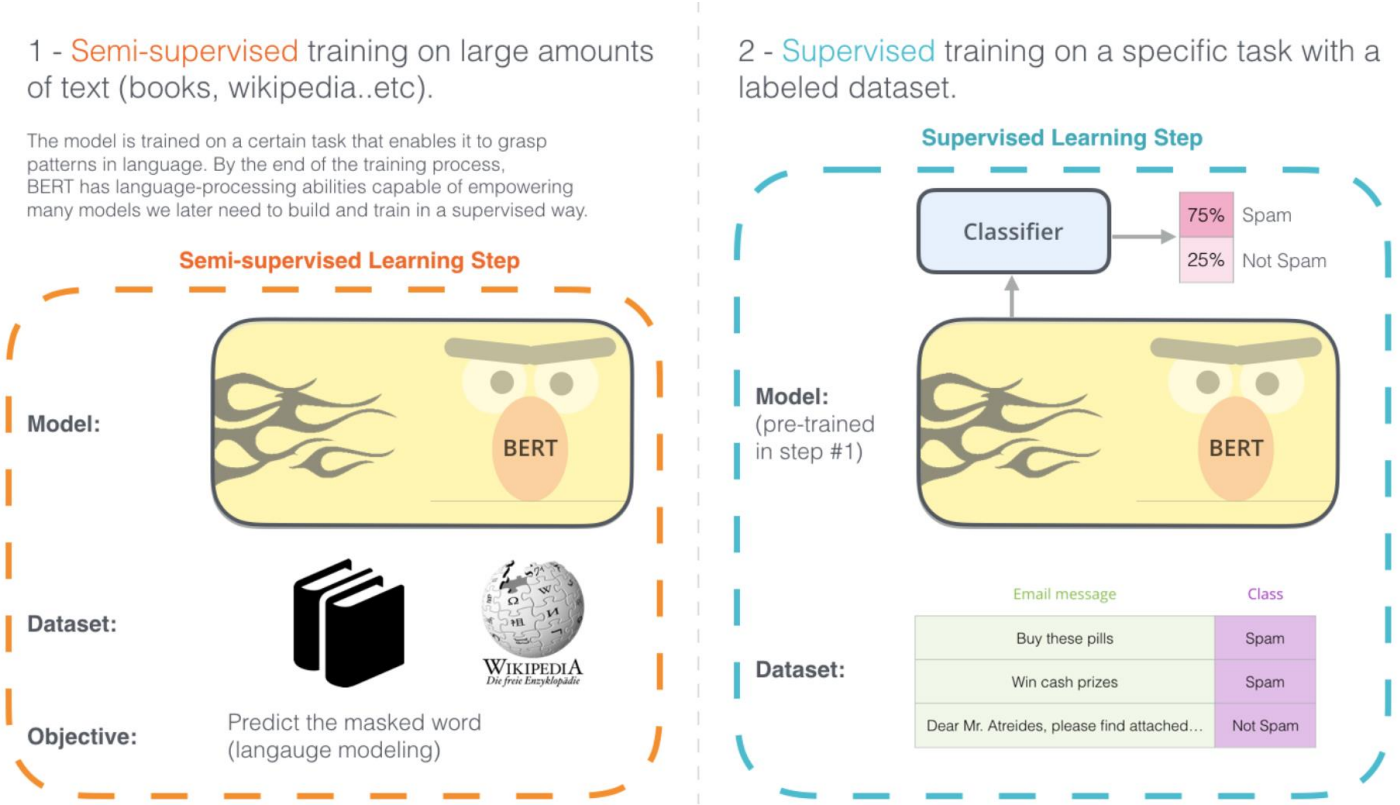
Transfer Learning with BERT Transformer

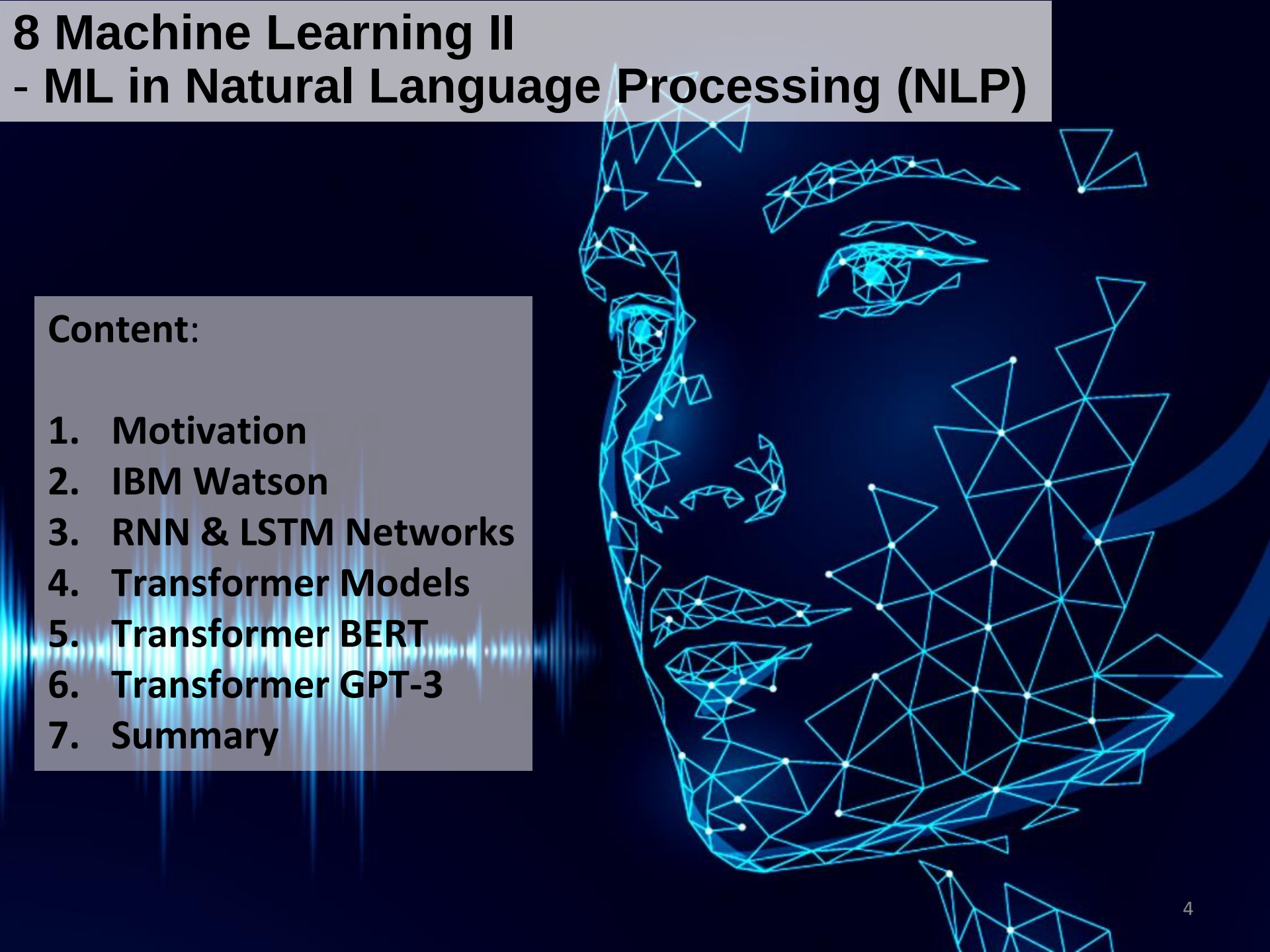


The two steps of how BERT is developed. You can download the model pre-trained in step 1 (trained on un-annotated data), and only worry about fine-tuning it for step 2. [Source for book icon].

Eine Form von **Transfer Learning:**
Fokus auf **Fine-Tuning, Pre-Training** übernehmen

The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning), 2018
jalammar.github.io/illustrated-bert/

8 Machine Learning II
- ML in Natural Language Processing (NLP)

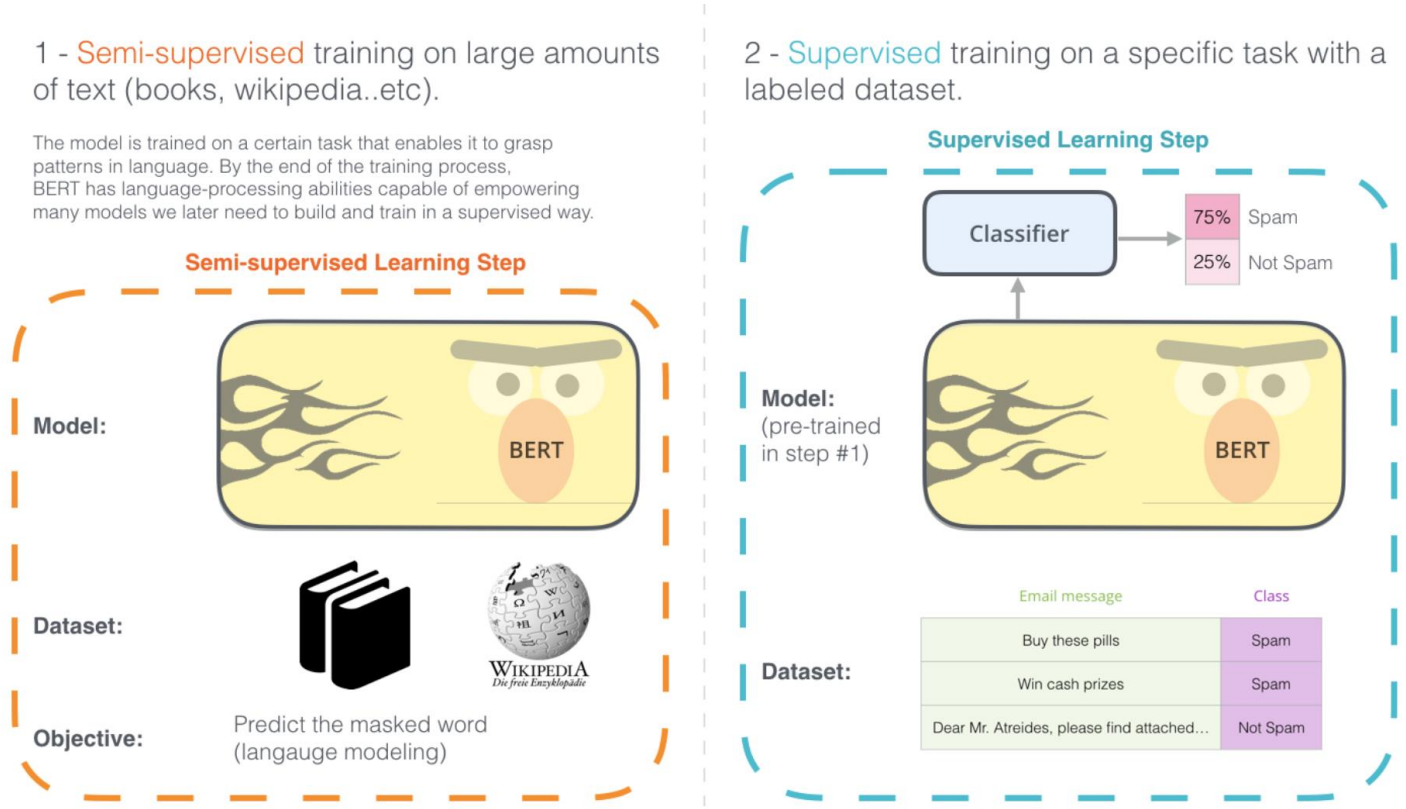


Content:

1. Motivation
2. IBM Watson
3. RNN & LSTM Networks
4. Transformer Models
5. Transformer BERT
6. Transformer GPT-3
7. Summary

8 Machine Learning II
- ML in Natural Language Processing (NLP)
(5) Transformer BERT, Google

BERT Transformer, Google 2018

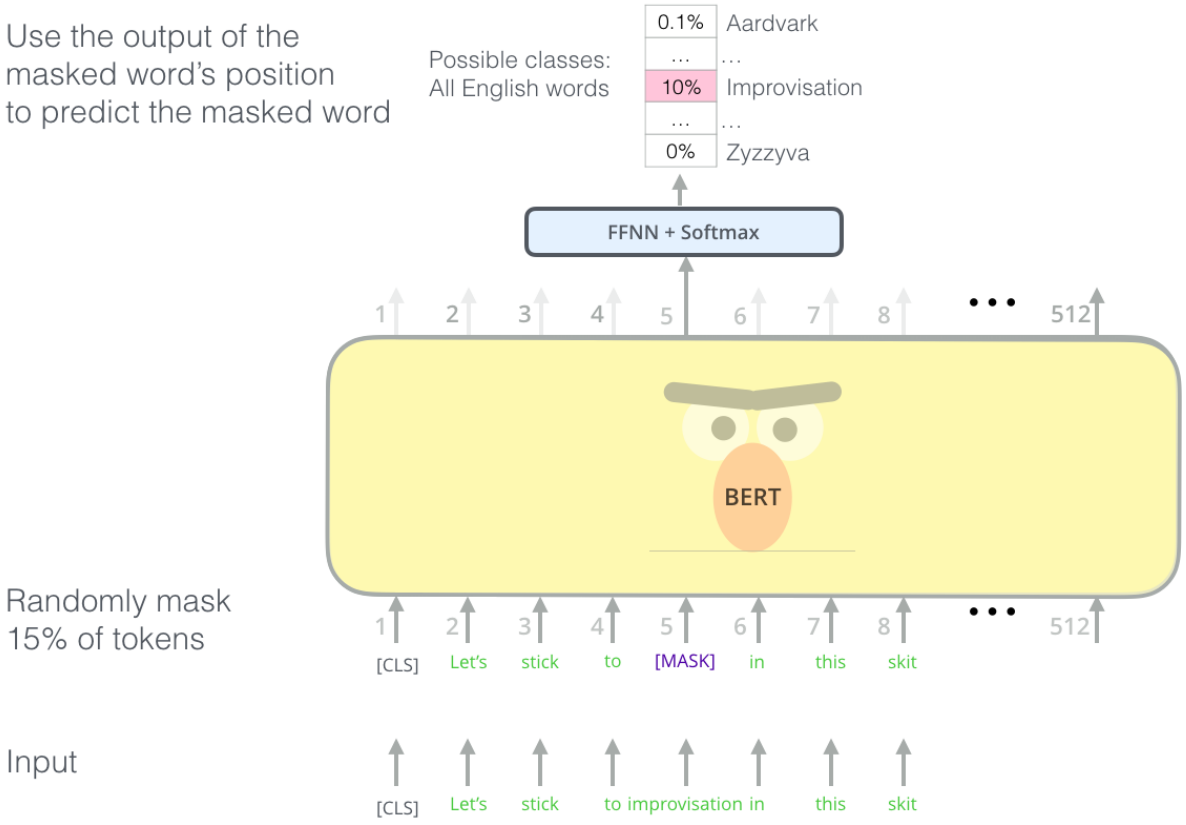


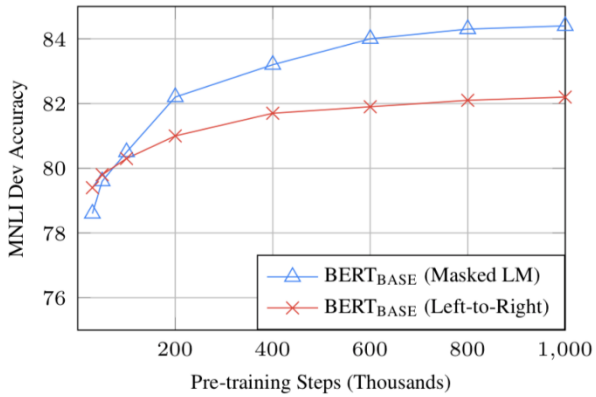
The two steps of how BERT is developed. You can download the model pre-trained in step 1 (trained on un-annotated data), and only worry about fine-tuning it for step 2. [Source for book icon].

Das **Training** läuft in zwei Schritten: **Pre-Training** und **Fine-Tuning** (eine Form von **Transfer Learning**)

The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning), 2018
jalammar.github.io/illustrated-bert/

BERT Training using Masked Language Model

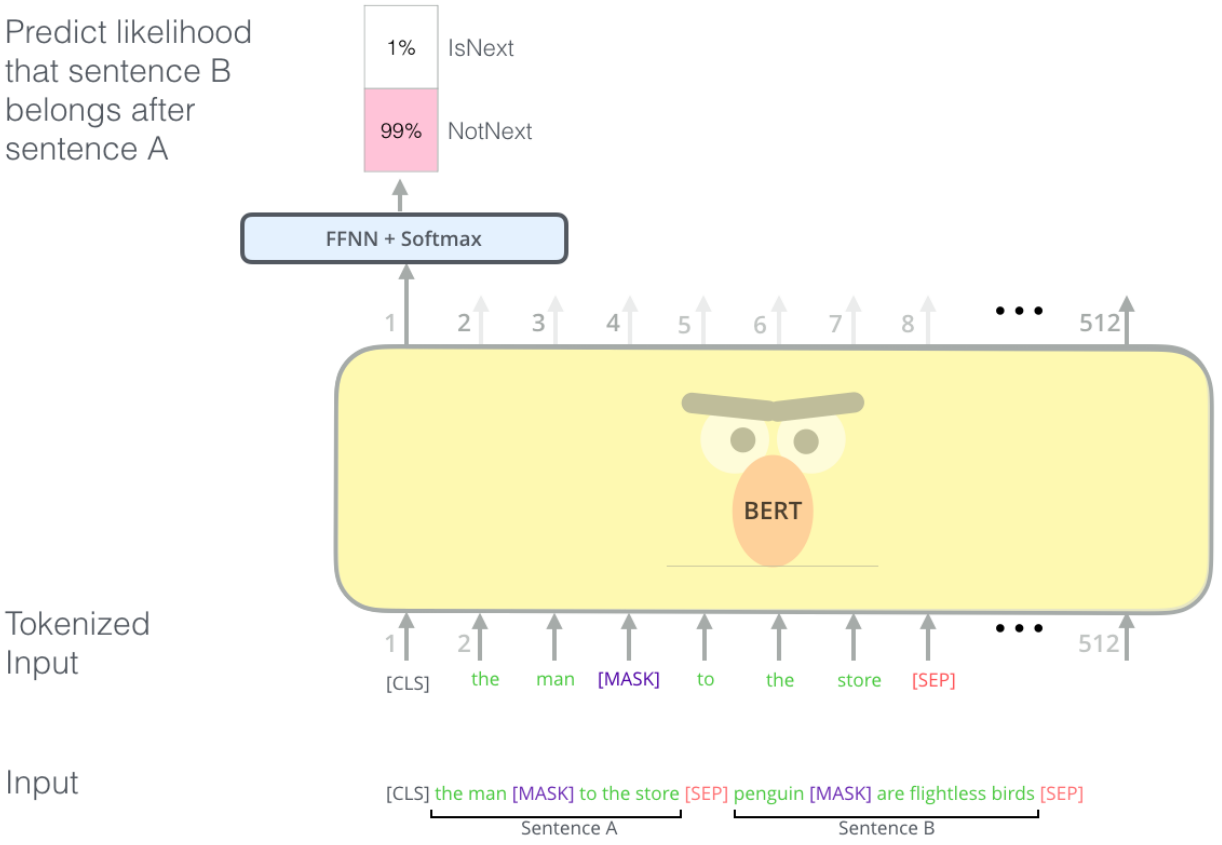




Masked Language Model: **BERT** maskiert **15% der Worte** im **Input** und lässt das fehlende Wort **vorhersagen (Bidirektionale Suche)**.

The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning), 2018
jalammar.github.io/illustrated-bert/

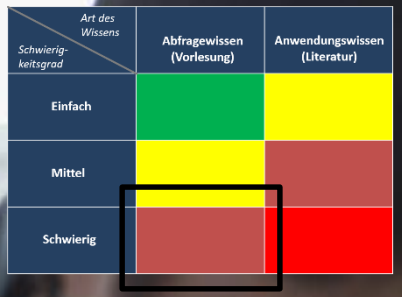
BERT Training using Next Sentence Prediction



Next Sentence Prediction: **BERT** bekommt **2 Sätze** im **Input** und sagt die Reihenfolge (ja/nein) voraus.

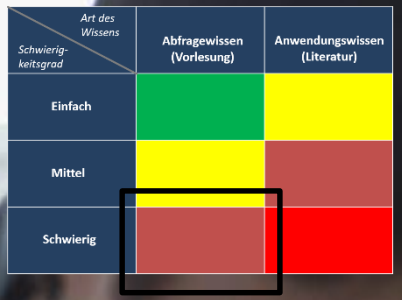
The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning), 2018
jalammar.github.io/illustrated-bert/

About **BERT: Which statement is FALSE?**



a) BERT consists of two stages: **Pre-Training** and **Fine-Tuning**
b) BERT comes with two types: a **Base Model** and **Large Model**
c) **Masked Language Model** is used to **predict Missing Words**
d) **Next Sentence Prediction** allows to **learn Sentence Relationships**
e) The **quality of the text generated by BERT** is **so high** that it is **difficult to distinguish** from that written by a human

Which statement on GPT3 is FALSE?



a) **GPT-3** shows **characteristics** of **zero-shot learning**.
b) **GPT-3** was an **open-source project acquired by Microsoft**.
c) **BERT** and **GPT-3** are widely used **pretrained Transformers**.
d) In terms of **parameters**, **GPT-3** is **smaller than BERT**.
e) GPT-3 uses a **Context Window** of size 2048.

GPT-3 vs BERT?

Both, **GPT-3 and BERT** have been **relatively new** for the industry. Their state-of-the-art performance has made them the **winners among other models** in the **natural language processing field**.
Being trained on 175 billion parameters, **GPT-3** becomes **470 times bigger in size** than **BERT-Large**.

While **BERT** requires an **elaborated fine-tuning process**, **GPT-3's** allows the users to **reprogram it using instructions and access it**.

Case in point — for **sentiment analysis** or **question answering** tasks, to use **BERT**, the users **have to train the model** on a separate layer on sentence encodings. However, **GPT-3** uses a **few-shot learning process** on the input token to predict the output result.

GPT-3 Vs BERT For NLP Tasks
analyticsindiamag.com/gpt-3-vs-bert-for-nlp-tasks/

GPT-3 vs BERT?

On **general NLP tasks** like machine translation, answering questions, complicated arithmetic calculations or learning new words, **GPT-3** works perfectly by **conditioning it with a few examples — few-shot learning**. Similarly, for **text generation** as well, **GPT-3** works on a few prompts to quickly churn out relevant outputs, with an accuracy of approximately 52%. OpenAI, simply, **by increasing the size of the model and its training parameters** created a mighty monster of a model.

While **transformer** includes two separate mechanisms — **encoder** and **decoder**, the **BERT** model only works on **encoding mechanisms** to generate a language model; however, the **GPT-3** combines encoding as well as decoding process to get a **transformer decoder** for producing text.

While **GPT-3** is commercially available **via an API**, but **not open-sourced**, **BERT** has been an **open-source model** since its inception that allows **users to fine-tune it according to their needs**. While **GPT3 generates output one token at a time**, **BERT**, on the other hand, is not autoregressive, thus **uses deep bidirectional context for predicting outcome on sentiment analysis and question answering**.

GPT-3 Vs BERT For NLP Tasks
analyticsindiamag.com/gpt-3-vs-bert-for-nlp-tasks/

OpenAI: Image GPT, Juni 2020

Generative Pretraining from Pixels

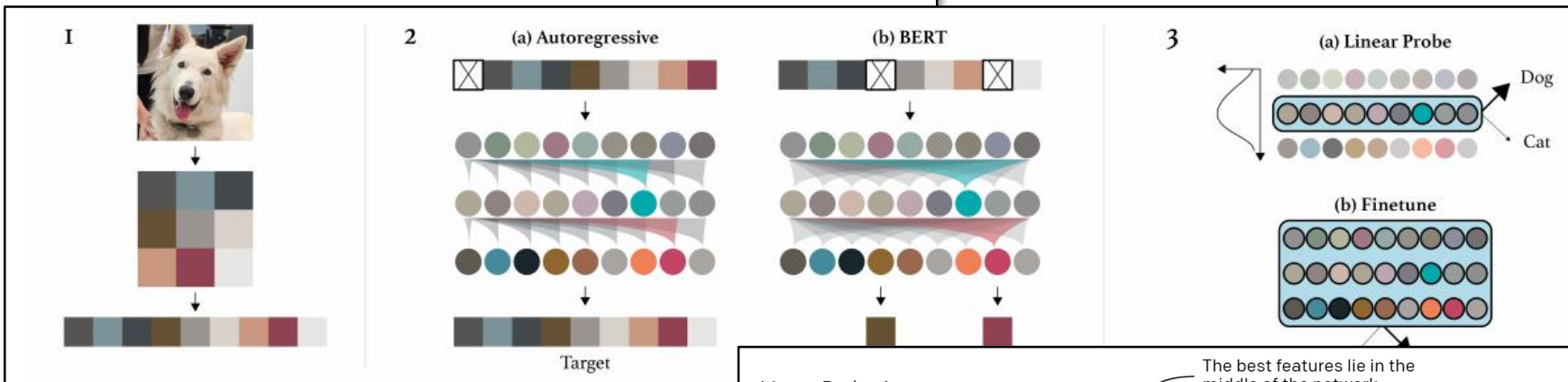


Figure 1. An overview of our approach. First, we pre-process r
We then chose one of two pre-training objectives, auto-regres
the representations learned by these objectives with linear pro

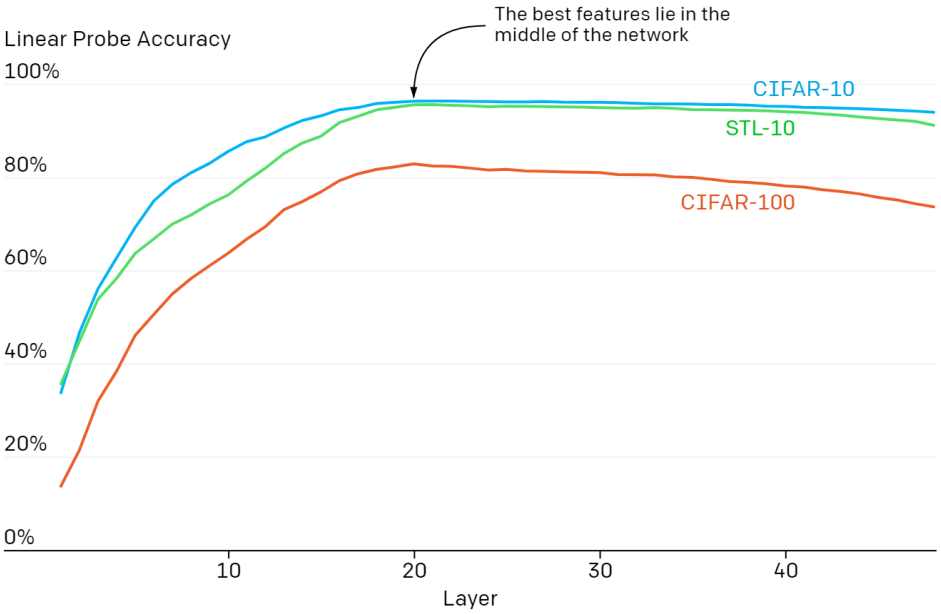
ture of ImageNet and web images is competitive with self-supervised benchmarks on ImageNet, achieving 72.0% top-1 accuracy on a linear probe of our features.

1. Introduction

Unsupervised pre-training played a central role in the resurgence of deep learning. Starting in the mid 2000's, approaches such as the Deep Belief Network (Hinton et al., 2006) and Denoising Autoencoder (Vincent et al., 2008) were commonly used in neural networks for computer vision (Lee et al., 2009) and speech recognition (Mohamed et al., 2009). It was believed that a model which learned the data distribution $P(X)$ would also learn beneficial fea-

prediction of corrupted inputs, clos
Denoising Autoencoder, which was
images.

As a higher dimensional, noisier, ar
ity than text, images are believed t
tive modeling. Here, self-supervise
encourage the modeling of more g
et al., 2015) have shown significant
of new training objectives (Oord e
architectures (Gomez et al., 2017),
pacity (Kolesnikov et al., 2019) ha
to achieve state of the art performa
(Hénaff et al., 2019) and sometime
vised representations in transfer le
2019; Misra & van der Maaten, 20



Feature quality depends heavily on the layer we choose to evaluate. In contrast with supervised models, the best features for these generative models lie in the middle of the network.

cdn.openai.com/papers/Generative_Pretraining_from_Pixels_V2.pdf

8 Machine Learning II
- ML in Natural Language Processing (NLP)
(7) Summary

Content:

1. **Motivation**
2. **IBM Watson**
3. **RNN & LSTM Networks**
4. **Transformer Models**
5. **Transformer BERT**
6. **Transformer GPT-3**
7. **Summary**

- Vor **Transformer** wurden für **NLP RNNs** und **LSTMs** eingesetzt.
- **Transformer** bauen auf dem **Attention-Mechanismus** auf und erlauben Parallel-verarbeitung.
- Die aktuell führenden **Transformer** sind **BERT** und **GPT**.

Michael Amberg

Todays Content:

1. **Motivation**
2. **IBM Watson**
3. **RNN & LSTM Networks**
4. **Transformer Models**
5. **Transformer BERT**
6. **Transformer GPT-3**
7. **Summary**

Transformer Networks (machine learning model)

Transformer (machine learning model)

From Wikipedia, the free encyclopedia

The **Transformer** is a deep learning model introduced in 2017, used primarily in the field of natural language processing (NLP).[1]

Like recurrent neural networks (RNNs), Transformers are designed to handle sequential data, such as natural language, for tasks such as translation and text summarization. However, unlike RNNs, Transformers do not require that the sequential data be processed in order. For example, if the input data is a natural language sentence, the Transformer does not need to process the beginning of it before the end. Due to this feature, the Transformer allows for much more parallelization than RNNs and therefore reduced training times.[1]

Transformers have rapidly become the model of choice for NLP problems,[2] replacing older recurrent neural network models such as the long short-term memory (LSTM). Since the Transformer model facilitates more parallelization during training, it has enabled training on larger datasets than was possible before it was introduced. This has led to the development of pretrained systems such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), which have been trained with huge general language datasets, such as Wikipedia Corpus, and can be fine-tuned to specific language tasks.[3][4]

5. Transformer BERT, Google

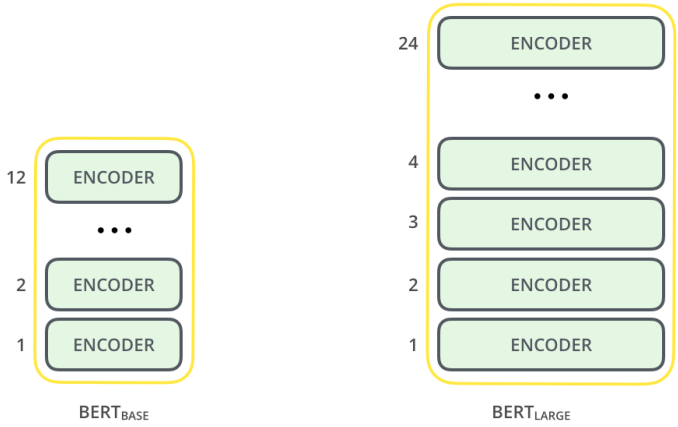
BERT Transformer, Google 2018

BERT (language model)

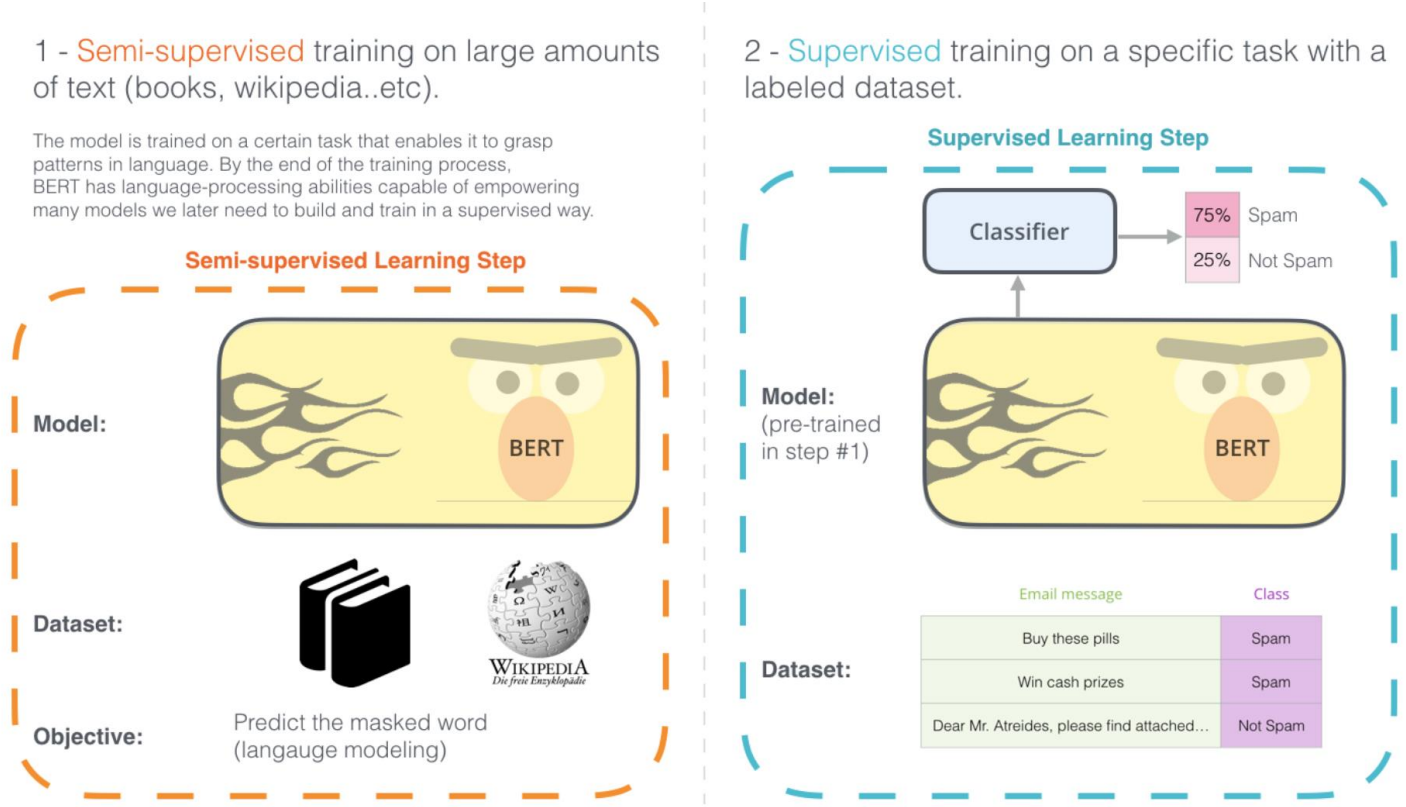
From Wikipedia, the free encyclopedia

Bidirectional Encoder Representations from Transformers (**BERT**) is a Transformer-based machine learning technique for natural language processing (NLP) pre-training developed by Google. BERT was created and published in 2018 by Jacob Devlin and his colleagues from Google.[1][2] As of 2019, Google has been leveraging BERT to better understand user searches.[3]

The original English-language BERT model comes with two pre-trained general types:[1] (1) the $\text{BERT}_{\text{BASE}}$ model, a 12-layer, 768-hidden, 12-heads, 110M parameter neural network architecture, and (2) the $\text{BERT}_{\text{LARGE}}$ model, a 24-layer, 1024-hidden, 16-heads, 340M parameter neural network architecture; both of which were trained on the BooksCorpus[4] with 800M words, and a version of the English Wikipedia with 2,500M words.



BERT Transformer, Google 2018



The two steps of how BERT is developed. You can download the model pre-trained in step 1 (trained on un-annotated data), and only worry about fine-tuning it for step 2. [Source for book icon].

Das **Training** läuft in zwei Schritten: **Pre-Training** und **Fine-Tuning** (eine Form von **Transfer Learning**)

The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning), 2018
jalammar.github.io/illustrated-bert/

BERT Transformer, Training

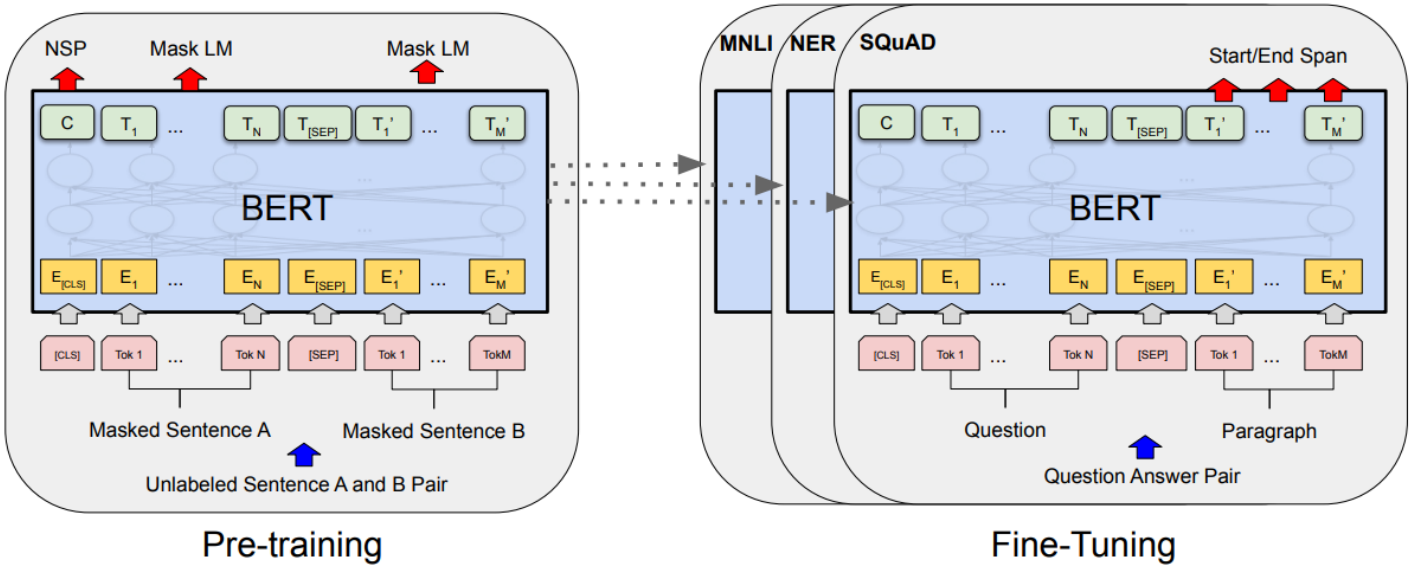
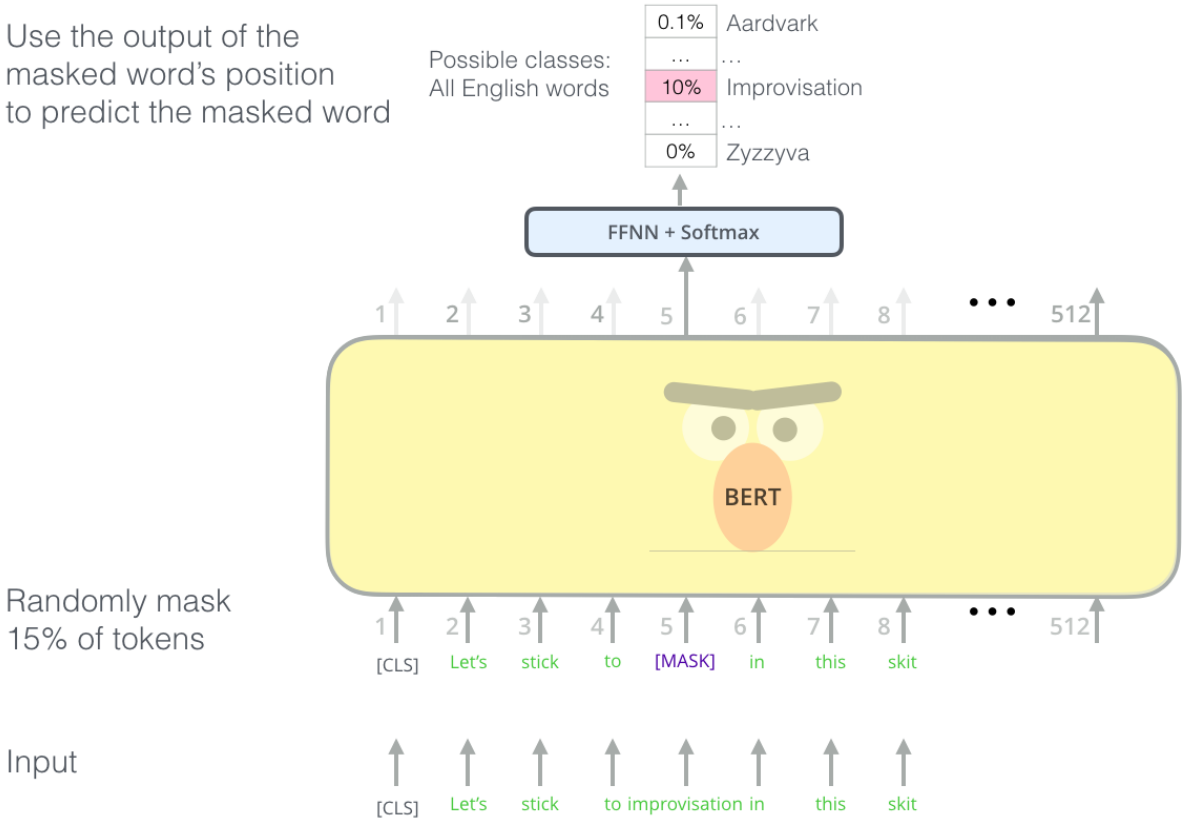


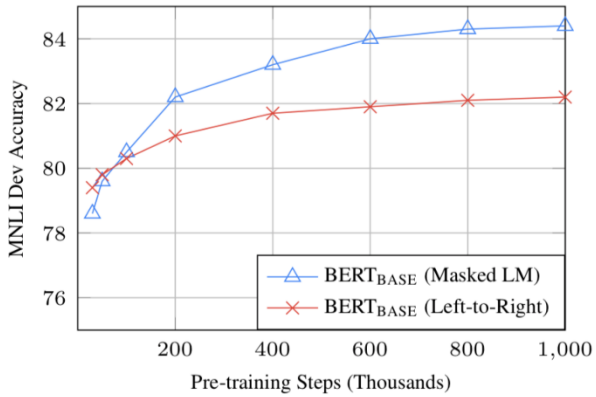
Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers).

Das **Training** läuft in zwei Schritten: **Pre-Training** und **Fine-Tuning**
(es können verschiedene Anwendungen im 2. Schritt gelernt werden)

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019
arxiv.org/pdf/1810.04805.pdf

BERT Training using Masked Language Model

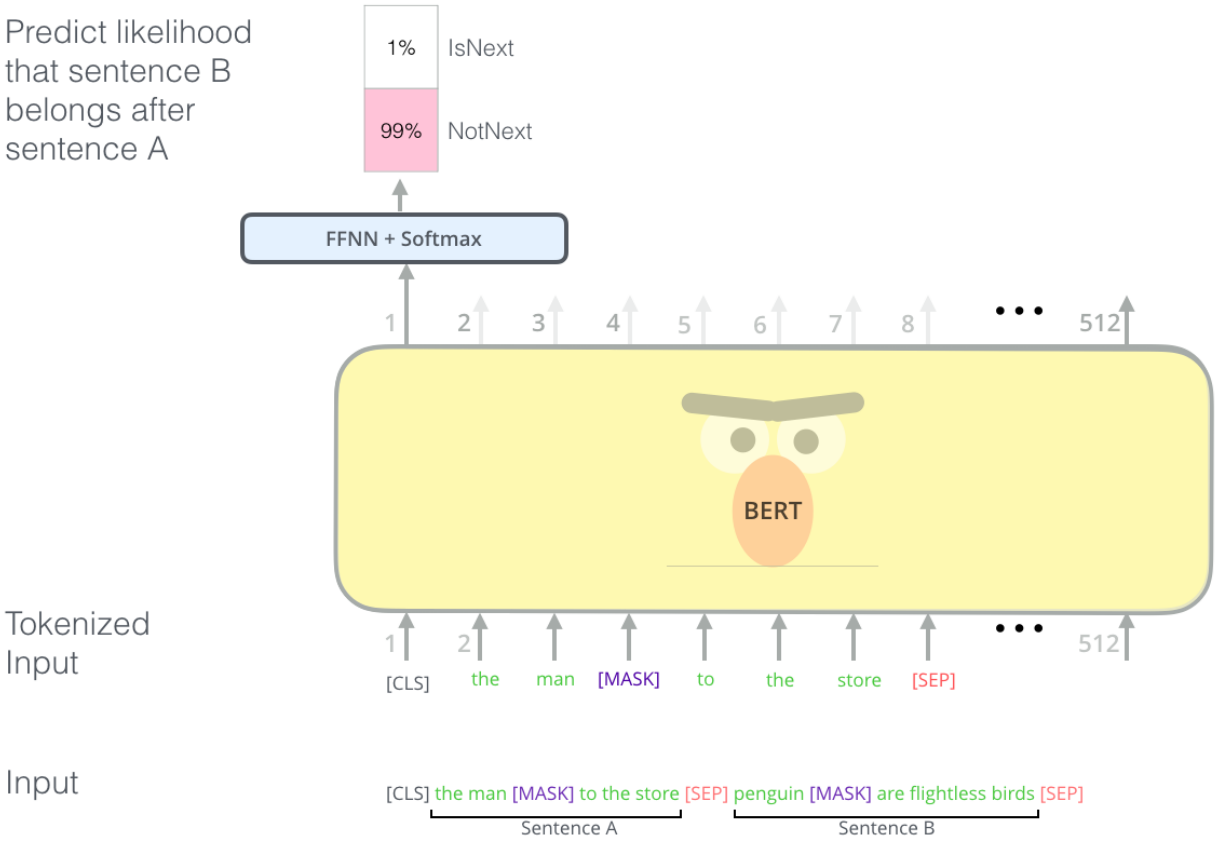




Masked Language Model: **BERT** maskiert **15% der Worte** im **Input** und lässt das fehlende Wort **vorhersagen (Bidirektionale Suche)**.

The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning), 2018
jalammar.github.io/illustrated-bert/

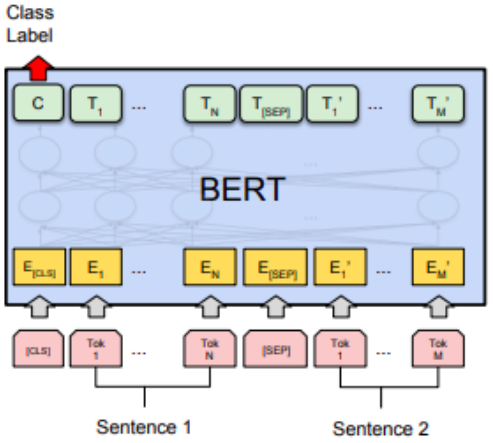
BERT Training using Next Sentence Prediction



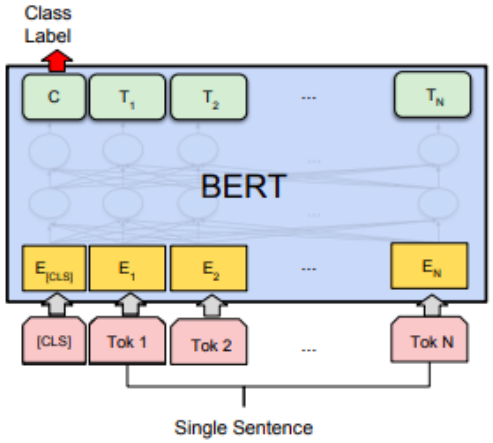
Next Sentence Prediction: **BERT** bekommt **2 Sätze** im **Input** und sagt die Reihenfolge (ja/nein) voraus.

The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning), 2018
jalammar.github.io/illustrated-bert/

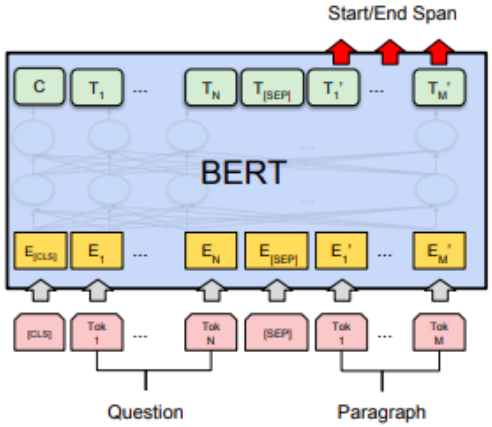
BERT Fine-Tuning Tasks (bzgl. GLUE Benchmark)



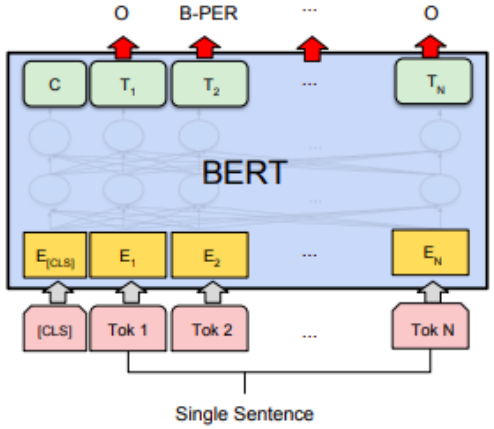
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning), 2018
jalammar.github.io/illustrated-bert/

BERT Transformer, GLUE Benchmark

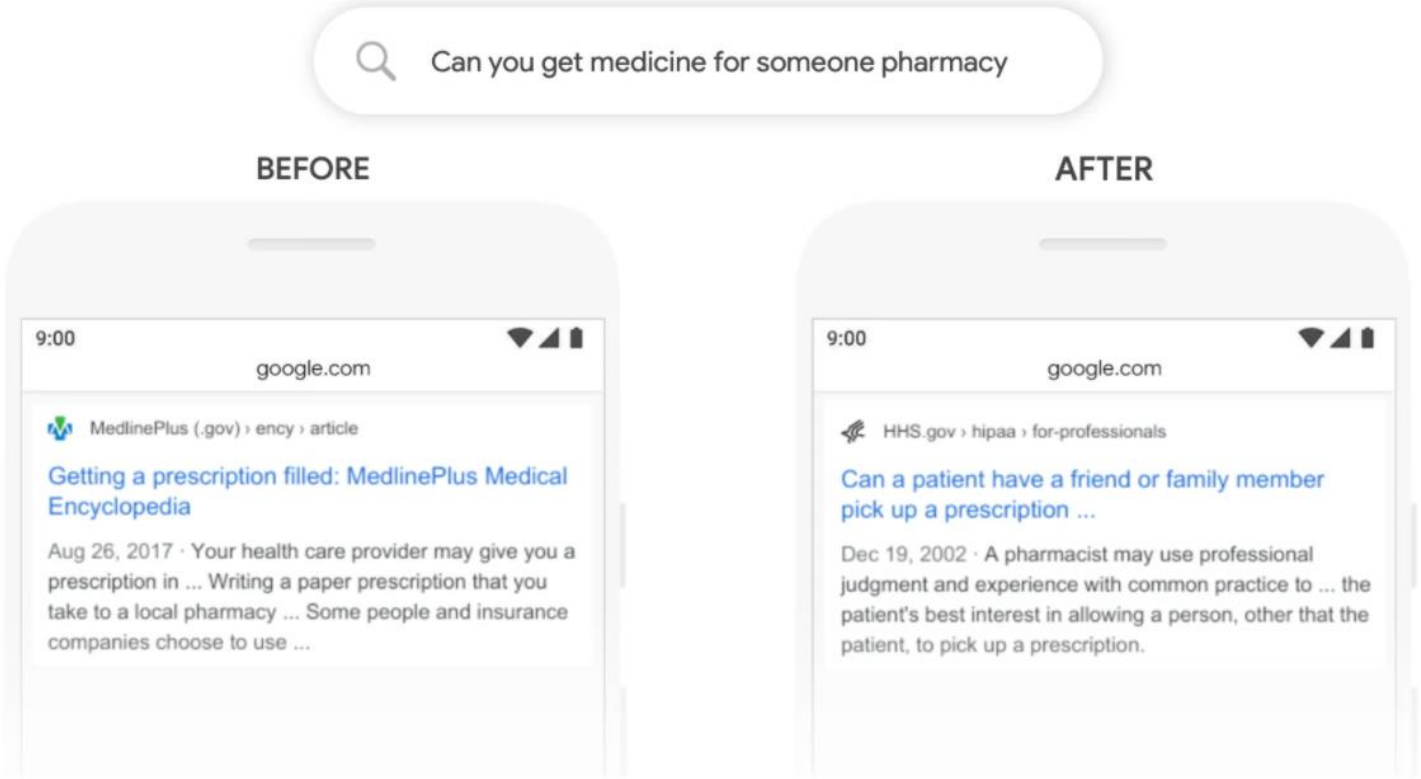
The General Language Understanding Evaluation (GLUE) benchmark is a collection of resources for training, evaluating, and analyzing natural language understanding systems. GLUE consists of:

- A benchmark of nine sentence- or sentence-pair language understanding tasks built on established existing datasets and selected to cover a diverse range of dataset sizes, text genres, and degrees of difficulty,
- A diagnostic dataset designed to evaluate and analyze model performance with respect to a wide range of linguistic phenomena found in natural language, and
- A public leaderboard for tracking performance on the benchmark and a dashboard for visualizing the performance of models on the diagnostic set.

	Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	M
	1	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.8	71.5	97.5	94.0/92.0	92.9/92.6	76.2/90.8	91.9	
	2	HFL iFLYTEK	MacALBERT + DKM		90.7	74.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6	91.3	
+	3	Alibaba DAMO NLP	StructBERT + TAPT		90.6	75.3	97.3	93.9/91.9	93.2/92.7	74.8/91.0	90.9	
+	4	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	91.6	
	5	ERNIE Team - Baidu	ERNIE		90.4	74.4	97.5	93.5/91.4	93.0/92.6	75.2/90.9	91.4	
	6	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	
	7	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0	
+	8	Huawei Noah's Ark Lab	NEZHA-Large		89.8	71.7	97.3	93.3/91.0	92.4/91.9	75.2/90.7	91.5	

https://gluebenchmark.com/leaderboard

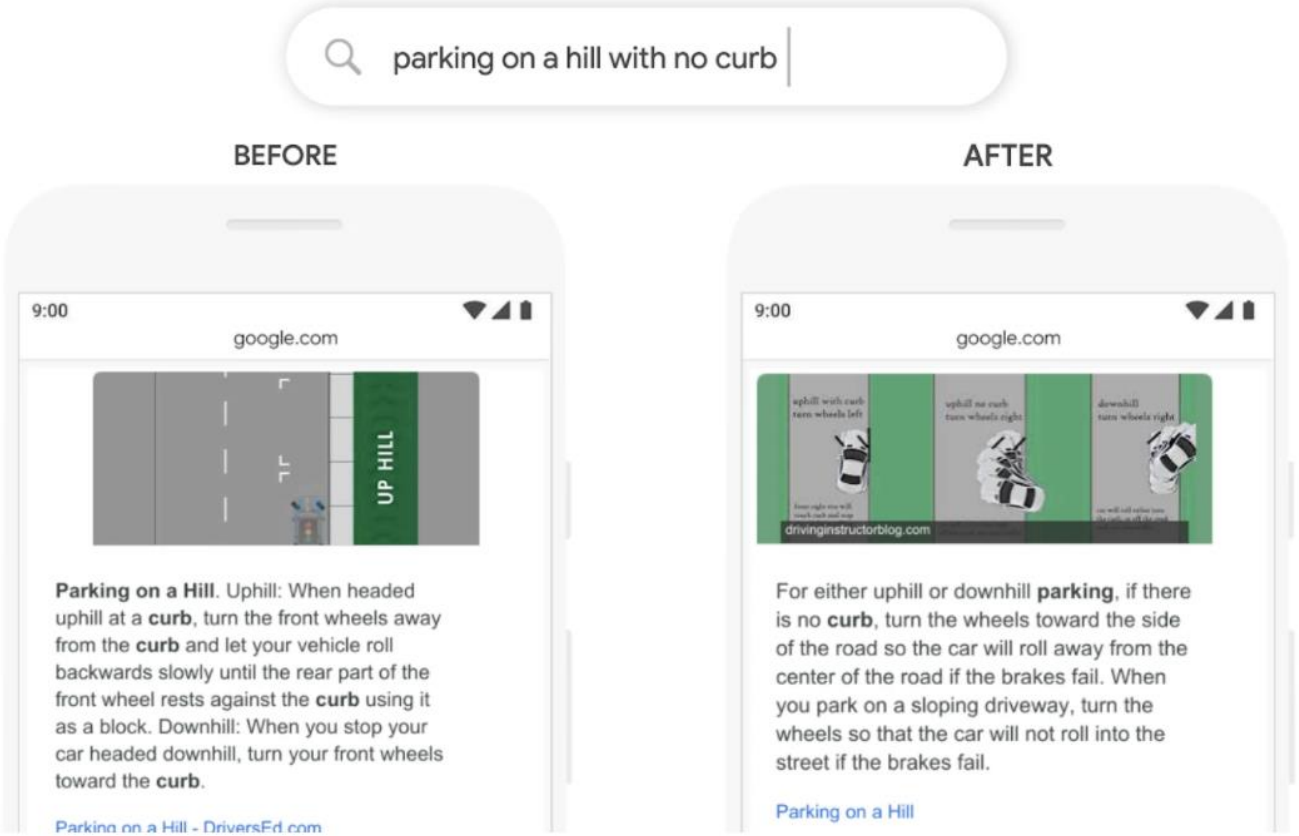
Google Search with BERT (1/3)



"Can you get medicine for someone pharmacy": With the BERT model, we can better understand that "for someone" is an important part of this query, whereas previously we missed the meaning, showing general results about filling prescriptions.

Google: Understanding searches better than ever before, 2019
blog.google/products/search/search-language-understanding-bert/

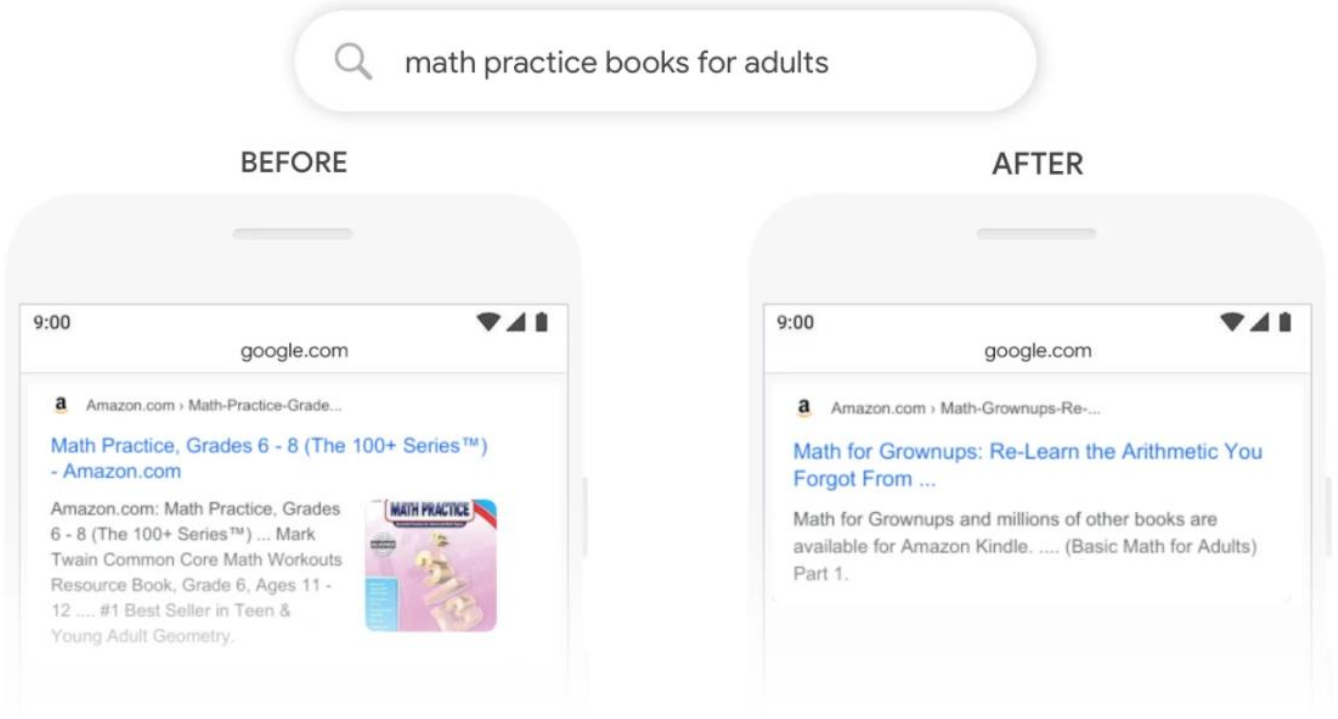
Google Search with BERT (2/3)



"Parking on a hill with no curb": In the past, a query like this would confuse our systems–we placed too much importance on the word "curb" and ignored the word "no", not understanding how critical that word was to appropriately responding to this query. So we'd return results for parking on a hill with a curb!

Google: Understanding searches better than ever before, 2019
blog.google/products/search/search-language-understanding-bert/

Google Search with BERT (3/3)



"math practice books for adults": While the previous results page included a book in the "Young Adult" category, BERT can better understand that "adult" is being matched out of context, and pick out a more helpful result.

Google: Understanding searches better than ever before, 2019
blog.google/products/search/search-language-understanding-bert/