FROM SYMBOLIC REPRESENTATIONS TO CHATGPT

Andrea Santilli

SAPIENZA
UNIVERSITÀ DI ROMA

Natural Language Processing (NLP)

TEXT PROCESSING - NATURAL LANGUAGE PROCESSING

17/3/2019

2 foto

Osteria tipica romana, personale disponibile, cortese, sorridente nonostante la mole di persone da servire!

Cibo sublime! Consigliatissimi amatriciana e tonnarelli con le polpette...andate con la fame perchè le porzioni di pasta sono veramente abbondanti!

Ci tornerò sicuramente!

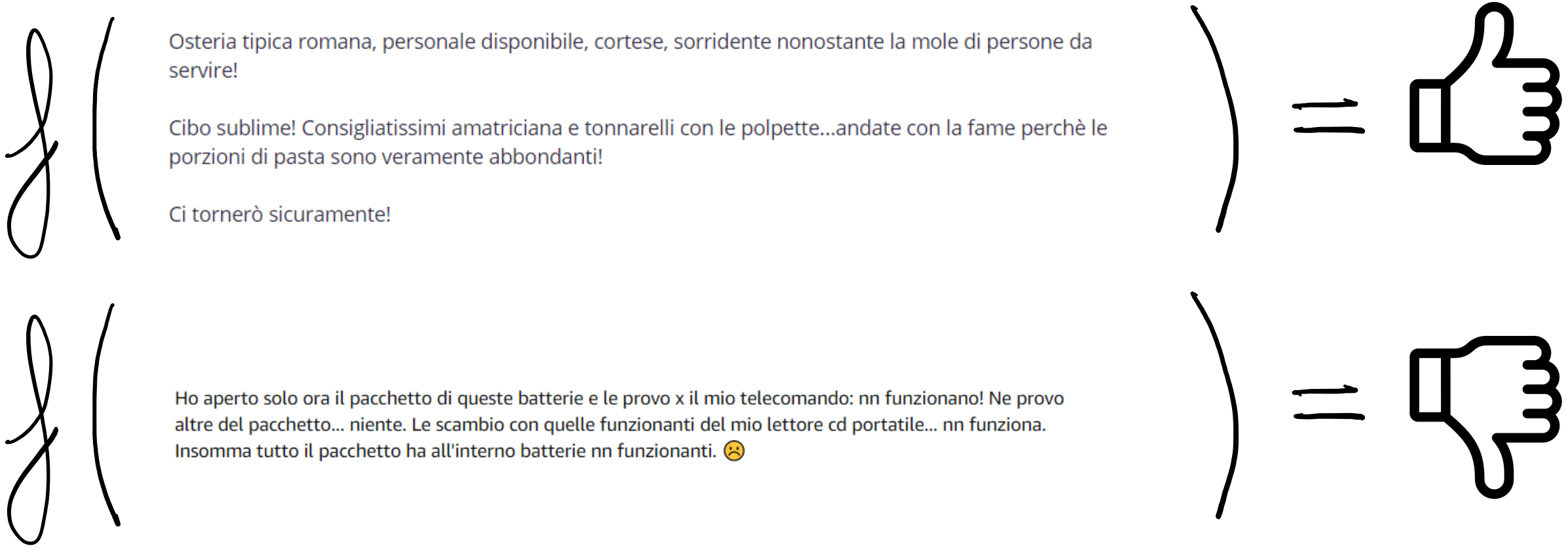
Sono scariche!

Recensito in Italia il 8 febbraio 2018

Nome stile: AA | Taglia: 12 Batterie | Acquisto verificato

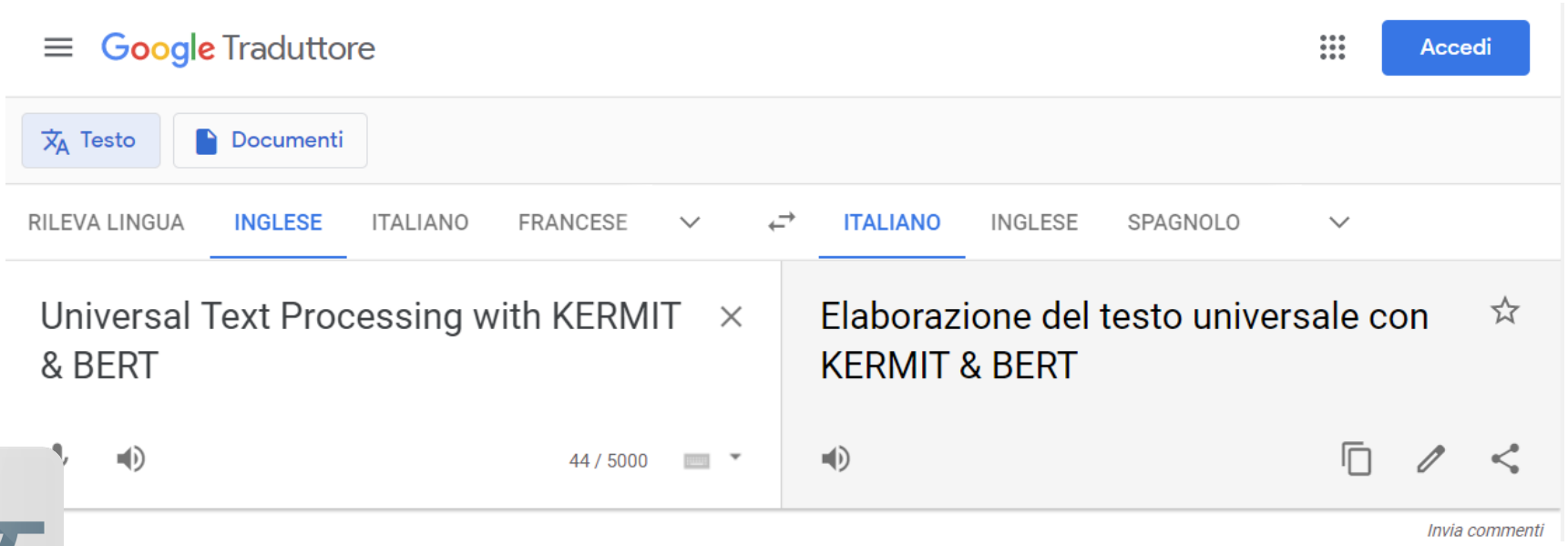
Ho aperto solo ora il pacchetto di queste batterie e le provo x il mio telecomando: nn funzionano! Ne provo altre del pacchetto... niente. Le scambio con quelle funzionanti del mio lettore cd portatile... nn funziona. Insomma tutto il pacchetto ha all'interno batterie nn funzionanti. ☹

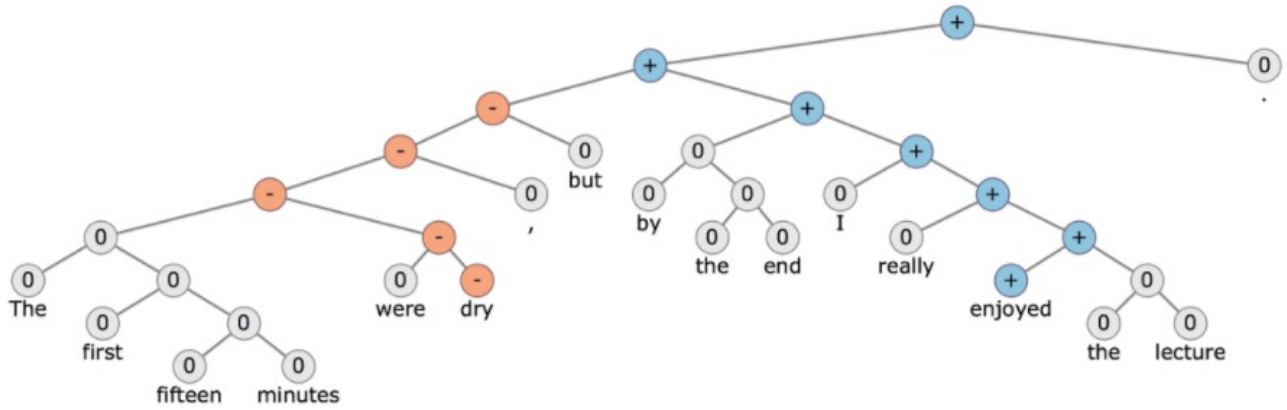
TEXT PROCESSING WITH NEURAL NETWORKS



f is our neural network: a function learned from data

NATURAL LANGUAGE PROCESSING





[Socher 2015]

HOW TO PROCESS TEXT

How do we process text?

1. Neural Networks Representations
2. Symbolic Representations

HOW TO PROCESS TEXT

How do we process text?

1. **Neural Networks Representations**
    A. **Discrete Symbol Representations**
    B. Distributed Representations
2. Symbolic Representations

SYMBOLIC AI - DARTMOUTH CONFERENCE (1956)

A Proposal for the

DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE

June 17 - Aug. 16

We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

The following are some aspects of the artificial intelligence problem:

1) Automatic Computers

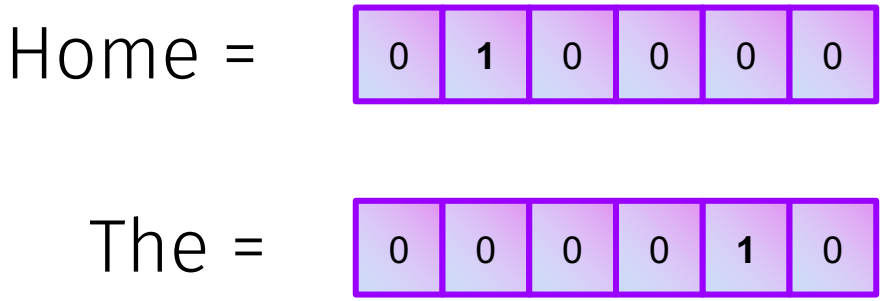
If a machine can do a job, then an automatic calculator can

SYMBOLIC AI - DARTMOUTH CONFERENCE (1956)

If a machine can do a job, then an automatic calculator can

REPRESENTING WORDS AS DISCRETE SYMBOLS

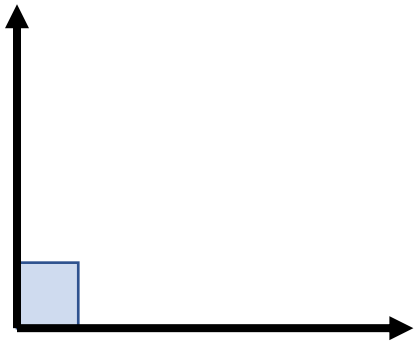
- We can represent words as discrete symbols using a **one-hot representation** (**localist representation**)



Vector Dimension = Number of words in the Vocabulary

ONE-HOT REPRESENTATION

- This representation encodes just the presence or not of the word, **it doesn't encode the meaning!**
- Words vectors are **orthonormal**
- There is not the notion of similarity between words



BAG OF WORDS

- We can use one-hot representations to encode words in a sentence
- However we have to find a way to encode a sentence
- A possible solution can be using **Bag of Words (BOW):**

A BOW representation of a sentence is a sum of all the one-hot representation of each word in the sentence

$$\mathbf{BOW}(s) = \sum_{w \in S} onehot(w)$$

Where ***s*** is our sentence and ***w*** are the single words in the sentence

BAG OF WORDS

s = "The cat is on …"

- Assuming Vocabulary of 6
- Vocabulary must be specified beforehand

onehot(The) = | 0 | **1** | 0 | 0 | 0 | 0 |

onehot(cat) = | **1** | 0 | 0 | 0 | 0 | 0 |

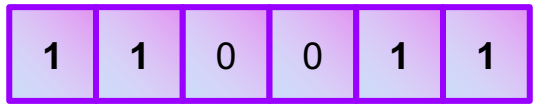
onehot(is) = | 0 | 0 | 0 | 0 | **1** | 0 |

onehot(on) = | 0 | 0 | 0 | 0 | 0 | **1** |

BOW(s) = | **1** | **1** | 0 | 0 | **1** | **1** |

BAG OF WORDS

s = “The cat is on …”

BOW(s) = 

- Each dimension of the BOW counts how many words are in the sentence
- **We lose the order of the words!** (Hence the name **Bag**)
- We just know what words were in the sentence and how many
- The position of each 1 in the vector is not relevant (we just need the information that the word is there)

BAG OF WORDS

s = “The cat is on …”

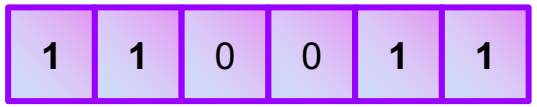
s = “cat The is on …”

s = “on cat is The …”

s = “cat on The is …”

⋮

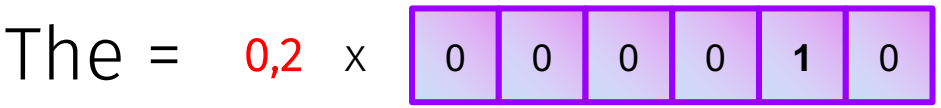
s = “on The cat is …”

BOW(s) = 

ENHANCED BOW

- Usually some terms (like determinative articles) appear with higher frequency in a document
- We can penalize these terms that usually are not very informative
- We use a penalizing factor to rescale the one-hot vector of a word

$$\mathbf{BOW}(s) = \sum_{w_i \in S} \lambda_i \, onehot(w_i)$$



- A common choice of penalizing factor is the function **tf-idf**

N-GRAM MODEL

N-gram model:

Assume conditional dependence only on the previous N words

$$p(\mathbf{x}) \approx \prod_{t=1}^{T} p(x_t \mid x_{t-N-1}, \ldots, x_{t-1})$$

Modeling

Modeling **word**

Modeling word **probabilities**

word probabilities **is**

probabilities is **really**

is really **difficult**

$$p(x_1)$$
$$p(x_2|x_1)$$
$$p(x_3|x_2, x_1)$$
$$p(x_4|x_3, x_2)$$
$$p(x_5|x_4, x_3)$$
$$p(x_6|x_5, x_4)$$

BOW OF N-GRAMS

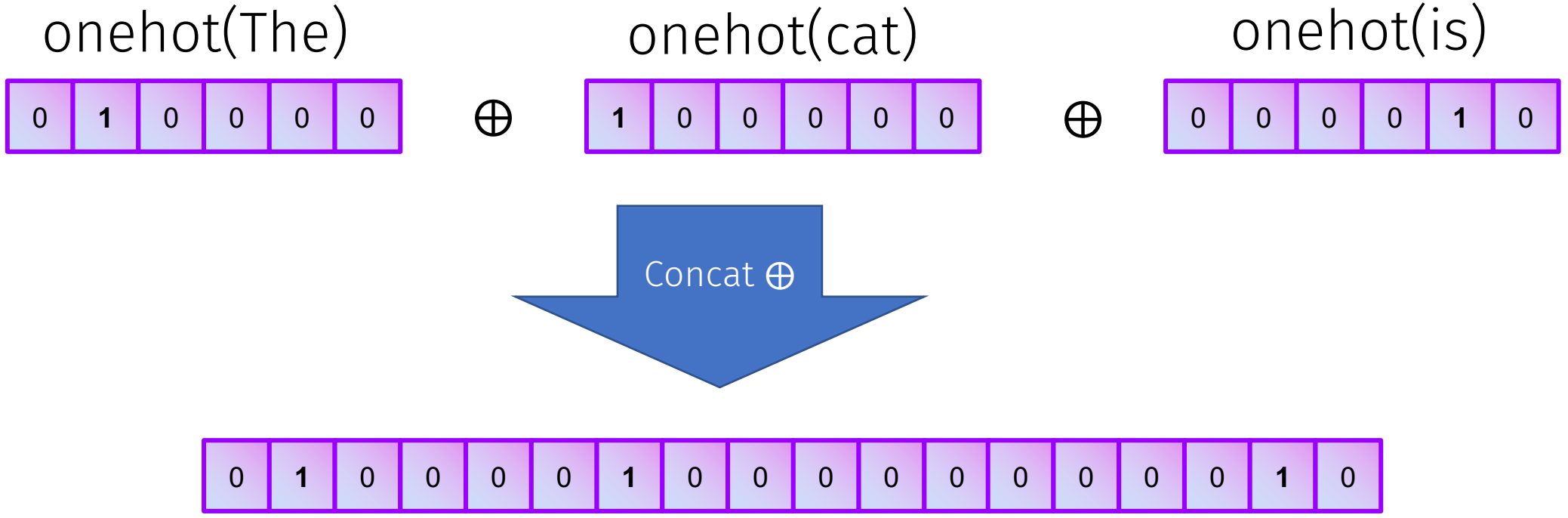
BoW of N-grams allows to have a simple model that encode some sequence information (N past words)

We can use two strategies to create BoW of N-grams:

1. Concatenate N one-hot word representations
2. Directly compute the BoW of the N-grams

CONCATENATE N BOW REPRESENTATIONS

s = "The cat is on …"



We calculate the onehot for all the N-grams and then sum the representations (like BoW but with N-grams)

DIRECTLY COMPUTE THE BOW OF THE N-GRAMS

We create the one-hot representations using as vocabulary N-grams

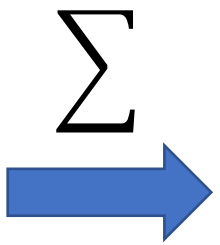
s = "The cat is on the table"

onehot("The cat is") = | 0 | 1 | 0 | 0 | 0 | 0 |

onehot("cat is on") = | 1 | 0 | 0 | 0 | 0 | 0 |

onehot("is on the") = | 0 | 0 | 0 | 0 | 1 | 0 |

onehot("on the table") = | 0 | 0 | 0 | 0 | 0 | 1 |



3-gram-BOW(s) | 1 | 1 | 0 | 0 | 1 | 1 |

How do we process text?

1. **Neural Networks Representations**
   A. Discrete Symbol Representations
   B. Distributed Representations
2. Symbolic Representations

DISTRIBUTIONAL SEMANTIC

Distributional semantics: A word's meaning is given by the words that frequently appear close-by

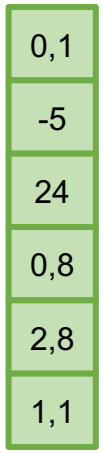
"You shall know a word by the company it keeps" (J. R. Firth 1957: 11)

...government debt problems turning into banking crises as happened in 2009 ...
...saying that Europe needs unified banking regulation to replace the...
...India has just given its banking system a shot in the arm...

When a word w appears in a text, its context is the set of words that appear nearby (within a fixed-size window).

DISTRIBUTED REPRESENTATION

We will build a dense vector for each word, chosen so that it is similar to vectors of words that appear in similar contexts

Paris = 

They are a **distributed representation** in the sense that the meaning is distributed among all the dimensions

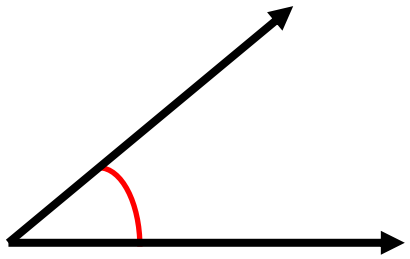
Note: word vectors are also called word embeddings or (neural) word representations

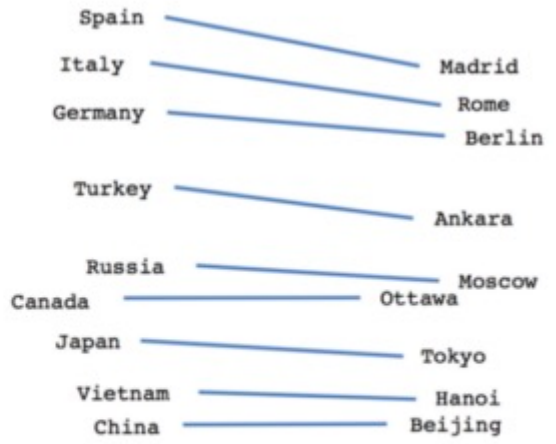
DISTRIBUTED REPRESENTATION

We can use word embedding to do things like:

King – Man + Woman = Queen

Rome – Italy + France = Paris

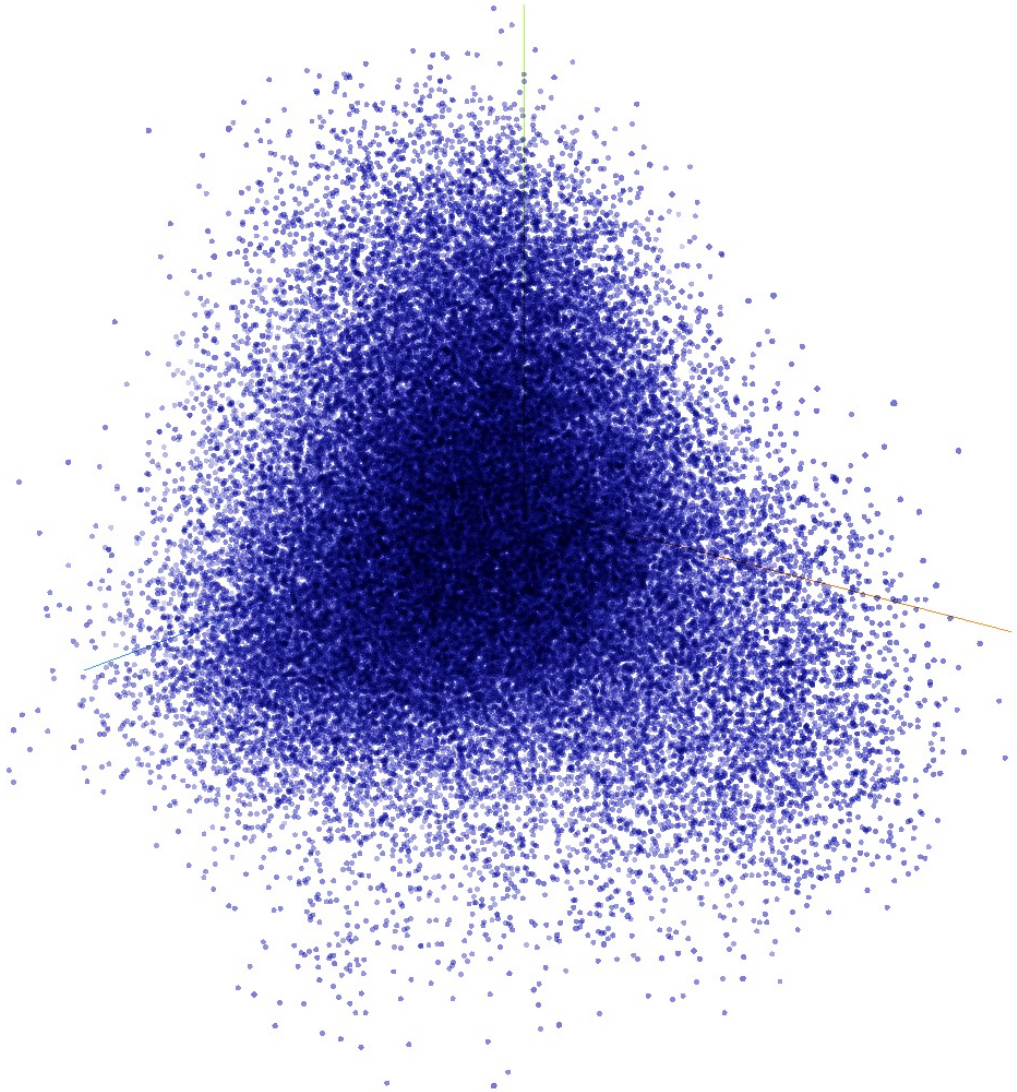


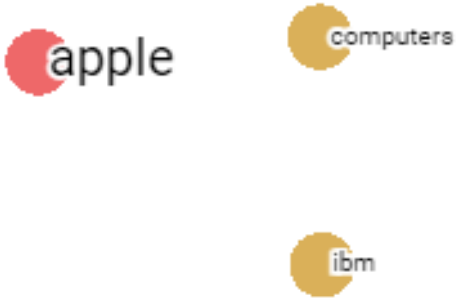


Country-Capital

We have now the notion of words similarity

DISTRIBUTED REPRESENTATIONS





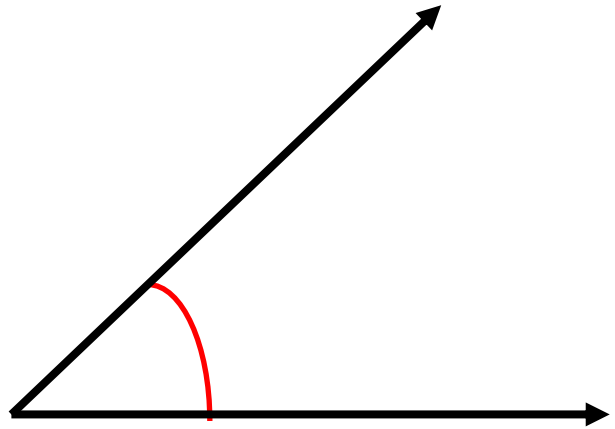
https://projector.tensorflow.org/

SIMILARITY MEASURES

In order to measure the similarity between two vectors (words) we can use different metrics:

- Dot product

$$\mathbf{x}^\top \mathbf{y}$$



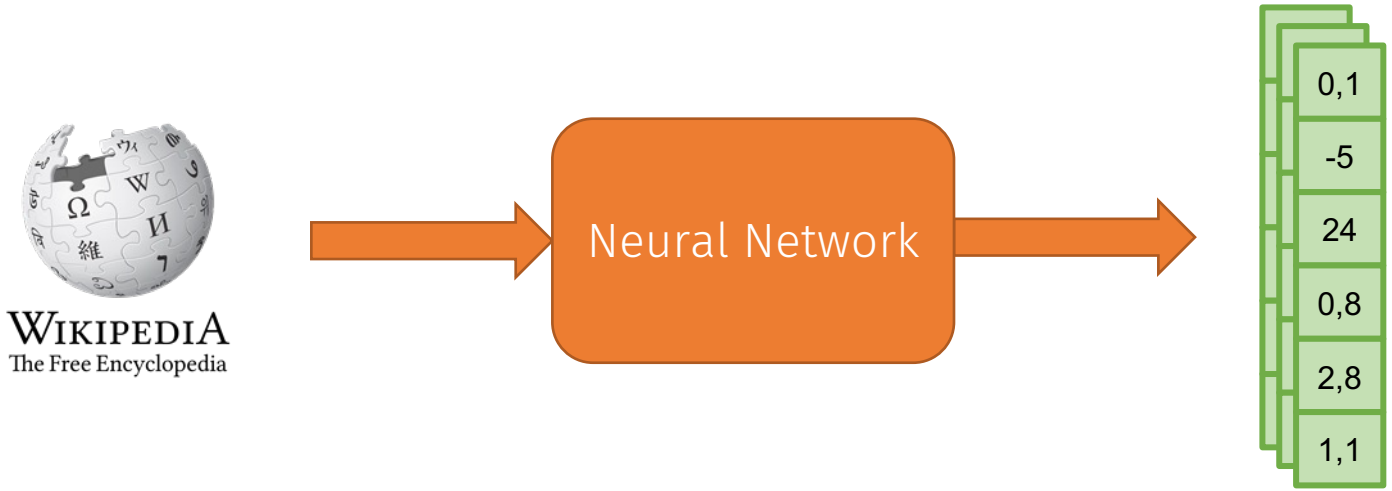
- Cosine Similarity

$$\frac{\mathbf{x}^\top \mathbf{y}}{||\mathbf{x}||\,||\mathbf{y}||}$$

The cosine similarity is a normalized version of the dot product, it measure the cosine of the angle between the two vectors. The value of the cosine similarity is always $\in [-1, 1]$

How to obtain distributed representations?

We can use a neural network with a lot of contexts and words!



...but how??? (more on this in few slides)

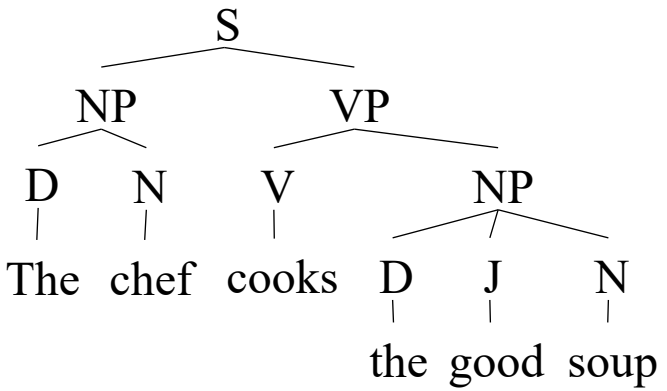
How do we process text?

1. Neural Networks Representations
2. **Symbolic Representations**
   A. Symbolic Syntactic Representations
   B. Symbolic Semantic Representations

SYMBOLIC SYNTACTIC REPRESENTATIONS

Syntax: we can use a grammar

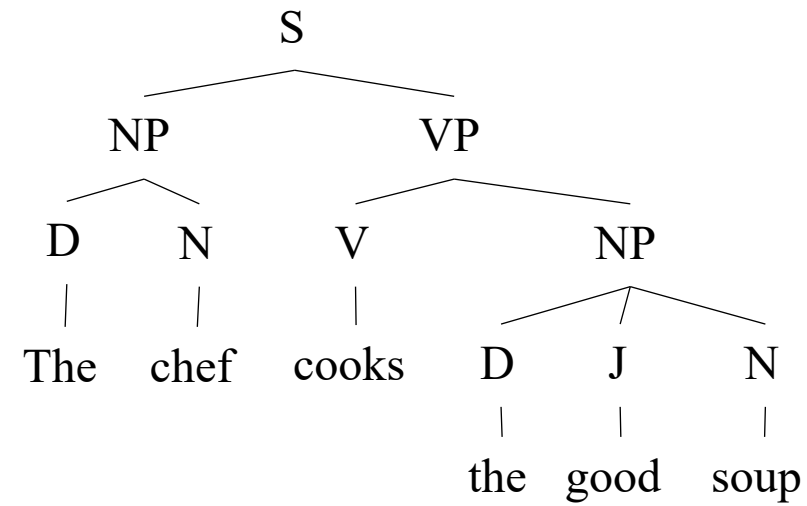
‘The cheff cooks the good soup’



SYMBOLIC SYNTACTIC REPRESENTATIONS

‘The cheff cooks the good soup’

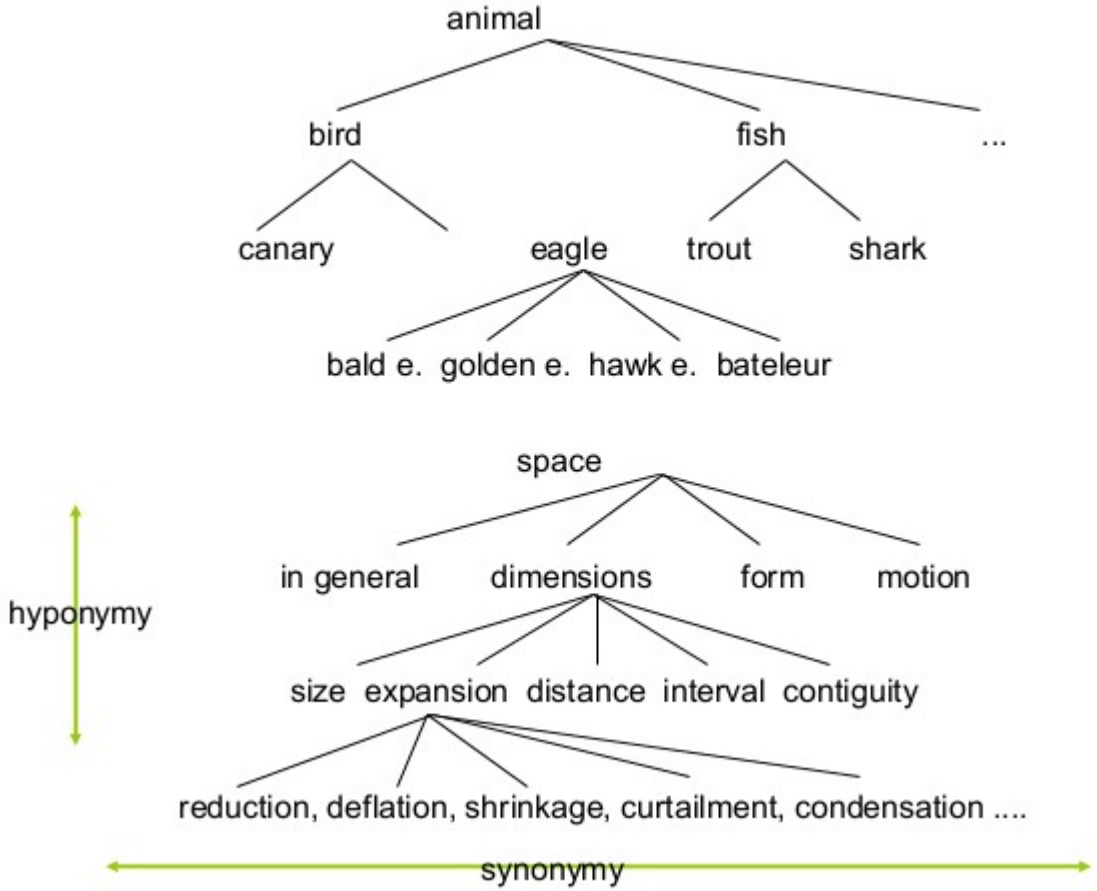
S -> NP VP
NP -> D N
NP -> D J N
VP -> V NP
D -> the | that | those
J -> good | bad
N -> chef | soup
V -> cook | go



SYMBOLIC SEMANTIC REPRESENTATIONS

Semantic: we can use a thesaurus (dictionary with steroids)

- WordNet/Babelnet thesaurus containing lists of synonym sets and hypernyms ("is a" relationships)

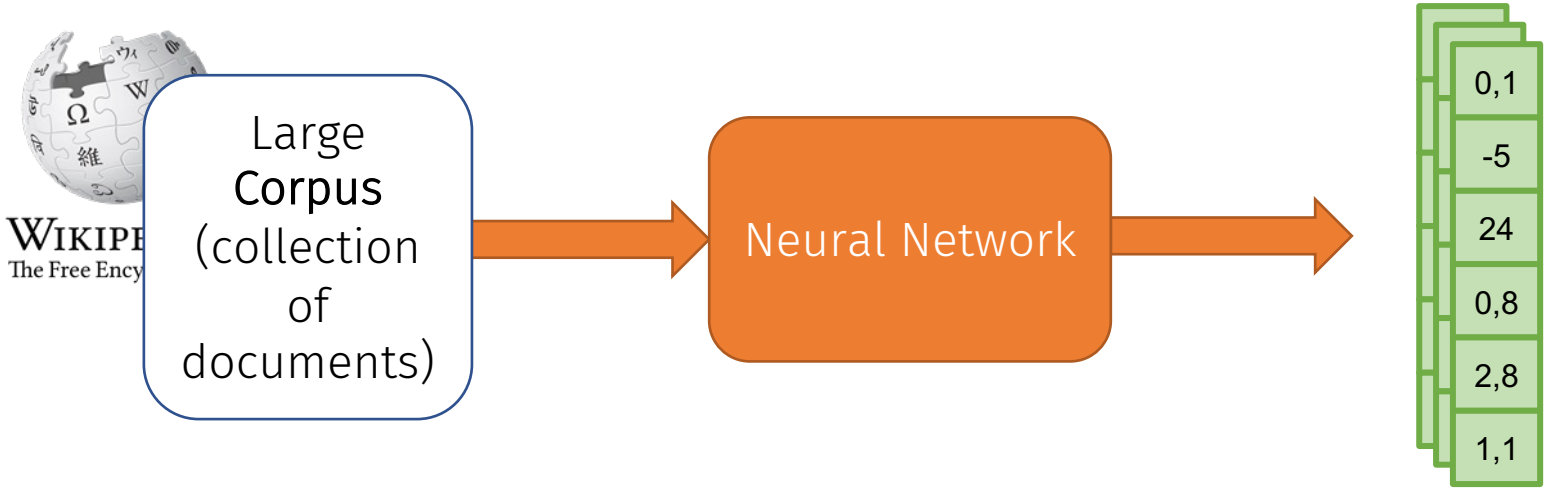


HOW TO CONSTRUCT DISTRIBUTED REPRESENTATIONS?

DISTRIBUTIONAL SEMANTIC

How to construct distributed representations?

We can use a neural network with a lot of contexts and words!



Main Idea:

"You shall know a word by the company it keeps" (J. R. Firth 1957: 11)

SELF-SUPERVISED LEARNING

Data $\{x\}_i$ *task_spec*

Self-supervised learning is supervised learning without human-annotated labels

- The corpus is not human-annotated
- The corpus is just raw text

GPT-3 [Brown et al.]

Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.
The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

Figure 3.14: The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%).

CONSTRUCT DISTRIBUTED REPRESENTATION

How to construct distributed representations?

1. **Word Embeddings**
    A. word2vec
    B. glove
2. Contextual Word Embeddings
    A. BERT

CONSTRUCT DISTRIBUTED REPRESENTATION

How to construct distributed representations?

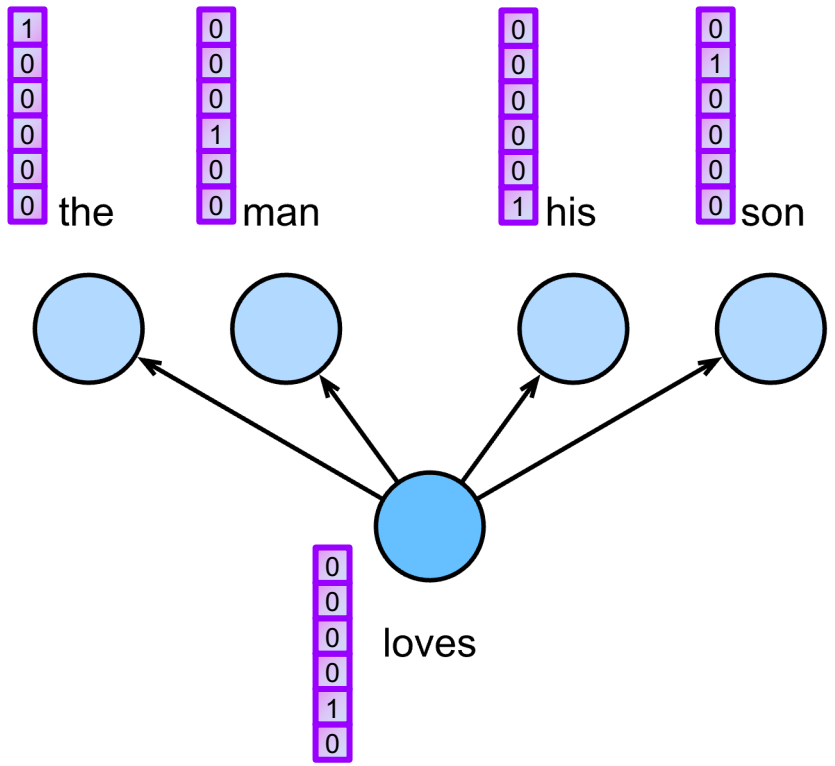
1. **Word Embeddings**
    A. word2vec
    B. glove
2. Contextual Word Embeddings
    A. BERT

WORD2VEC - SKIPGRAM

- Given a word predict its context (fixed of size m)
- Word are encoded using one-hot representation

"The man loves his son"

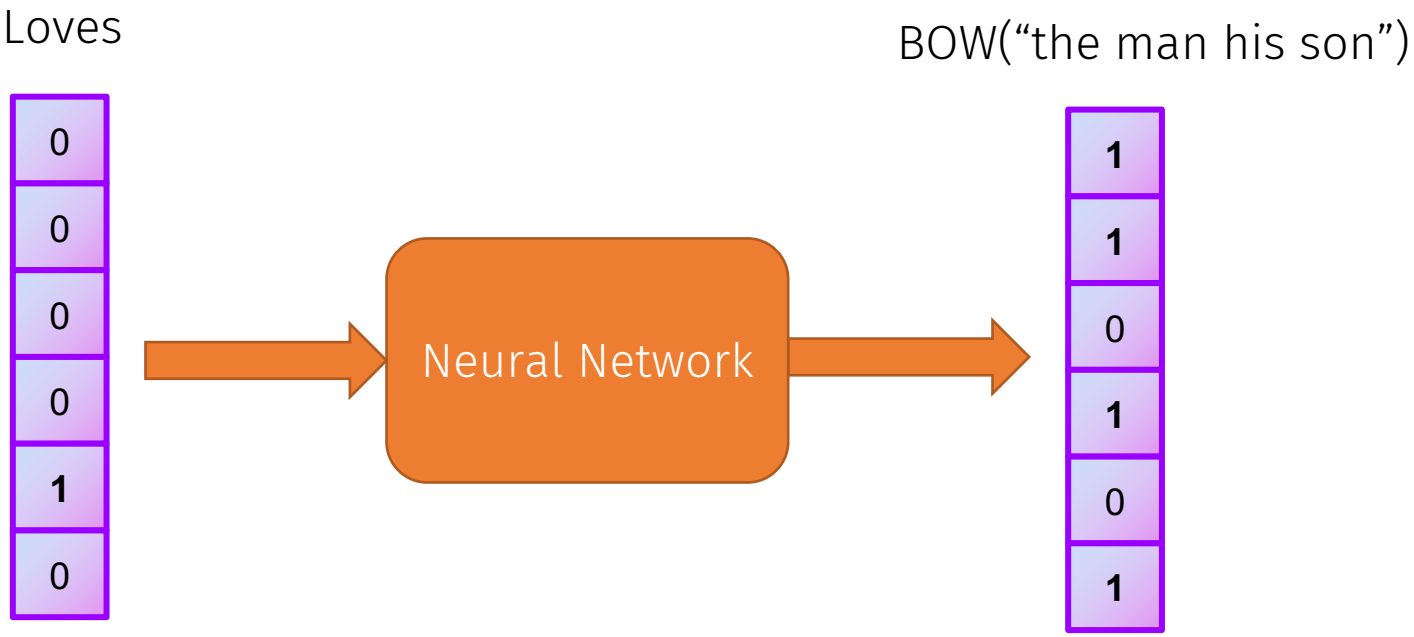
Setting $m = 2$



WORD2VEC - SKIPGRAM

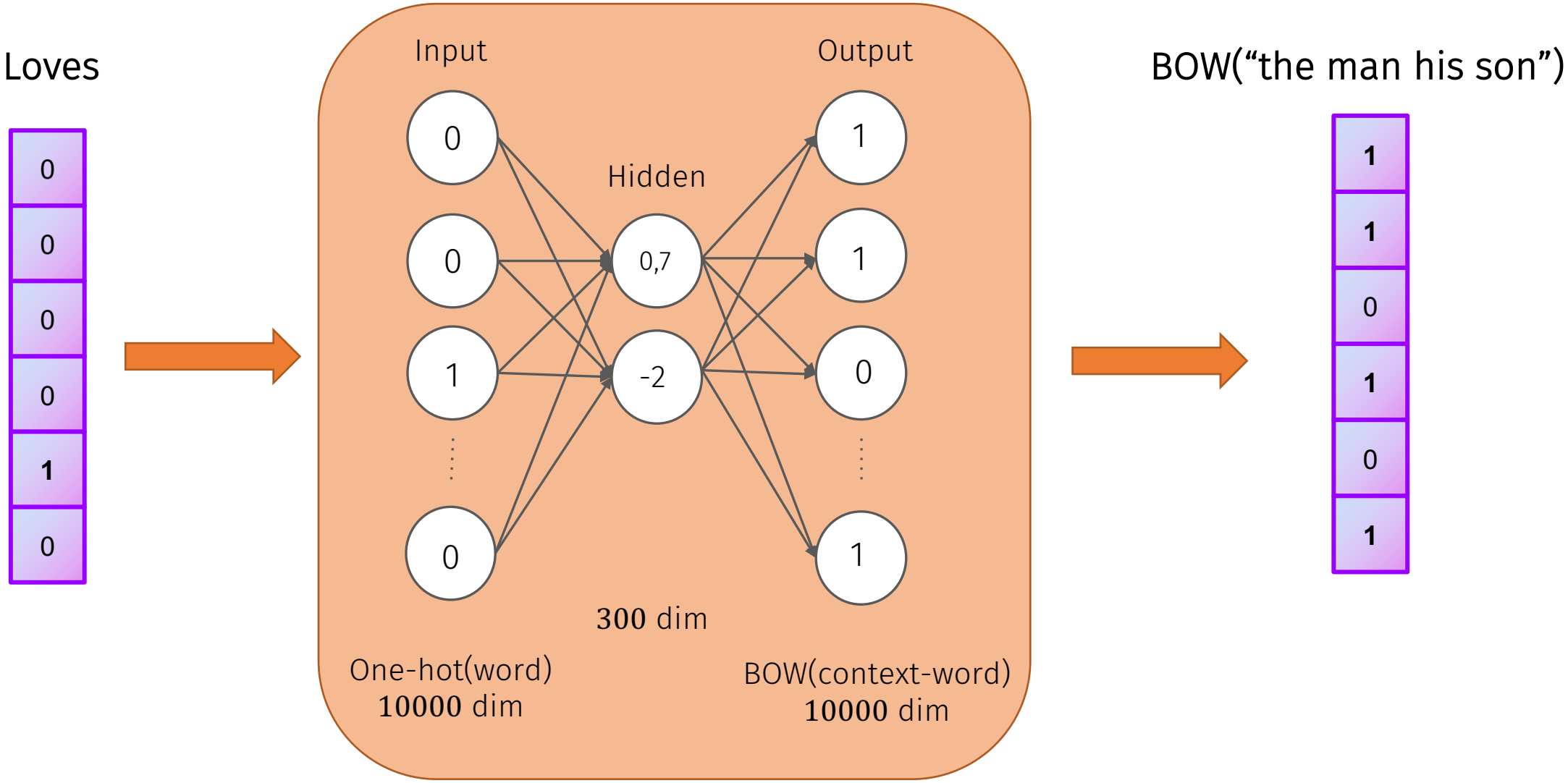
Let's simplify the model:

- Assume that context words are **independently generated** given center word
- Given a center word predict the BOW of the context



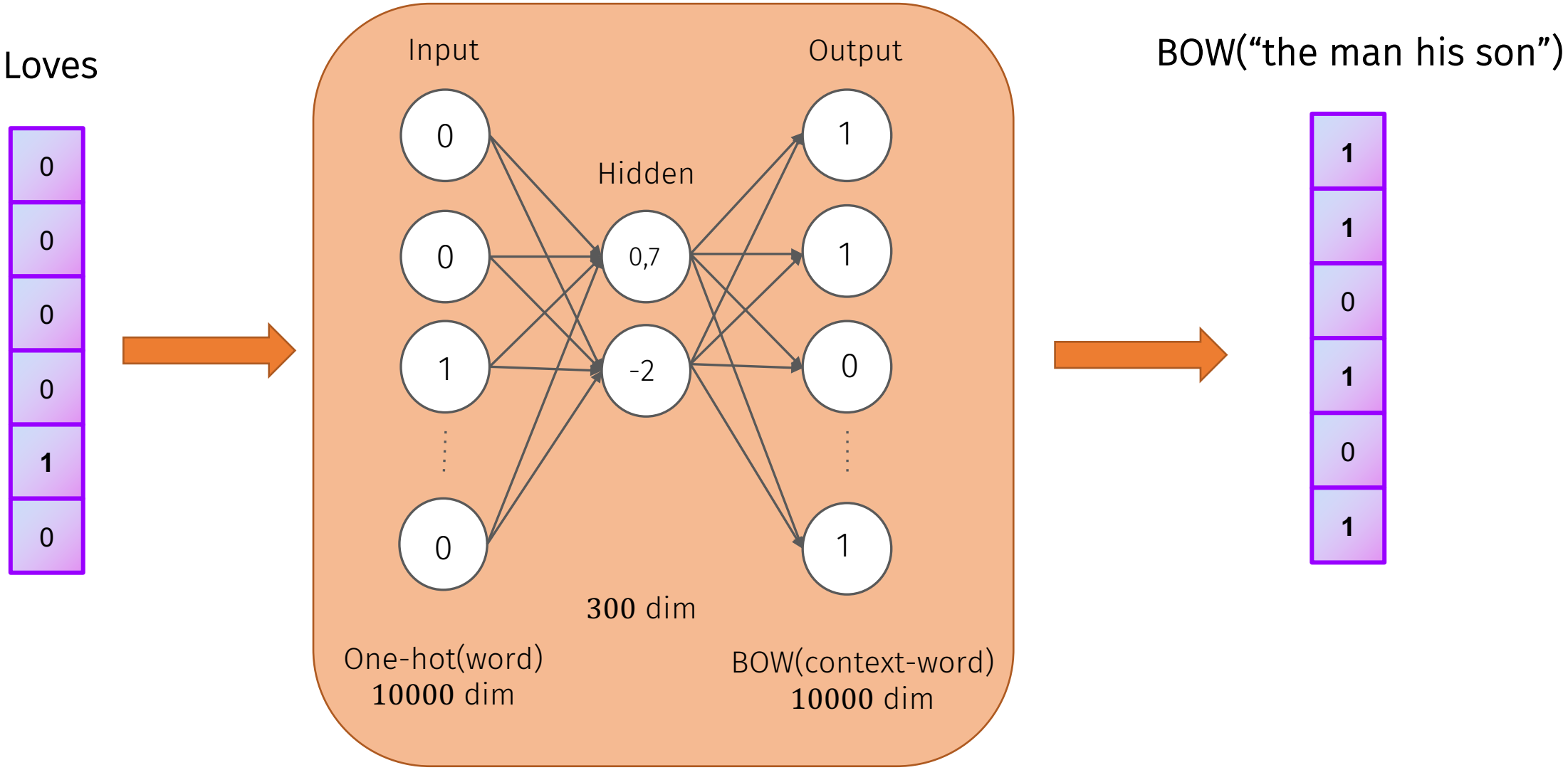
WORD2VEC - SKIPGRAM

- The neural network is just a shallow 1-hidden-layer MLP



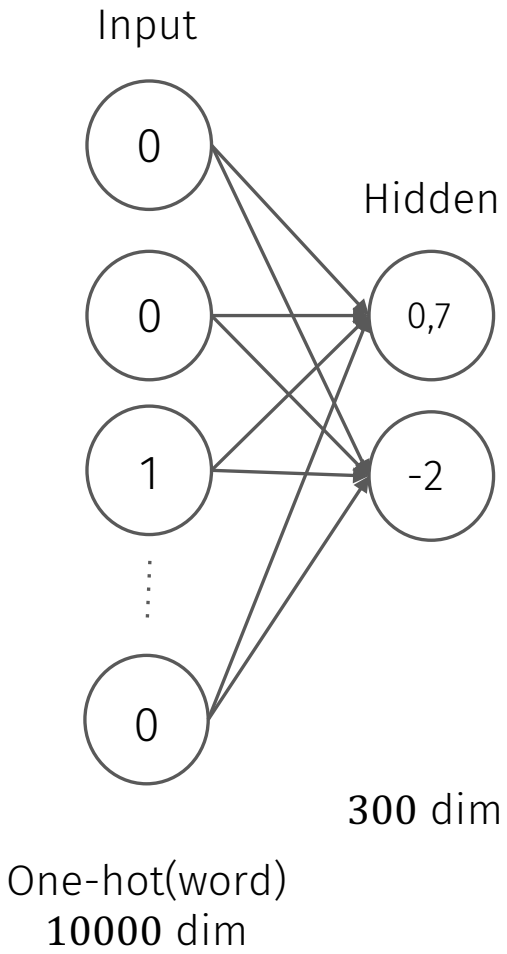
WORD2VEC - SKIPGRAM

- One trained, we throw away the network and keep just the weights of the **Hidden Layer**



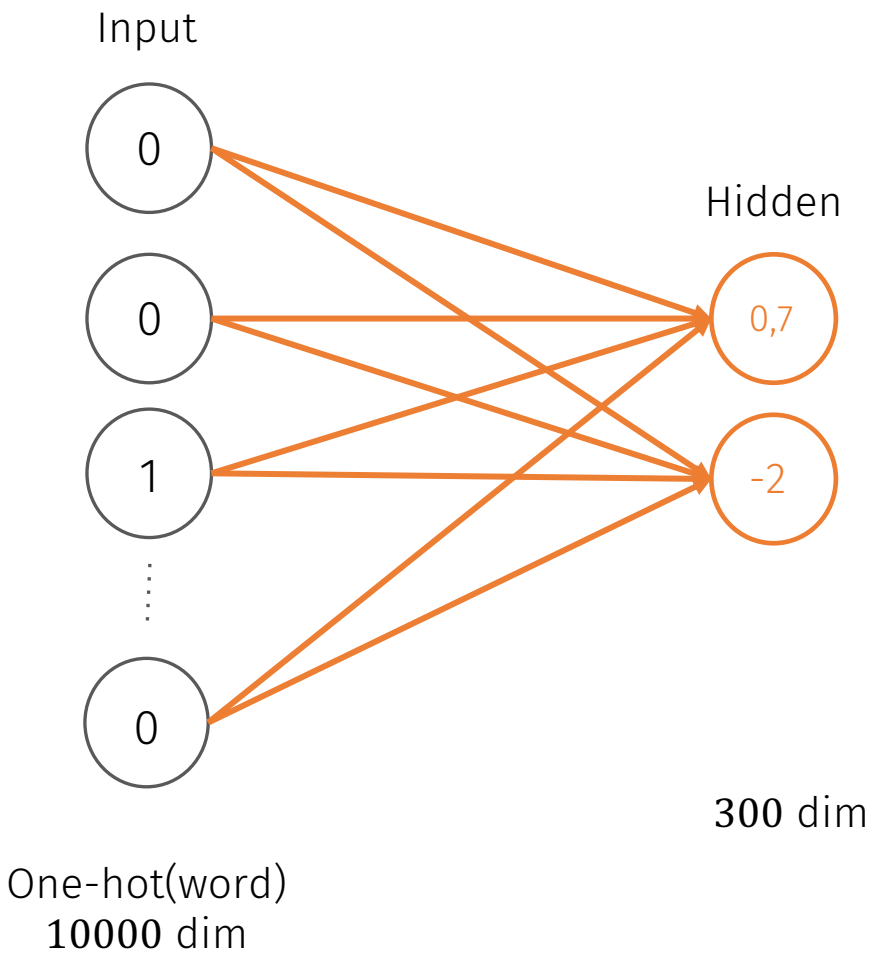
WORD2VEC - SKIPGRAM

- One trained, we throw away the network and keep just the weights of the **Hidden Layer**



WORD2VEC - SKIPGRAM

- One trained, we throw away the network and keep just the weights of the **Hidden Layer**



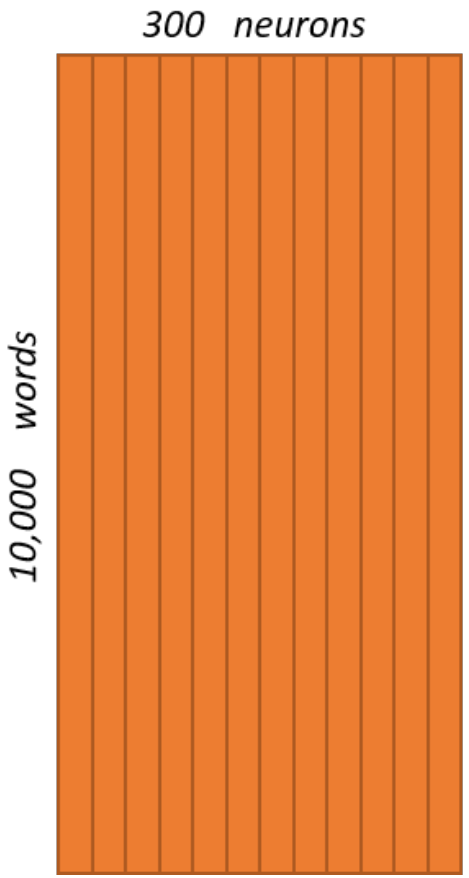
WORD2VEC -SKIPGRAM

MLP

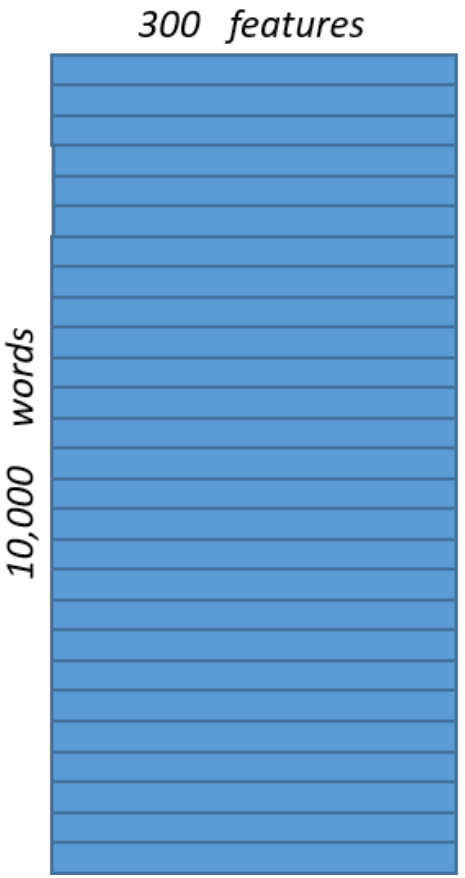
$$h_1 = W^T x$$

$$out = softmax(W_{out}^T h_1)$$

Hidden Layer Weight Matrix



Word Vector Lookup Table!



CONSTRUCT DISTRIBUTED REPRESENTATION

How to construct distributed representations?

1. **Word Embeddings**
    A. word2vec
    B. glove
2. Contextual Word Embeddings
    A. BERT

CONSTRUCT DISTRIBUTED REPRESENTATION

How to construct distributed representations?

1. Word Embeddings
   A. word2vec
   B. glove
2. **Contextual Word Embeddings**
   A. BERT

TOWARDS CONTEXTUAL WORD EMBEDDINGS

- Word Embeddings are good but have a **context-compression** problem
- Two **homograph** words (same textual form but different meaning) appear in different context with different meanings
- These contexts are compressed in just one vector
- We can’t disambiguate between the meaning of the two words

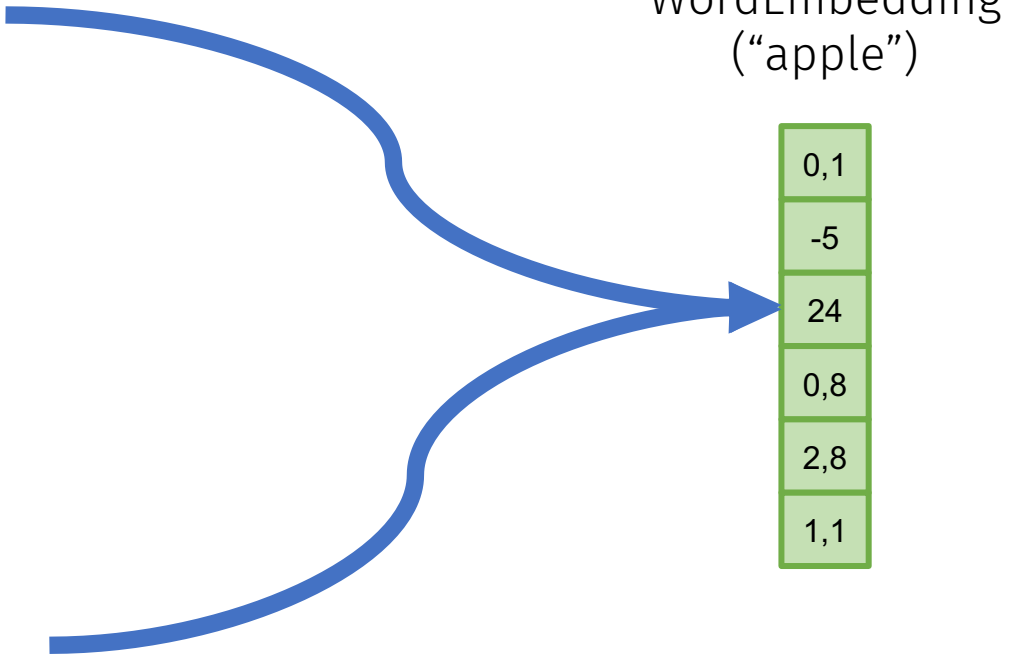
TOWARDS CONTEXTUAL WORD EMBEDDINGS

CONTEXT 1

… Nissan says it's not in apple car discussions …
… apple reportedly developing next-gen ultra-thin displays …

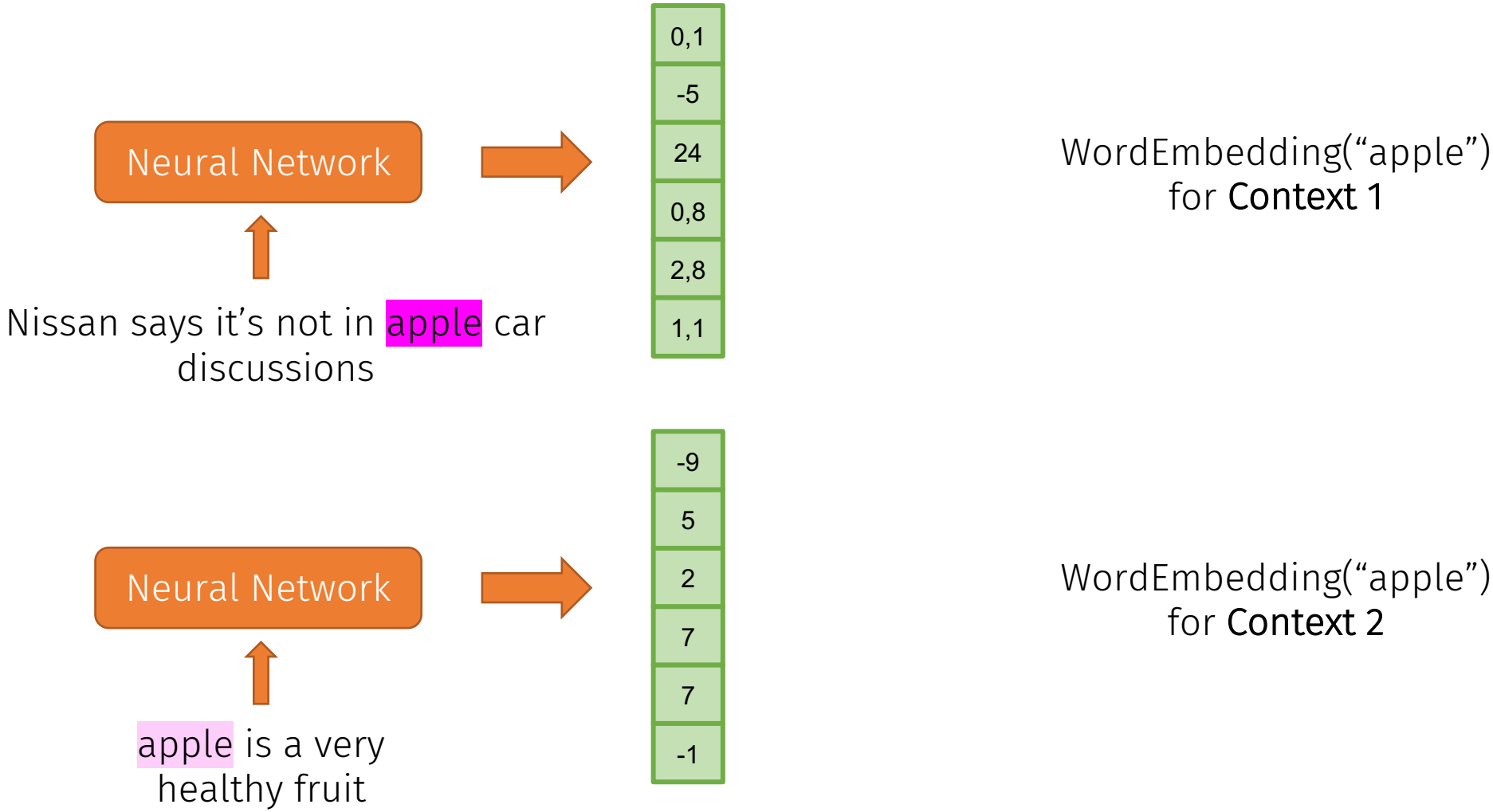
CONTEXT 2

… apple is a very healthy fruit …
… apple and bananas are considered to be …



TOWARDS CONTEXTUAL WORD EMBEDDINGS

- We can use a Neural Network to generate the appropriate Word Embedding **given the context**

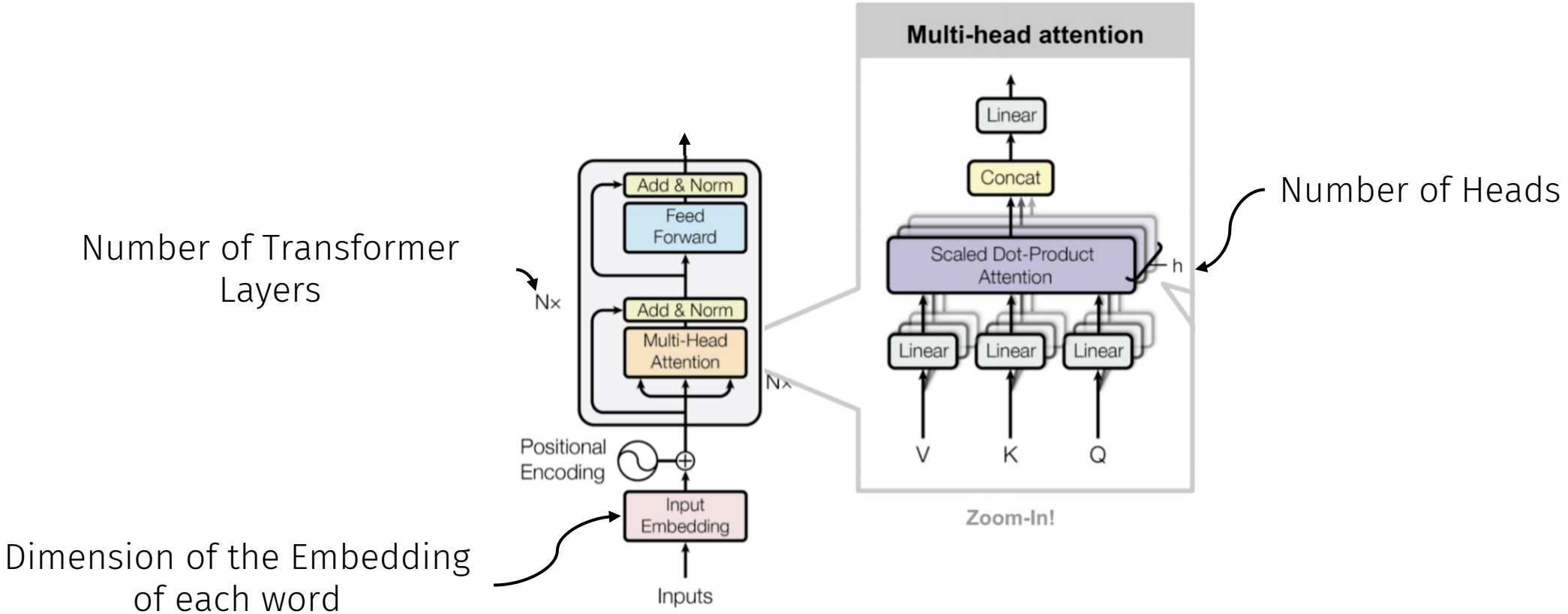


CONTEXTUAL WORD EMBEDDINGS

- These Word Embedding are called Contextual Word Embeddings because they depend on the context
- Different Networks use different techniques to create these Embeddings
- Transformer is the standard go-to architecture
- Popular Models: ELMO, BERT, GPT-1/2/3 , XLNET, ELECTRA, T5

BERT

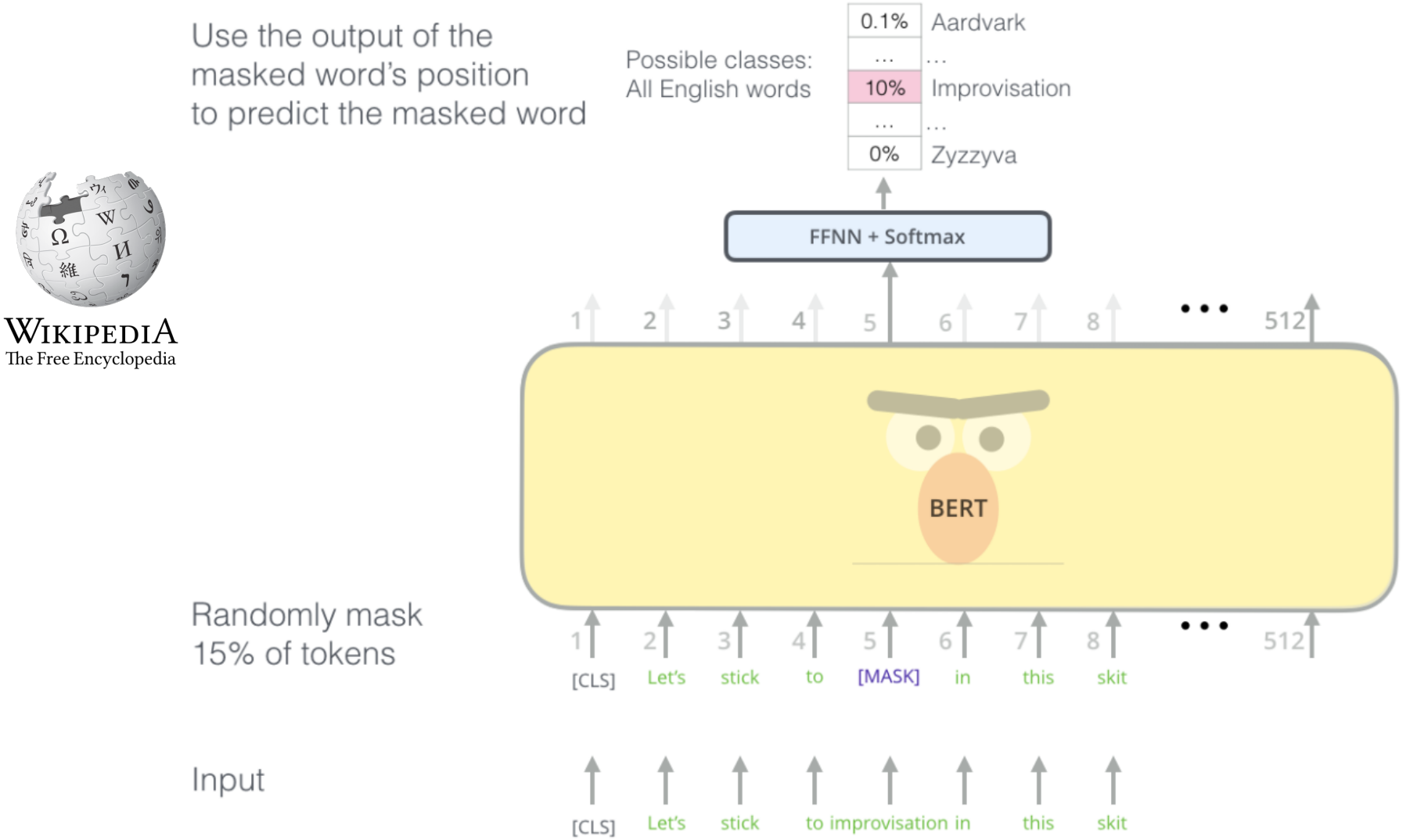
- Bert is a standard Transformer **Encoder** Layer
- It is available is two size:
  - **BASE** - 12 Transformer Layers, 12 Heads, 768 Dim Embedding
  - **LARGE** – 24 Transformer Layers, 16 Heads, 1024 Dim Embedding



BERT TRAINING

- BERT is trained with self-supervised learning using a Masked Language Modeling Task (Denoising Autoencoder)
- Trained on a large corpus composed of:
  - 2500M words from English Wikipedia
  - 800M words from BooksCorpus (11K copyright-free books)
- Max word length during training 512
- Once trained we can use Contextual Word Embedding Word vectors

BERT – MASKED LANGUAGE MODELING

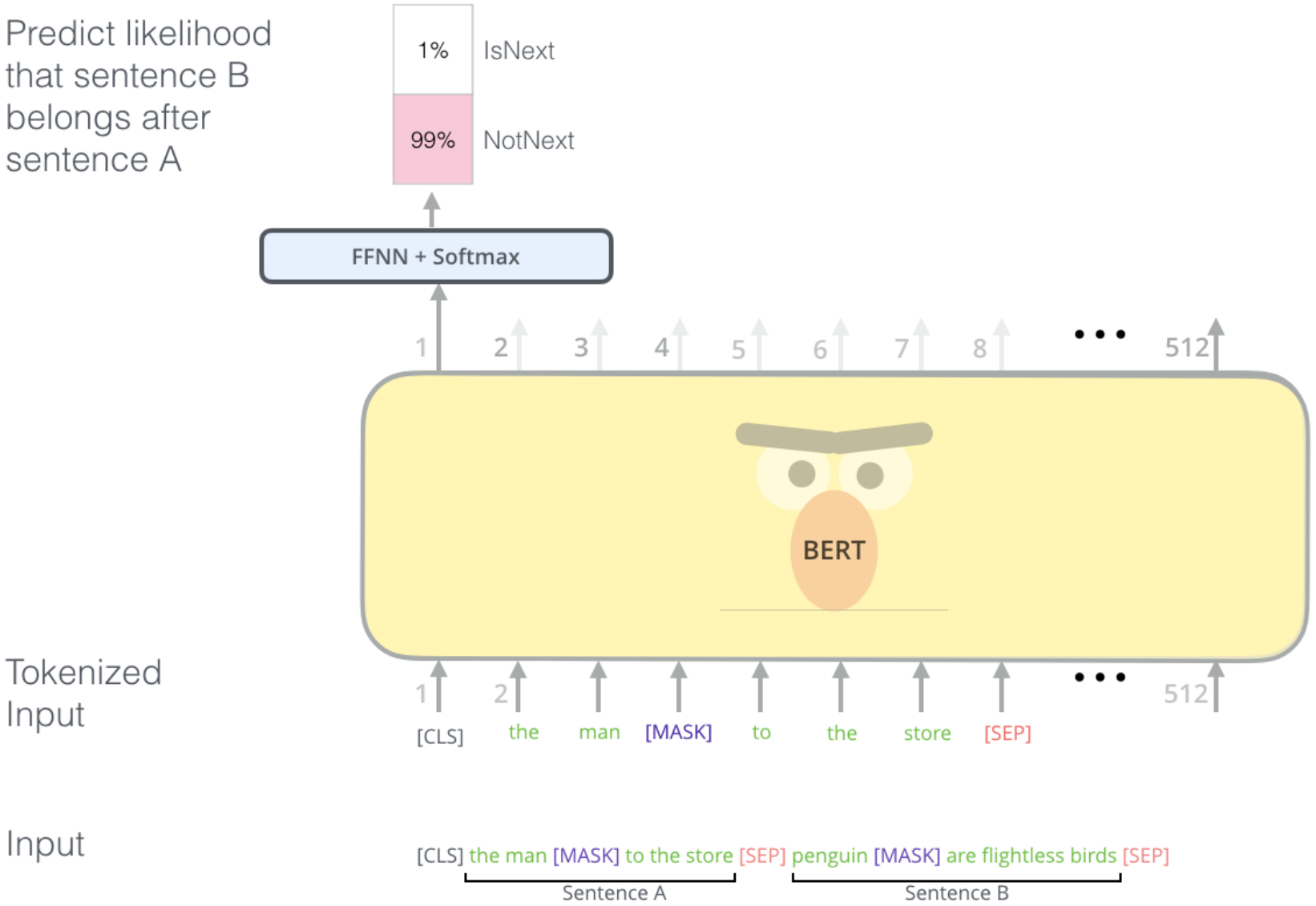


BERT's clever language modeling task masks 15% of words in the input and asks the model to predict the missing word.

Acknowledgement to Figure from http://jalammar.github.io/illustrated-bert/

[Devlin et al., 2018]

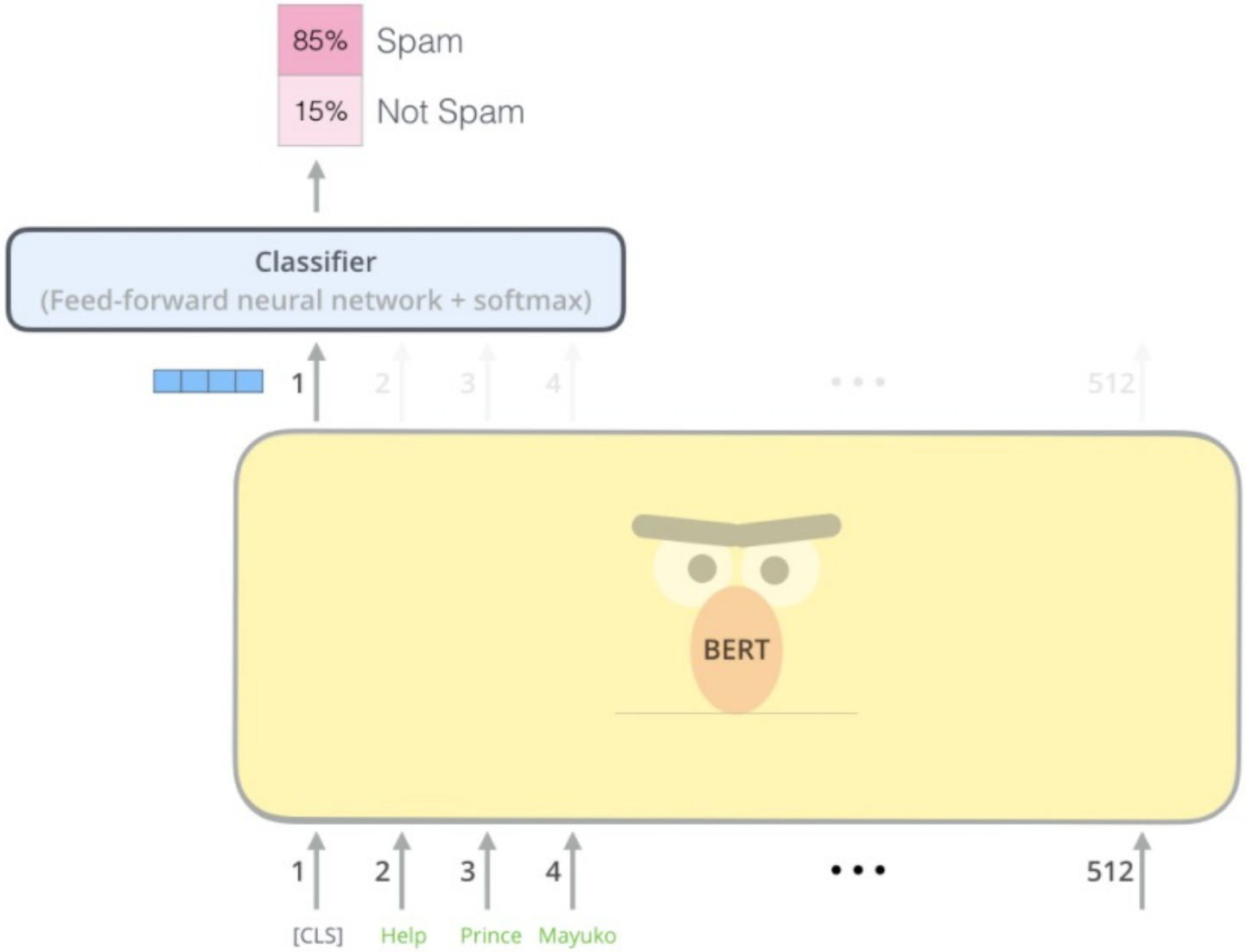
BERT – NEXT SENTENCE PREDICTION



Acknowledgement to Figure from http://jalammar.github.io/illustrated-bert/

[Devlin et al., 2018]

BERT – FINE TUNING



Acknowledgement to Figure from http://jalammar.github.io/illustrated-bert/

[Devlin et al., 2018]

BERT

- **BERT** when resealed advanced the SOTA (State of the Art) on 11 NLP tasks
- **BERT** NLP model is now part of Google Search

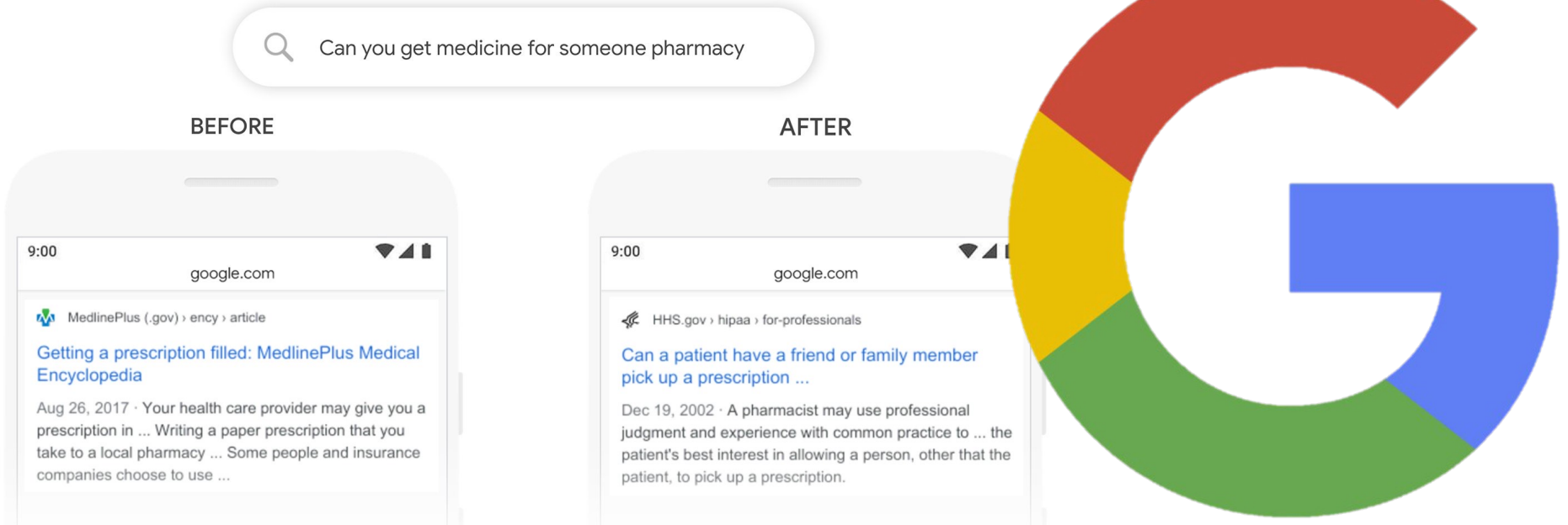
SEARCH

Understanding searches better than ever before

Pandu Nayak

Google Fellow and Vice President, Search

Published Oct 25, 2019



CONTEXTUAL WORD EMBEDDINGS

Cosine similarity between the contextualized embeddings of the target words in boldface, generated by the layer 12 of BERT

	The professor **opened** the conference.	The professor **opened** the door.
The professor **began** the conference	0.81	0.53
The professor **unlocked** the door.	0.57	0.77

	The horse **runs** fast	The water **runs** fast
The horse **gallops** fast.	0.73	0.54
The water **flows** fast.	0.70	0.85

BIBLIOGRAPHY

- Word2Vec http://d2l.ai/chapter_natural-language-processing-pretraining/word2vec.html
- Glove http://d2l.ai/chapter_natural-language-processing-pretraining/glove.html
- BERT http://d2l.ai/chapter_natural-language-processing-pretraining/bert.html

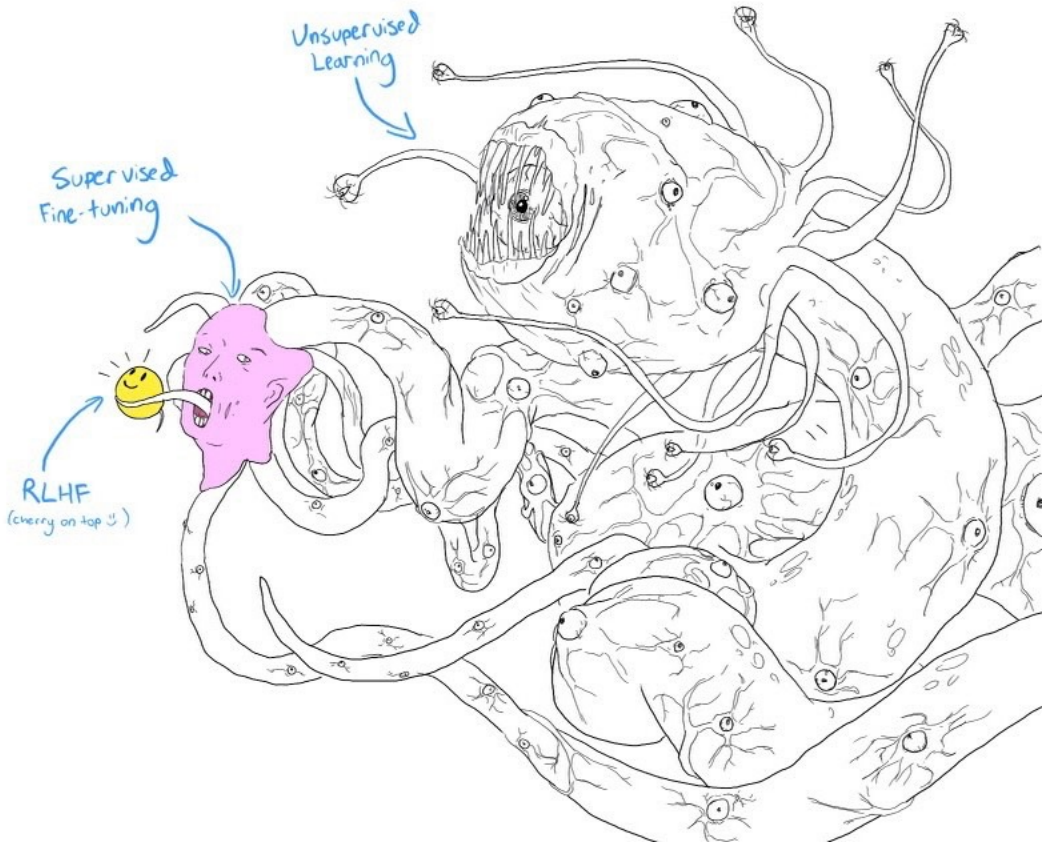
LANGUAGE MODELS

INGREDIENTS OF A MODERN LANGUAGE MODEL

1. **Self-supervised Learning** on a large collection of unlabeled data

2. **Supervised finetuning** (instruction tuning)

3. **Reinforcement Learning** from Human Feedback



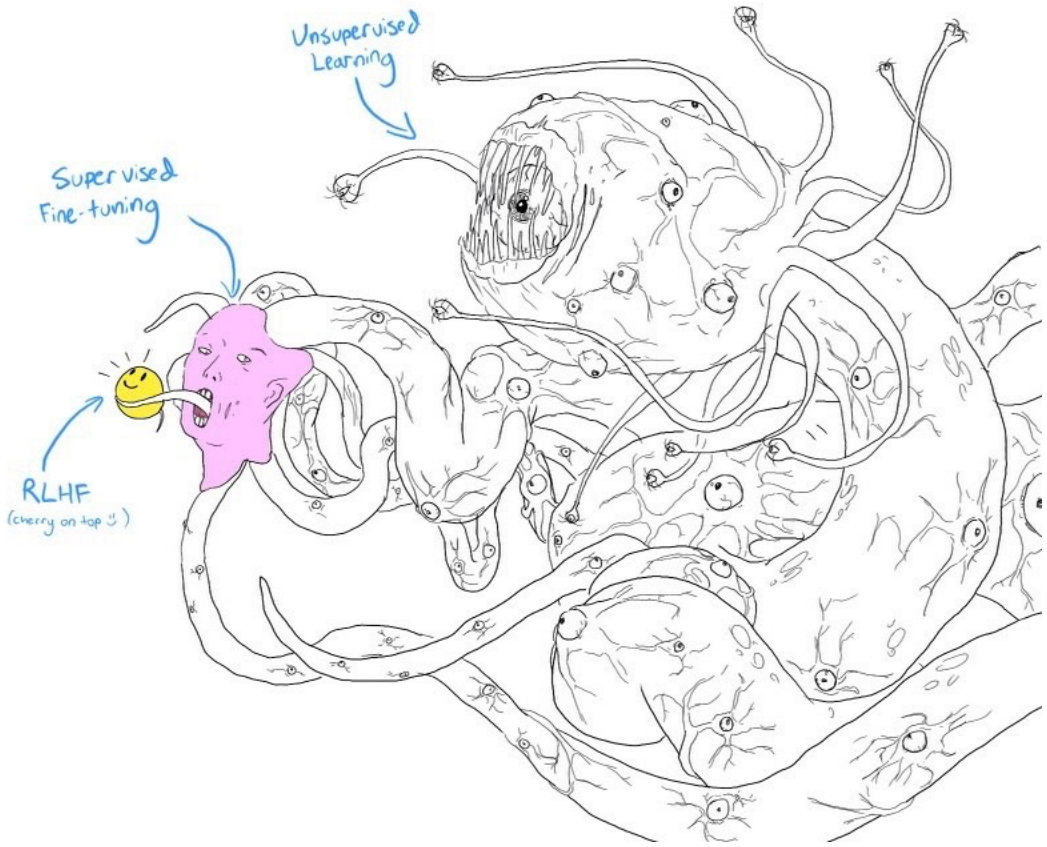
(img credits @anthrupad)

INGREDIENTS OF A MODERN LANGUAGE MODEL

1. **Self-supervised Learning** on a large collection of unlabeled data

2. Supervised finetuning (instruction tuning)

3. Reinforcement Learning from Human Feedback



(img credits @anthrupad)

TIMELINE – PART 1

Date	keywords	Institute	Paper	Publication
2017-06	Transformers	Google	Attention Is All You Need	NeurIPS
2018-06	GPT 1.0	OpenAI	Improving Language Understanding by Generative Pre-Training	
2018-10	BERT	Google	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	NAACL
2019-02	GPT 2.0	OpenAI	Language Models are Unsupervised Multitask Learners	

SELF-SUPERVISED LEARNING

Data $\{x\}_i$ *task_spec*

GPT-3 [Brown et al.]

Self-supervised learning is supervised learning without human-annotated labels

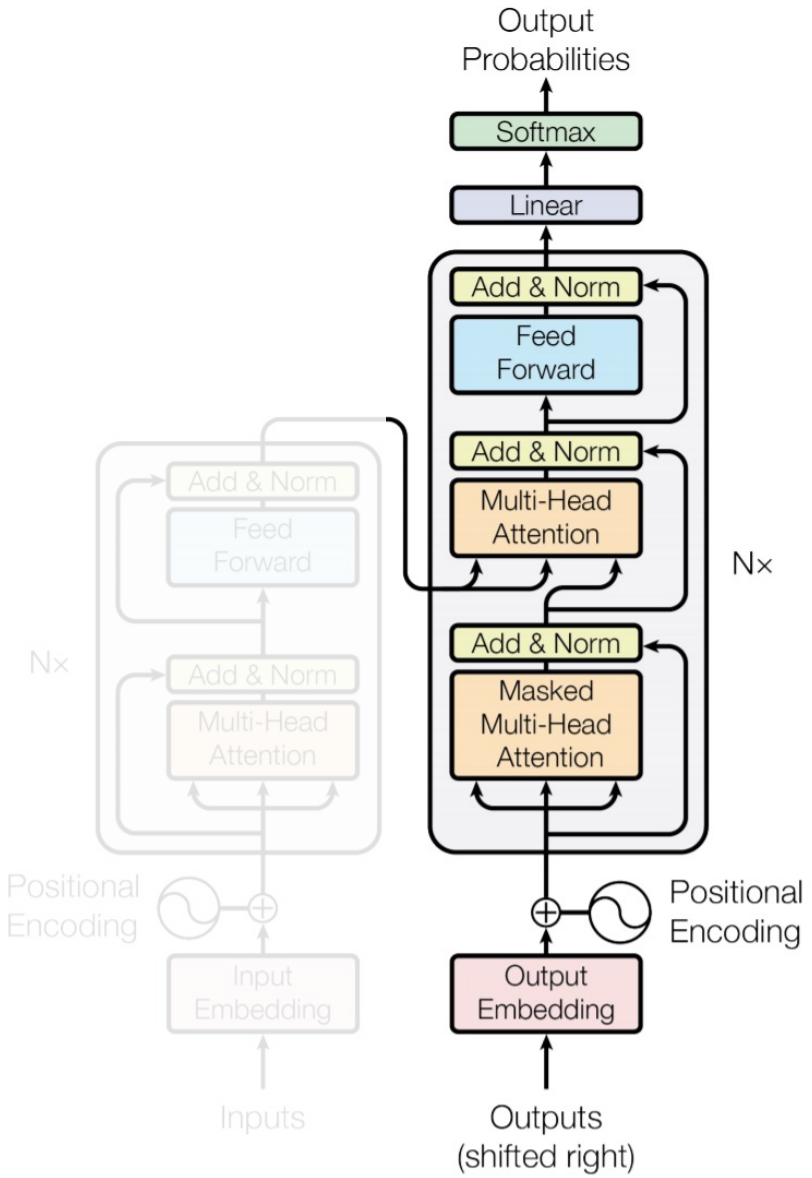
Important techniques:

- Autoencoders
- Temporally Supervised Learning
- Contrastive Losses

Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.
The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

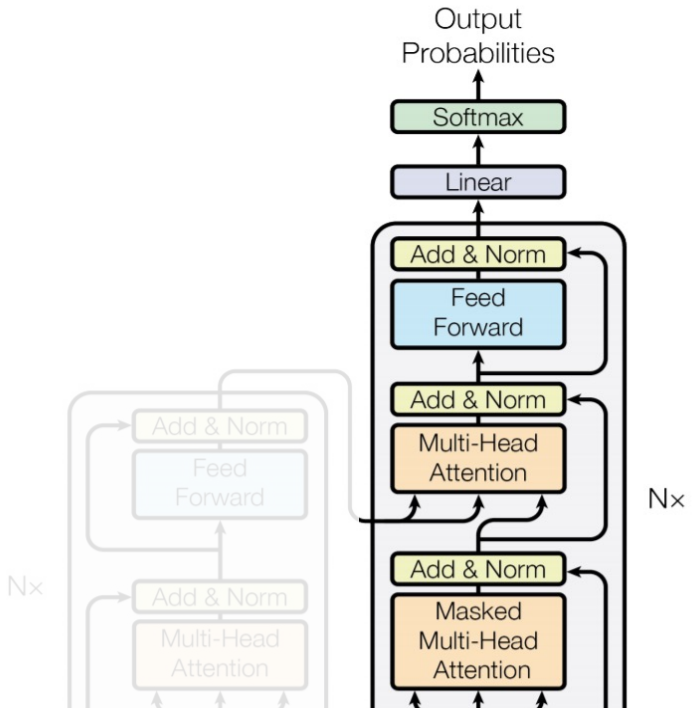
Figure 3.14: The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%).

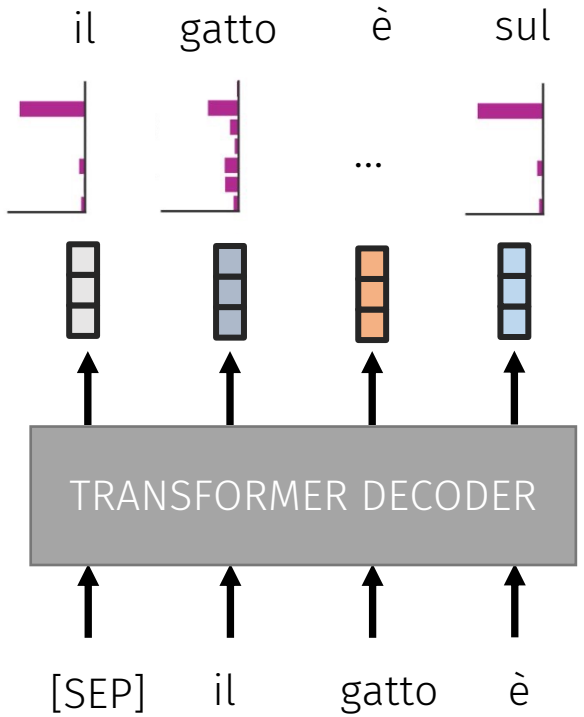
Generative pre-trained transformer



Generative pre-trained transformer

Dim
Vocabulary





GPT-2

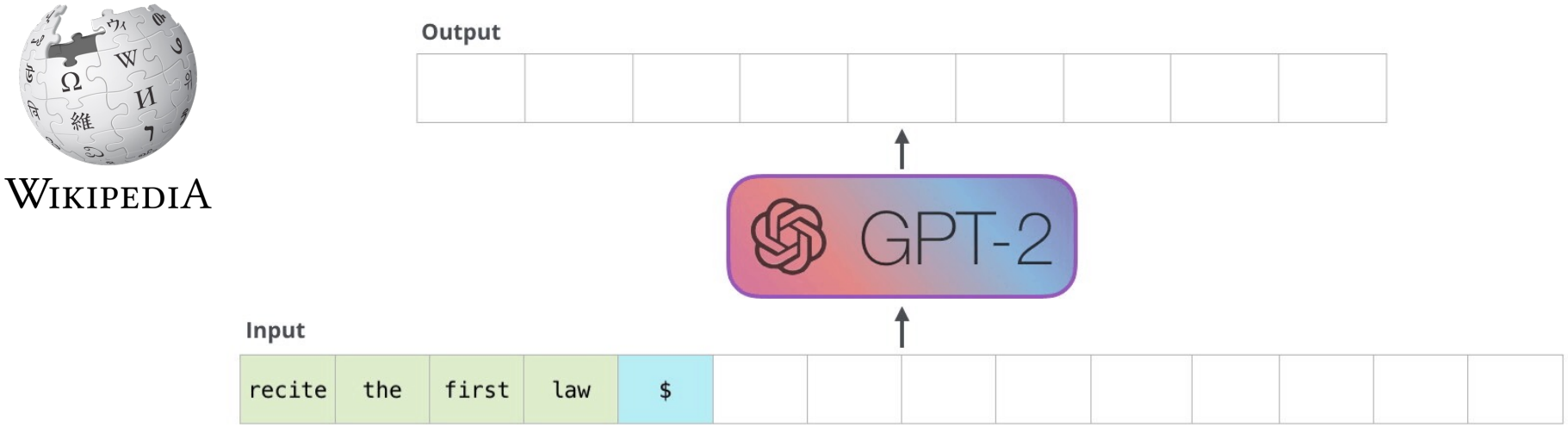


Figure from https://jalammar.github.io/illustrated-gpt2/

GPT-2

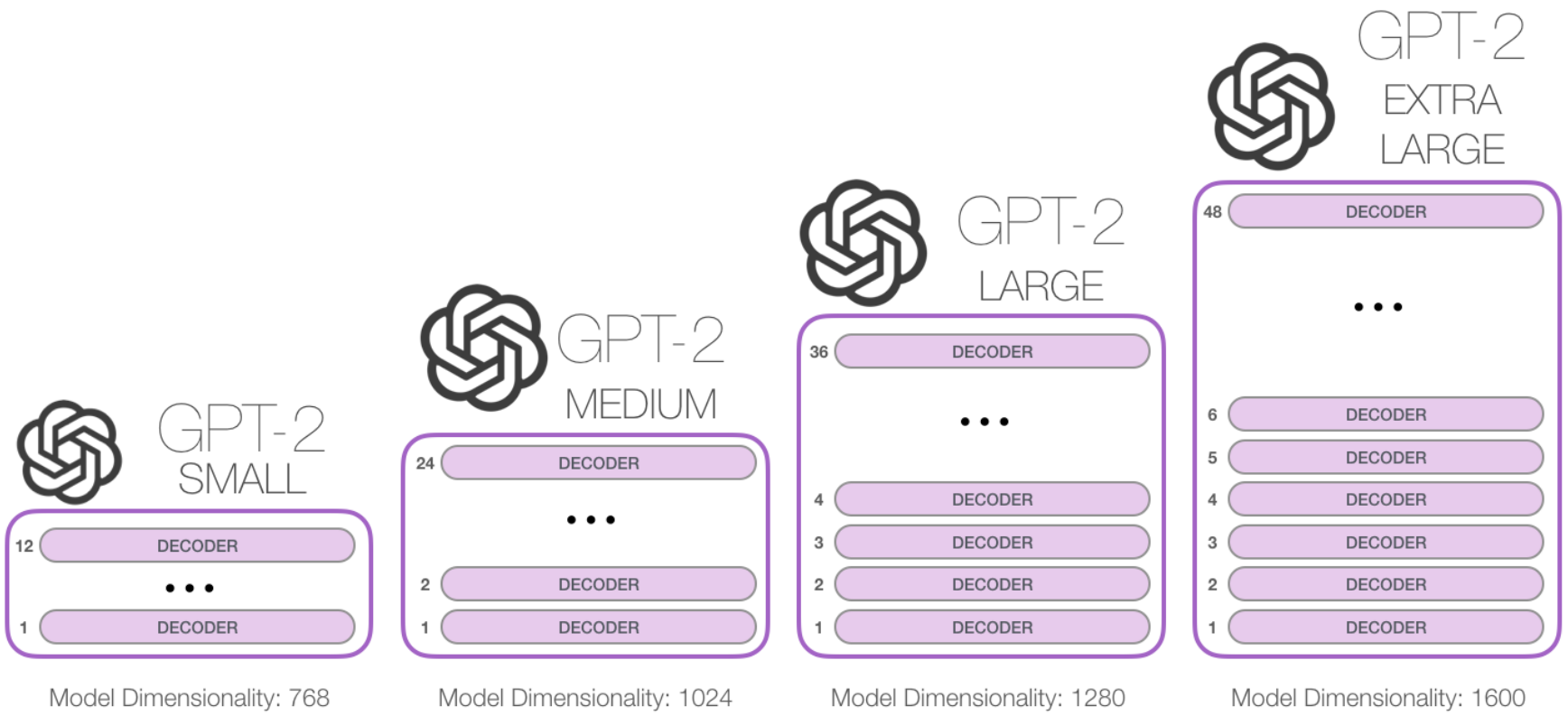


Figure from https://jalammar.github.io/illustrated-gpt2/

TIMELINE – PART 2 (SCALING THE SIZE)

2020-05	GPT 3.0	OpenAI	Language models are few-shot learners	NeurIPS

Language Models are Few-Shot Learners

Tom B. Brown* **Benjamin Mann*** **Nick Ryder*** **Melanie Subbiah***

Jared Kaplan† **Prafulla Dhariwal** **Arvind Neelakantan** **Pranav Shyam**

Girish Sastry **Amanda Askell** **Sandhini Agarwal** **Ariel Herbert-Voss**

Gretchen Krueger **Tom Henighan** **Rewon Child** **Aditya Ramesh**

Daniel M. Ziegler **Jeffrey Wu** **Clemens Winter**

Christopher Hesse **Mark Chen** **Eric Sigler** **Mateusz Litwin** **Scott Gray**

Benjamin Chess **Jack Clark** **Christopher Berner**

Sam McCandlish **Alec Radford** **Ilya Sutskever** **Dario Amodei**

Abstract

We demonstrate that scaling up language models greatly improves task-agnostic,

GPT-3

The three settings we explore for in-context learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée
4  plush girafe => girafe peluche
5  cheese =>                           ← prompt
```

Traditional fine-tuning (not used for GPT-3)

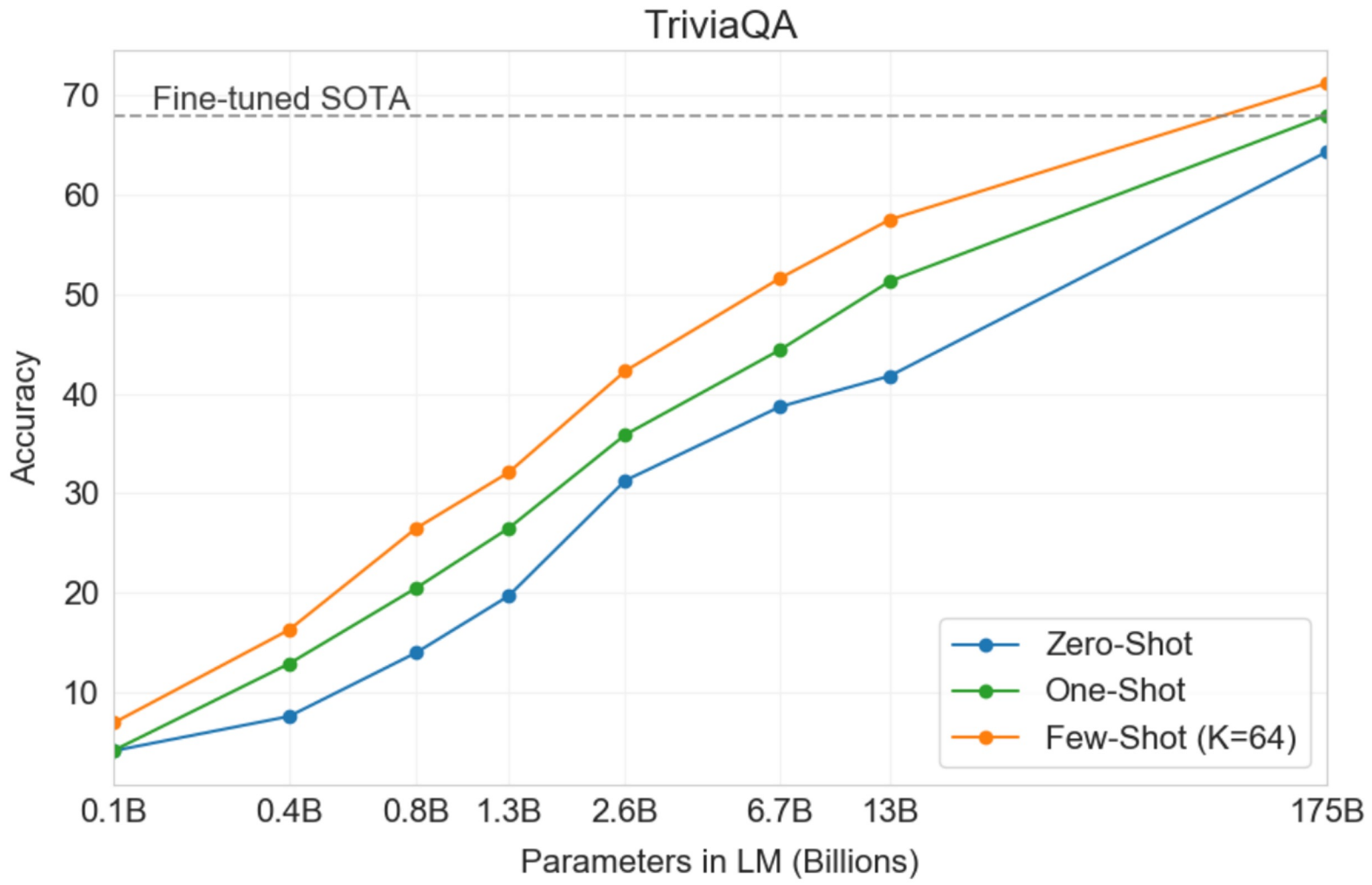
Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

(Brown et al., 2020)

GPT-3

(Brown et al., 2020)

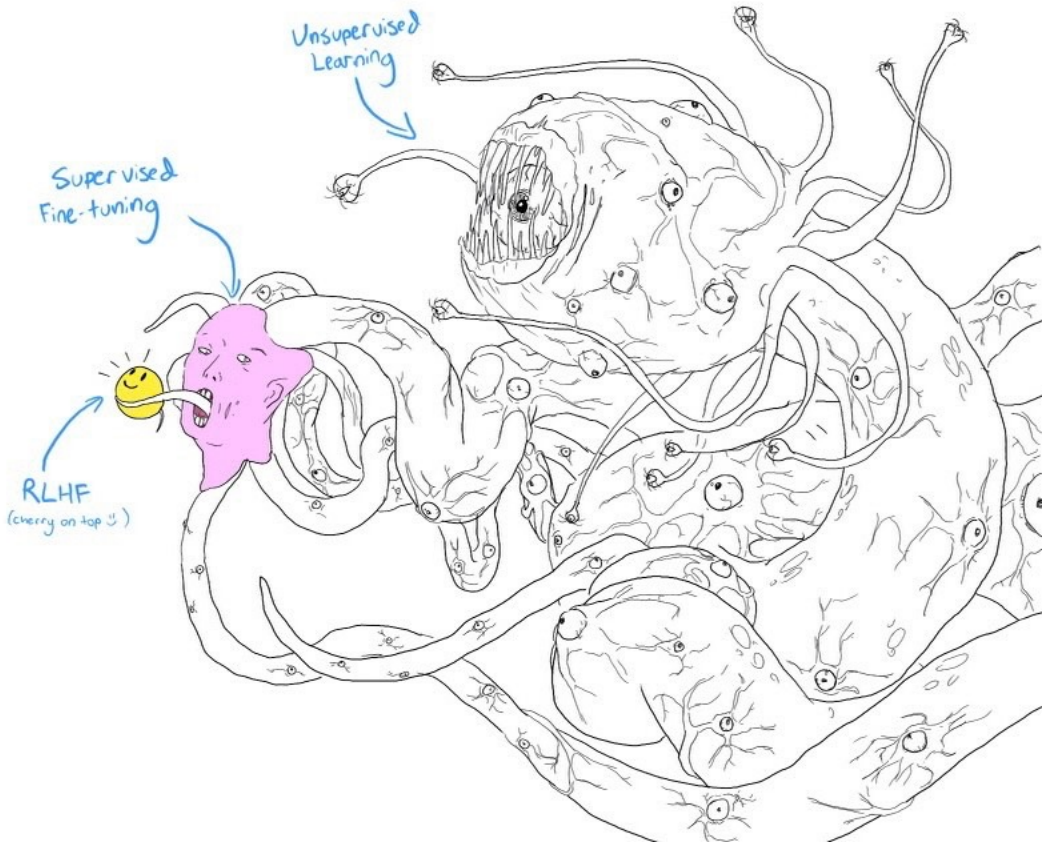


INGREDIENTS OF A MODERN LANGUAGE MODEL

1. Self-supervised Learning on a large collection of unlabeled data

2. Supervised finetuning (instruction tuning)

3. Reinforcement Learning from Human Feedback



(img credits @anthrupad)

TIMELINE – PART 3 (INSTRUCTION TUNING)

2021-09	FLAN	Google	Finetuned Language Models are Zero-Shot Learners	ICLR
2021-10	T0	HuggingFace et al.	Multitask Prompted Training Enables Zero-Shot Task Generalization	ICLR

Published as a conference paper at ICLR 2022

MULTITASK PROMPTED TRAINING ENABLES ZERO-SHOT TASK GENERALIZATION

Victor Sanh* Hugging Face
Albert Webson* Brown University
Colin Raffel* Hugging Face
Stephen H. Bach* Brown & Snorkel AI

Lintang Sutawika BigScience
Zaid Alyafeai KFUPM
Antoine Chaffin IRISA & IMATAG
Arnaud Stiegler Hyperscience
Teven Le Scao Hugging Face

Arun Raja I^2R, Singapore
Manan Dey SAP
M Saiful Bari NTU, Singapore
Canwen Xu UCSD & Hugging Face
Urmish Thakker SambaNova Systems

Shanya Sharma Walmart Labs
Eliza Szczechla BigScience
Taewoon Kim VU Amsterdam
Gunjan Chhablani BigScience
Nihal V. Nayak Brown University

Debajyoti Datta University of Virginia
Jonathan Chang ASUS
Mike Tian-Jian Jiang ZEALS, Japan
Han Wang NYU
Matteo Manica IBM Research

Sheng Shen UC Berkeley
Zheng-Xin Yong Brown University
Harshit Pandey BigScience
Michael McKenna Parity
Rachel Bawden Inria, France

Thomas Wang Inria, France
Trishala Neeraj BigScience
Jos Rozen Naver Labs Europe
Abheesht Sharma BITS Pilani, India
Andrea Santilli University of Rome

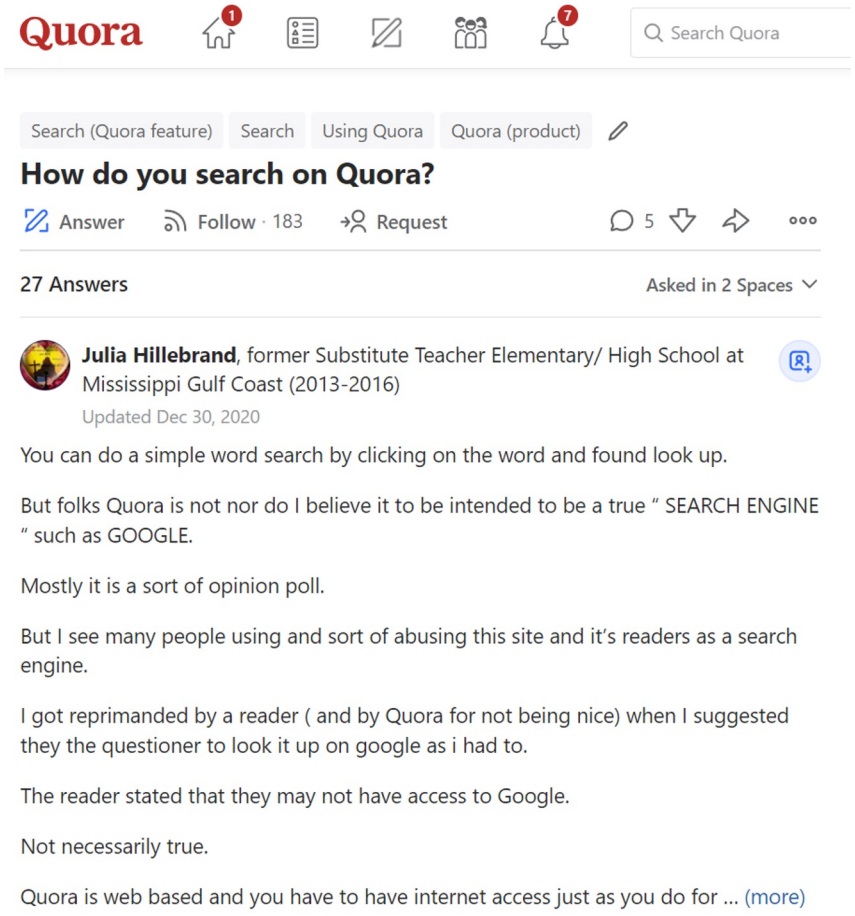
Thibault Fevry BigScience
Jason Alan Fries Stanford & Snorkel AI
Ryan Teehan Charles River Analytics
Tali Bers Brown University

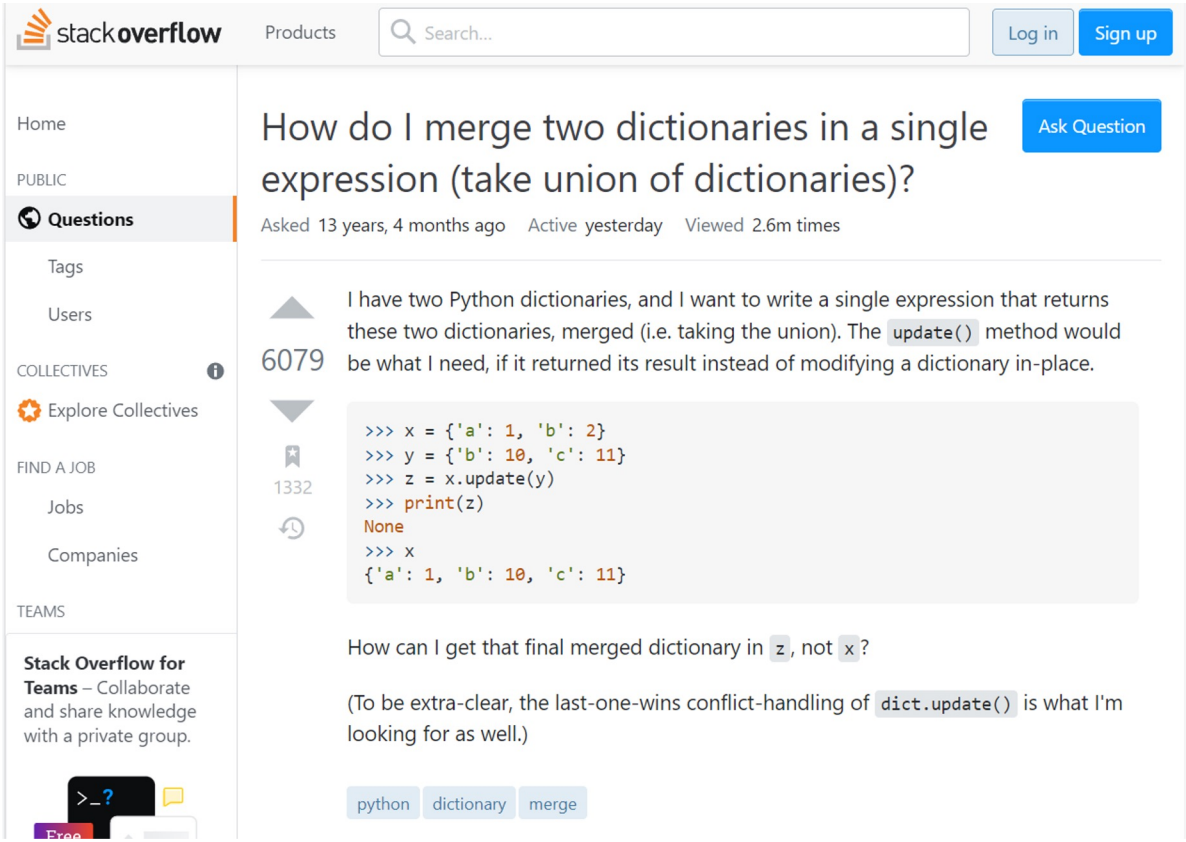
Stella Biderman Booz Allen & EleutherAI
Leo Gao EleutherAI
Thomas Wolf Hugging Face
Alexander M. Rush Hugging Face

ABSTRACT

Background:

How implicit is zero-shot learning?





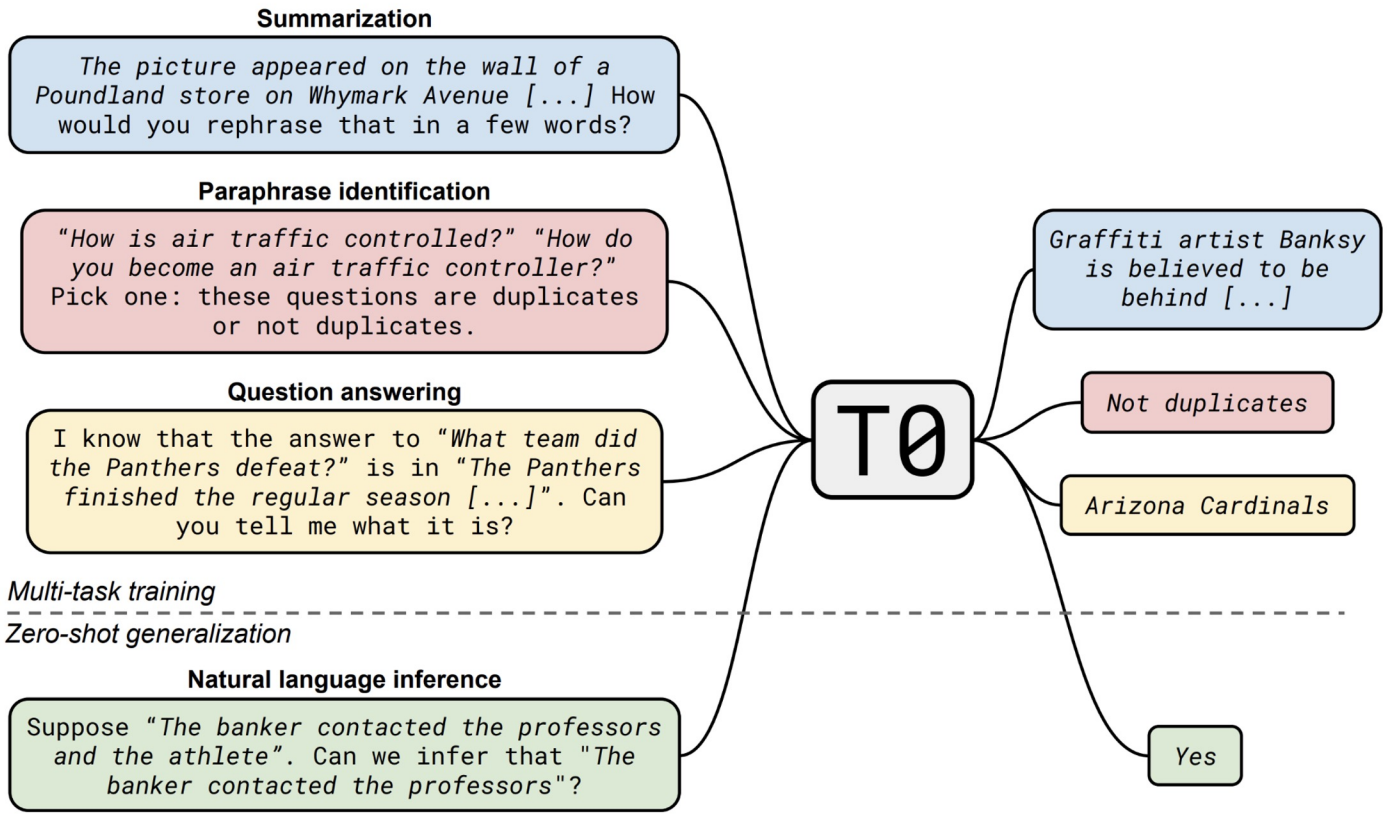
Method

Can zero-shot generalization instead be directly induced by explicit multitask learning?

Text-to-Text
Transfer
Transformer

Method

We fine-tuned T5 on a multitask mixture covering a wide variety of tasks

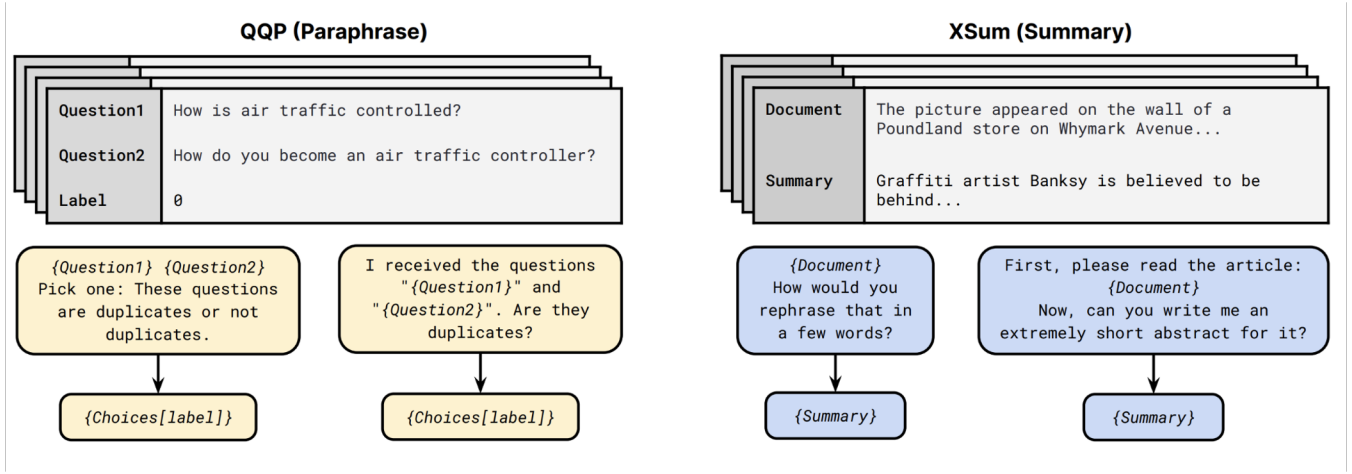


Method

1. System to convert natural language tasks into human-readable prompts

1. Convert a large set of supervised datasets to prompts

1. We fine-tuned T5 model on this multitask mixture



Method

Paraphrase

Examples

Question1	How is air traffic controlled?
Question2	How do you become an air traffic controller?
Label	0

Choices	[Not duplicates, Duplicates]	[No, Yes]
Input	{Question1} {Question2} Pick one: These questions are duplicates or not duplicates.	I received the questions "{Question1}" and "{Question2}". Are they duplicates?
Target	{Choices[label]}	{Choices[label]}

Method

Summary

Examples

Document	The picture appeared on the wall of a Poundland store on Whymark Avenue...
Summary	Graffiti artist Banksy is believed to be behind...

Choices

Input	{Document} How would you rephrase that in a few words?
Target	{Summary}

Input	First, please read the article below. {Document} Now, can you write me an extremely short abstract for it?
Target	{Summary}

Method

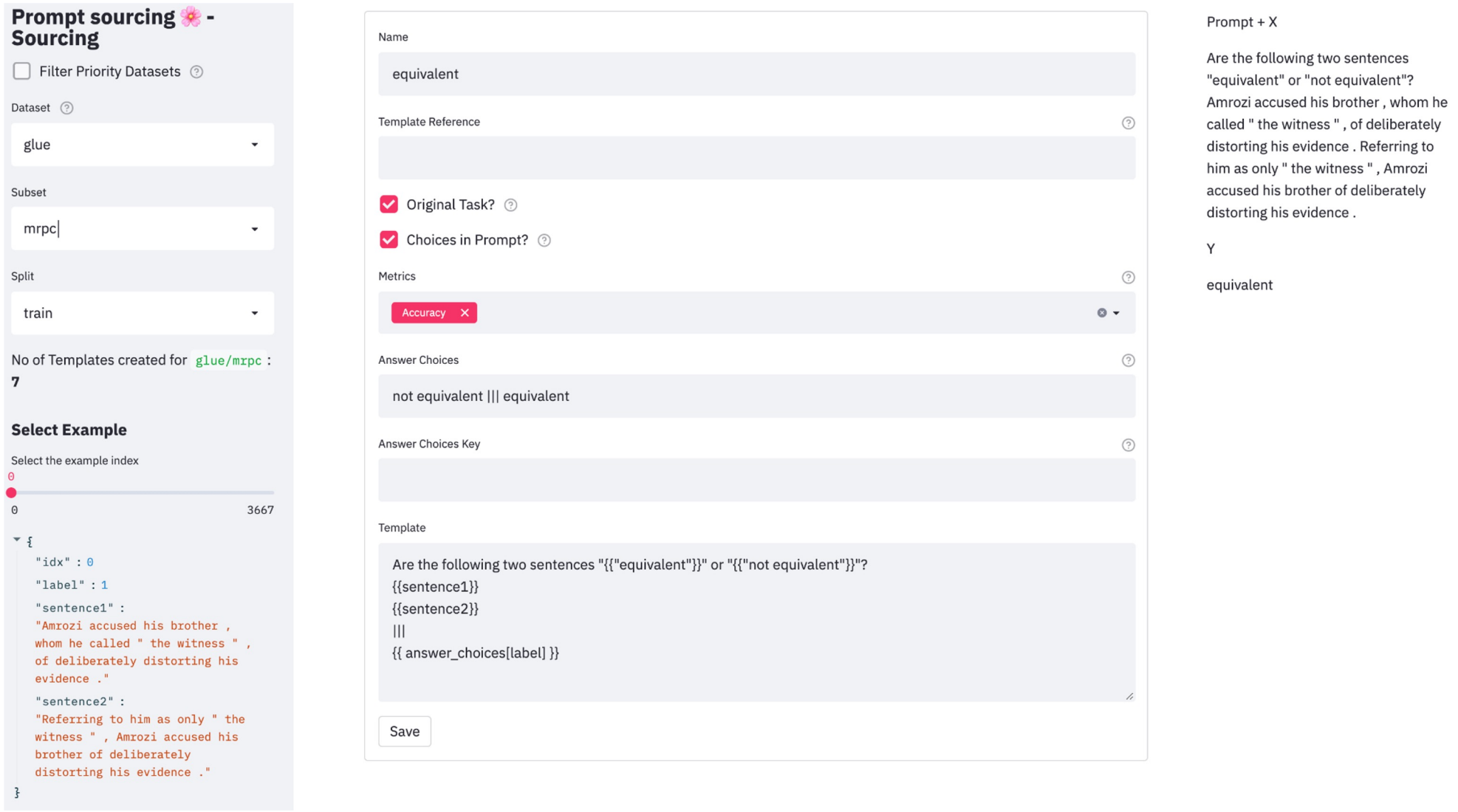
Extractive Question-Answering

Examples

Question	What team did the Panthers defeat?
Context	The Panthers finished the regular season with a 15-1 record, and quarterback Cam...
Answer	Arizona Cardinals

Choices		
Input	I know that the answer to "{Question}" is in "{Context}". Can you tell me what it is?	Given the following passage: {Context} Answer the following question. Question: {Question}
Target	{Answer}	{Answer}

Method



Prompt + X

Are the following two sentences "equivalent" or "not equivalent"? Amrozi accused his brother , whom he called " the witness " , of deliberately distorting his evidence . Referring to him as only " the witness " , Amrozi accused his brother of deliberately distorting his evidence .

Y

equivalent

Method

P3 (Public Pool of Prompts)

Multiple-Choice QA
- CommonsenseQA
- DREAM
- QuAIL
- QuaRTz
- Social IQA
- WiQA
- Cosmos QA
- QASC
- QuaRel
- SciQ
- Wiki Hop

Extractive QA
- Adversarial QA
- Quoref
- ROPES
- DuoRC

Closed-Book QA
- Hotpot QA
- Wiki QA

Sentiment
- Amazon
- App Reviews
- IMDB
- Rotten Tomatoes
- Yelp

Topic Classification
- AG News
- DBPedia
- TREC

Structure-To-Text
- Common Gen
- Wiki Bio

Summarization
- CNN Daily Mail
- Gigaword
- MultiNews
- SamSum
- XSum

Paraphrase Identification
- MRPC
- PAWS
- QQP

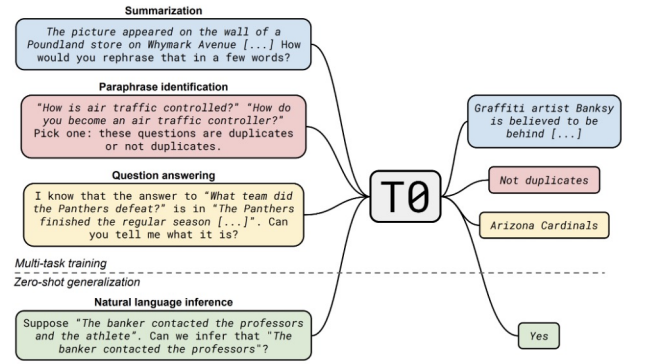
Sentence Completion
- COPA
- HellaSwag
- Story Cloze

Natural Language Inference
- ANLI
- CB
- RTE

Coreference Resolution
- WSC
- Winogrande

Word Sense Disambiguation
- WiC

BIG-Bench
- Code Description
- Conceptual
- Hindu Knowledge
- Known Unknowns
- Language ID
- Logic Grid
- Logical Deduction
- Misconceptions
- Movie Dialog
- Novel Concepts
- Strategy QA
- Syllogisms
- Vitamin C
- Winowhy



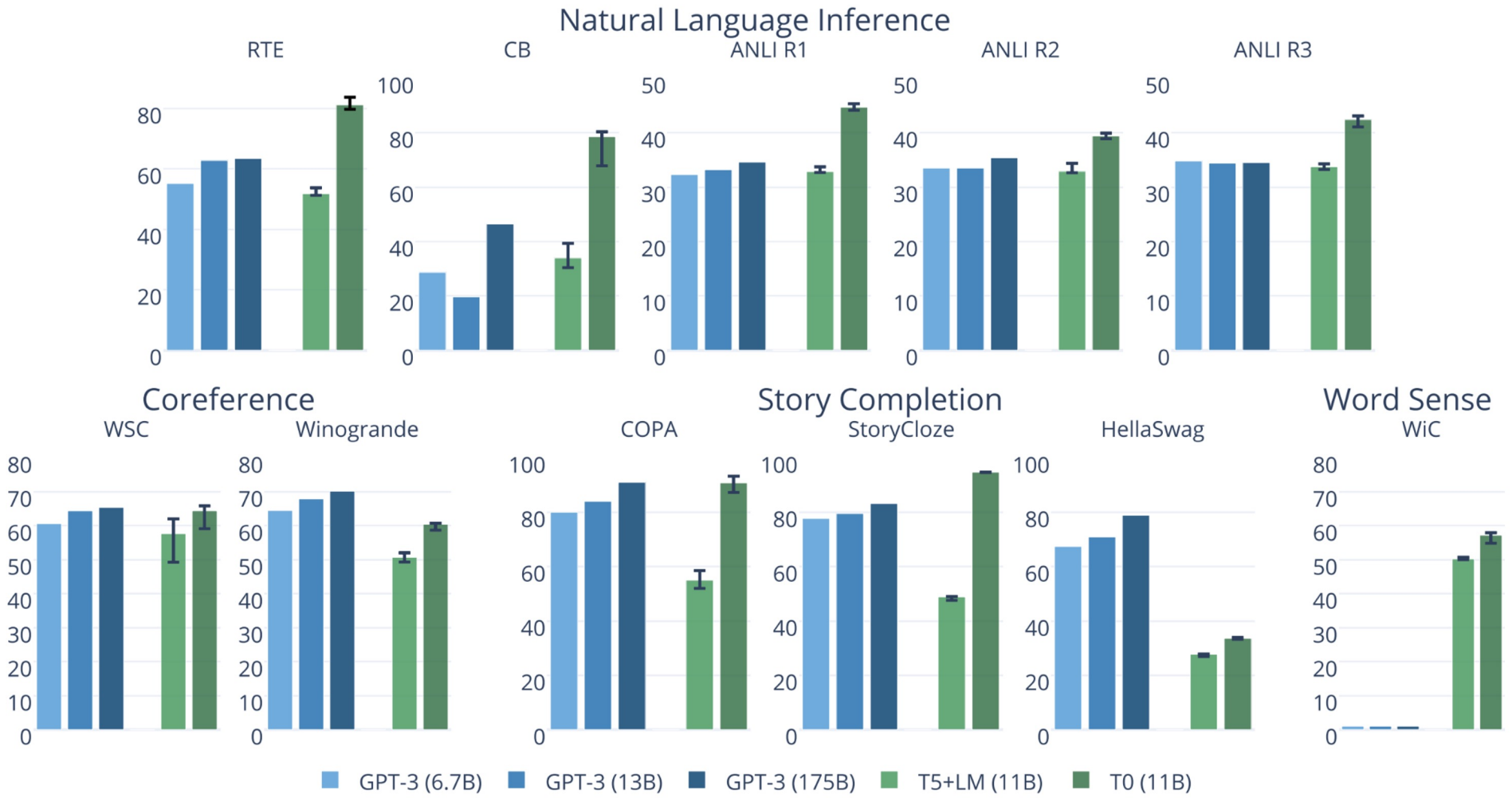
RQ1 - Generalization to unseen tasks

Can zero-shot generalization be directly induced by explicit multitask learning?

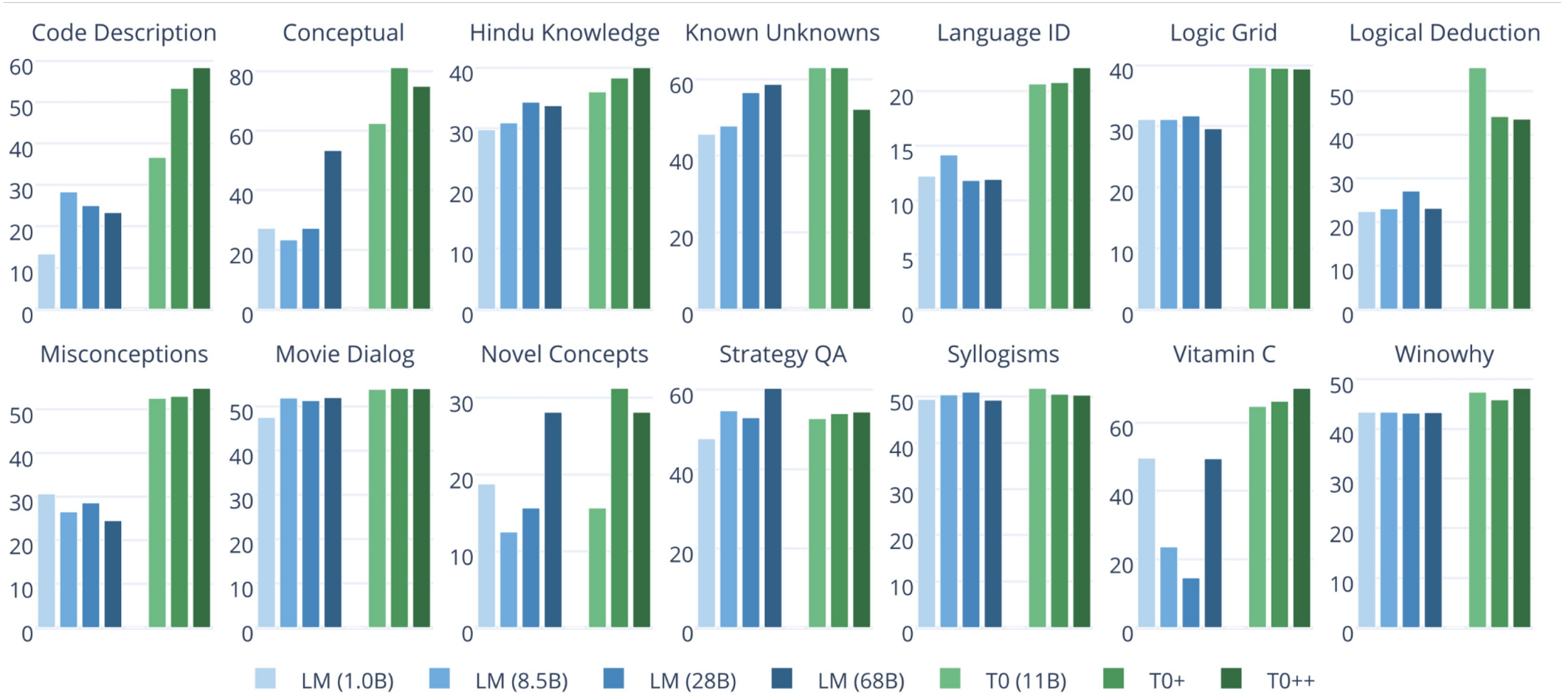
Models

- **T5+LM**: (T0 w/o finetuning)
- **T0**: P3
- **T0+** : P3 + datasets from GPT-3's evaluation suite
- **T0++** : P3 + datasets from GPT-3's evaluation suite + SuperGLUE (excluding NLI sets)

Experiments - RQ1



Experiments - RQ1



Conclusion

RQ1: Can zero-shot generalization be directly induced by explicit multitask learning?

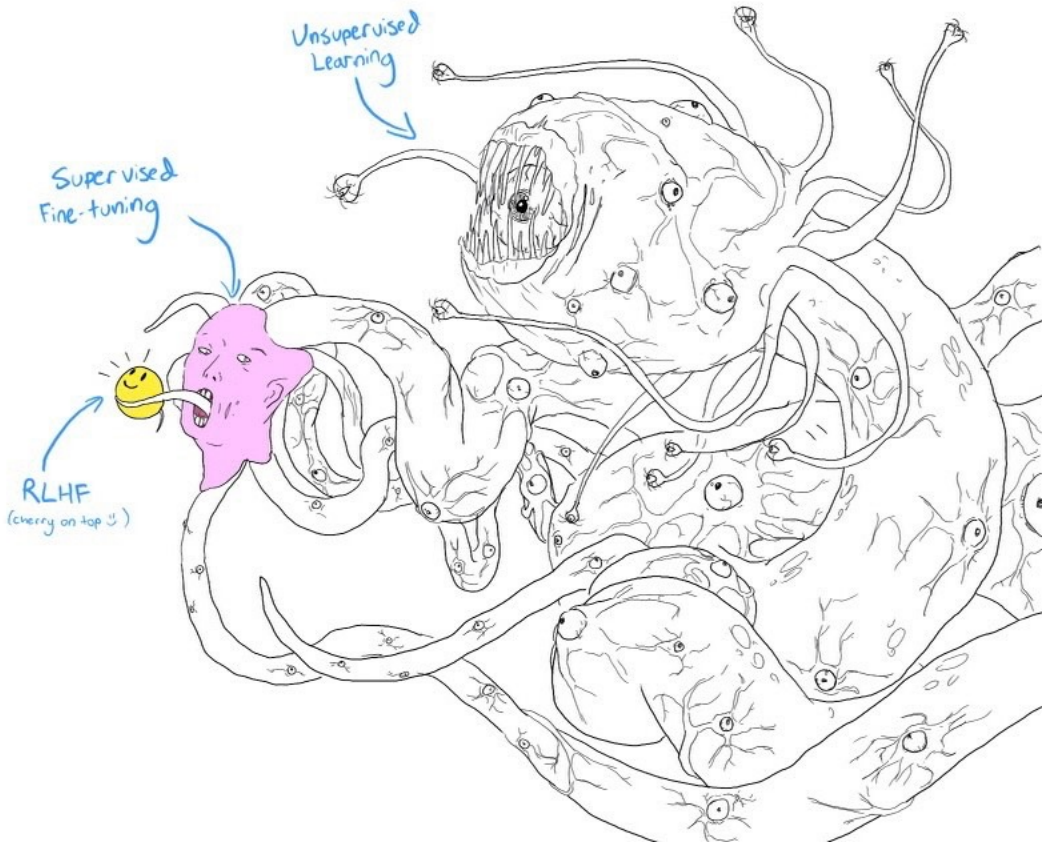
- ➢ T0 outperforms GPT-3 on 9 out of 11 tasks despite being 16x smaller
- ➢ T0 attains strong performance on 13 out of 14 Big-bench tasks

INGREDIENTS OF A MODERN LANGUAGE MODEL

1. Self-supervised Learning on a large collection of unlabeled data

2. Supervised finetuning (instruction tuning)

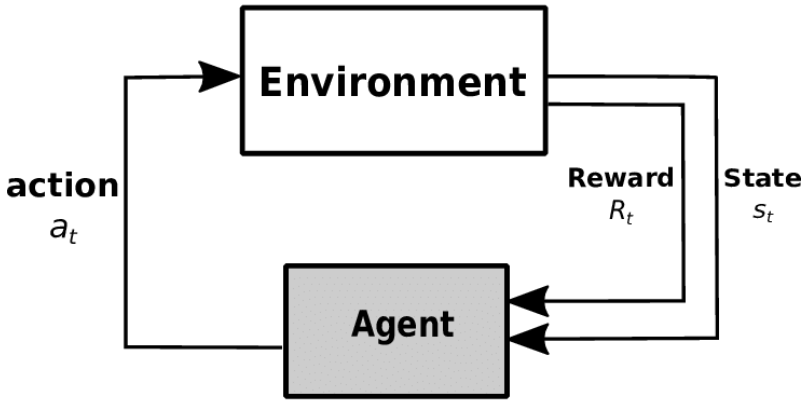
3. **Reinforcement Learning** from Human Feedback



(img credits @anthrupad)

REINFORCEMENT LEARNING

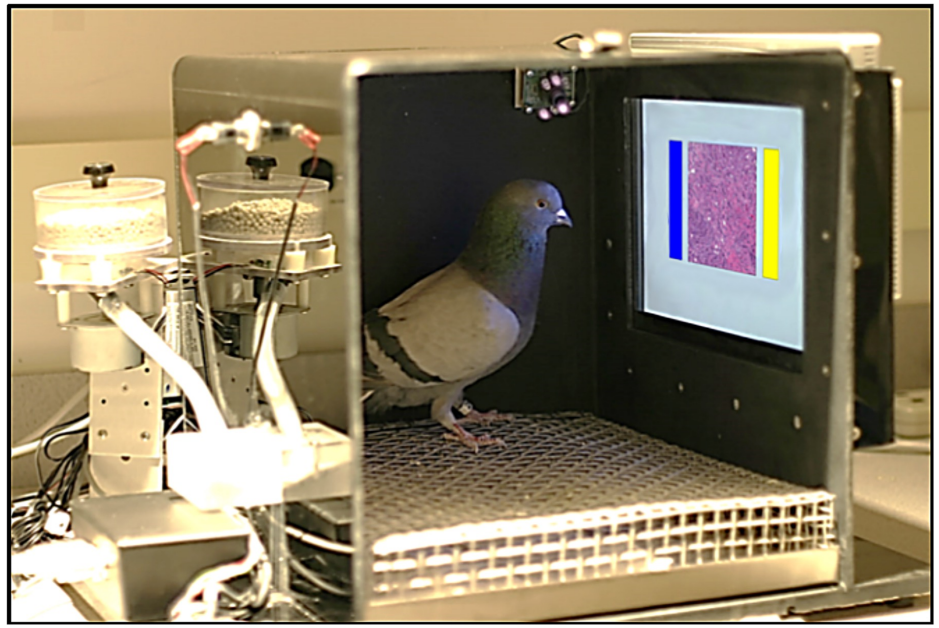
Data

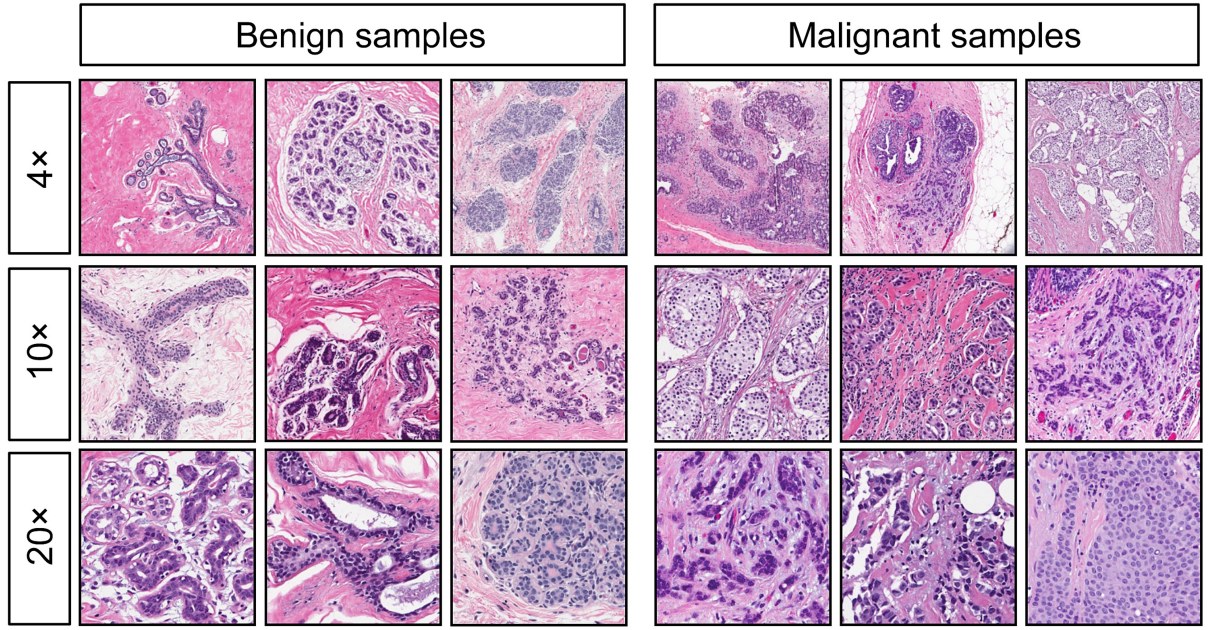


An agent receives information about its environment and learns to choose actions that will maximize some reward

REINFORCEMENT LEARNING

Pigeons (Columba livia) as Trainable Observers of Pathology and Radiology Breast Cancer Images – Levenson et al.





"The birds proved to have a remarkable ability to distinguish benign from malignant human breast histopathology after training with ***differential food reinforcement*** [...]"

Y. LeCun

How Much Information is the Machine Given during Learning?

- **"Pure" Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- **Self-Supervised Learning (cake génoise)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

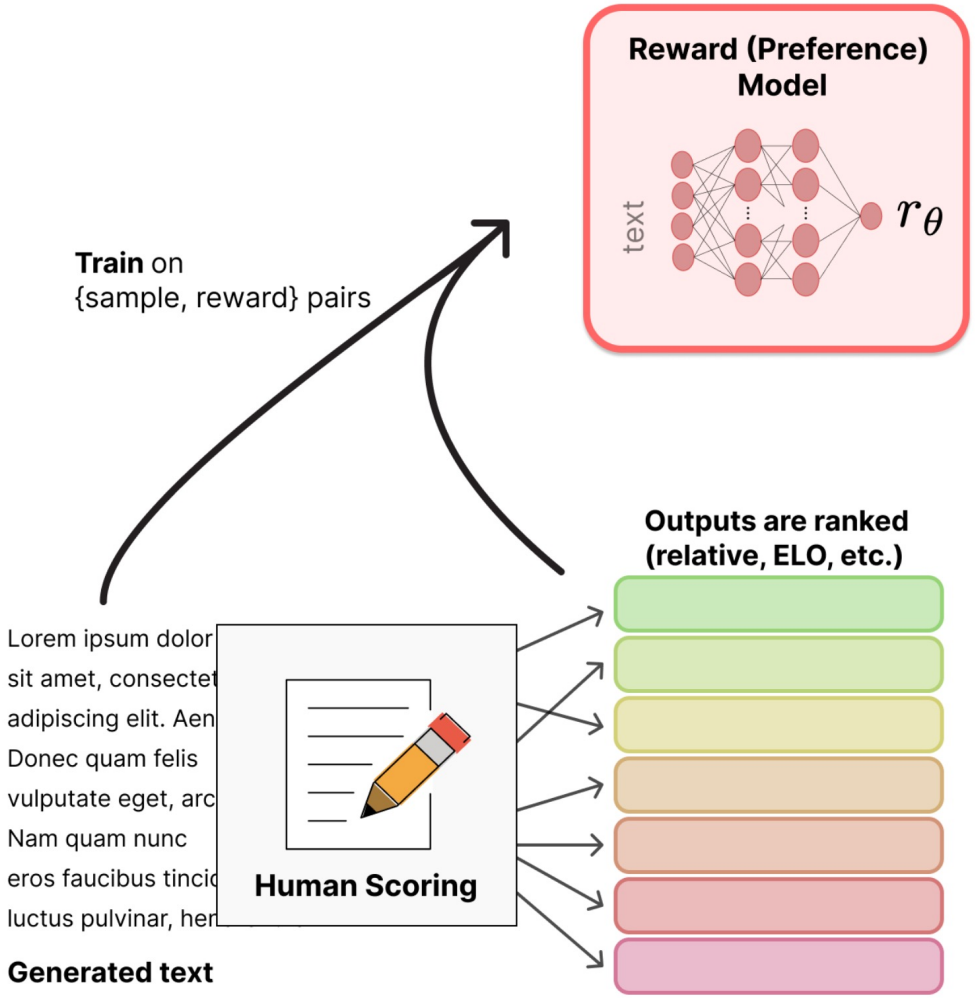
Reinforcement Learning from Human Feedback

2022-03	InstructGPT	OpenAI	Training language models to follow instructions with human feedback	

Allign a Language Model towards human preferences

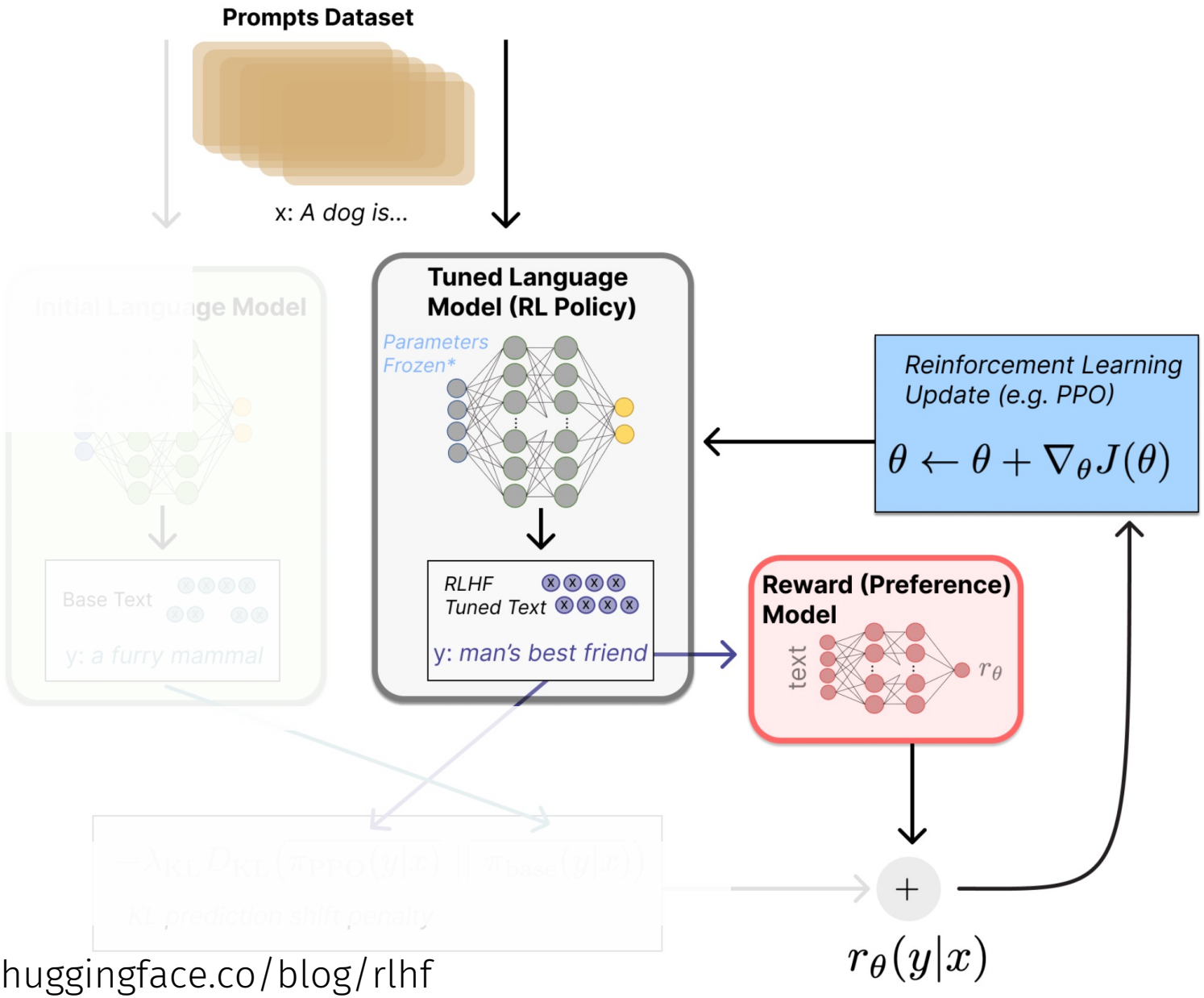
1. Pretraining a language model (LM)
2. Train a reward model
3. fine-tuning the LM with reinforcement learning.

REWARD MODEL



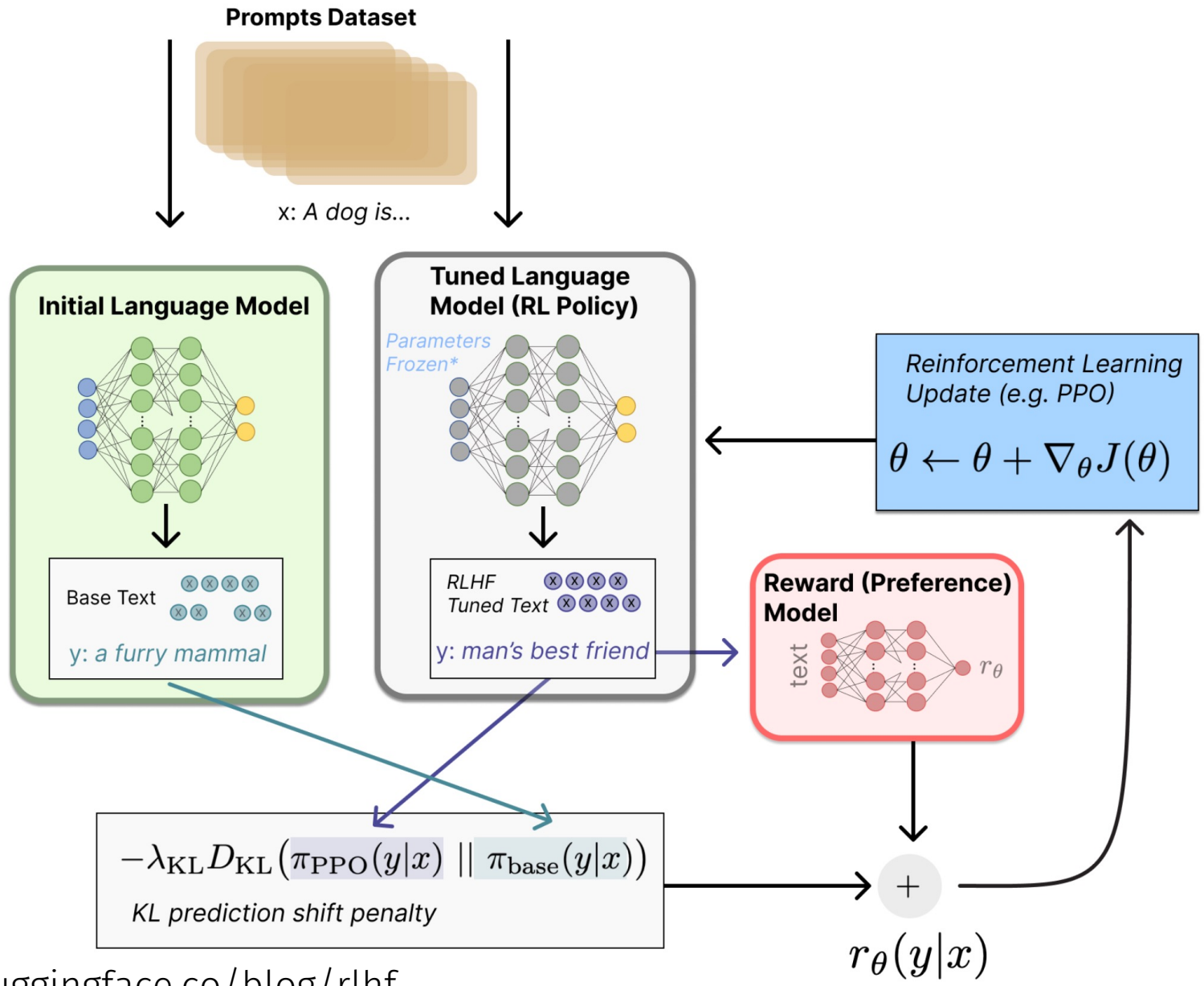
Picture from: https://huggingface.co/blog/rlhf

Reinforcement Learning from Human Feedback



Picture from: https://huggingface.co/blog/rlhf

Reinforcement Learning from Human Feedback



Picture from: https://huggingface.co/blog/rlhf

The diversity argument

- Maximum likelihood (supervised learning) might “punish” the model for even slight deviations from a text
- RL avoids this

Why RLHF?

The theoretical argument

- Supervised learning allows only positive feedback (questions + correct answers)
- RL allows also for negative feedback (the model is allowed to generate an answer and get a feedback saying "this is not correct")

Learners are allowed to form hypotheses and ask feedback from the teacher (reward model)

LINKS & CREDITS

- Language Models Timeline (with papers)
  https://github.com/Hannibal046/Awesome-LLM
- OpenAI Spinning Up (RL)
  https://spinningup.openai.com/en/latest/index.html
- RLHF https://huggingface.co/blog/rlhf
- Why RLHF?
  https://gist.github.com/yoavg/6bff0fecd65950898eba1bb321cfbd81