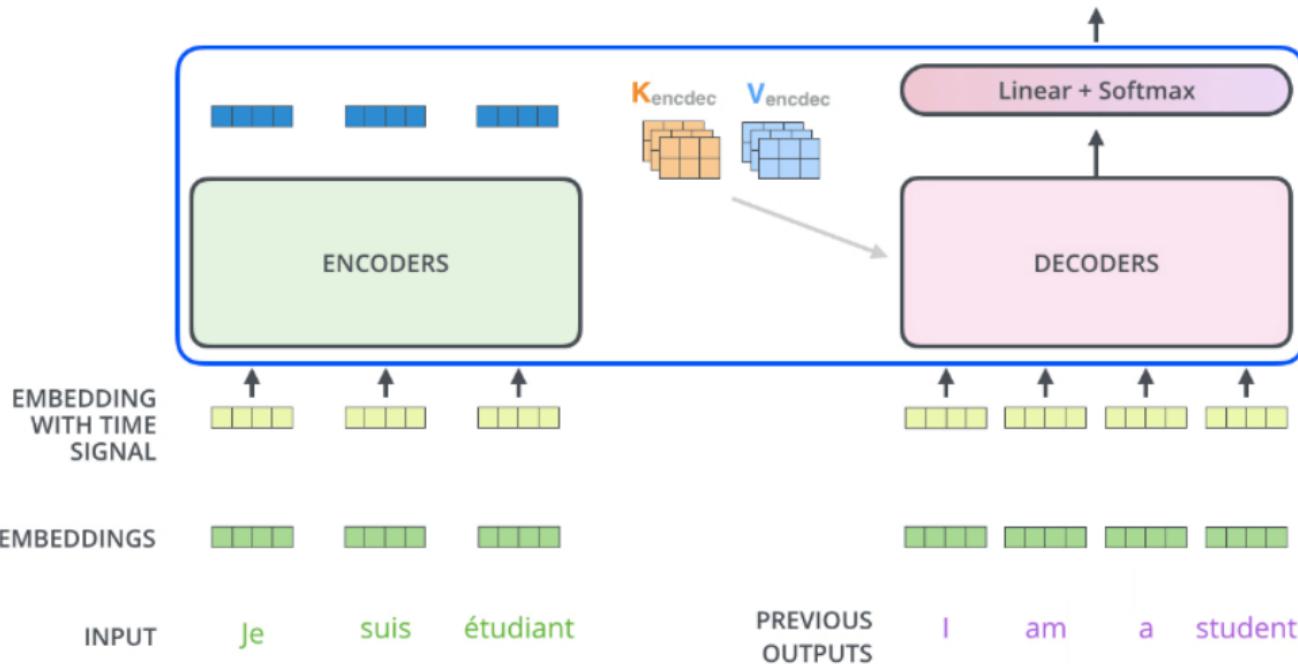
René Witte



Decoding time step: 1 2 3 4 5 6

OUTPUT

I am a student



Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks (RNNs)

Long Short-Term Memory Networks (LSTMs)

Sequence-to-Sequence Models

Transformers

BERT

GPT-2

GPT-3

Notes and Further Reading

[The Illustrated Transformer, <http://jalammar.github.io/illustrated-transformer/>]

Human Input

“In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.”

GPT-2 Generated Output

“The scientist named the population, after their distinctive horn, Ovid’s Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. “By the time we reached the top of one peak, the water looked blue, with some crystals on top,” said Pérez.”

...

See <https://openai.com/blog/better-language-models/#sample1>

Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks (RNNs)

Long Short-Term Memory Networks (LSTMs)

Sequence-to-Sequence Models

Transformers

BERT

GPT-2

GPT-3

Notes and Further Reading

Another example...

René Witte



Human Input

"For today's homework assignment, please describe the reasons for the US Civil War."

GPT-2 Generated Output

"It is easy to identify why the Civil War happened, because so many people and so many books and so much television and films tell us that it was the cause, that it has something to do with race or economics or religion. There is much agreement that it was essentially a war of slavery on behalf of capitalism, about a century of slavery. But that's not what most people think of when they think of the Civil War. Many people think of the war as about states rights. You're not wrong about that, but it's kind of misleading to say that the Civil War was a conflict between states' rights and federalism. So let's try again. What's the idea that has inspired so many people to go onto the battlefields of Gettysburg and Antietam and Gettysberg and Petersburg and Fredericksburg? The American idea of the republic—a notion of limited government—is a great part of the history...."

See <https://openai.com/blog/better-language-models/#sample6>

Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks (RNNs)

Long Short-Term Memory Networks (LSTMs)

Sequence-to-Sequence Models

Transformers

BERT

GPT-2

GPT-3

Notes and Further Reading

GPT-2 Release Controversy

René Witte



Initial Announcement

- OpenAI announced GPT-2 in February 14, 2019
<https://openai.com/blog/better-language-models/>
- They initially refused to make the model or training data public, saying that it is **too dangerous**, with potential misuse for generating fake news or for phishing purposes, identity theft, manipulation of social media content, etc.
- This was harshly criticized by many members of the research community



Anima Anandkumar @AnimaAnandkumar · Feb 14

What you are doing is opposite of open. It is unfortunate that you hype up +propagate fear + thwart reproducibility+scientific endeavor. There is active research from other groups in unsupervised language models. You hype it up like it has never been done before. [@jackclarkSF](#)

4

31

223

||

<https://anima-ai.org/2019/02/18/an-open-and-shut-case-on-openai/>

GPT-2 Release

GPT-2 was finally released on November 5, 2019:

<https://openai.com/blog/gpt-2-1-5b-release/>

Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks (RNNs)

Long Short-Term Memory Networks (LSTMs)

Sequence-to-Sequence Models

Transformers

BERT

GPT-2

GPT-3

Notes and Further Reading

OpenAI's current model: GPT-3

Introduced in a July 2020 paper (<https://arxiv.org/pdf/2005.14165.pdf>)

- 96 layers and 175 billion parameters
- cost to train: est. \$5 million
- cost to run (yearly cloud costs): est. \$87,000

See <https://bdtechtalks.com/2020/09/21/gpt-3-economy-business-model/>

Now fully commercial

- OpenAI moved from non-profit to for-profit in 2019
- GPT-3 is only available via paid cloud API
- Microsoft exclusively licensed the GPT-3 source code
<https://openai.com/blog/openai-licenses-gpt-3-technology-to-microsoft/>

[Introduction](#)

[ImageNet](#)

[AlexNet](#)

[RNNs and LSTMs](#)

[Recurrent Neural Networks \(RNNs\)](#)

[Long Short-Term Memory Networks \(LSTMs\)](#)

[Sequence-to-Sequence Models](#)

[Transformers](#)

[BERT](#)

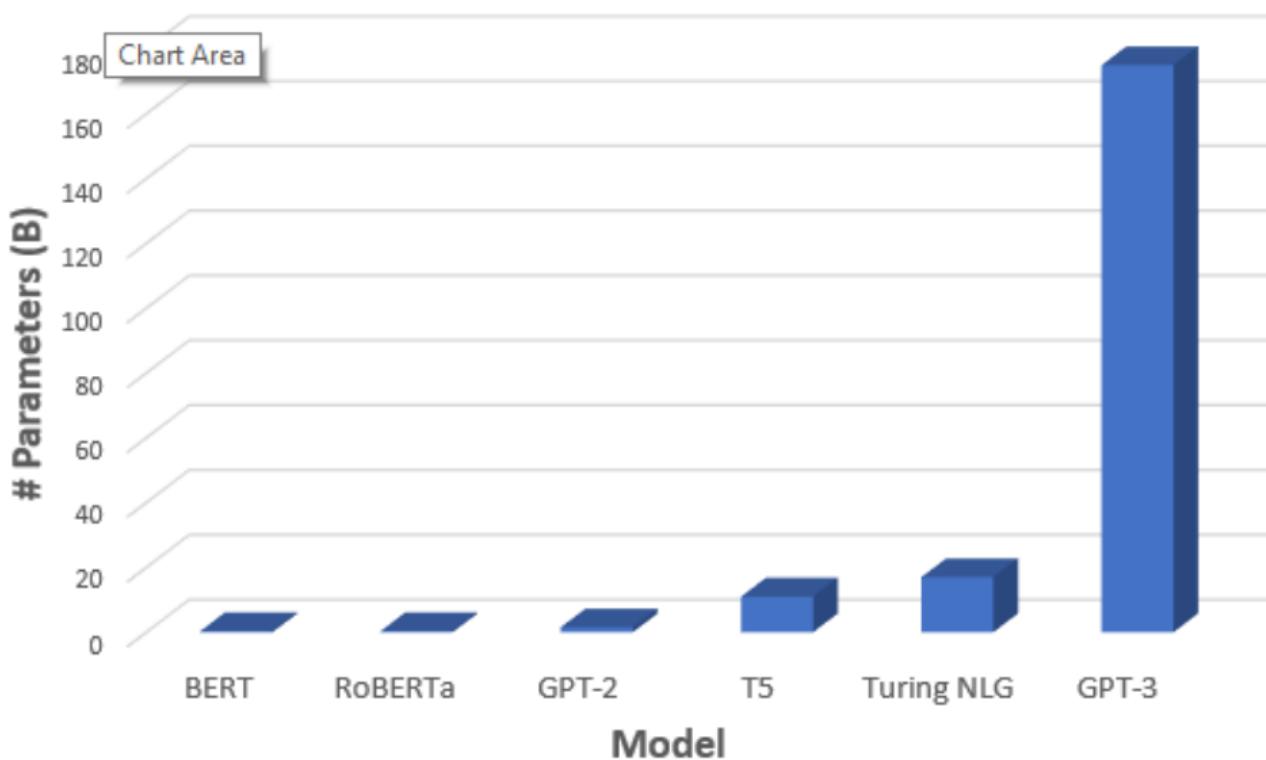
[GPT-2](#)

[GPT-3](#)

[Notes and Further Reading](#)

Model size comparison

René Witte



Introduction

ImageNet
AlexNet

RNNs and LSTMs

Recurrent Neural Networks (RNNs)
Long Short-Term Memory Networks (LSTMs)
Sequence-to-Sequence Models

Transformers

BERT
GPT-2
GPT-3

Notes and Further Reading

GPT-3 can basically do all NLP tasks

- Generate text
- Generate code
- Answer questions
- Chat
- Summarization
- Translate between languages
- ...

See some examples at

<https://www.twilio.com/blog/ultimate-guide-openai-gpt-3-language-model>

Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

Transformers

BERT

GPT-2

GPT-3

Notes and Further
Reading

Example application: Generating press releases

René Witte



Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

Transformers

BERT

GPT-2

GPT-3

Notes and Further
Reading

TESLA UNVEILS 'PROJECT MORPHEUS' FLYING CAR CAPABLE OF REACHING FIGHTER JET LEVEL SPEEDS

BAKZ.T.
FUTURE



📍 TORONTO

Generating Press Releases with GPT-3

<https://www.youtube.com/watch?v=rFZ75HkY9bo>

GPT-3: A billion times more complex than Eliza, but does it understand anything?

René Witte



The screenshot shows an Emacs terminal window titled "emacs@matrix <2>". The window contains the following text:

```
File Edit Options Buffers Tools Text Help
I am the psychotherapist. Please, describe your problems. Each time you are
finished talking, type RET twice.

I am scared of GPT-3!

Why do you say you are afraid of gpt 3?

Guess this is the end...

What makes you believe guess this is the end?

U:**- *doctor*      All (12,0)      (Doctor Fill)
```

[Introduction](#)

[ImageNet](#)

[AlexNet](#)

[RNNs and LSTMs](#)

[Recurrent Neural Networks \(RNNs\)](#)

[Long Short-Term Memory Networks \(LSTMs\)](#)

[Sequence-to-Sequence Models](#)

[Transformers](#)

[BERT](#)

[GPT-2](#)

[GPT-3](#)

[Notes and Further Reading](#)

Outline

René Witte



1 Introduction

Introduction

ImageNet

AlexNet

2 RNNs and LSTMs

RNNs and LSTMs

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

3 Transformers

Transformers

BERT

GPT-2

GPT-3

4 Notes and Further Reading

Notes and Further
Reading

Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

Transformers

BERT
GPT-2
GPT-3

Notes and Further Reading

Supplemental

- [Fos19, Chapter 6] (LSTM)
- [LHH19, Chapters 8–10] (RNN, LSTM, Seq2Seq)

References

René Witte



[Fos19] David Foster.

Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play.
O'Reilly, 2019.

<https://concordiauniversity.on.worldcat.org/oclc/1136155457>.

[LHH19] Hobson Lane, Cole Howard, and Hannes Max Hapke.

Natural Language Processing in Action.
Manning Publications Co., 2019.

<https://concordiauniversity.on.worldcat.org/oclc/1102387045>.

Introduction

ImageNet

AlexNet

RNNs and LSTMs

Recurrent Neural Networks
(RNNs)

Long Short-Term Memory
Networks (LSTMs)

Sequence-to-Sequence
Models

Transformers

BERT

GPT-2

GPT-3

Notes and Further
Reading