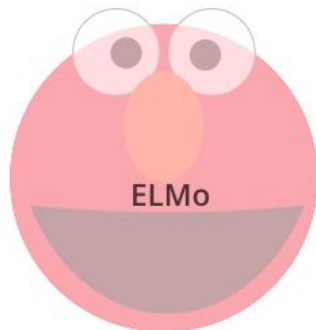# Lecture 04: Context based models & BERT overview

Anastasia Yanina

Outline

1. OpenAI Transformer
2. ELMo
3. BERT
4. GPT
5. Q & A

Based on: http://web.stanford.edu/class/cs224n/slides/cs224n-2019-lecture13-contextual-representations.pdf
https://jalammar.github.io/illustrated-transformer/
http://jalammar.github.io/illustrated-bert/
https://medium.com/mlreview/understanding-building-blocks-of-ulmfit-818d3775325b

Image source: http://jalammar.github.io/illustrated-bert/

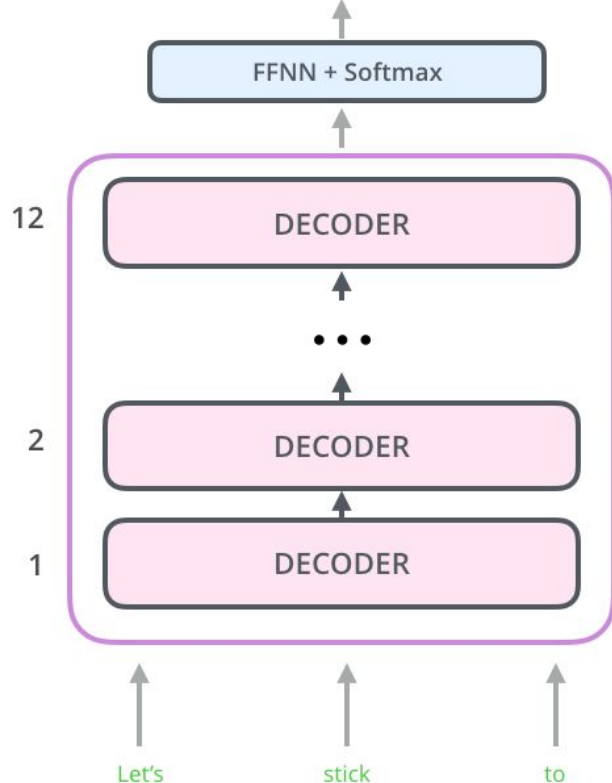# **OpenAI Transformer:** Pre-training Decoder for Language Modeling

- The Encoder-Decoder structure of the transformer made it perfect for machine translation
- But what about sentence classification?
- **Main goal: pre-train a language model that can be fine-tuned for other tasks**
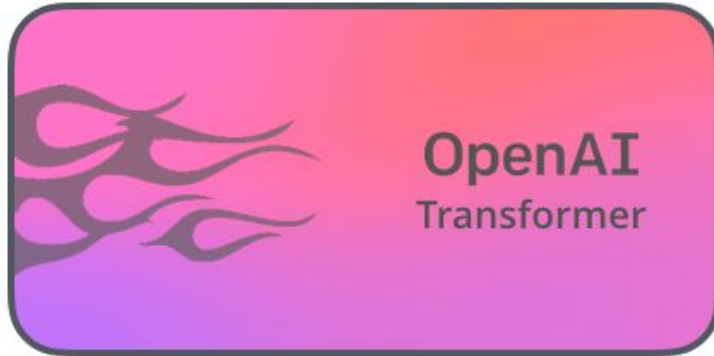
Image source: http://jalammar.github.io/illustrated-bert/
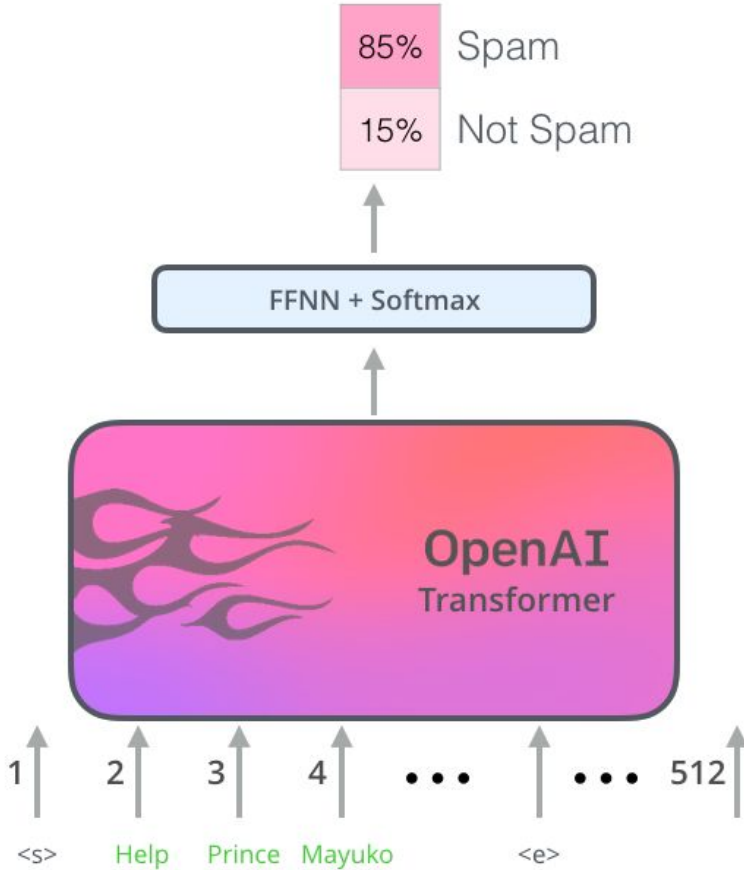
OpenAI Transformer
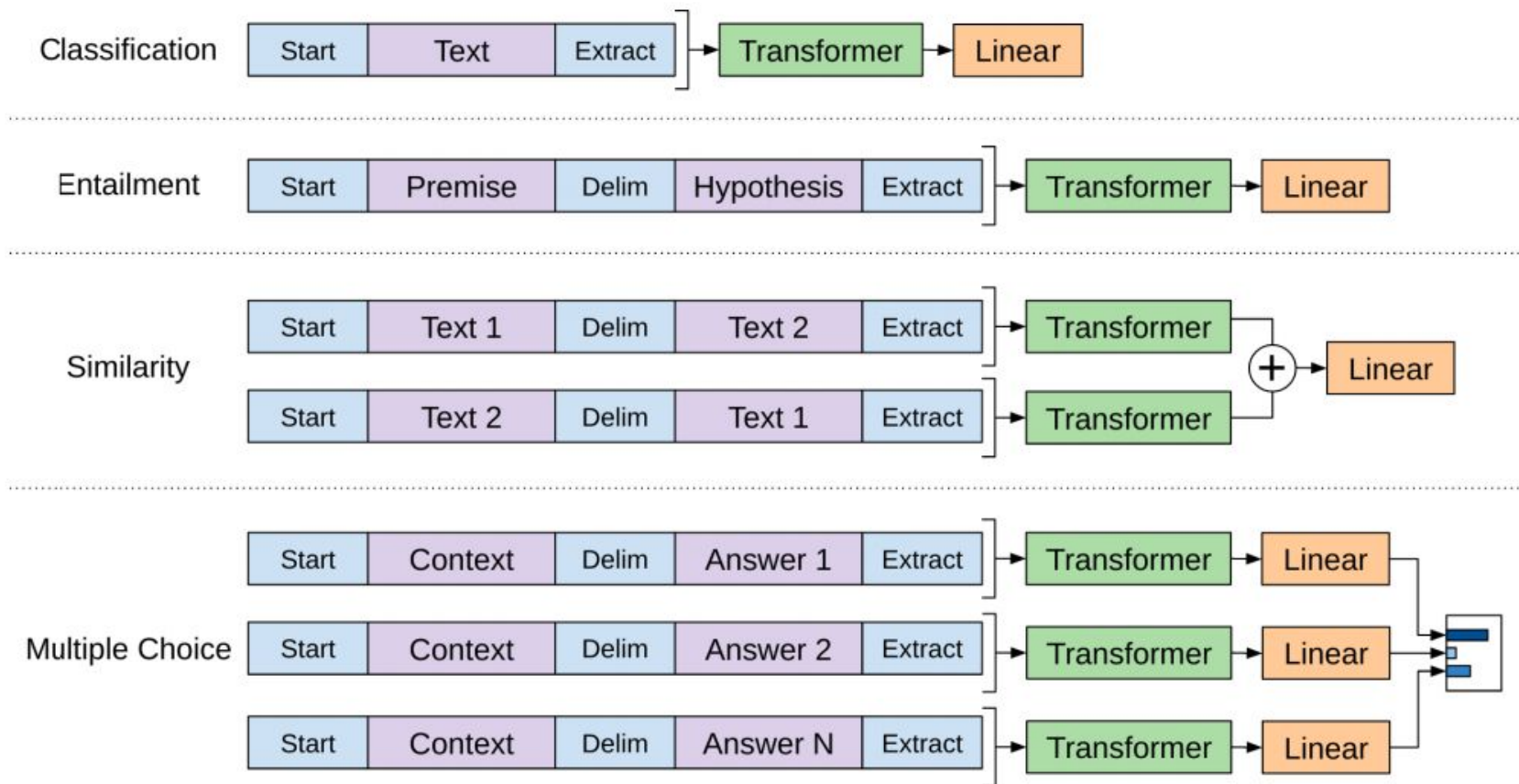
Differences from vanilla Transformer:

- no encoder
- decoder layers would not have the encoder-decoder attention sublayer
- Pre-train the model on predicting the next word using massive (unlabeled) datasets
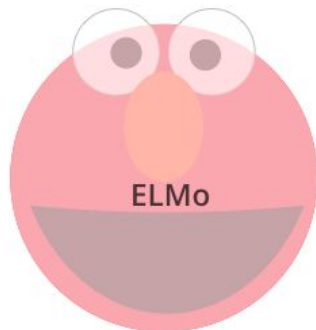
OpenAI Transformer

- During pre-training phase layers have been tuned to reasonably handle language
- Now let's use it for downstream tasks (e.g. sentence classification)

Input transformations for different tasks

THE TRANSFORMER

ULM-FiT

OpenAI Transformer

ELMo

BERT

Image source: http://jalammar.github.io/illustrated-bert/

ELMo: context that matters

ELMo: contextualized word embeddings

"Why not give it an embedding based on the context it's used in – to both capture the word meaning in that context as well as other contextual information?"

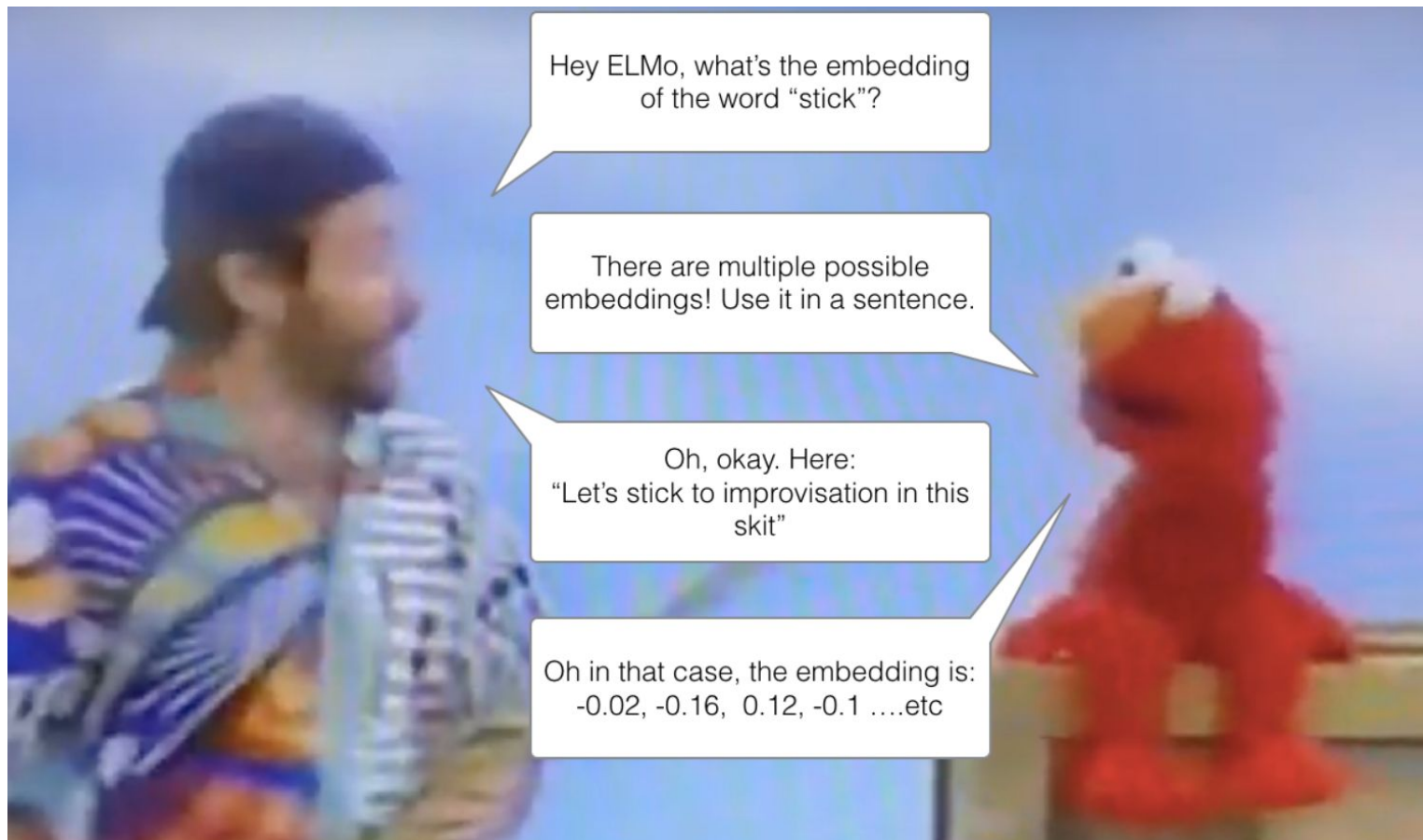Peters et. al., 2017, McCann et. al., 2017, Peters et. al., 2018 in the ELMo paper

ELMo – deep contextualized word representations
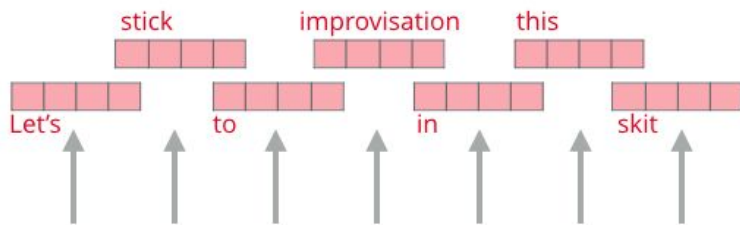
What does it stand for?

1. **E**xpedited **L**abour **M**arket **O**pinion
2. **E**lectric **L**ight **M**achine **O**rganization
3. **E**nough **L**et's **M**ove **O**n

4. **Embeddings from Language Models**

ELMo: contextualized word embeddings

ELMo: Contextualized word embeddings

ELMo Embeddings

stick improvisation this

Let's to in skit

- uses a bi-directional LSTM trained on Language Modeling task
- a model can learn without labels

ELMo

Words to embed

Let's stick to improvisation in this skit

Bidirectional Language Models (biLMs)

Possible classes:
All English words

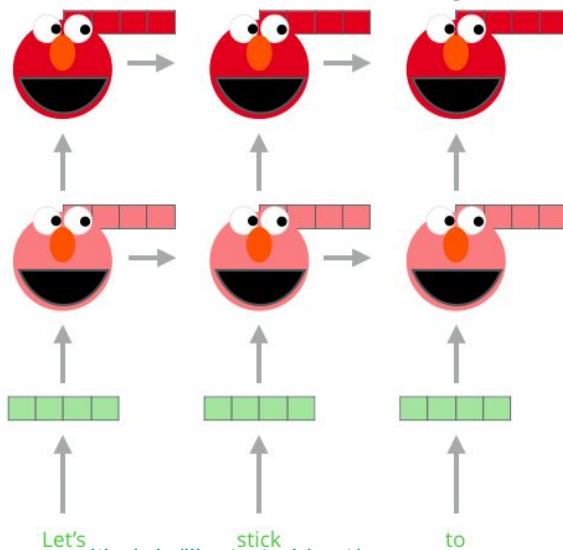| 0.1% | Aardvark |
...	...
10%	Improvisation
...	...
0%	Zyzzyva

Output Layer

FFNN + Softmax

LSTM Layer #2

LSTM Layer #1

Embedding

Let's   stick   to

biLMs consist of forward and backward LMs:

- Forward:

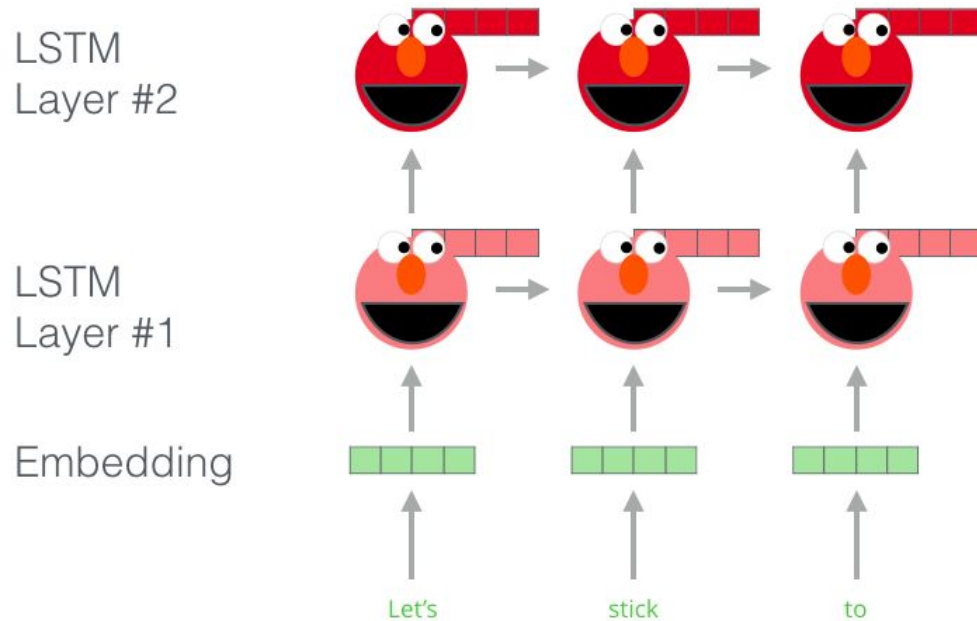$$p(t_1,t_2,...,t_N)=\prod_{k=1}^{N} p(t_k|t_1,t_2,...,t_{k-1})$$

- Backward:

$$p(t_1,t_2,...,t_N)=\prod_{k=1}^{N} p(t_k|t_{k+1},t_{k+2},...,t_N)$$

LSTM predicts next word in both directions to build biLMs

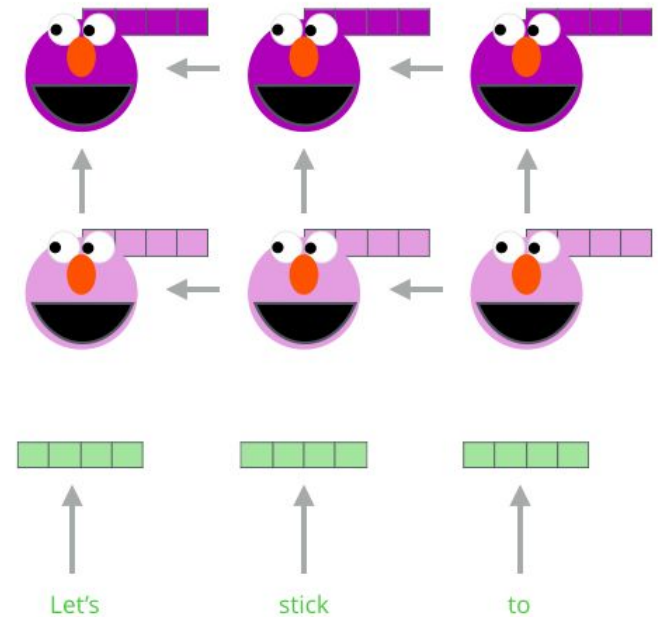# ELMo: main pipeline

Embedding of "stick" in "Let's stick to" - Step #1

ELMo: main pipeline
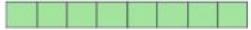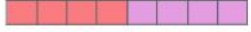
Embedding of "stick" in "Let's stick to" - Step #2

1- Concatenate hidden layers

Forward Language Model

Backward Language Model

2- Multiply each vector by a weight based on the task

$\times \ s_2$

$\times \ s_1$

$\times \ s_0$

Let's   stick   to

Let's   stick   to

3- Sum the (now weighted) vectors

ELMo embedding of "stick" for this task in this context
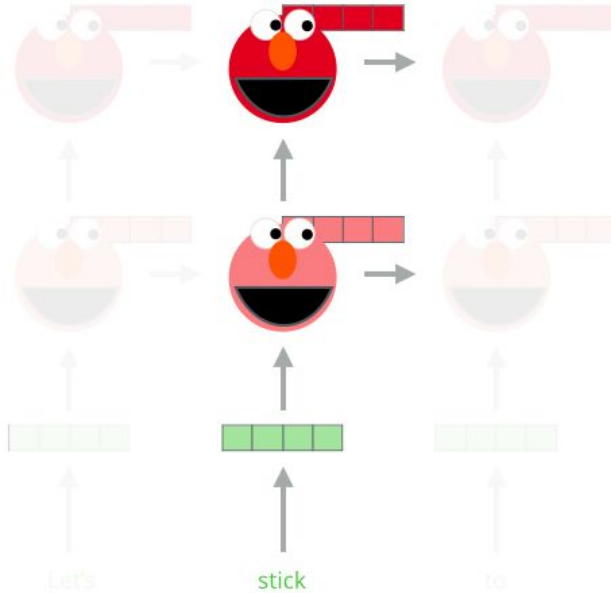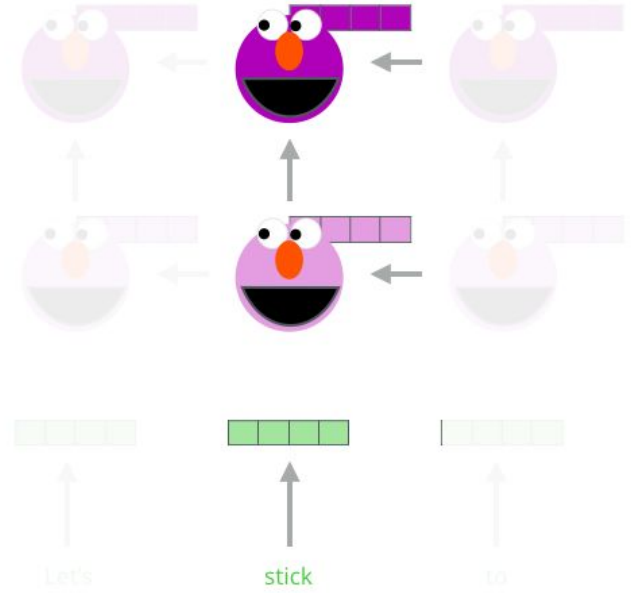
ELMo represents a word as a linear combination of corresponding hidden layers

ELMo can be integrated to almost all neural NLP tasks with simple concatenation to the embedding layer
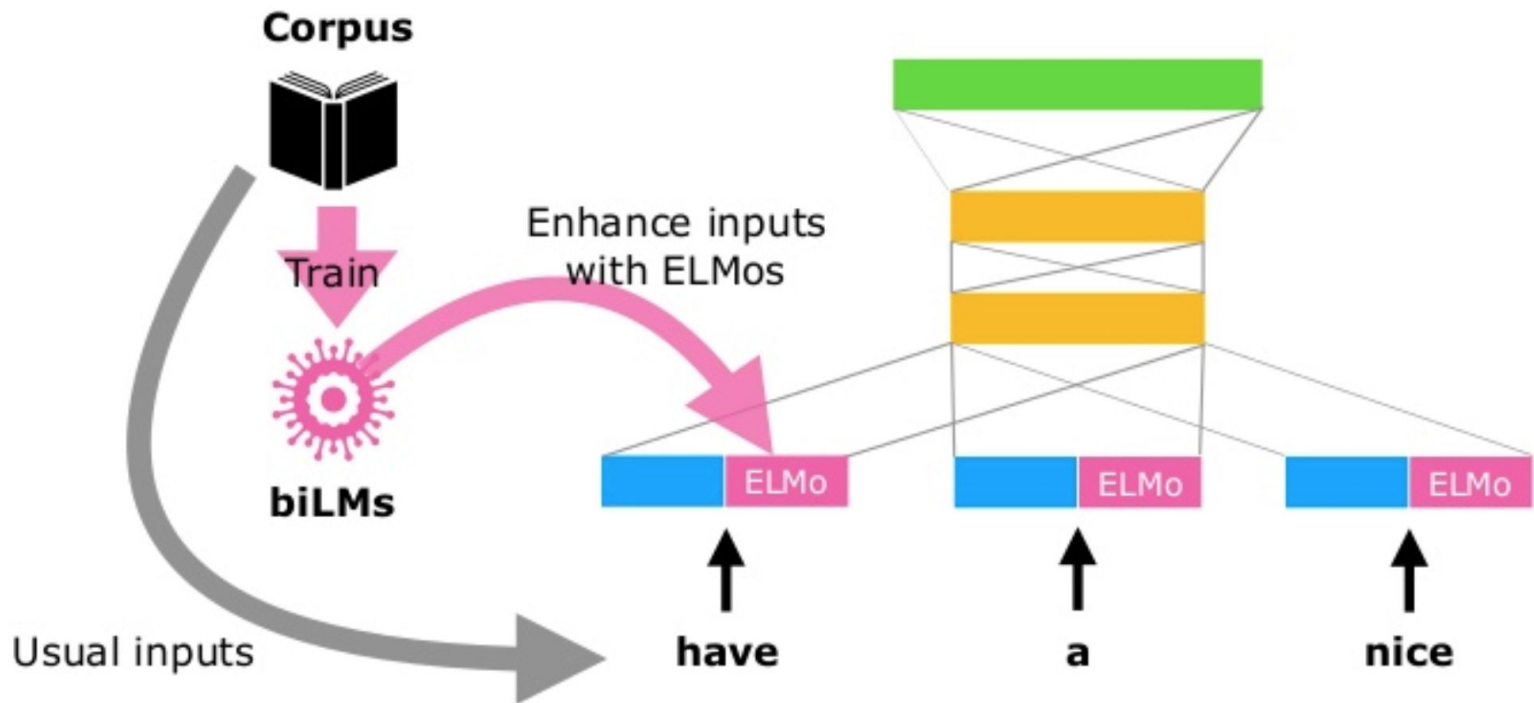
- Pretrained ELMo models: http://allennlp.org/elmo

- AllenNLP is a library on the top of PyTorch

- Higher levels seems to catch semantics while lower layer probably capture syntactic features

THE TRANSFORMER

ULM-FiT

OpenAI Transformer

ELMo

BERT

Image source: http://jalammar.github.io/illustrated-bert/
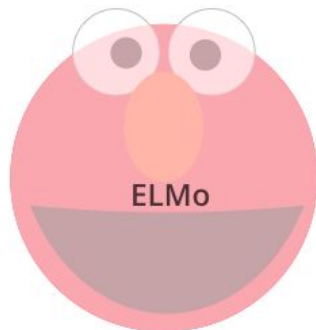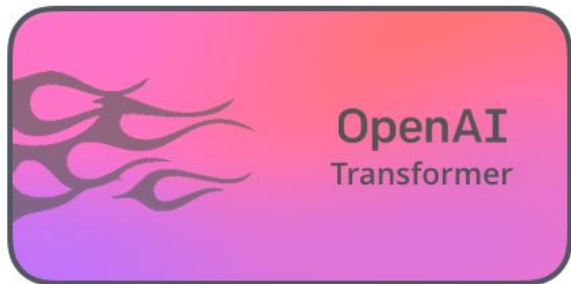
BERT

Bidirectional Encoder Representations from Transformers

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.

Semi-supervised Learning Step

Model:

BERT

Dataset:

WIKIPEDIA
Die freie Enzyklopädie

Objective: Predict the masked word (langauge modeling)

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

Supervised Learning Step

Classifier → 75% Spam / 25% Not Spam

Model: (pre-trained in step #1)
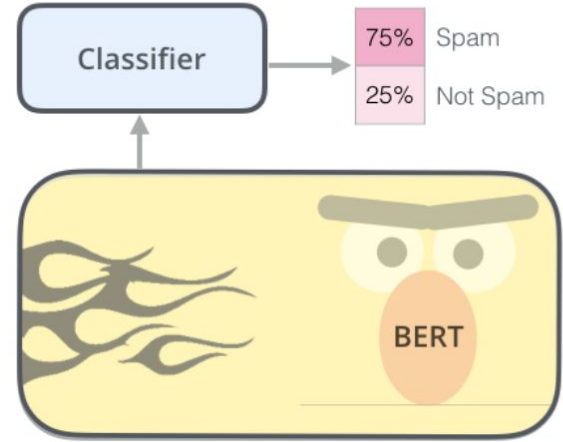
BERT

Dataset:

Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached…	Not Spam

Image source: http://jalammar.github.io/illustrated-bert/

Input
Features

Output
Prediction

Help Prince Mayuko Transfer
Huge Inheritance

→ BERT →

Classifier
(Feed-forward
neural network +
softmax)

→

| 85% | Spam |
| 15% | Not Spam |

Image source: http://jalammar.github.io/illustrated-bert/
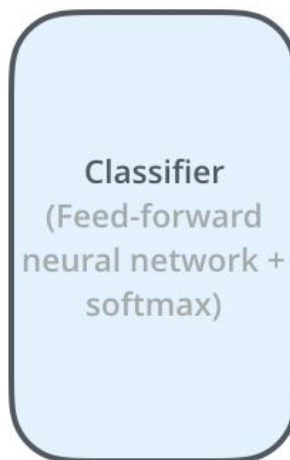
BERT: base and large
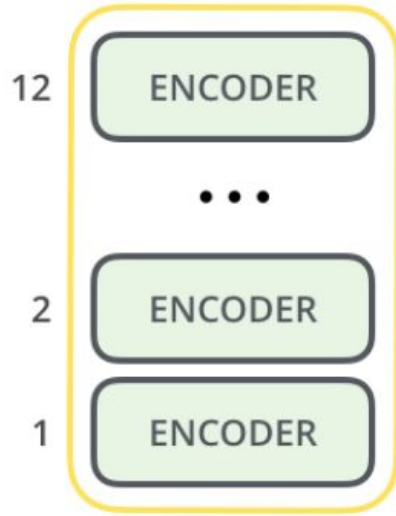
BERT_{BASE}
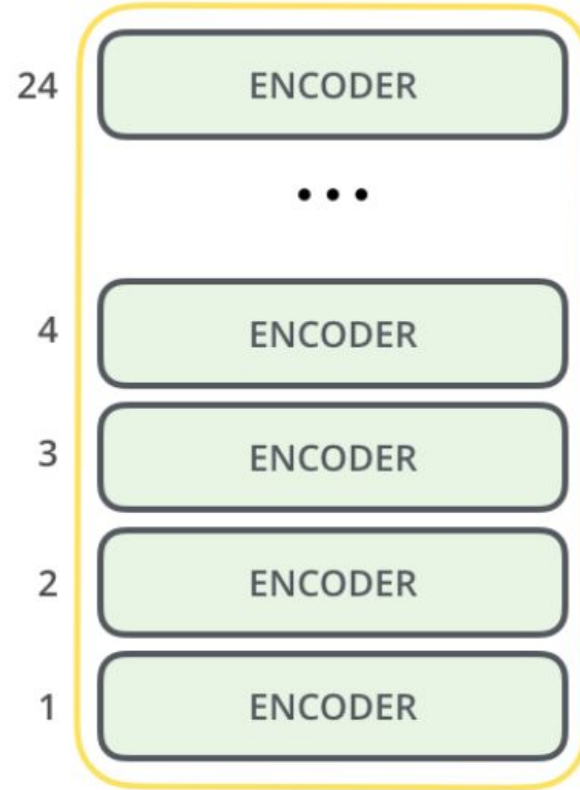
BERT_{LARGE}

BERT vs. Transformer

	THE TRANSFORMER	Base BERT	Large BERT
Encoders	6	12	24
Units in FFN	512	768	1024
Attention Heads	8	12	16

Image source: http://jalammar.github.io/illustrated-bert/

Model inputs

ENCODER

Feed Forward

z_1 z_2 z_3

Self-Attention

x_1 x_2 x_3

Je   suis   étudiant

Can be parallelized
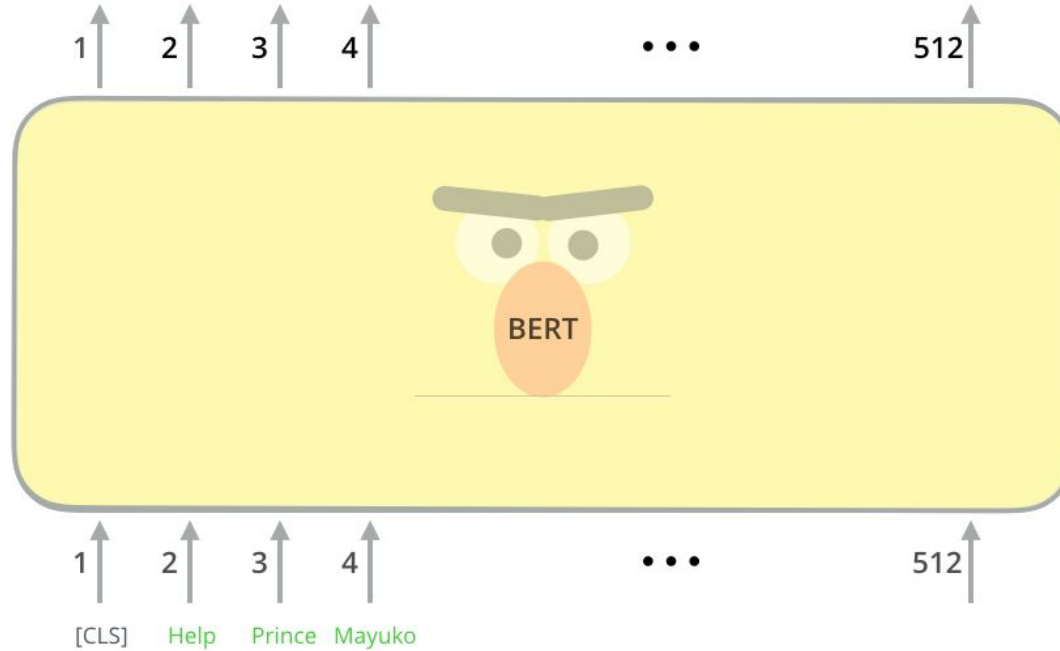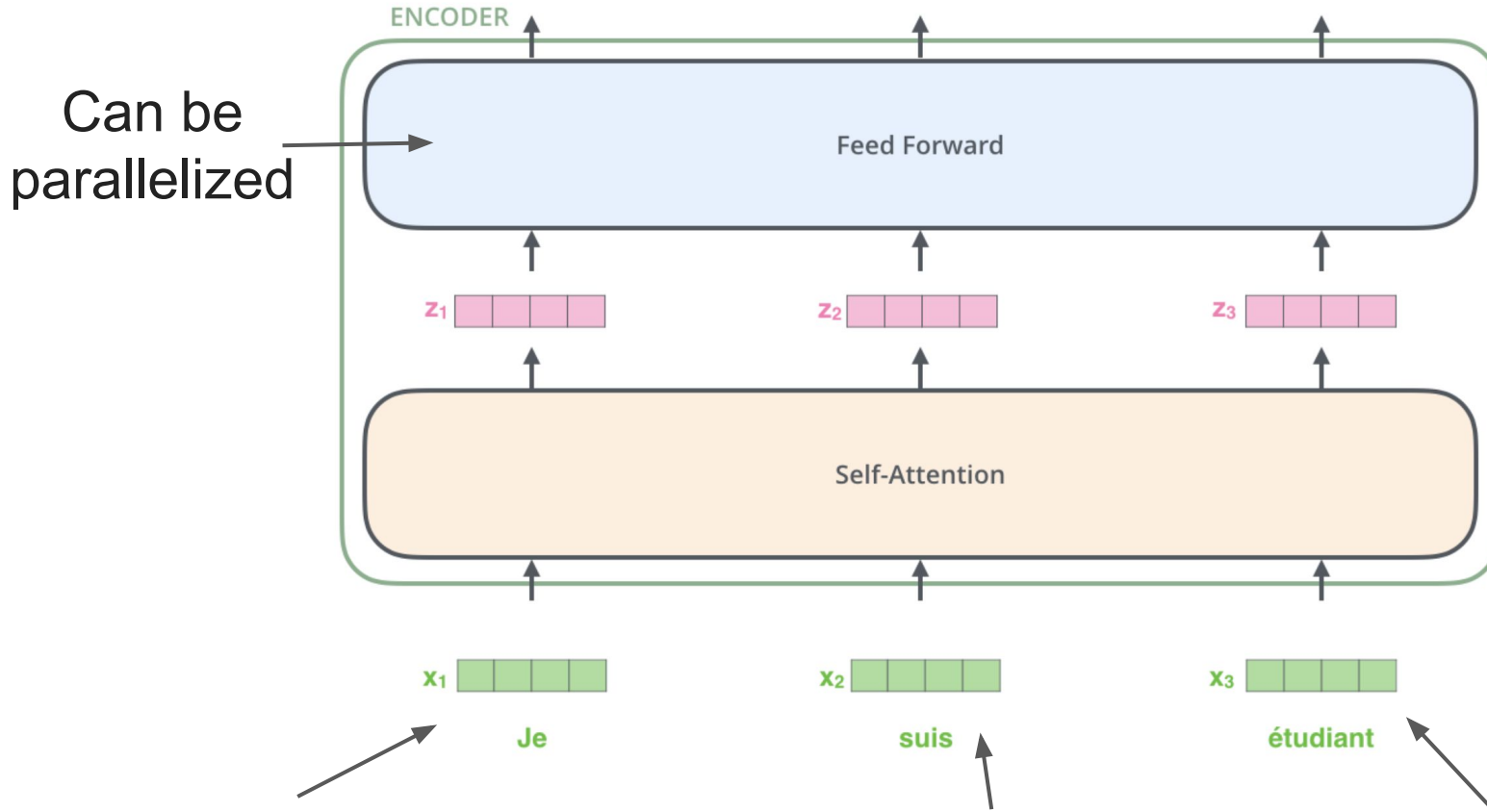
the word in each position flows through its own path in the encoder

1 2 3 4 ••• 512

1 2 3 4 ••• 512

[CLS] Help Prince Mayuko
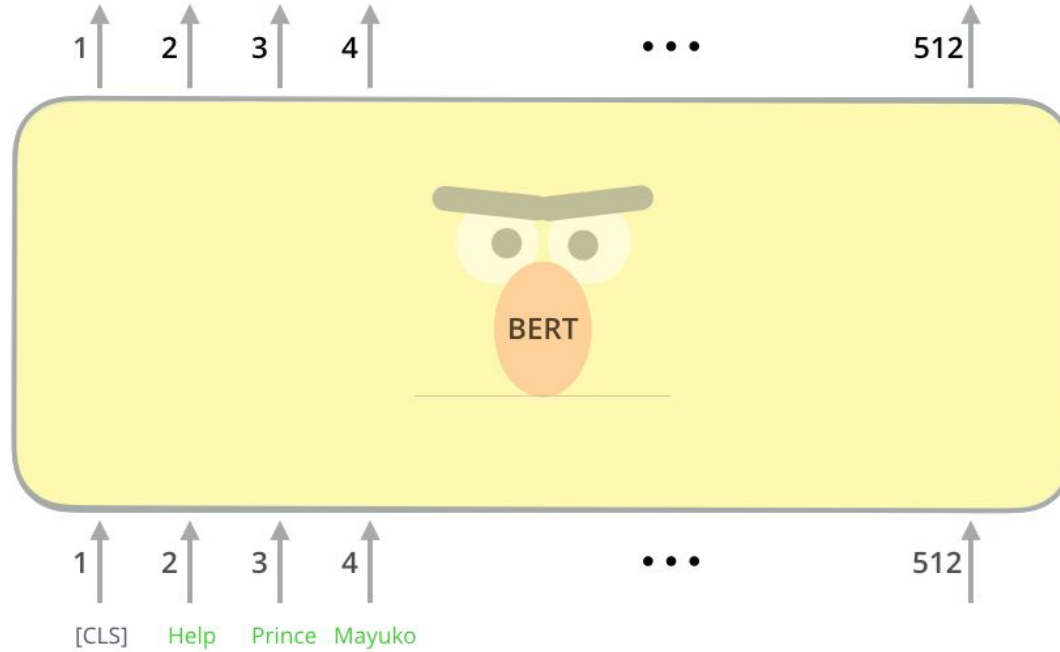
Identical to the Transformer up until this point

Why is BERT so special?

Model outputs

Each position outputs a vector
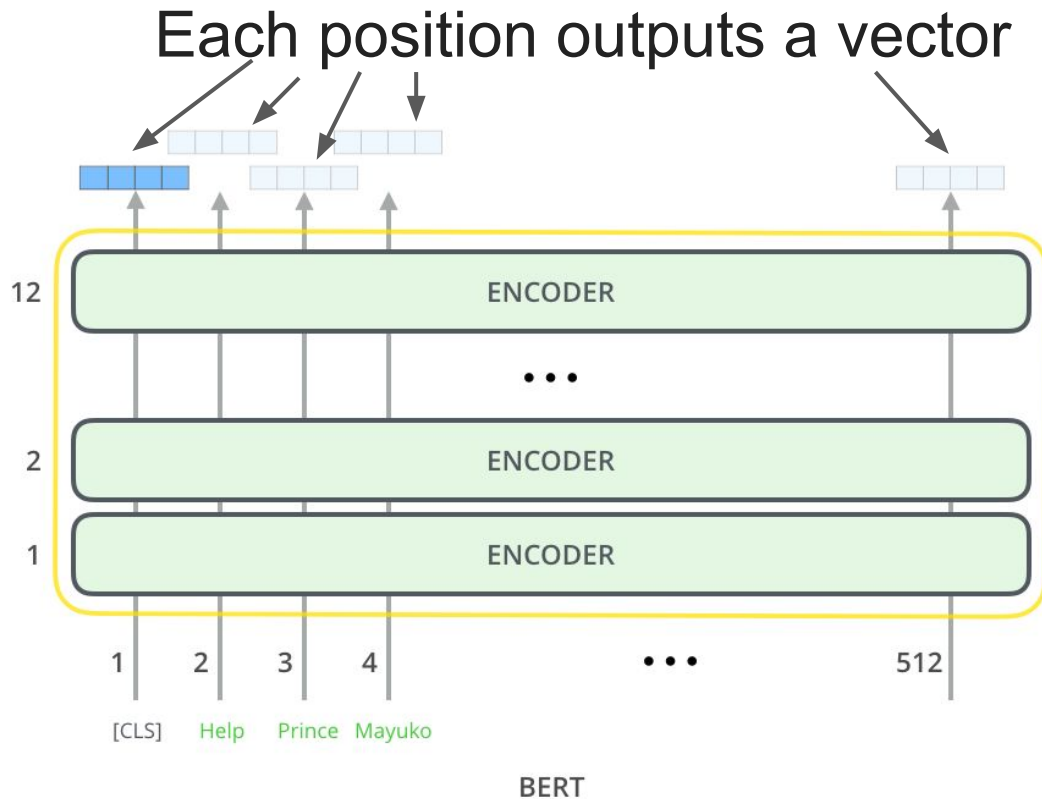
For sentence classification we focus on the first position
(that we passed [CLS] token to)

Model inputs
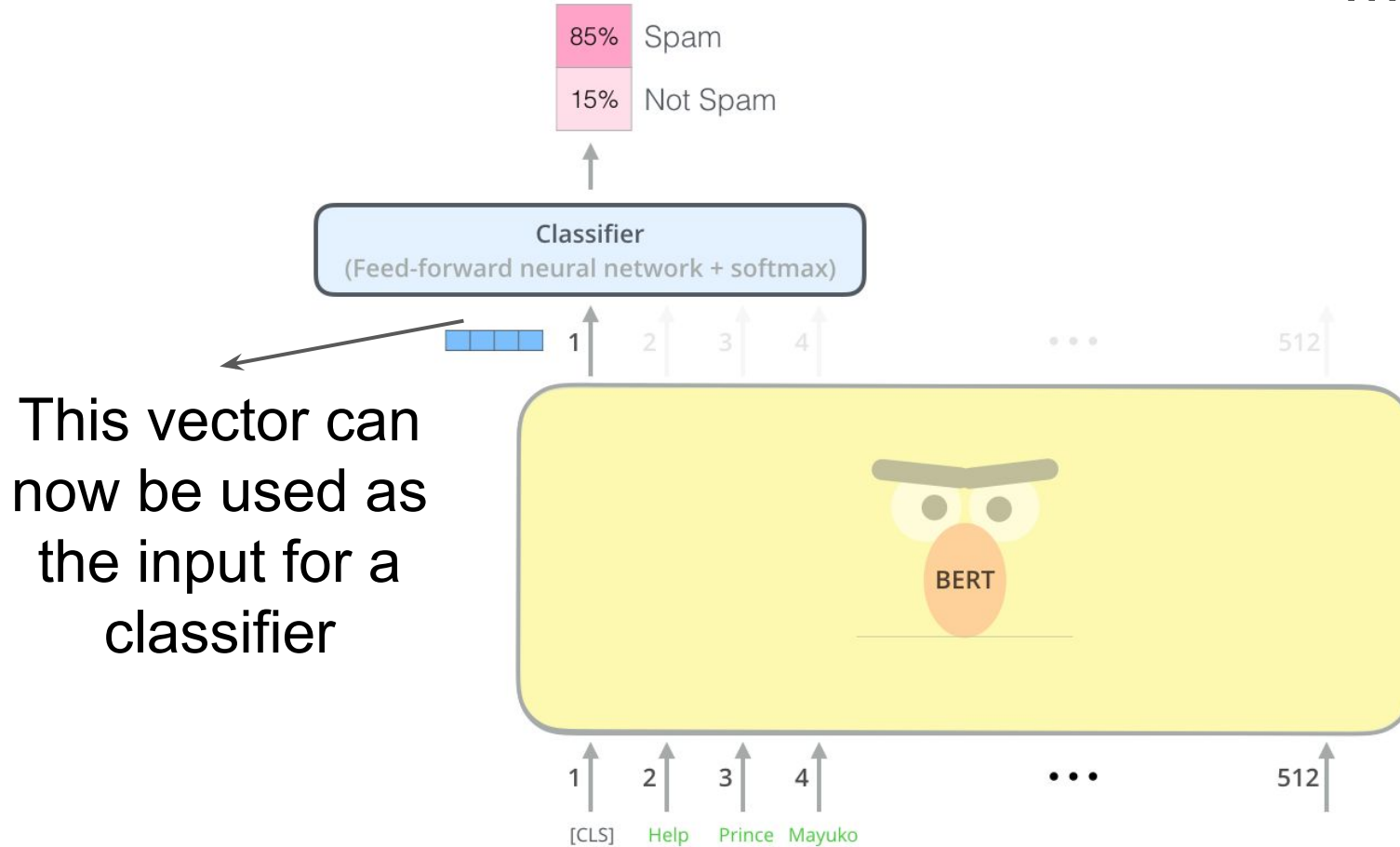
This vector can now be used as the input for a classifier

Similar to CNN concept!

BERT: pre-training

Use the output of the masked word's position to predict the masked word

Possible classes:
All English words

0.1%	Aardvark
...	...
10%	Improvisation
...	...
0%	Zyzzyva

FFNN + Softmax

1 2 3 4 5 6 7 8 ... 512

BERT

Randomly mask 15% of tokens

1 2 3 4 5 6 7 8 ... 512

[CLS] Let's stick to [MASK] in this skit

Input

[CLS] Let's stick to improvisation in this skit

- "Masked Language Model" approach

- To make BERT better at handling relationships between multiple sentences, the pre-training process includes an additional task:

  "Given two sentences (A and B), is B likely to be the sentence that follows A, or not?"

BERT: pre-training

Predict likelihood that sentence B belongs after sentence A

1% IsNext
99% NotNext

FFNN + Softmax

1 2 3 4 5 6 7 8 ••• 512

BERT

Tokenized Input

1 2 the man [MASK] to the store [SEP] ••• 512
[CLS]

Input

[CLS] the man [MASK] to the store [SEP] penguin [MASK] are flightless birds [SEP]

Sentence A   Sentence B

Image source: http://jalammar.github.io/illustrated-bert/
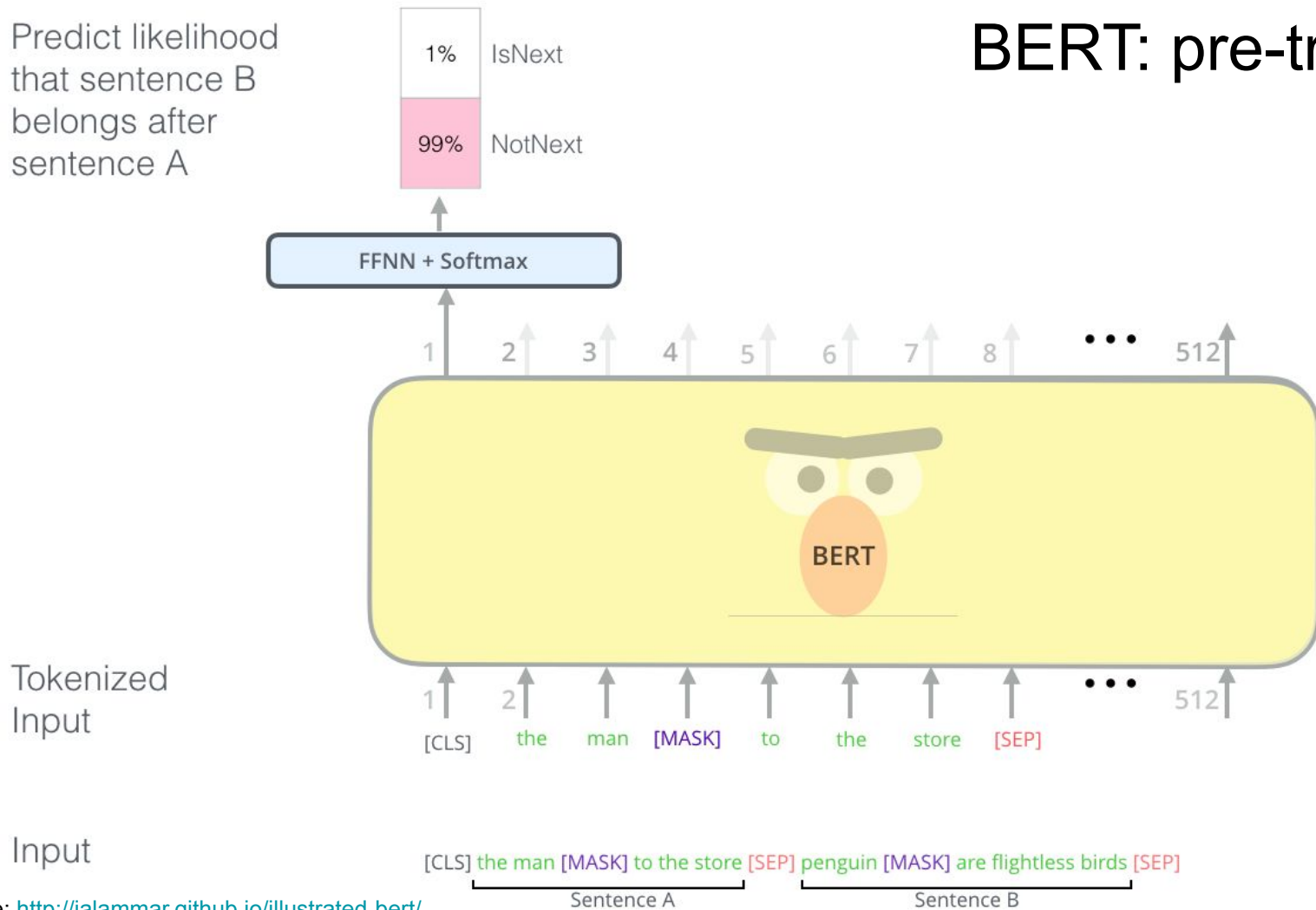
BERT: input data format

For each tokenized input sentence, we need to create:

- **input ids**: a sequence of integers identifying each input token to its index number in the BERT tokenizer vocabulary
- **segment mask**: a sequence of 1s and 0s used to identify whether the input is one sentence or two sentences long. For one sentence inputs, this is simply a sequence of 0s. For two sentence inputs, there is a 0 for each token of the first sentence, followed by a 1 for each token of the second sentence
- **attention mask**: a sequence of 1s and 0s, with 1s for all input tokens and 0s for all padding tokens

BERT: fine-tuning for different tasks

BERT for feature extraction
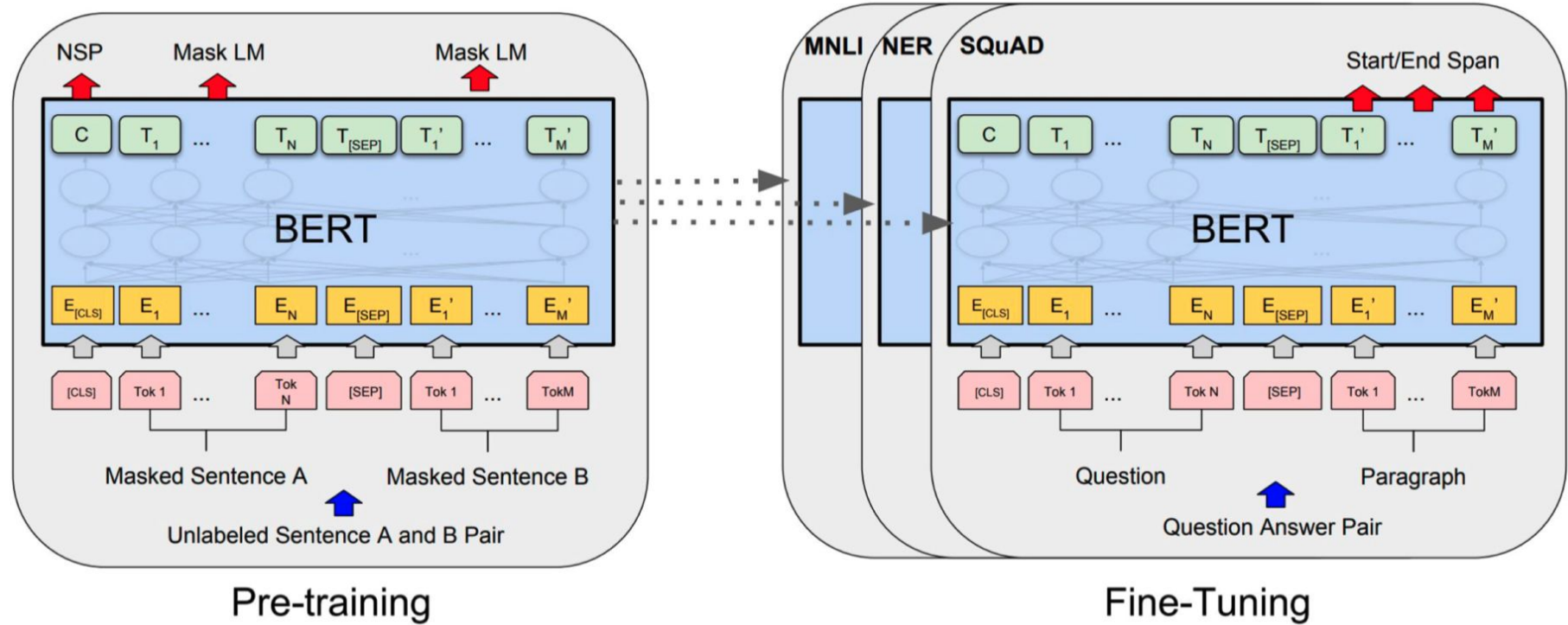
Generate Contexualized Embeddings

The output of each encoder layer along each token's path can be used as a feature representing that token.
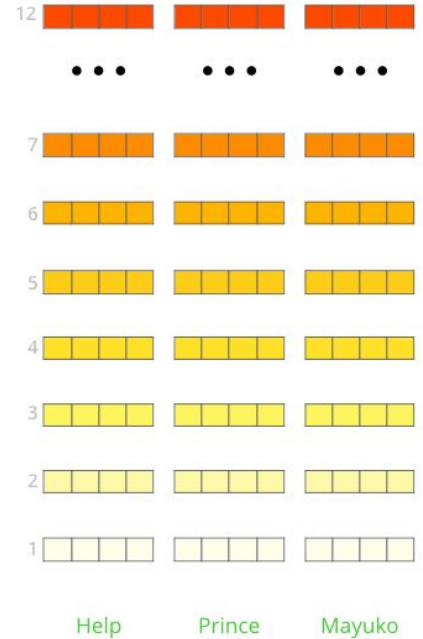
But which one should we use?

BERT for feature extraction

What is the best contextualized embedding for "Help" in that context?
For named-entity recognition task CoNLL-2003 NER

	Dev F1 Score
First Layer	91.0
Last Hidden Layer	94.9
Sum All 12 Layers	95.5
Second-to-Last Hidden Layer	95.6
Sum Last Four Hidden	95.9
Concat Last Four Hidden	96.1
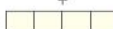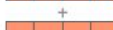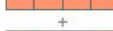
<u>Example:</u> Unaffable -> un, ##aff, ##able
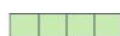
- Single model for 104 languages with a large shared vocabulary (119,547 WordPiece model)
- Non-word-initial units are prefixed with ##
- The first 106 symbols: constants like PAD and UNK
- 36.5% of the vocabulary are non-initial word pieces
- The alphabet consists of 9,997 unique characters that are defined as word-initial (C) and continuation symbols (##C), which together make up 19,994 word pieces
- The rest are multi character word pieces of various length.

WordPiece length distribution

Corpora/Data

Youtube videos

BooksCorpus

SWAG,IMDB,Twitter

Knowledge Graph

Monolingual Corpora (104 languages)

Biomedical corpus

Scientific publications

Clinical notes/EHR

Hierarchical diagnostic codes

English Wikipedia

Graph Neural Net

Pre-trained models

SciBERT

ERNIE(1)

M-BERT

ClinicalBERT

ERNIE(2)

G-BERT

BioBERT

VideoBERT

BERT

TransBERT

Fine-tuned models

Generic & Domain specific NLP tasks (e.g. NER)

DocBERT

Relation classification

PatentBERT

Code Switching
(e.g. English/Hindi mix sentences)

Prediction tasks
(e.g. Hospital readmission)

Video captioning
(classification)

Classification tasks
(e.g. medication recommendation)

- [BERT repo](#)

- [Try out BERT on TPU](#)

- [WordPieces Tokenizer](#)

- [PyTorch Implementation of BERT](#)

GPT-2 & GPT-3

- Transformer-based architecture
- trained to predict the **next** word
- 1.5 billion parameters
- Trained on 8 million web-pages

Output

On language tasks (question answering, reading comprehension, summarization, translation) works well **WITHOUT** fine-tuning

EXAMPLES

Who wrote the book the origin of species?

Correct answer: *Charles Darwin*
Model answer: Charles Darwin

What is the largest state in the U.S. by land mass?

Correct answer: *Alaska*
Model answer: California

GPT-2: language modeling

Both its sun-speckled shade and the cool grass beneath were a welcome respite after the stifling kitchen, and I was glad to relax against the tree's rough, brittle bark and begin my breakfast of buttery, toasted bread and fresh fruit. Even the water was tasty, it was so clean and cold. It almost made up for the lack of...

Correct answer: *coffee*

Model answer: food

EXAMPLE

French sentence:

Un homme a expliqué que l'opération gratuite qu'il avait subie pour soigner une hernie lui permettrait de travailler à nouveau.

Reference translation:

One man explained that the free hernia surgery he'd received will allow him to work again.

Model translation:

```
A man told me that the operation gratuity he had been promised would not allow him to travel.
```

GPT-2: fake news and hype

New AI fake text generator may be too dangerous to ... - The Guardian
https://www.theguardian.com/.../elon-musk-backed-ai-writes-convincing-news-fiction
4 days ago - The Elon Musk-backed nonprofit company OpenAI declines to release research publicly for fear of misuse. The creators of a revolutionary AI system that can write news stories and works of fiction – dubbed "deepfakes for text" – have taken the unusual step of not releasing ...

OpenAI built a text generator so good, it's considered too dangerous to ...
https://techcrunch.com/2019/02/17/openai-text-generator-dangerous/ ▾
12 hours ago - A storm is brewing over a new language model, built by non-profit artificial intelligence research company OpenAI, which it says is so good at ...

The AI Text Generator That's Too Dangerous to Make Public | WIRED
https://www.wired.com/story/ai-text-generator-too-dangerous-to-make-public/ ▾
4 days ago - In 2015, car-and-rocket man Elon Musk joined with influential startup backer Sam Altman to put artificial intelligence on a new, more open ...

Elon Musk-backed AI Company Claims It Made a Text Generator ...
https://gizmodo.com/elon-musk-backed-ai-company-claims-it-made-a-text-gener-183... ▾
Elon Musk-backed AI Company Claims It Made a Text Generator That's **Too Dangerous to** Release · Rhett Jones · Friday 12:15pm · Filed to: OpenAI Filed to: ...

Scientists have made an AI that they think is too dangerous to ...
https://www.weforum.org/.../amazing-new-ai-churns-out-coherent-paragraphs-of-text/ ▾
3 days ago - Sample outputs suggest that the AI system is an extraordinary step forward, producing text rich with context, nuance and even something ...

New AI Fake Text Generator May Be Too Dangerous To ... - Slashdot
https://news.slashdot.org/.../new-ai-fake-text-generator-may-be-too-dangerous-to-rele... ▾
3 days ago - An anonymous reader shares a report: The creators of a revolutionary AI system that can write news stories and works of fiction -- dubbed ...

Top stories

OpenAI built a text generator so good, it's considered too dangerous to release
TechCrunch
11 hours ago

Elon Musk's AI company created a fake news generator it's too scared to make public
BGR.com
9 hours ago

The AI That Can Write A Fake News Story From A Handful Of Words
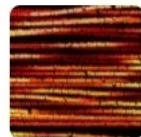NDTV.com
2 hours ago

When Is Technology Too Dangerous to Release to the Public?
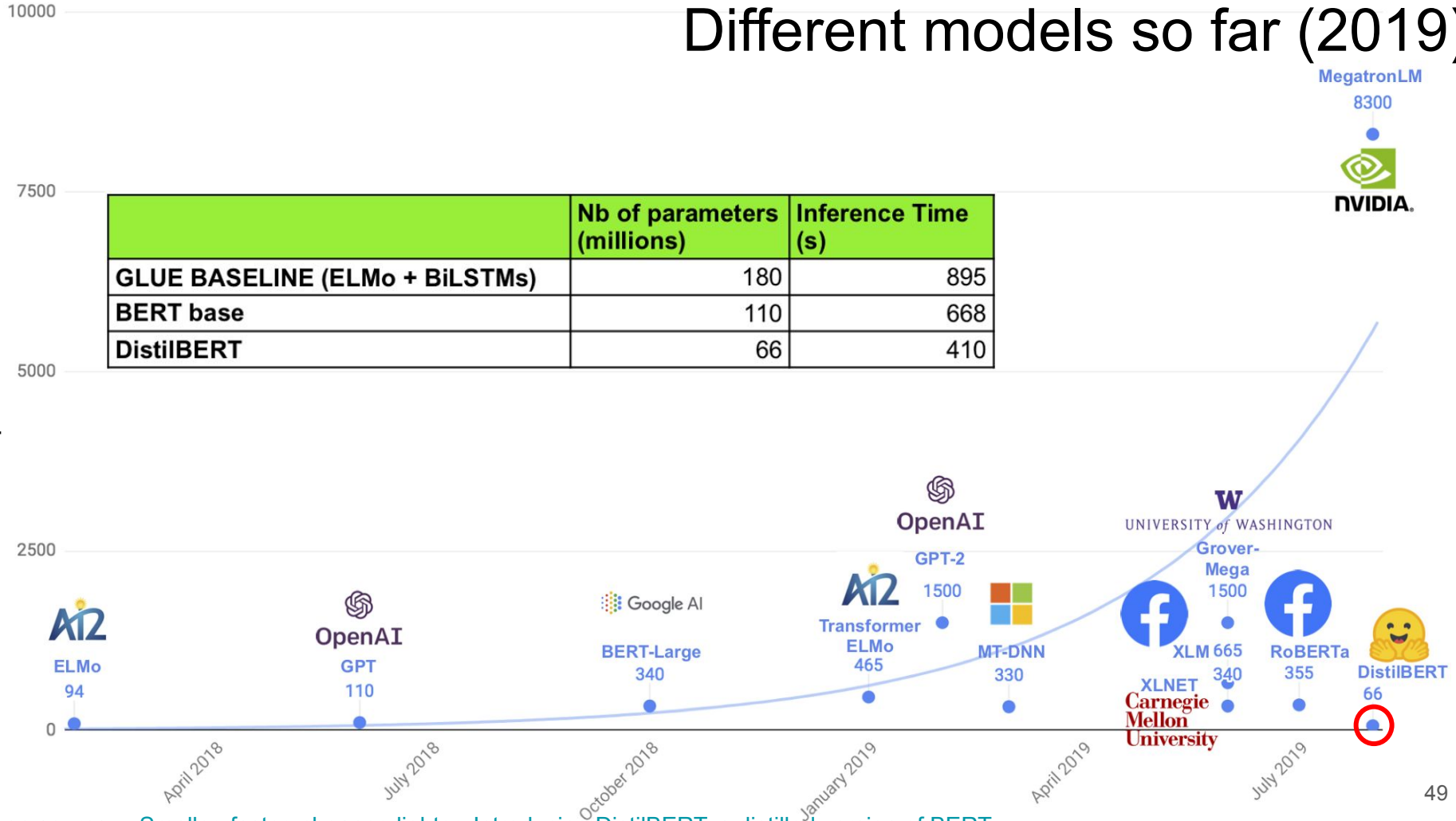Slate · 2 days ago

Scientists Developed an AI So Advanced They Say It's Too Dangerous to Release
ScienceAlert · 6 days ago

Different models so far (2019)
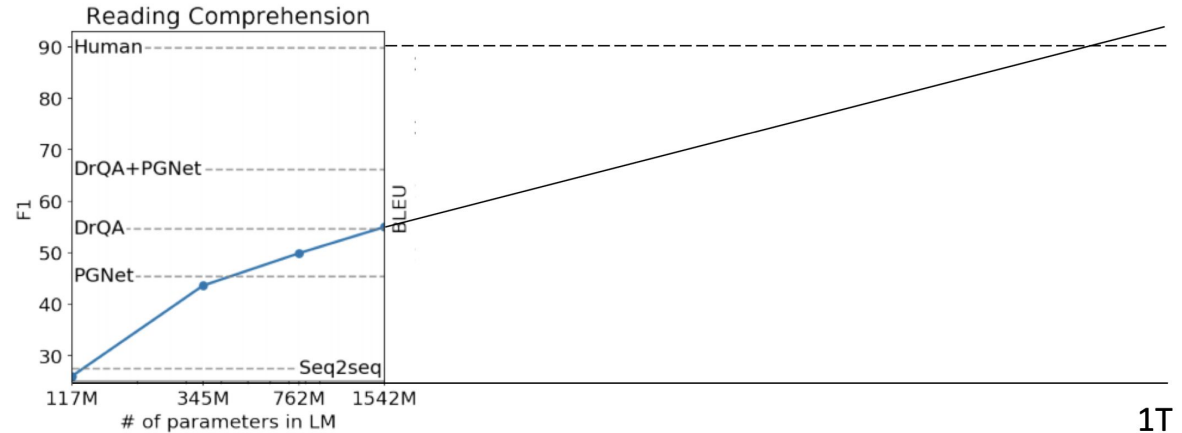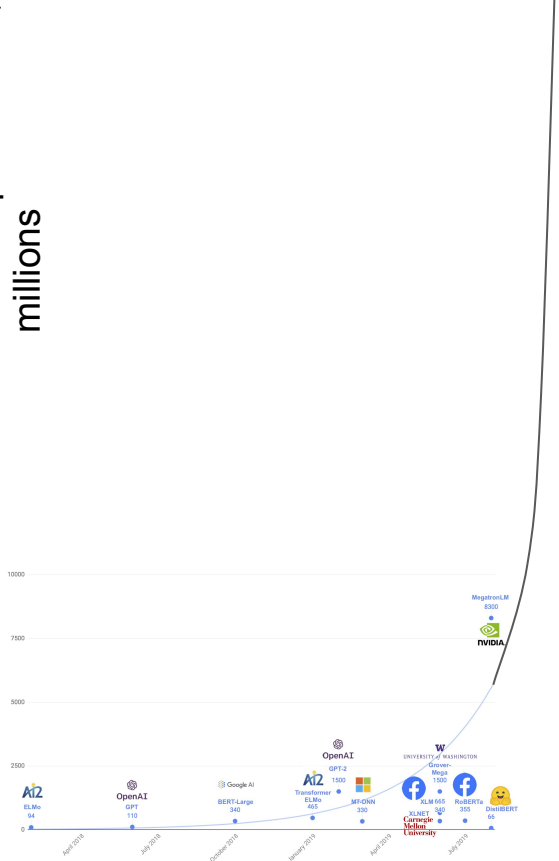
	Nb of parameters (millions)	Inference Time (s)
GLUE BASELINE (ELMo + BiLSTMs)	180	895
BERT base	110	668
DistilBERT	66	410

number of parameters, millions

MegatronLM
8300

nVIDIA.

OpenAI

UNIVERSITY of WASHINGTON

GPT-2
1500

Grover-
Mega
1500

Google AI

Transformer
ELMo
465

MT-DNN
330

OpenAI

XLM 665

RoBERTa
355

DistilBERT
66

ELMo
94

GPT
110

BERT-Large
340

XLNET
340

Carnegie
Mellon
University

April 2018   July 2018   October 2018   January 2019   April 2019   July 2019

49

Image source: <u>Smaller, faster, cheaper, lighter: Introducing DistilBERT, a distilled version of BERT</u>

Latest achievements: GPT-3

GPT-3, May 2020

Proportions are not preserved for visual sake

Hypothesis from Stanford CS224N Lecture 20 (2019)

- GPT-2: 1.5 billion parameters
- GPT-3: **175 billion** parameters

Geoffrey Hinton @geoffreyhinton · Jun 10
Extrapolating the spectacular performance of GPT3 into the future suggests that the answer to life, the universe and everything is just 4.398 trillion parameters.
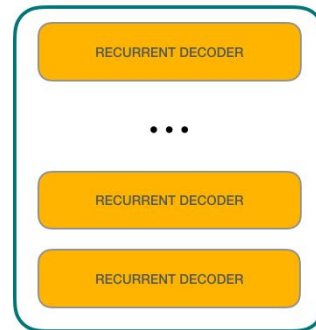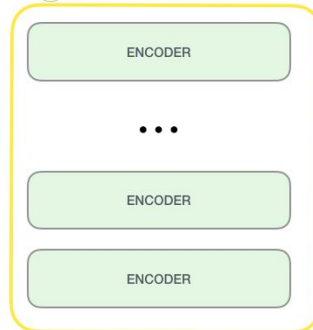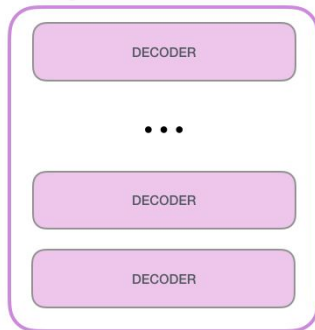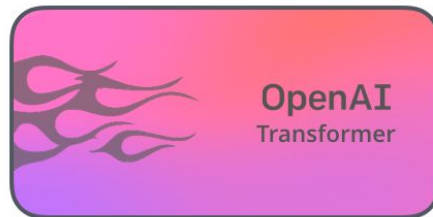
💬 62   🔁 643   ❤️ 3.4K   ⬆️

- [Transformer](#)

- [OpenAI Transformer](#)

- [ELMO](#)

- [BERT](#)

- [BERTology](#)

- [GPT](#)

- [GPT-2](#)

- [GPT-3](#)

Image sources: http://jalammar.github.io/illustrated-bert/
http://jalammar.github.io/illustrated-gpt2/

- Transfer learning caused a giant leap in Computer Vision and Natural Language Processing
  - Which domain is next?
- Using pre-trained BERT might be a good idea in many tasks
  - Or even using DistillBERT
- Better problem statement leads to better results
  - What information is hidden within the data and can be retrieved?