INTRODUZIONE ALL’AI E AL MACHINE LEARNING PER SPECIALISTI DELL’INGEGNERIA – QUARTO INCONTRO

AI GENERATIVA E LARGE SCALE LANGUAGE MODELS

AGENDA

CONSIGLIO NAZIONALE
DEGLI INGEGNERI

CENTENARIO
DELL'ALBO
DEGLI INGEGNERI

Con la collaborazione incondizionata della
Associazione Italiana di Intelligenza Artificiale

Associazione
Italiana per
l'Intelligenza
Artificiale

- **CONVEGNO ON LINE: Mercoledì 7 Maggio, ore 15.00 – 18.00**
- AI Generativa e Large Scale Language Models
- OVERVIEW
  - Foundation Models for Natural Language Processing.
  - Internals of Encoder-Decoder architectures.
  - Chat GPT.
  - Prompt Engineering e Few Shot Learning.
  - Tendenze recenti.
- USE CASES:
  - Process management in banking,
  - Information Extraction per la medicina,
  - Modelli di forecasting.

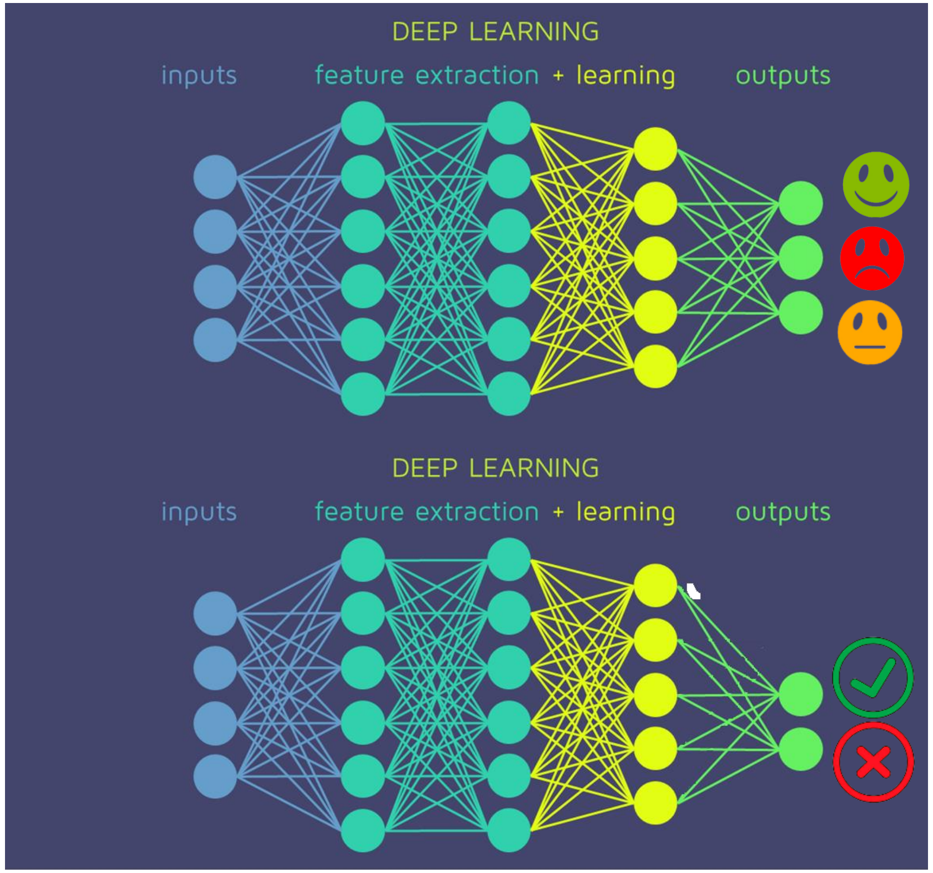
OVERVIEW

- Le Reti Neurali: dai percettroni ai Transfomers
  - Il ruolo dei Foundation Models in NLP
  - Internals of Encoder-decoder architectures
- Modelli Generativi e Large Language Models: la famiglia GPT, e chatGPT
  - Chat GPT: principi di funzionamento
- Few-shot Learning
  - 0-shot learning models
  - Prompt Engineering
- Use cases: process management nel sistema bancario, information extraction per la medicina, modelli previsionali (forecasting)

RETI NEURALI (RECAP)

PERCETTRONI E *MULTILAYER PERCEPTRONS, CONVOLUTIONAL NEURAL NETWORKS E RECURRENT NETWORKS*

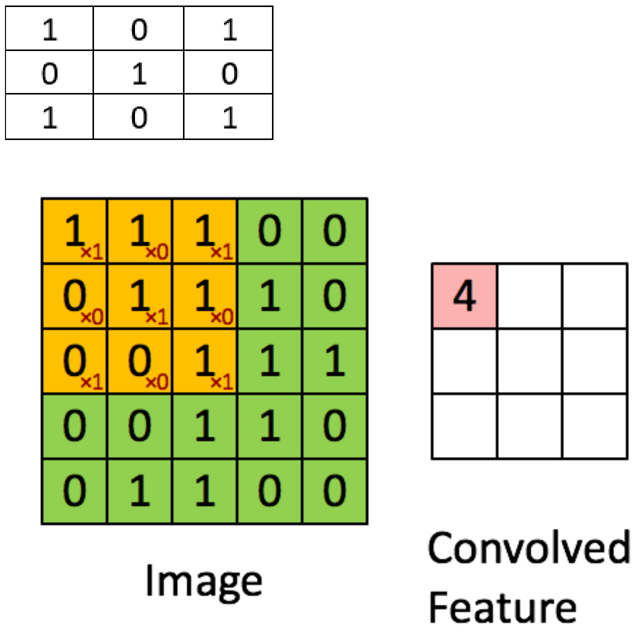
SOME CONSIDERATIONS (2)



- Multi-classification MLPs
  - there will be an output unit for each of the labels
  - *Ex:* *n*-way topic classification
    - 3 labels in Sentiment Analysis: Positive, Negative, Neutral
- Direct Classification MLPs
  - Binary TASK (True/False)

CONVOLUTIONAL NEURAL NETWORKS (LE CUN, 1998)

- Mainly used for images related tasks
  - image classification
  - face detection
  - etc…
- **Learn feature representations**
  - by ***convolving*** over the input
  - with a ***filter***, that slides over the input image
- **Compositionality** (local)
  - Each filter composes a local patch of lower-level features into a higher-level representation
- **Location Invariance**
  - the detection of specific patterns is independent of where it occurs



1	0	1
0	1	0
1	0	1

$1_{\times 1}$	$1_{\times 0}$	$1_{\times 1}$	0	0
$0_{\times 0}$	$1_{\times 1}$	$1_{\times 0}$	1	0
$0_{\times 1}$	$0_{\times 0}$	$1_{\times 1}$	1	1
0	0	1	1	0
0	1	1	0	0

Image

4		

Convolved Feature

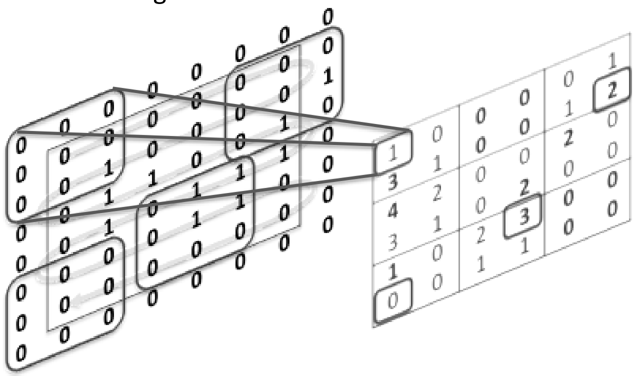
A FUTHER EXAMPLE OF: CONVOLUTION WITH POOLING, AND DECIMATION OPERATIONS

Filter

-1	0	1
-2	0	2
-1	0	1

Image

Convolution



Max Pooling

3	3	0	0	2	2
3	3	0	0	2	2
4	4	2	2	2	2
4	4	2	2	2	2
1	1	3	3	0	0
1	1	3	3	0	0

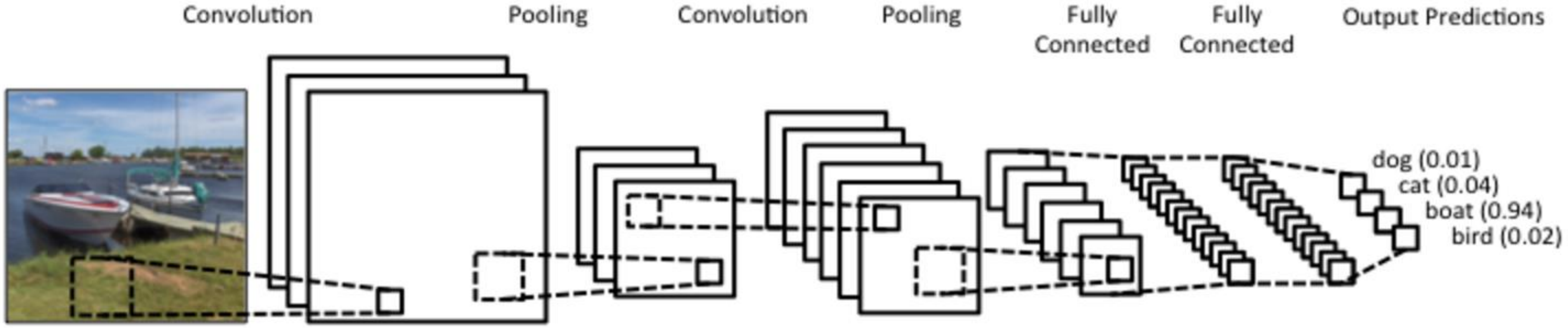
Decimation

3	0	2
4	2	2
1	3	0

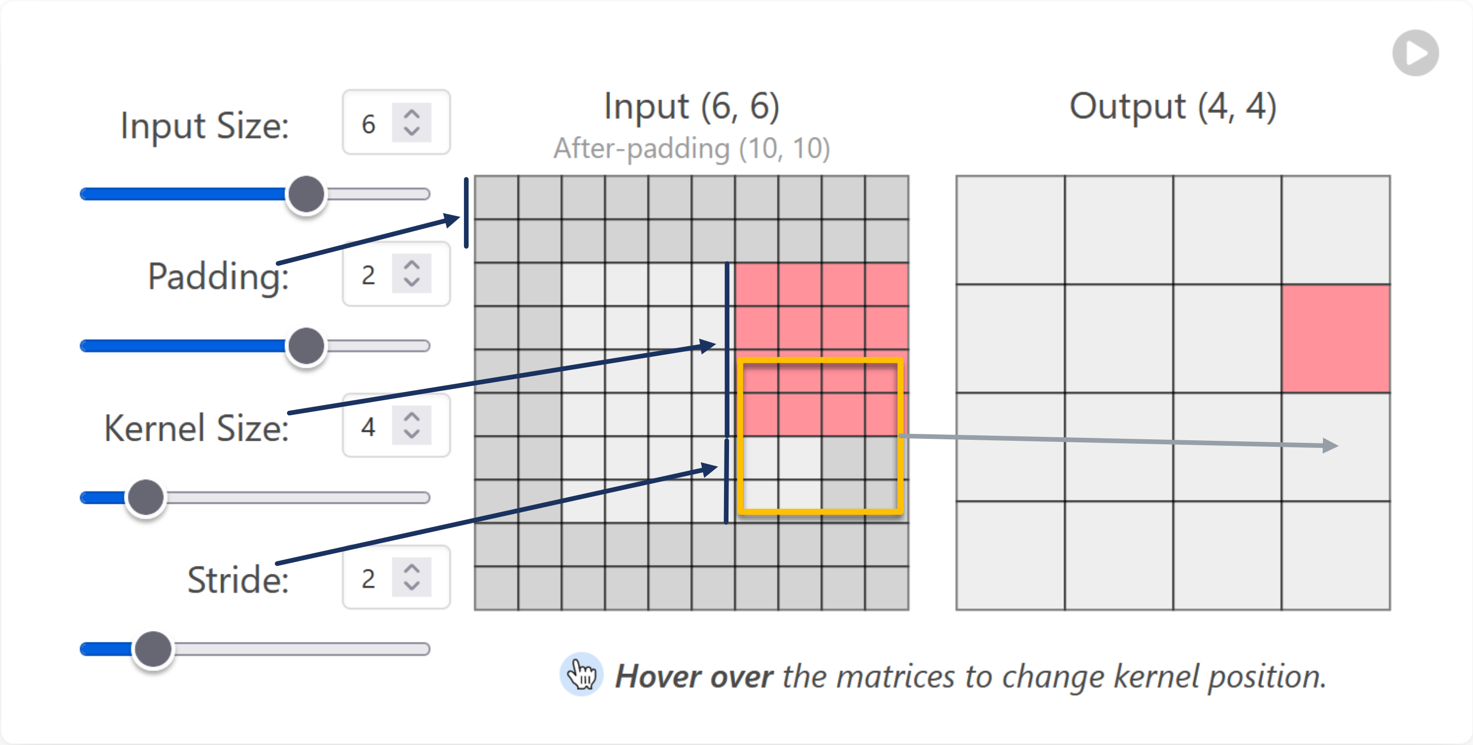
- An image is convolved with a filter; curved rectangular regions in the first large matrix depict a random set of image locations
- Maximum values within small 2×2 regions are indicated in bold in the central matrix
- The results are pooled, using max-pooling then decimated by a factor of two, to yield the final matrix

CONVOLUTIONAL NEURAL NETWORKS

- CNNs automatically learn the parameters of the filters
  - a filter is a matrix of parameters
  - the key aspect is that a filter is adopted for the whole image
- Convolution can be applied in **multiple** layers
  - a layer $l+1$ is computed by convolving over output produced in layer l
  - Pooling is an operation often adopted for taking the most informative features that are learned after a convolution step



TRAINING A CNN: TERMINOLOGY



Input Size: 6

Padding: 2

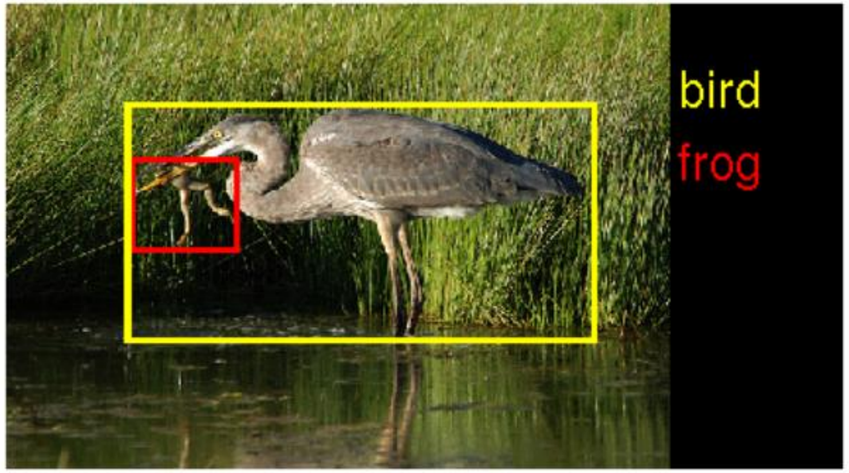
Kernel Size: 4

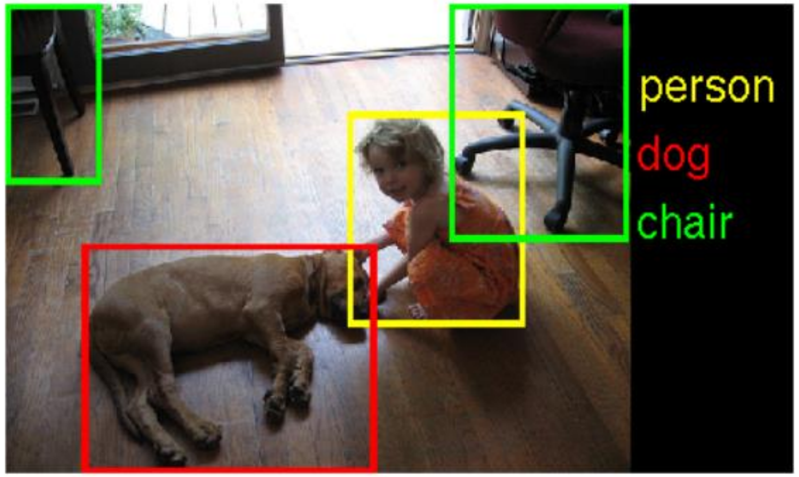
Stride: 2

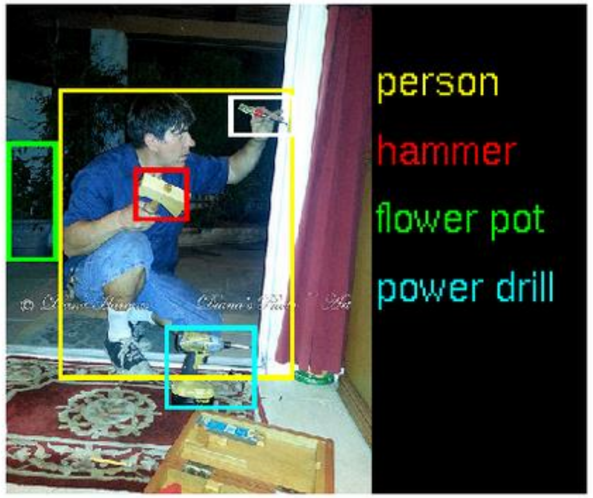
Hover over *the matrices to change kernel position.*

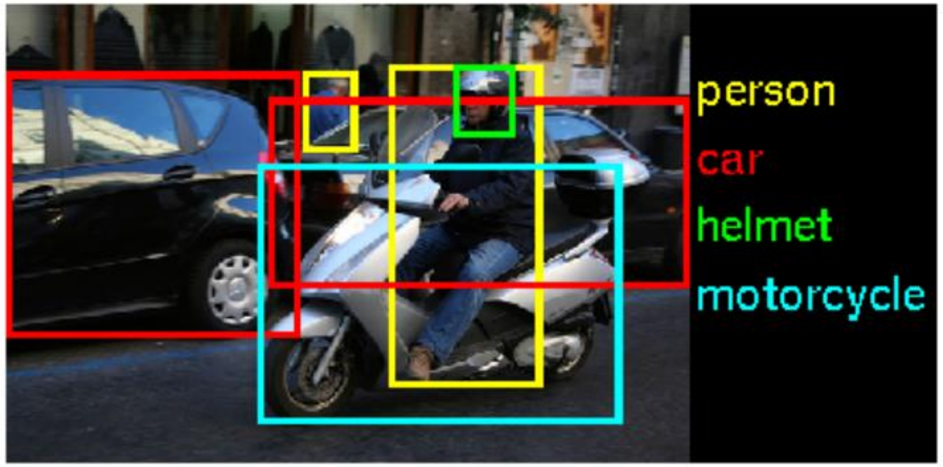
$$O = \frac{InputD - KernelD + 2PaddingD}{StrideD} + 1$$

ILSVRC2014 EXAMPLES









WHAT HAS BEEN LEARNT?

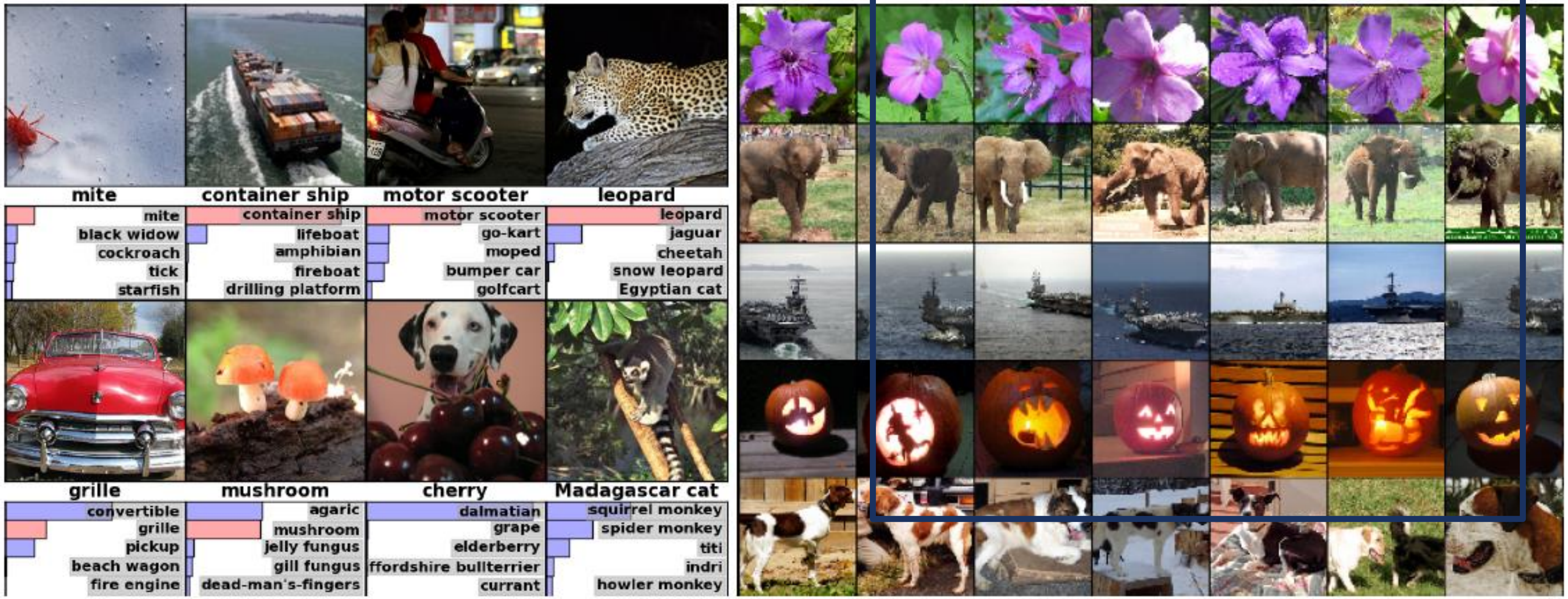
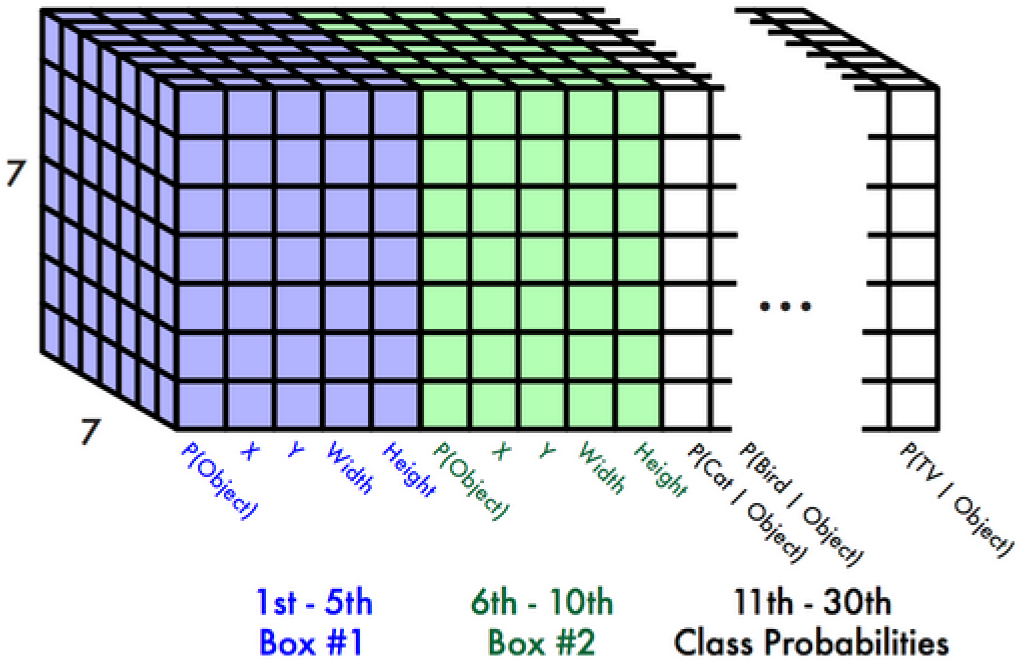


Figure 4: **(Left)** Eight ILSVRC-2010 test images and the five labels considered most probable by our model. The correct label is written under each image, and the probability assigned to the correct label is also shown with a red bar (if it happens to be in the top 5). **(Right)** Five ILSVRC-2010 test images in the first column. The remaining columns show the six training images that produce feature vectors in the last hidden layer with the smallest Euclidean distance from the feature vector for the test image.

YOLO: THE OUTPUT SIZE

Each cell predicts:

- For each bounding box:
  - 4 coordinates (x, y, w, h)
  - 1 confidence value
- Some number of class probabilities



For Pascal VOC:

- 7x7 grid
- 2 bounding boxes / cell
- 20 classes

7 x 7 x (2 x 5 + 20) = 7 x 7 x 30 tensor = **1470 outputs**

See also: https://pjreddie.com/publications/

CURRENT CNNS: YOLO

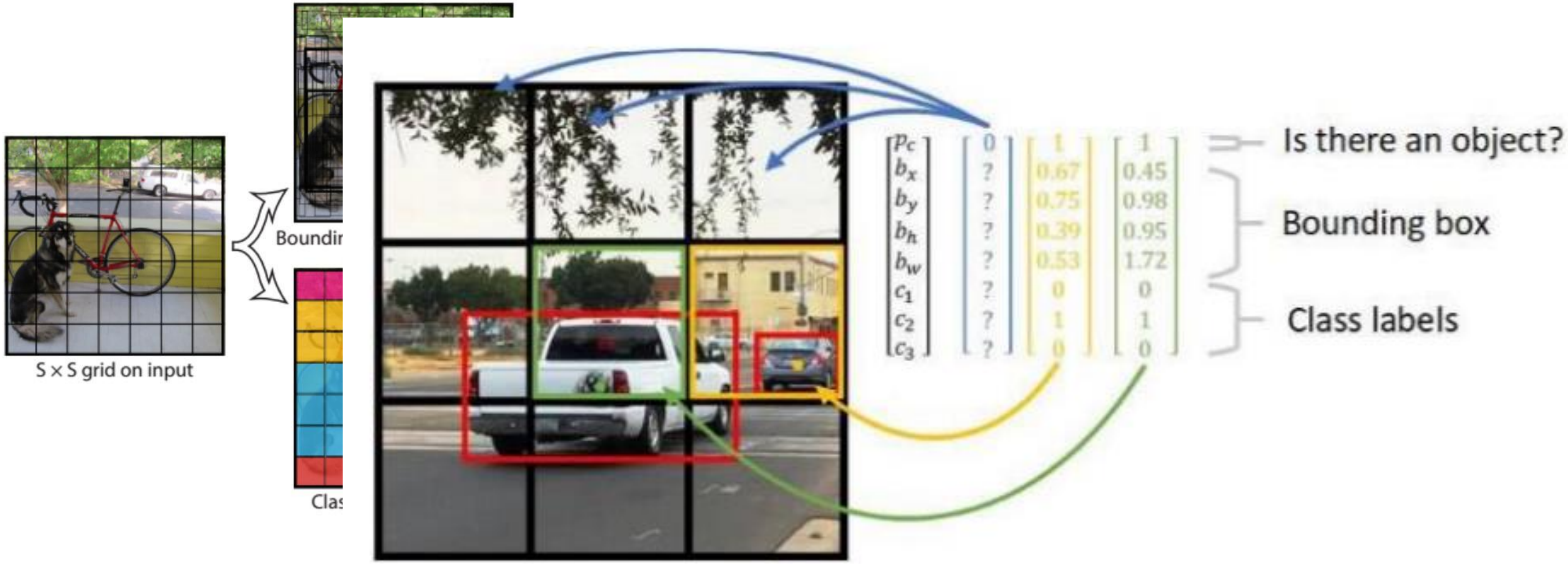
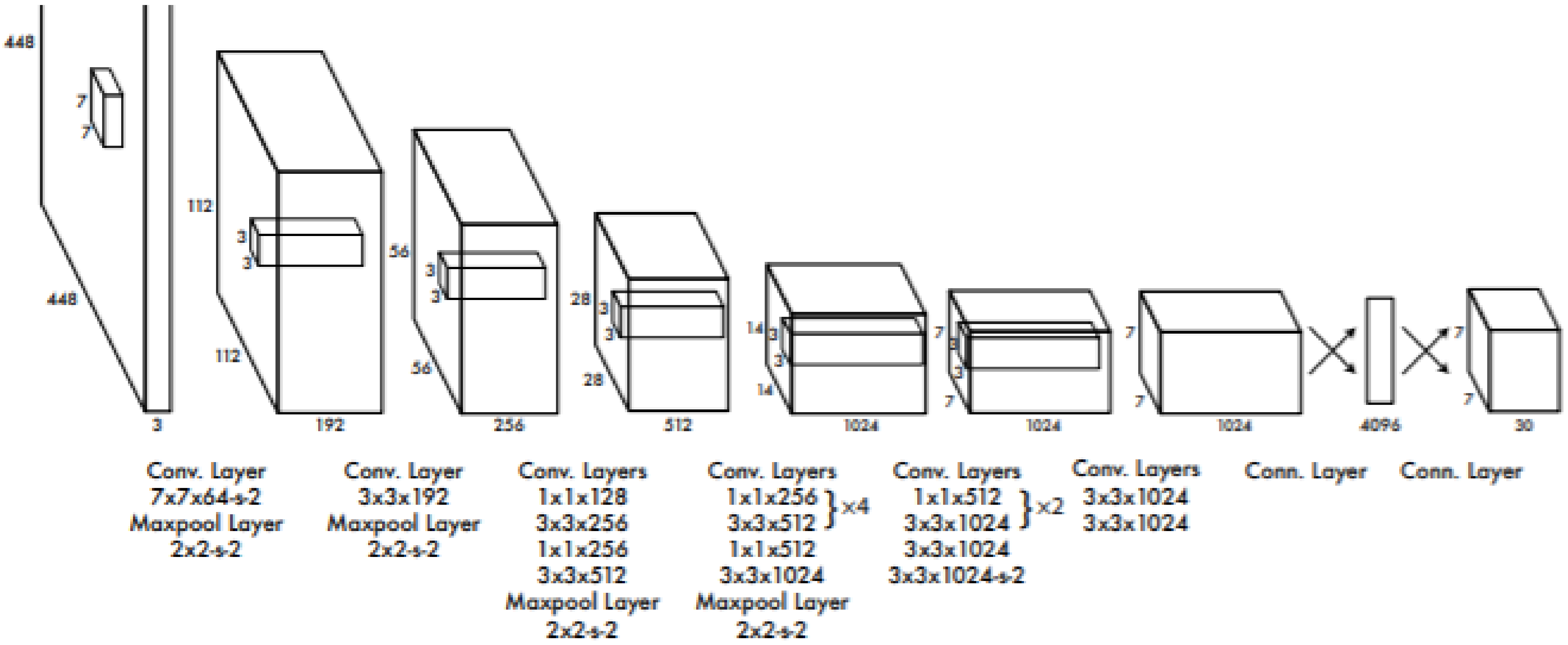


Figure 2: The Model. Our system models detection as a regression problem. It divides the image into an $S \times S$ grid and for each grid cell predicts B bounding boxes, confidence for those boxes, and C class probabilities. These predictions are encoded as an $S \times S \times (B * 5 + C)$ tensor.

YOLO: THE ARCHITECTURE



YOLO: RESULTS

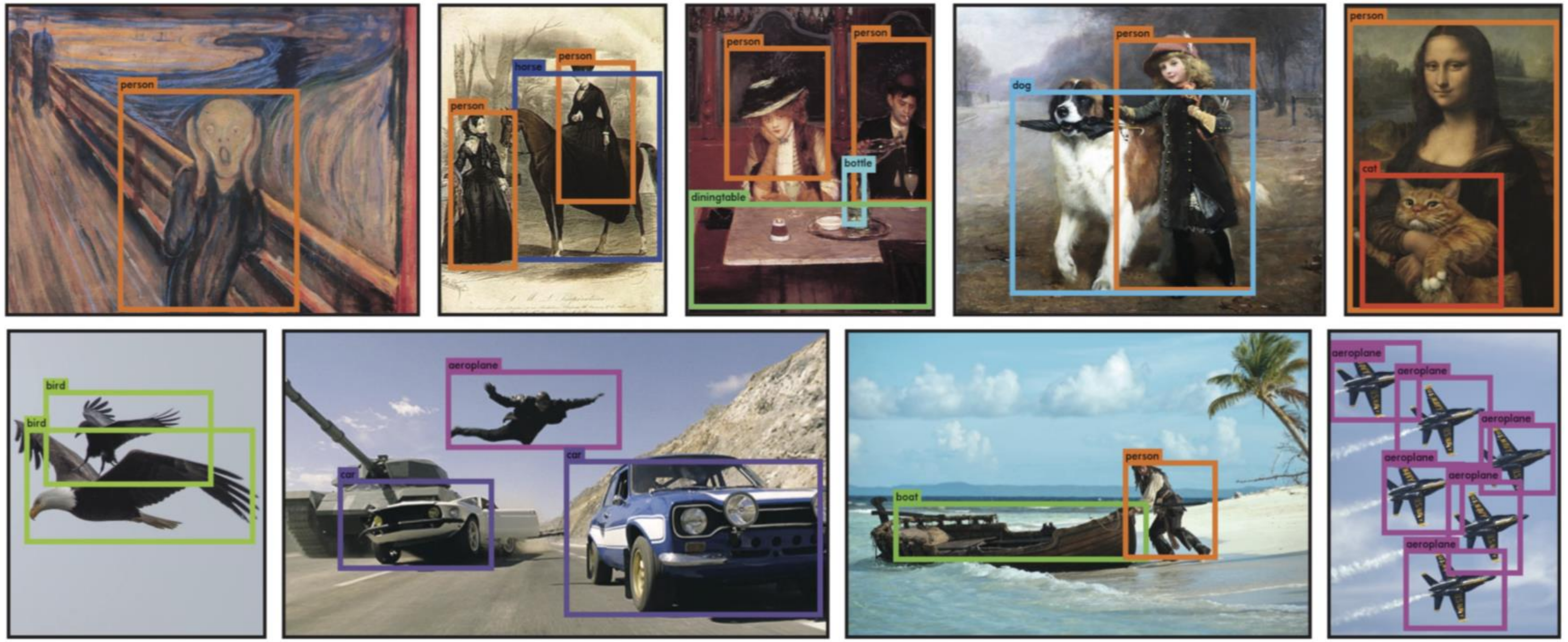
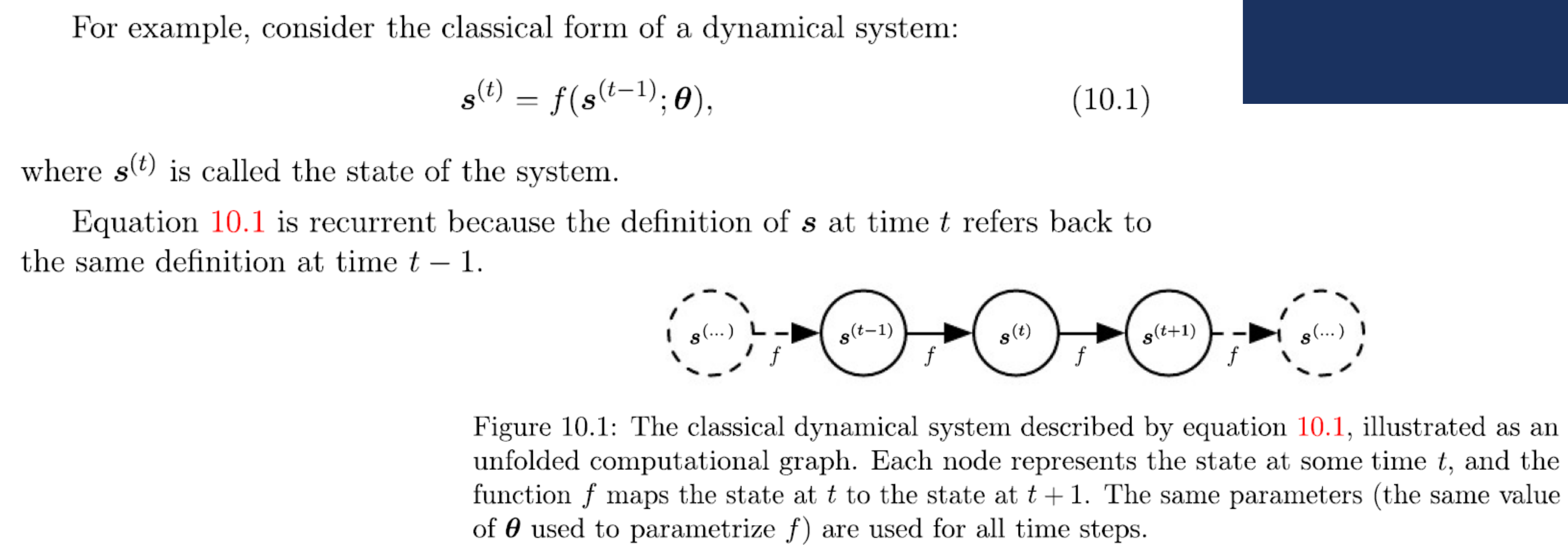


Figure 6: Qualitative Results. YOLO running on sample artwork and natural images from the internet. It is mostly accurate although it does think one person is an airplane.

RETI NEURALI RICORRENTI (RECAP)

LE RETI RICORRENTI

RECURRENT NEURAL NETWORKS



For example, consider the classical form of a dynamical system:

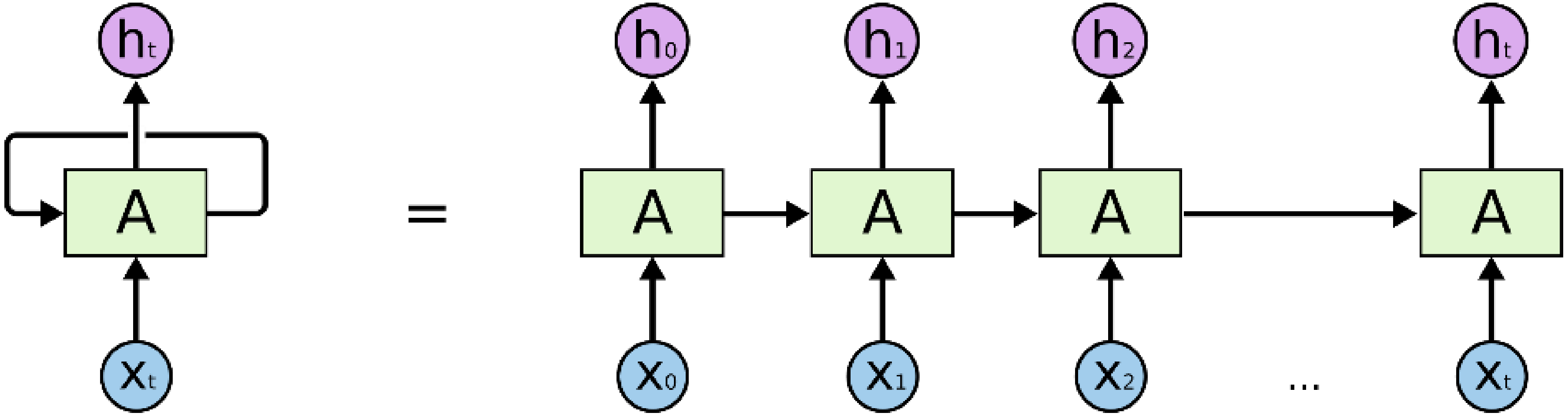
$$\boldsymbol{s}^{(t)} = f(\boldsymbol{s}^{(t-1)}; \boldsymbol{\theta}), \tag{10.1}$$

where $\boldsymbol{s}^{(t)}$ is called the state of the system.

Equation 10.1 is recurrent because the definition of $\boldsymbol{s}$ at time t refers back to the same definition at time $t-1$.

$\boldsymbol{s}^{(\dots)}$
f
$\boldsymbol{s}^{(t-1)}$
f
$\boldsymbol{s}^{(t)}$
f
$\boldsymbol{s}^{(t+1)}$
f
$\boldsymbol{s}^{(\dots)}$

Figure 10.1: The classical dynamical system described by equation 10.1, illustrated as an unfolded computational graph. Each node represents the state at some time t, and the function f maps the state at t to the state at $t+1$. The same parameters (the same value of $\boldsymbol{\theta}$ used to parametrize f) are used for all time steps.



TRAINING A RNN

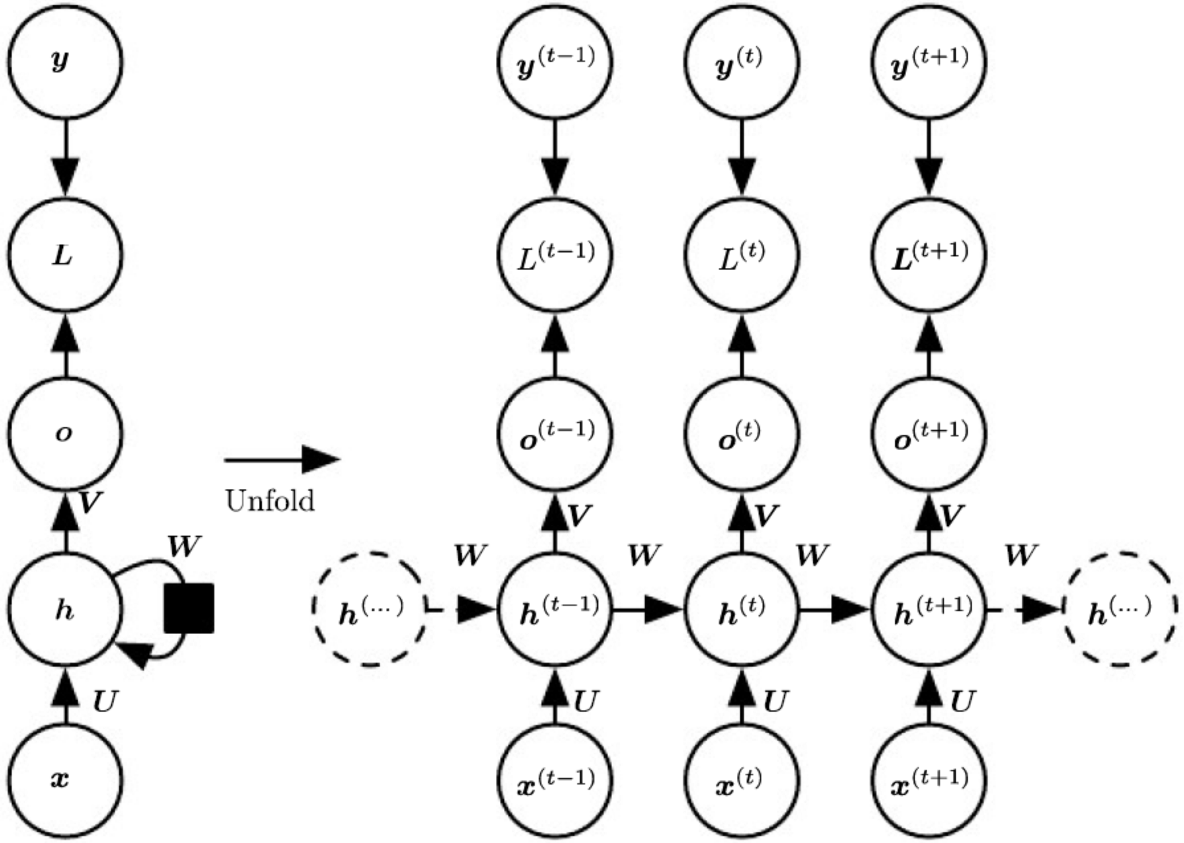


Figure 10.3: The computational graph to compute the training loss of a recurrent network that maps an input sequence of $\boldsymbol{x}$ values to a corresponding sequence of output $\boldsymbol{o}$ values. A loss L measures how far each $\boldsymbol{o}$ is from the corresponding training target $\boldsymbol{y}$. When using softmax outputs, we assume $\boldsymbol{o}$ is the unnormalized log probabilities. The loss L internally computes $\hat{\boldsymbol{y}} = \text{softmax}(\boldsymbol{o})$ and compares this to the target $\boldsymbol{y}$. The RNN has input to hidden connections parametrized by a weight matrix $\boldsymbol{U}$, hidden-to-hidden recurrent connections parametrized by a weight matrix $\boldsymbol{W}$, and hidden-to-output connections parametrized by a weight matrix $\boldsymbol{V}$. Equation 10.8 defines forward propagation in this model. *(Left)* The RNN and its loss drawn with recurrent connections. *(Right)* The same seen as a time-unfolded computational graph, where each node is now associated with one particular time instance.

TYPES OF RNNS

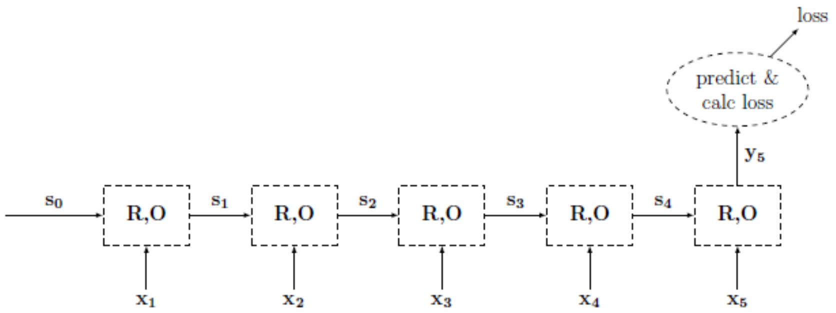


Figure 7: Acceptor RNN Training Graph.

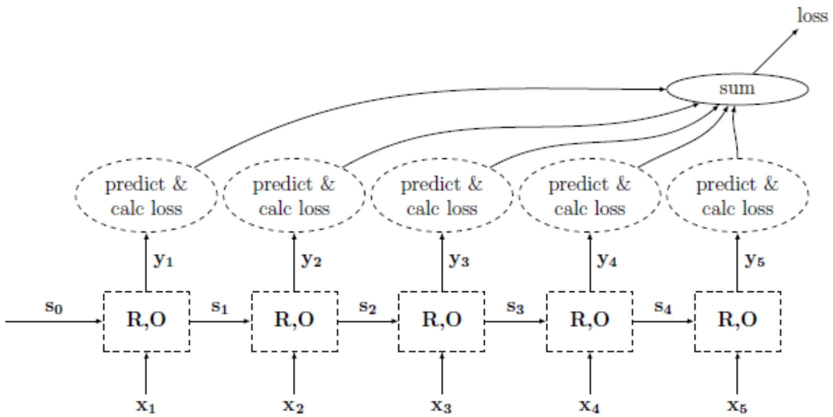


Figure 8: Transducer RNN Training Graph.

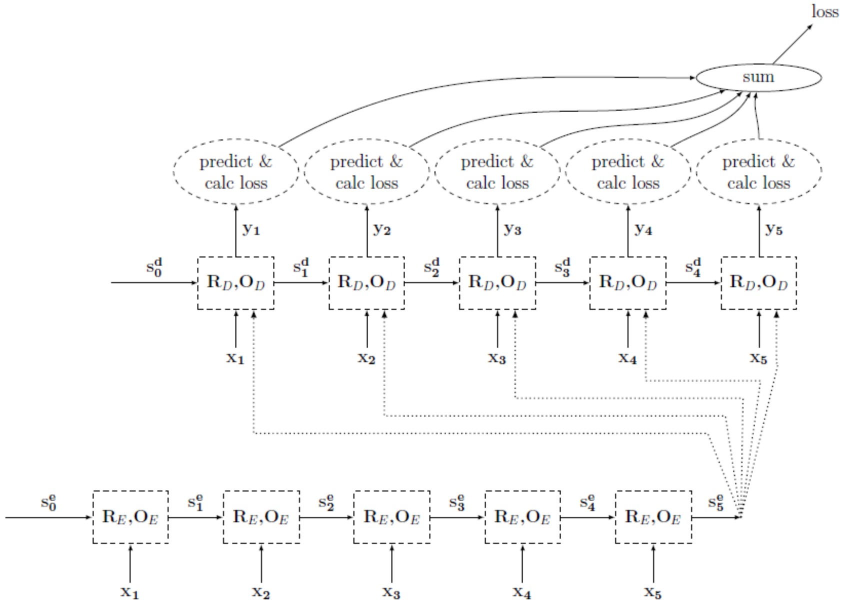


Figure 9: Encoder-Decoder RNN Training Graph.

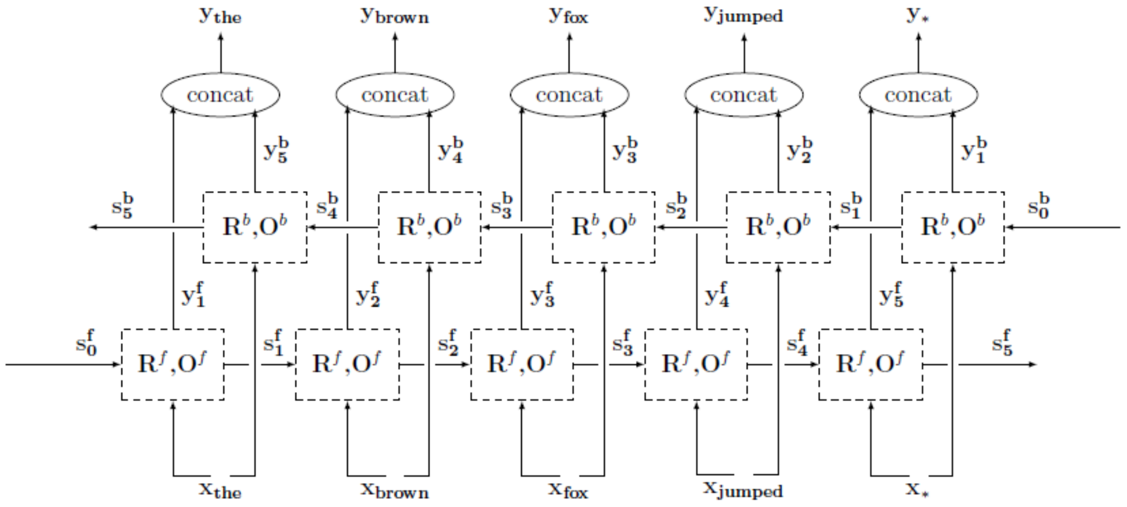
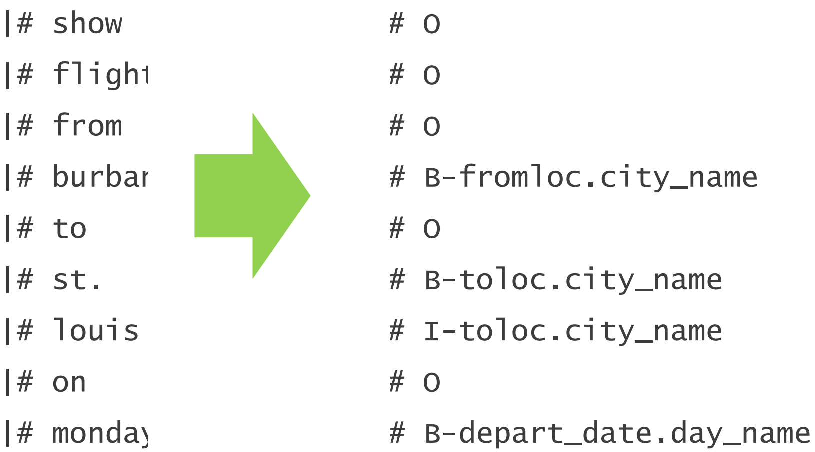


Figure 11: biRNN over the sentence "the brown fox jumped .".

EXAMPLES: LANGUAGE UNDERSTANDING

THE MS COGNITIVE TOOLKIT

Task: Slot tagging with an LSTM



```
|# show          # O
|# flight        # O
|# from          # O
|# burbar        # B-fromloc.city_name
|# to            # O
|# st.           # B-toloc.city_name
|# louis         # I-toloc.city_name
|# on            # O
|# monday        # B-depart_date.day_name
```

https://learn.microsoft.com/en-us/cognitive-toolkit/Hands-On-Labs-Language-Understanding

EXAMPLES: LANGUAGE UNDERSTANDING

THE MS COGNITIVE TOOLKIT

Task: Slot tagging with an LSTM

```
19  |x 178:1 |# BOS      |y 128:1 |# O
19  |x 770:1 |# show     |y 128:1 |# O
19  |x 429:1 |# flights  |y 128:1 |# O
19  |x 444:1 |# from     |y 128:1 |# O
19  |x 272:1 |# burbank  |y 48:1  |# B-fromloc.city_name
19  |x 851:1 |# to       |y 128:1 |# O
19  |x 789:1 |# st.      |y 78:1  |# B-toloc.city_name
19  |x 564:1 |# louis    |y 125:1 |# I-toloc.city_name
19  |x 654:1 |# on       |y 128:1 |# O
19  |x 601:1 |# monday   |y 26:1  |# B-depart_date.day_name
19  |x 179:1 |# EOS      |y 128:1 |# O
```

```
    ^
    |
+-------+
| Dense |
+-------+
    ^
    |
+------+
| LSTM |
+------+
    ^
    |
+-------+
| Embed |
+-------+
    ^
    |
```

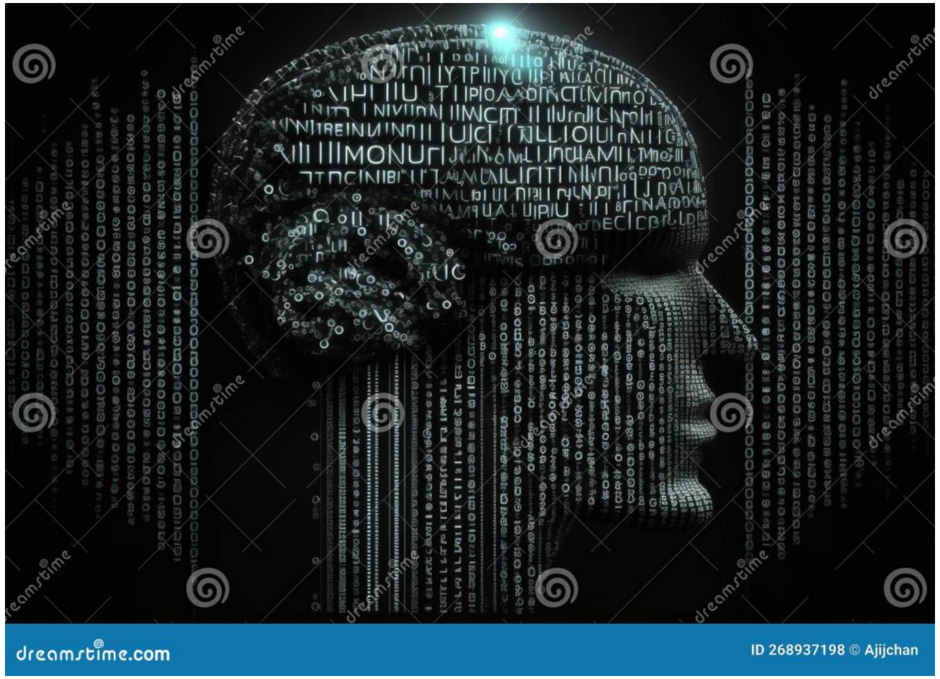
EXAMPLES: LANGUAGE UNDERSTANDING

THE MS COGNITIVE TOOLKIT

Task: Slot tagging with an LSTM

```
19   |x 178:1 |# BOS       |y 128:1 |# O
19   |x 770:1 |# show      |y 128:1 |# O
19   |x 429:1 |# flights   |y 128:1 |# O
19   |x 444:1 |# from      |y 128:1 |# O
19   |x 272:1 |# burbank   |y 48:1  |# B-fromloc.city_name
19   |x 851:1 |# to        |y 128:1 |# O
19   |x 789:1 |# st.       |y 78:1  |# B-toloc.city_name
19   |x 564:1 |# louis     |y 125:1 |# I-toloc.city_name
19   |x 654:1 |# on        |y 128:1 |# O
19   |x 601:1 |# monday    |y 26:1  |# B-depart_date.day_name
19   |x 179:1 |# EOS       |y 128:1 |# O
```

```
y       "O"        "O"        "O"        "O"  "B-fromloc.city_name"
         ^          ^          ^          ^          ^
         |          |          |          |          |
     +-------+  +-------+  +-------+  +-------+  +-------+
     | Dense |  | Dense |  | Dense |  | Dense |  | Dense |  ...
     +-------+  +-------+  +-------+  +-------+  +-------+
         ^          ^          ^          ^          ^
         |          |          |          |          |
     +------+   +------+   +------+   +------+   +------+
0 -->| LSTM |-->| LSTM |-->| LSTM |-->| LSTM |-->| LSTM |-->...
     +------+   +------+   +------+   +------+   +------+
         ^          ^          ^          ^          ^
         |          |          |          |          |
     +-------+  +-------+  +-------+  +-------+  +-------+
     | Embed |  | Embed |  | Embed |  | Embed |  | Embed |  ...
     +-------+  +-------+  +-------+  +-------+  +-------+
         ^          ^          ^          ^          ^
         |          |          |          |          |
x ------>+--------->+--------->+--------->+--------->+------...
        BOS      "show"    "flights"    "from"   "burbank"
```



dreamstime.com ID 268937198 © Ajijchan

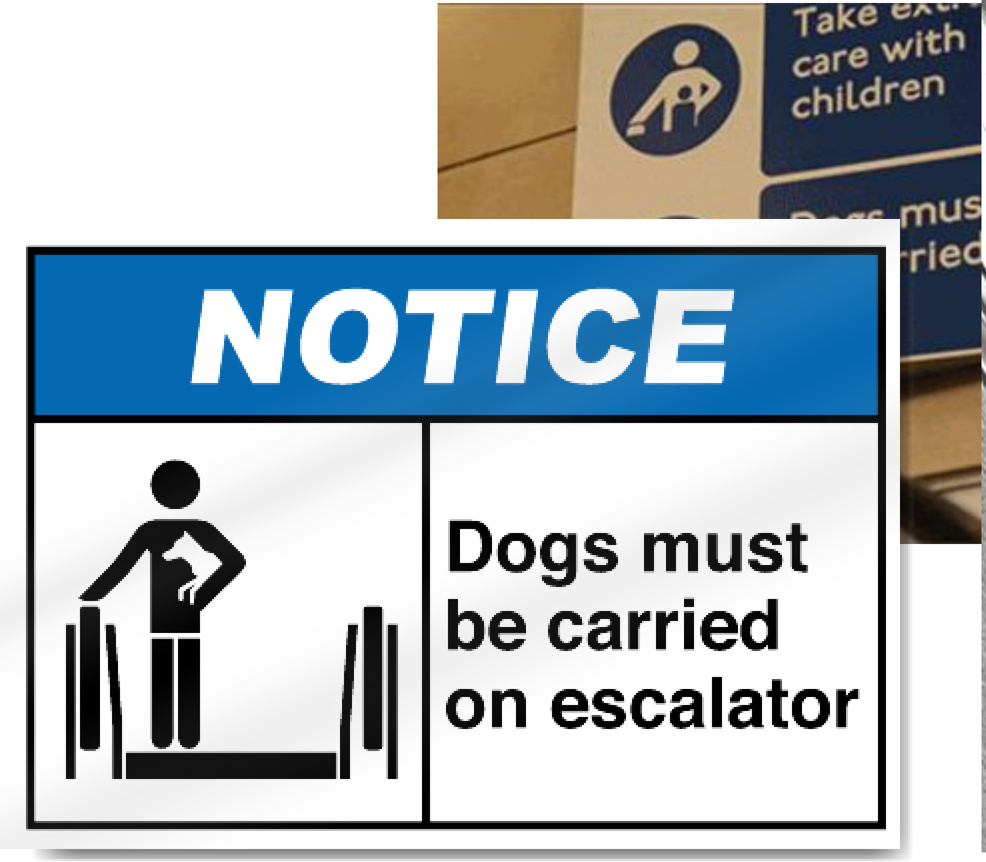
MODELLI FONDAZIONALI PER IL *NLP*

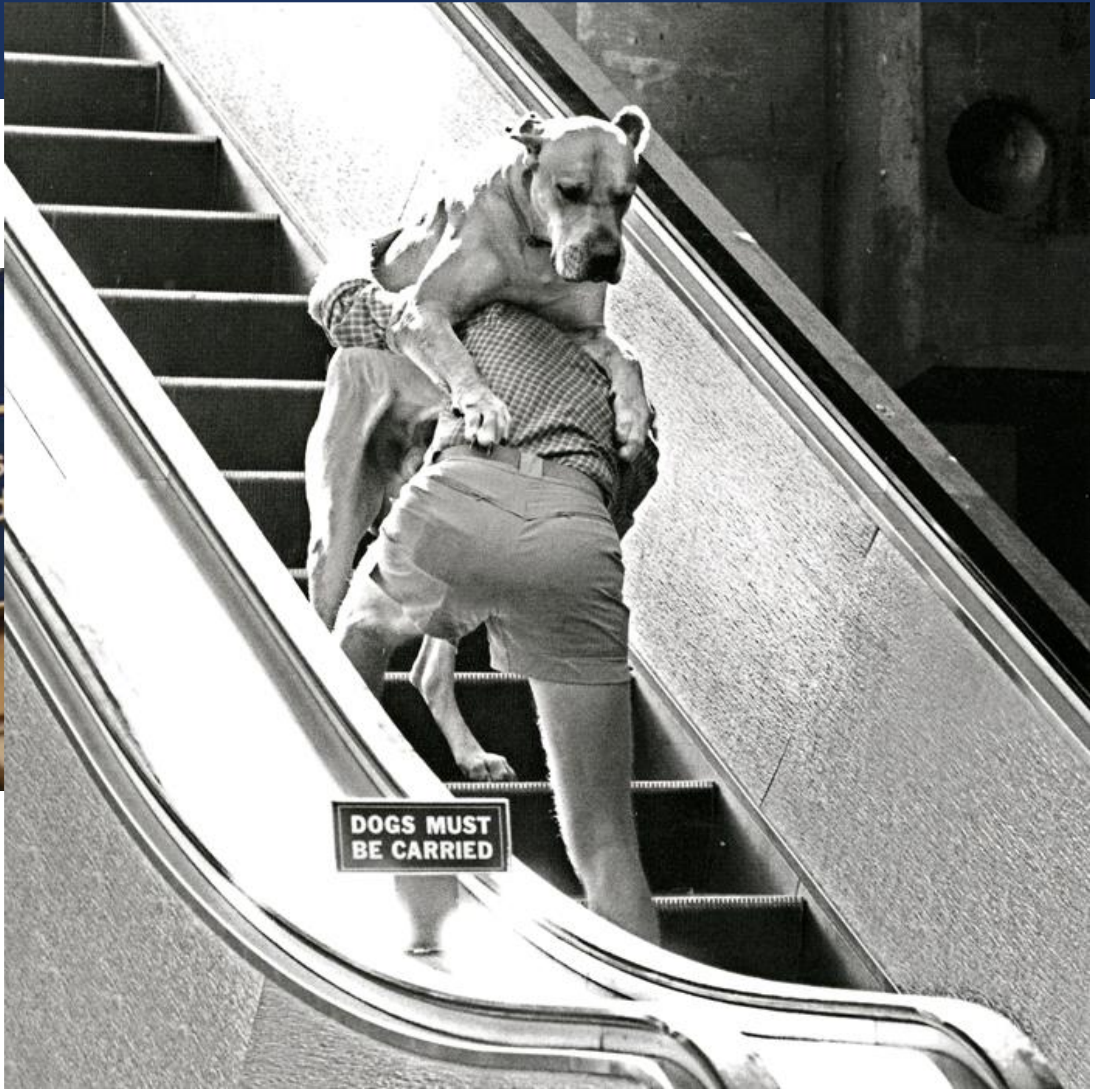
NATURAL LANGUAGE UNDERSTNDING, PROBABILISTIC LANGUAGE MODELS, TRASFORMERS

NATURAL LANGUAGE PROCESSING:
AT THE HEART OF GENERATIVE AI SYSTEMS

- Syntax, Semantics and Pragmatics in Artificial Intelligent Agents
- Language Modeling:
  - Statistical approaches
  - Neural approaches to NL semantics
  - Neural Probabilistic Language Models
- Encoder-Decoder architectures

NATURAL LANGUAGE & AMBIGUITY





AMBIGUITY: AN EXAMPLE

- "*Dogs must be carried on this escalator*"

can be consistently interpreted in a number of ways:

- *All dogs should have a chance to go on this wonderful escalator ride*
- *This escalator is for dog-holders only*
- *You can't carry your pet on the other escalators*
- *When riding with a pet, carry it*

THE NLP CHAIN: LEVELS OF LINGUISTIC ANALYSIS

- Given an **valid utterance** such as

  John, I am freezing

- vs.

  I, John, freezing am

Pragmatics: what does it do?

Semantics: what does it mean?

Syntax: what is grammatical?

ANALOGY WITH ARTIFICIAL LANGUAGES

Syntax: no compiler errors

Semantics: no implementation bugs

Pragmatics: implemented the right algorithm

Different syntax, same semantics (5):

$$2 + 3 \Leftrightarrow 3 + 2$$

Same syntax, different semantics (1 and 1.5):

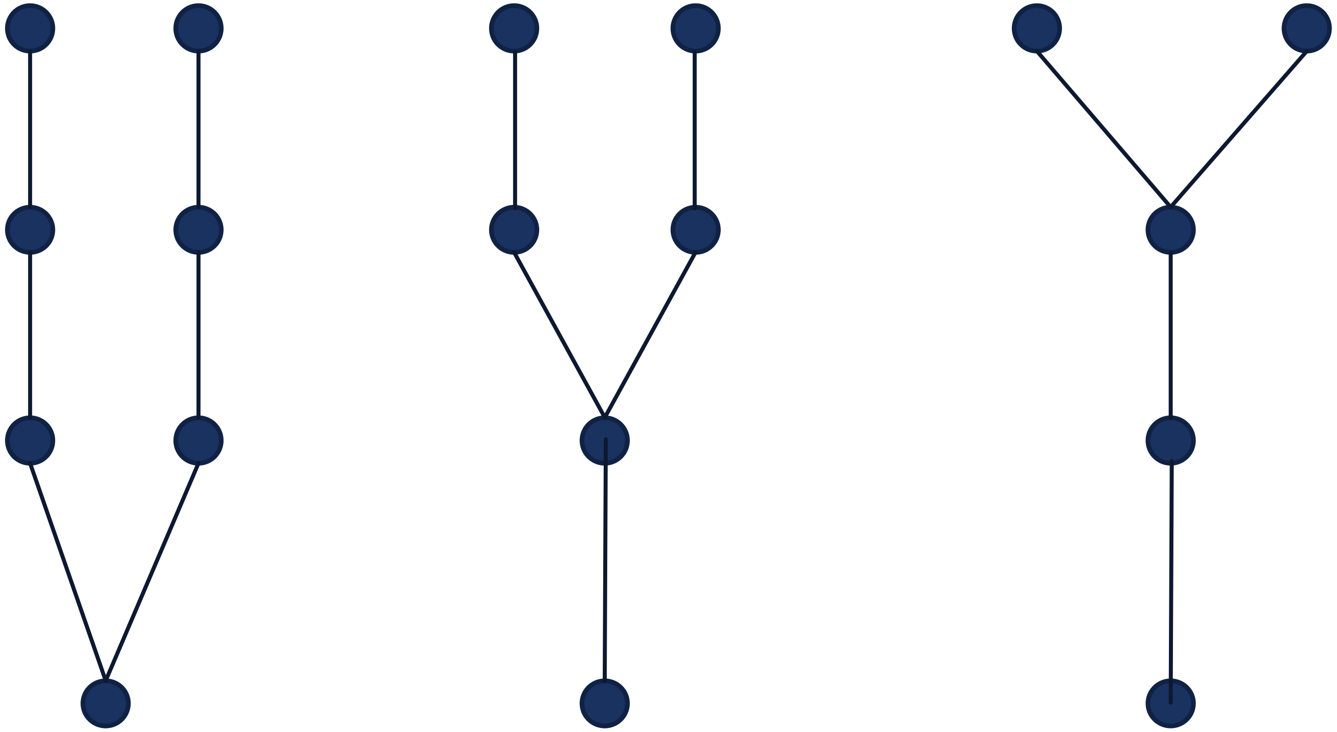
$$3 \;/\; 2 \text{ (Python 2.7)} \not\Leftrightarrow 3 \;/\; 2 \text{ (Python 3)}$$

Good semantics, bad pragmatics:

correct implementation of deep neural network
for estimating coin flip prob.

AMBIGUITY AND LINGUISTIC LEVELS

- Semantics
- Syntax
- Morphology
- Phonology

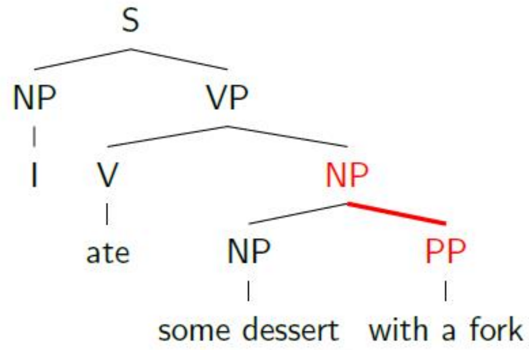


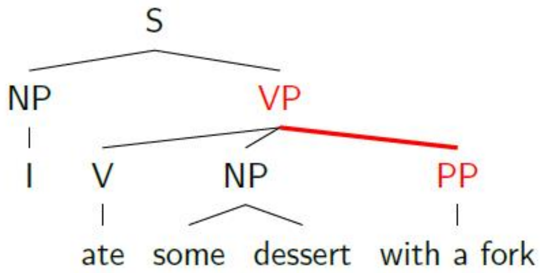
🇬🇧 *can/can* *eat cake with fork* *earth observation satellite* *Eco’s book*

🇮🇹 *del (pane) /del (libro)* *compro la borsa in pelle* *il timore dei manager*

GRAMMARS & AMBIGUITY

I ate some dessert with a fork.





PARSING & AMBIGUITY

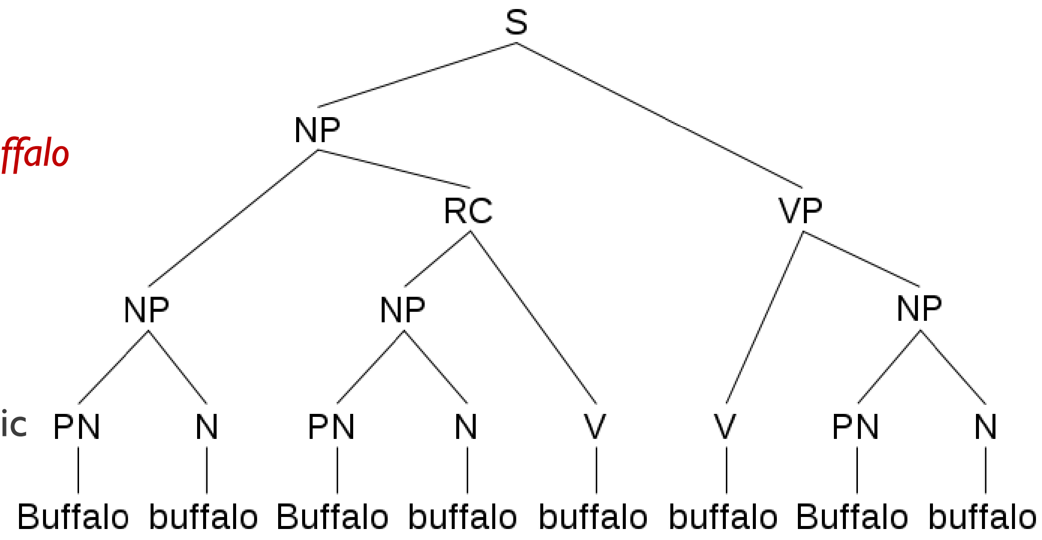
- The parser search space is huge as for the effect of several forms of ambiguity that interacts in a combinatorial way
  - e.g. *La vecchia porta la sbarra*,
  - or *Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo*
- Notice the strong relationship with semantics
  - Most of the ambiguities cannot be solved at the sole syntactic level
  - Lexical information (e.g. word senses) are crucial:

🇬🇧

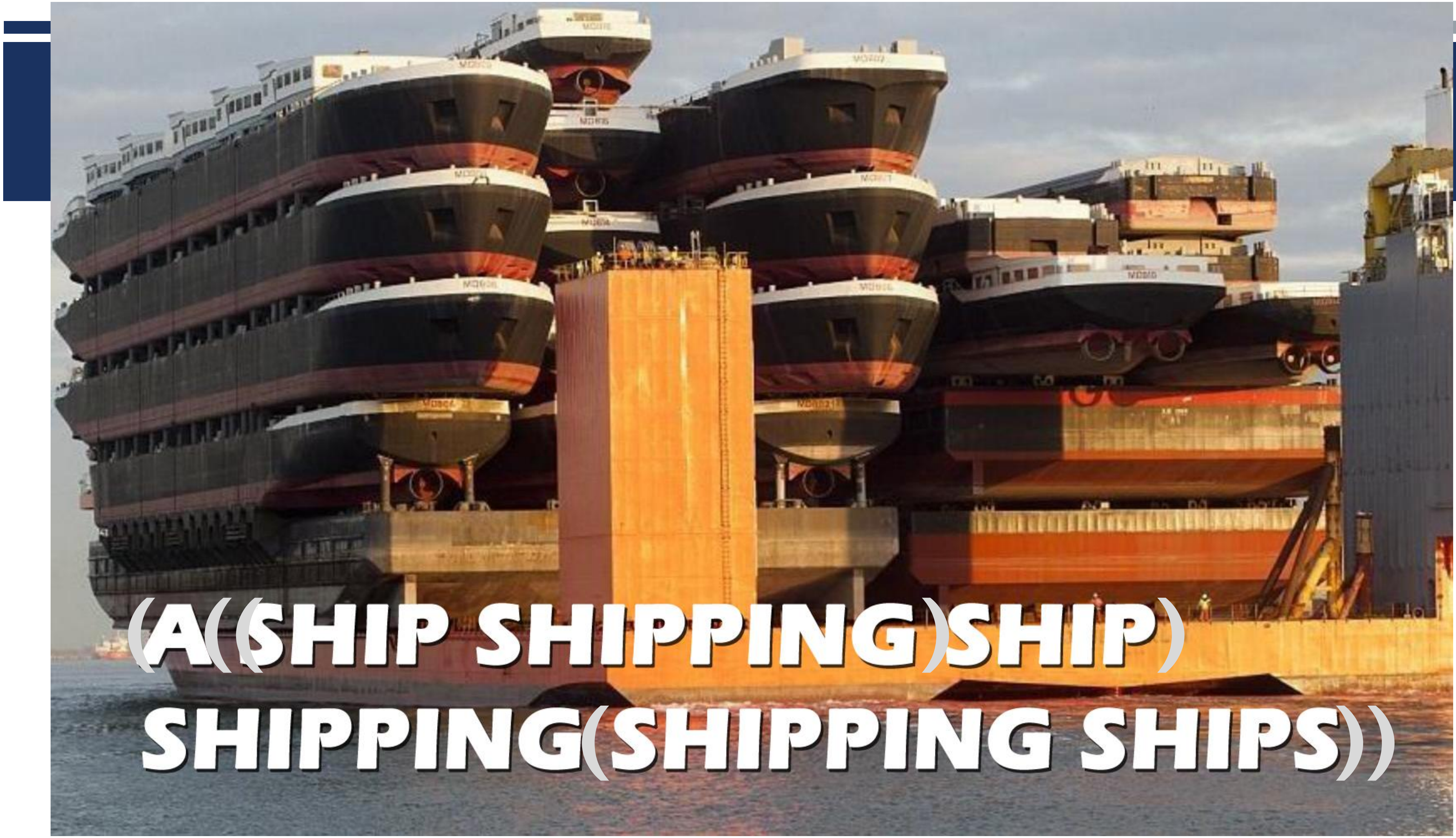
- *To operate in a market* viz. *To operate a body part*

🇮🇹

- *Operare in un mercato* ≠ *Operare un paziente*



Bison from Buffalo, New York who are intimidated by other bison in their community also happen to intimidate other bison in their community



(A((SHIP SHIPPING)SHIP) SHIPPING(SHIPPING SHIPS))

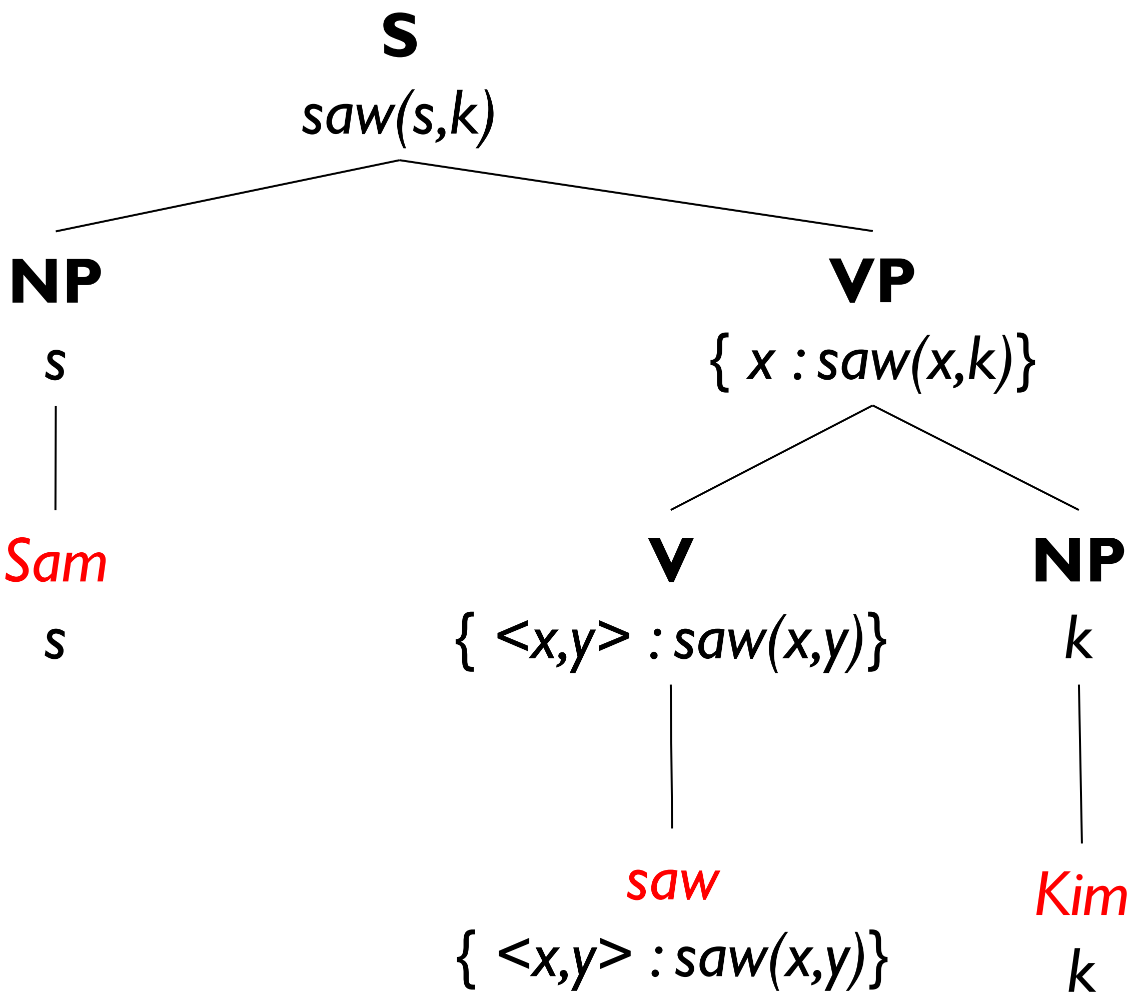
SEMANTICS

- What is the meaning of the sentence

  John saw Kim?

- Desirable Properties:
  - It should be derivable as a function of the individual constituent parts, i.e. the meanings of costituents such as *Kim*, *John* and *see*
  - Independent from syntactic phenomena, e.g. *Kim was seen by John* is a paraphrasis as *it has the same semantics*
  - It must be directy used to trigger some inferences:
    - *Who was seen by John?* Kim!
    - *John saw Kim. He started running to her.*

A TRUTH CONDITIONAL SEMANTICS



S
saw(s,k)

NP
s

Sam
s

VP
{ x : saw(x,k)}

V
{ <x,y> : saw(x,y)}

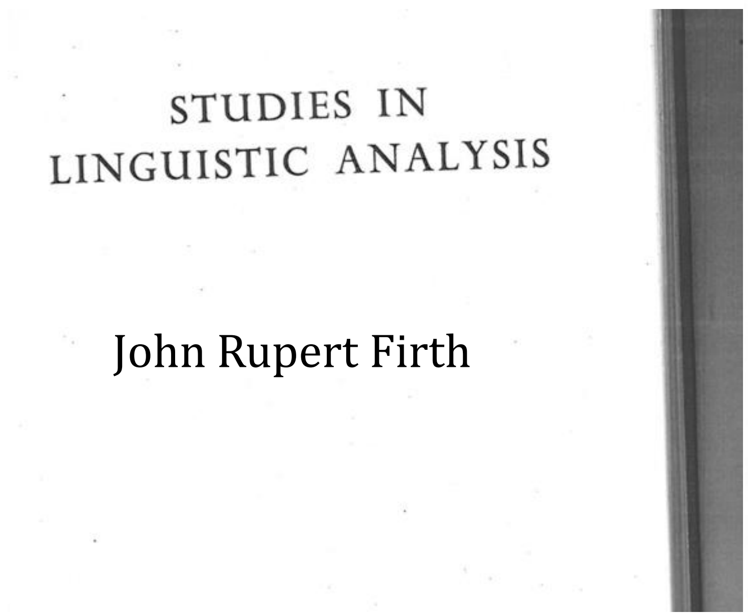
saw
{ <x,y> : saw(x,y)}

NP
k

Kim
k

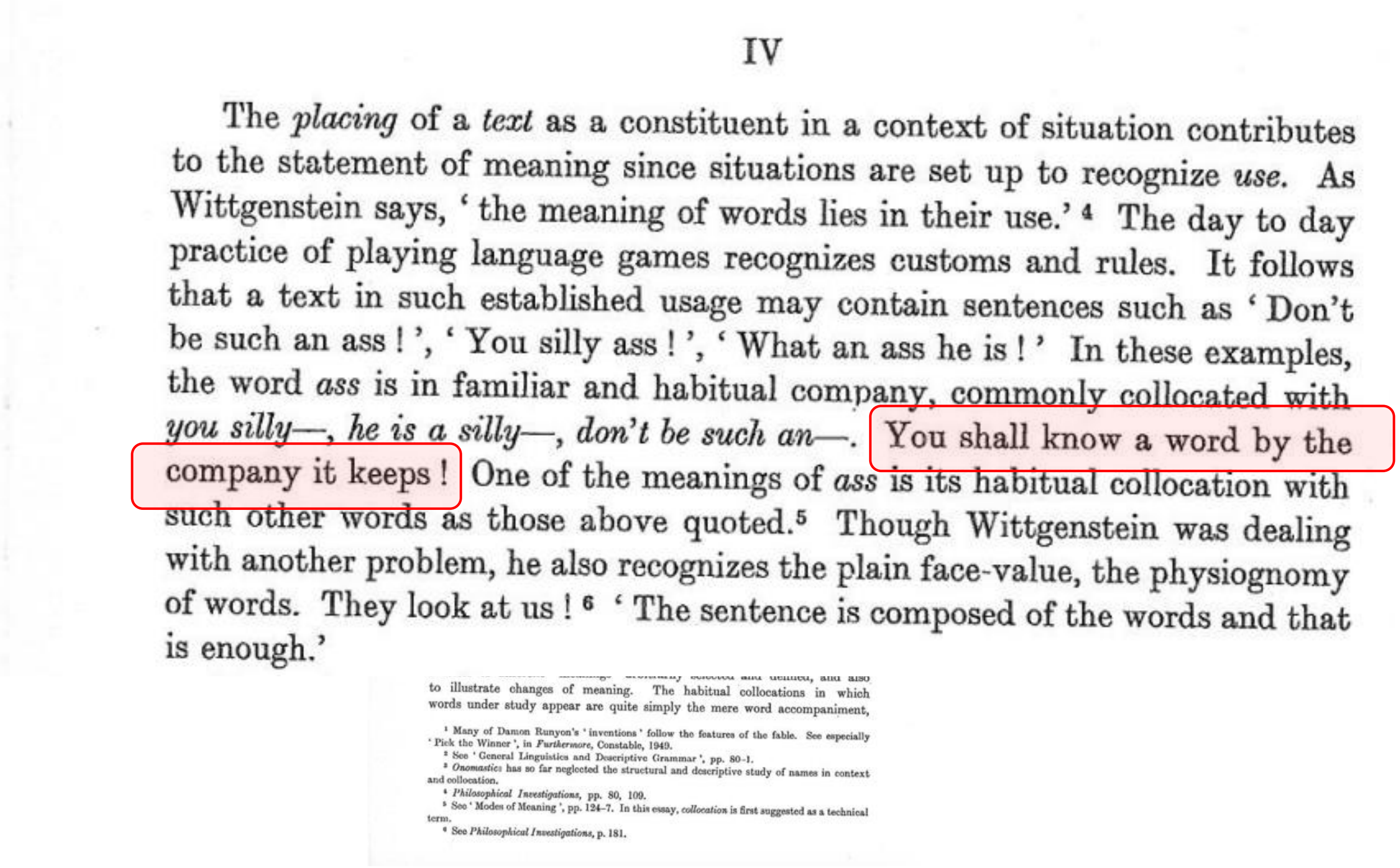
John saw Kim

THE DISTRIBUTIONAL HYPOTHESIS



STUDIES IN LINGUISTIC ANALYSIS

John Rupert Firth



IV

The *placing* of a *text* as a constituent in a context of situation contributes to the statement of meaning since situations are set up to recognize *use*. As Wittgenstein says, 'the meaning of words lies in their use.'[4] The day to day practice of playing language games recognizes customs and rules. It follows that a text in such established usage may contain sentences such as 'Don't be such an ass!', 'You silly ass!', 'What an ass he is!' In these examples, the word *ass* is in familiar and habitual company, commonly collocated with *you silly*—, *he is a silly*—, *don't be such an*—. You shall know a word by the company it keeps! One of the meanings of *ass* is its habitual collocation with such other words as those above quoted.[5] Though Wittgenstein was dealing with another problem, he also recognizes the plain face-value, the physiognomy of words. They look at us![6] 'The sentence is composed of the words and that is enough.'

to illustrate changes of meaning. The habitual collocations in which words under study appear are quite simply the mere word accompaniment,

[1] Many of Damon Runyon's 'inventions' follow the features of the fable. See especially 'Pick the Winner', in *Furthermore*, Constable, 1949.

[2] See 'General Linguistics and Descriptive Grammar', pp. 80-1.

[3] *Onomastics* has so far neglected the structural and descriptive study of names in context and collocation.

[4] *Philosophical Investigations*, pp. 80, 109.

[5] See 'Modes of Meaning', pp. 124–7. In this essay, *collocation* is first suggested as a technical term.

[6] See *Philosophical Investigations*, p. 181.

Firth, J.R. (1957). "A synopsis of linguistic theory 1930-1955". Studies in Linguistic Analysis: 1–32. Reprinted in *F.R. Palmer, ed. (1968). Selected Papers of J.R. Firth 1952-1959. London: Longman.*

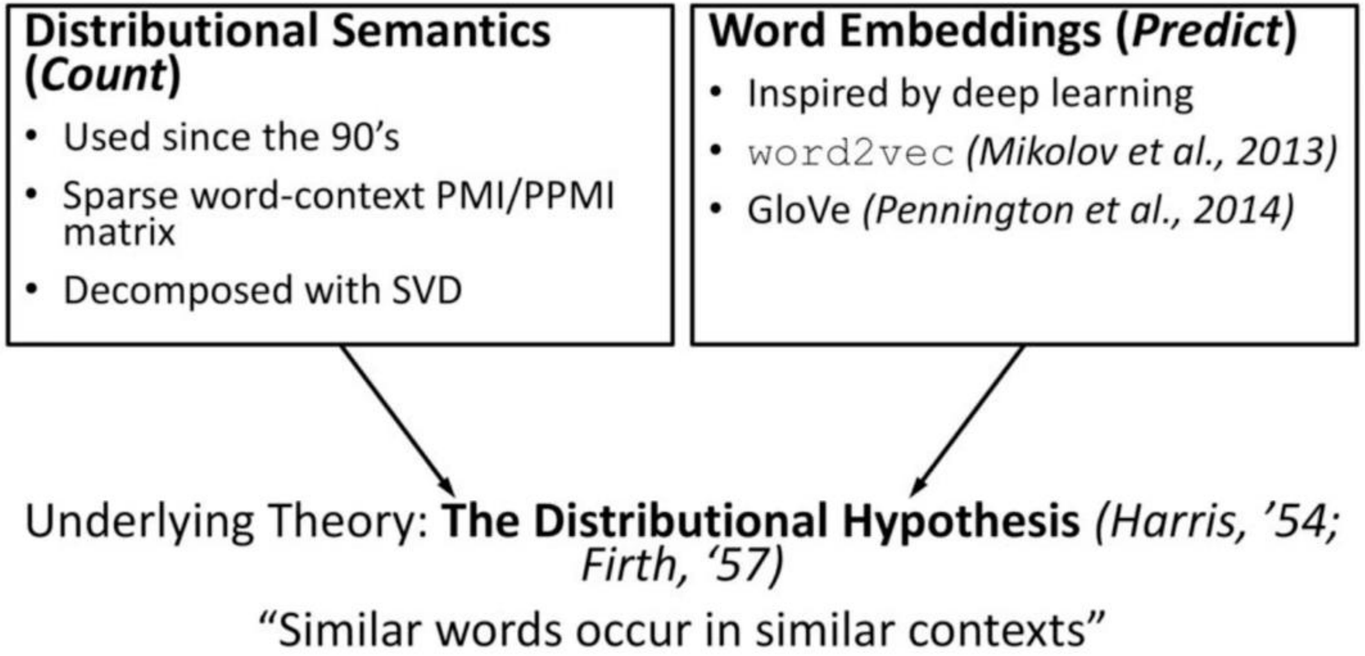
https://cs.brown.edu/courses/csci2952d/readings/lecture1-firth.pdf

LINGUISTICS AND COMPUTATIONAL SEMANTICS

- **Foundation**: Linguistic theory positing that **words with similar contexts have similar meanings**.
  - … and **representation** from out computational perspective

- **Computational Leap**: tied to the Vector Space Model (Salton, 1975); represents documents and words as **vectors in a metric space**.
  - **Key Idea**: Documents are characterized by their words, and words by the documents they appear in.
  - 👶 Initially a Bag of Words model

APPROACHES FOR REPRESENTING WORDS



Distributional Semantics (*Count*)

- Used since the 90's
- Sparse word-context PMI/PPMI matrix
- Decomposed with SVD

Word Embeddings (*Predict*)

- Inspired by deep learning
- `word2vec` *(Mikolov et al., 2013)*
- GloVe *(Pennington et al., 2014)*

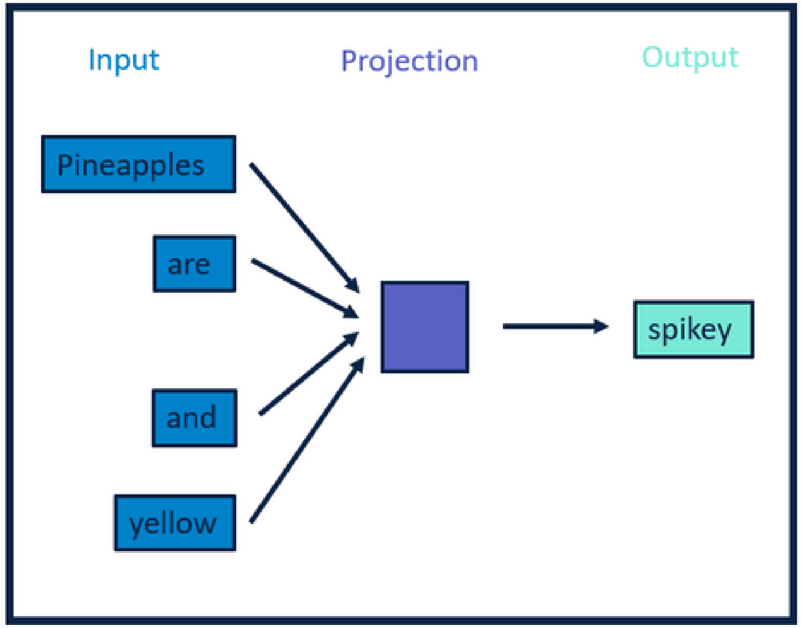
Underlying Theory: **The Distributional Hypothesis** *(Harris, '54; Firth, '57)*

"Similar words occur in similar contexts"

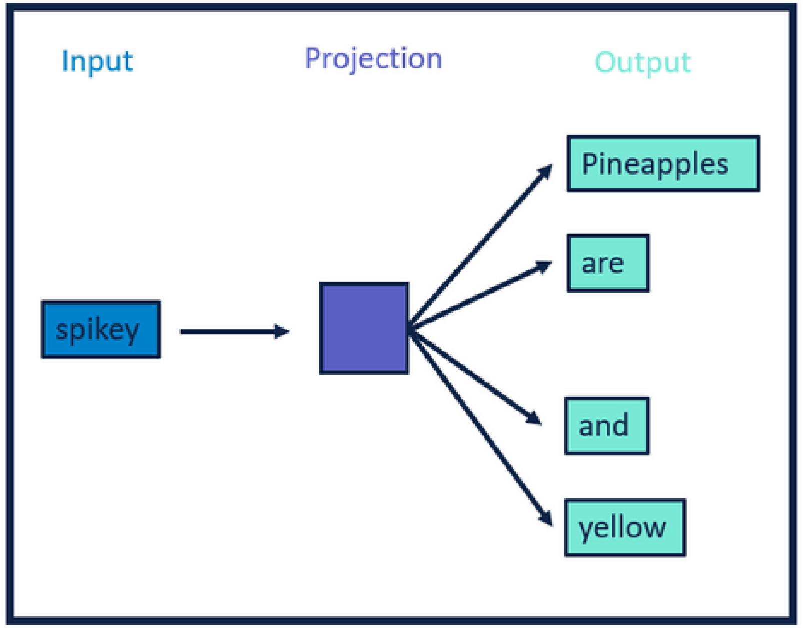
(Baroni et al, 2014) Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors – ACL
`https://aclanthology.org/P14-1023/`

THE TWO MODELS BEHIND WORD2VEC

Contextual Bag Of Word:
Predicts a target word based on context words.

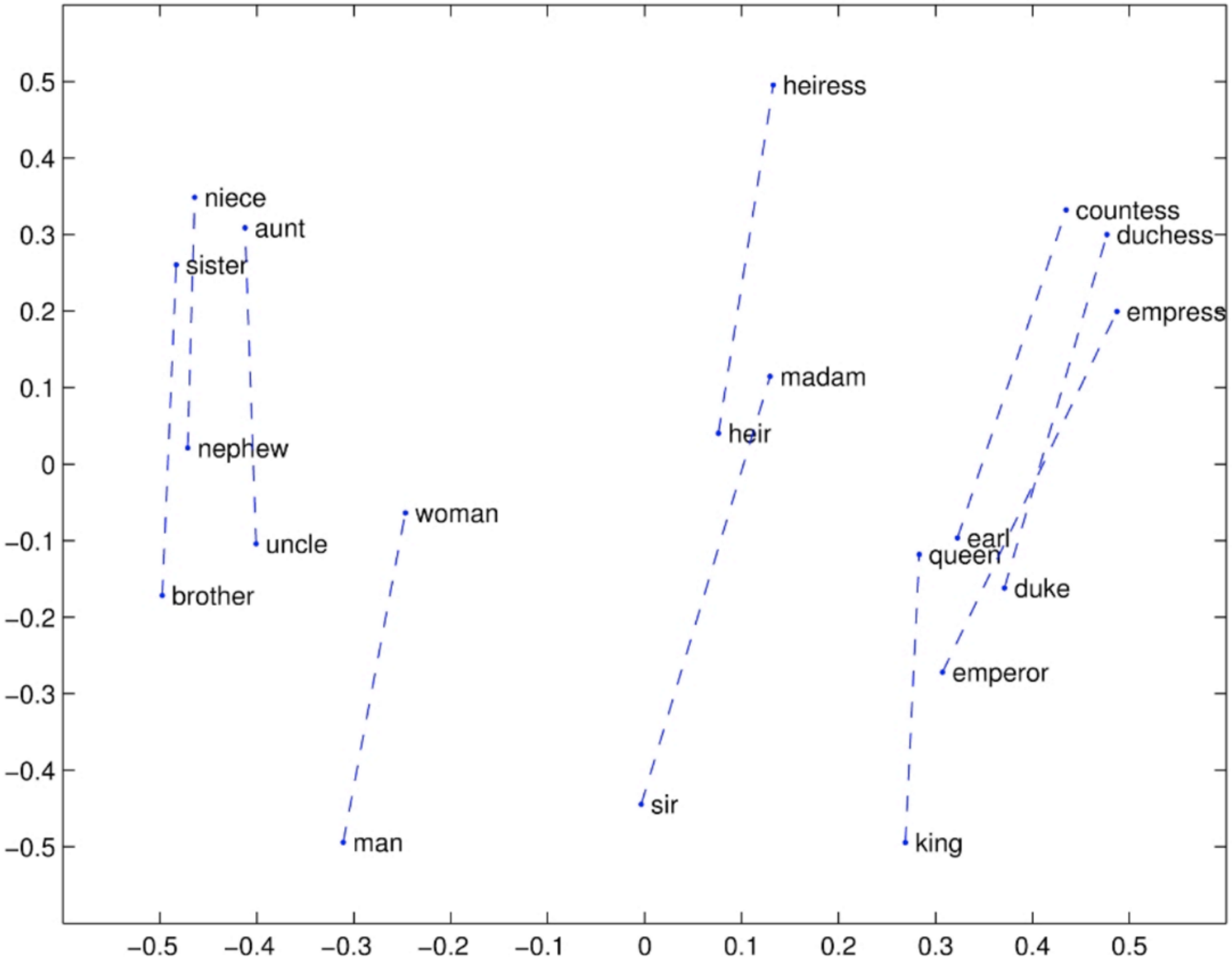


CBOW

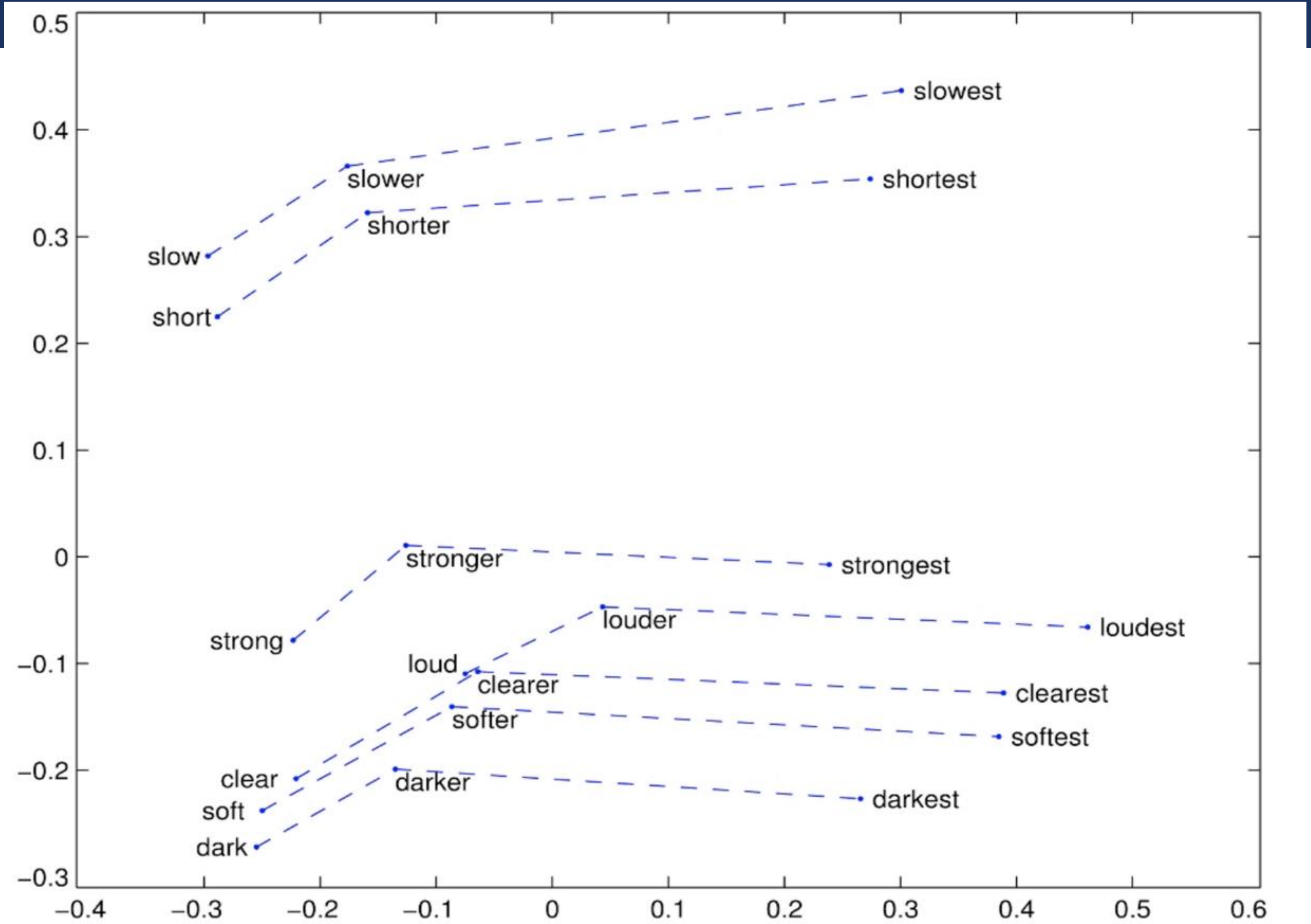


Skip-gram

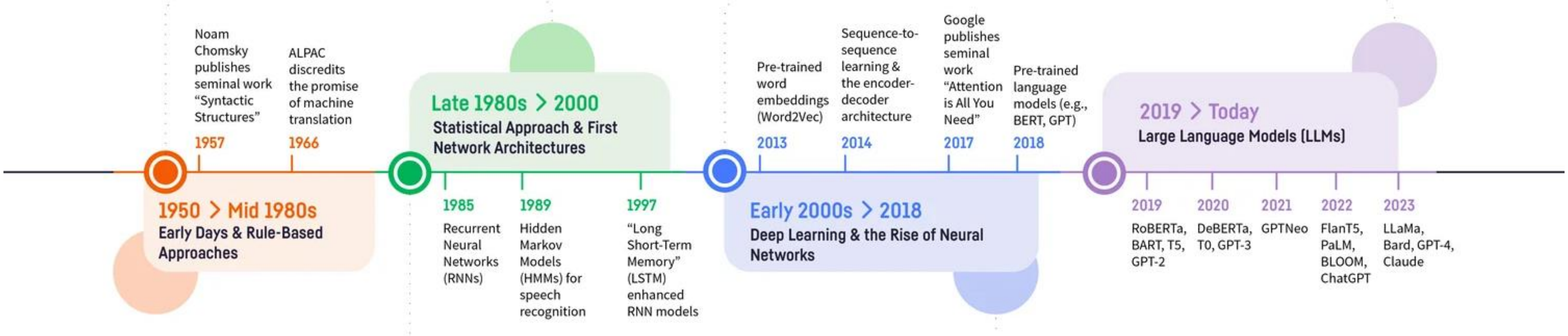
GEOMETRY AND MEANING …



GEOMETRY AND MEANING …



LANGUAGE MODELING



1950 > Mid 1980s
Early Days & Rule-Based Approaches

1957
Noam Chomsky publishes seminal work "Syntactic Structures"

1966
ALPAC discredits the promise of machine translation

Late 1980s > 2000
Statistical Approach & First Network Architectures

1985
Recurrent Neural Networks (RNNs)

1989
Hidden Markov Models (HMMs) for speech recognition

1997
"Long Short-Term Memory" (LSTM) enhanced RNN models

Early 2000s > 2018
Deep Learning & the Rise of Neural Networks

2013
Pre-trained word embeddings (Word2Vec)

2014
Sequence-to-sequence learning & the encoder-decoder architecture

2017
Google publishes seminal work "Attention is All You Need"

2018
Pre-trained language models (e.g., BERT, GPT)

2019 > Today
Large Language Models (LLMs)

2019
RoBERTa, BART, T5, GPT-2

2020
DeBERTa, T0, GPT-3

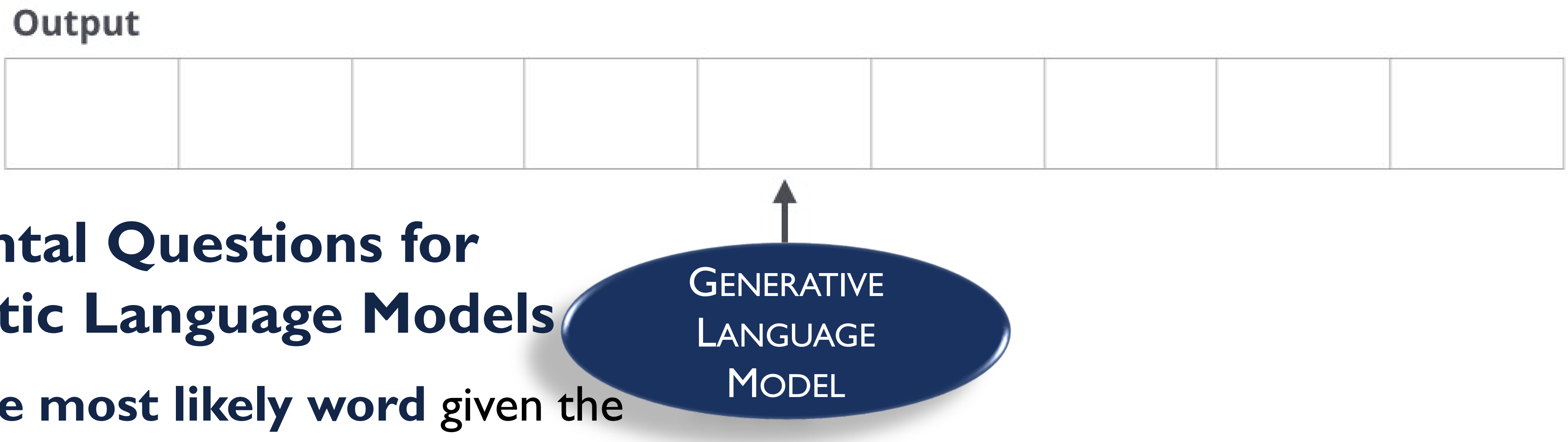
2021
GPTNeo

2022
FlanT5, PaLM, BLOOM, ChatGPT

2023
LLaMa, Bard, GPT-4, Claude

- Language Modeling:
  - Statistical approaches
  - Neural approaches to NL semantics

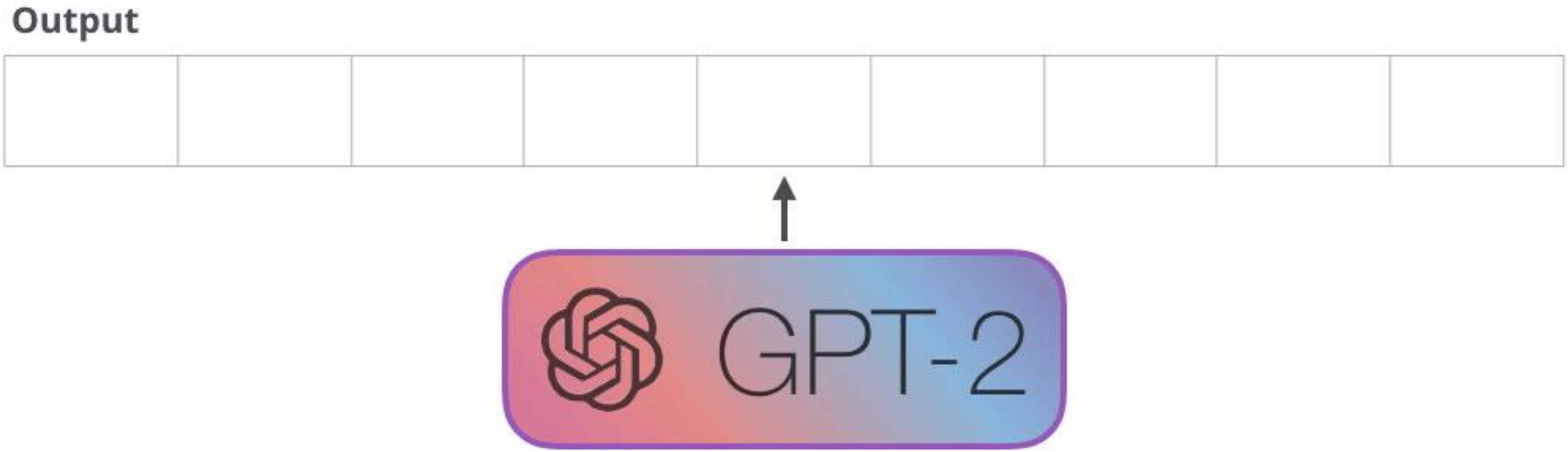
NATURAL LANGUAGE AS A MARKOV PROCESS



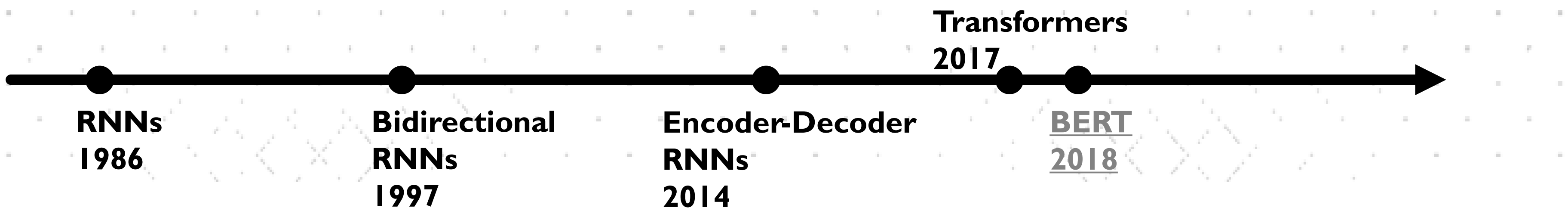
Fundamental Questions for Probabilistic Language Models

- What is **the most likely word** given the left most recent context?
- What is the **probabilty of an entire sentence**?
- What is the **most likely (inner/hidden) state sequence** given the (observable) sentence?

LANGUAGE MODELING AS A NEURAL DECODING



BERT: Encoding Natural Language Semantics through Trasformers

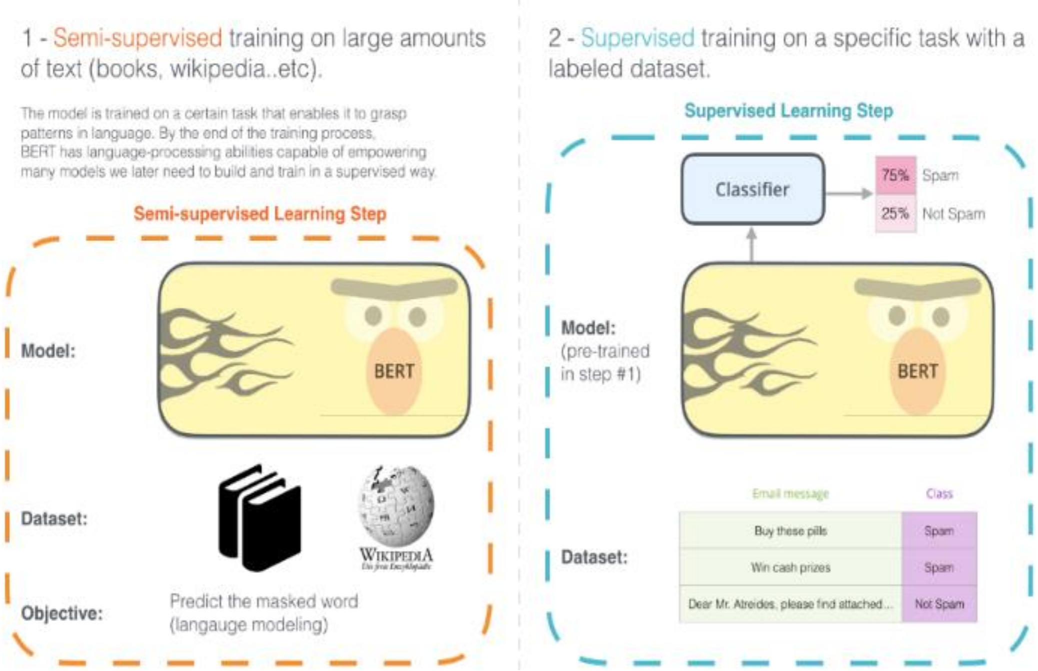

RNNs 1986

Bidirectional RNNs 1997

Encoder-Decoder RNNs 2014

Transformers 2017

BERT 2018


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:

Dataset:

Objective: Predict the masked word (langauge modeling)

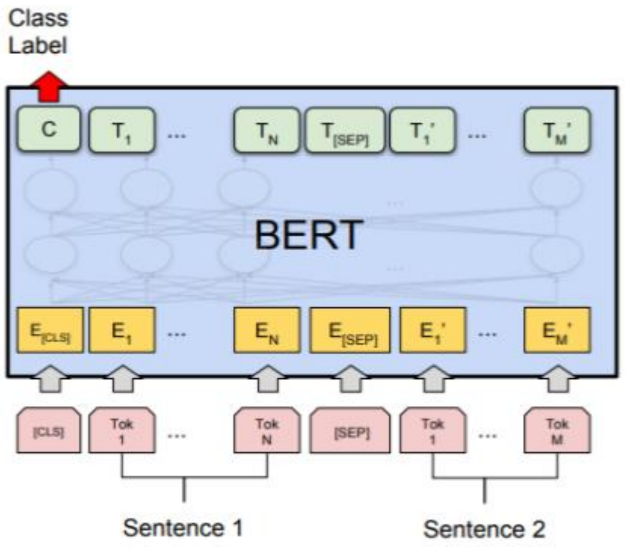
2 - Supervised training on a specific task with a labeled dataset.

Supervised Learning Step

Model: (pre-trained in step #1)

Dataset:

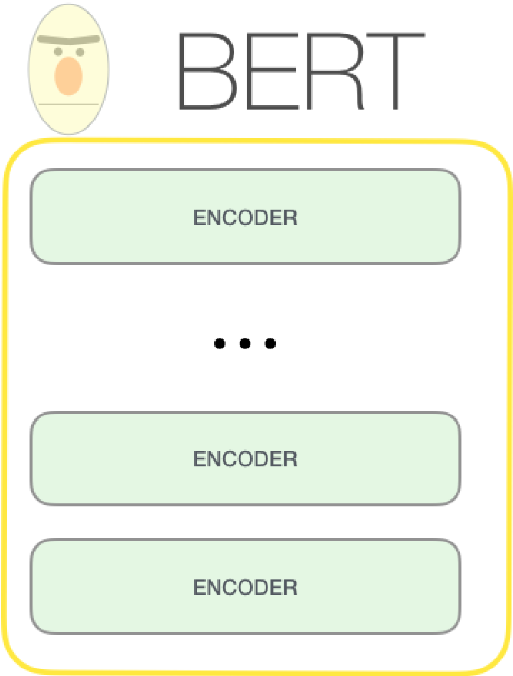
Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached...	Not Spam



BERT (DEVLIN ET AL, 2018)

Bidirectional Encoder Representations from Transformers

- Only the encoder is used
- Designed to generate **contextual meaningful representation** of input words
  - Representations are **context sensitive,** thanks to self-attention
  - Understand the context of a word in a sentence from **both left and right sides** (bidirectionally).
- Representations are embeddings
  - not suitable for text generation
  - … but for many other tasks



Images from `https://jalammar.github.io/illustrated-bert/`

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.

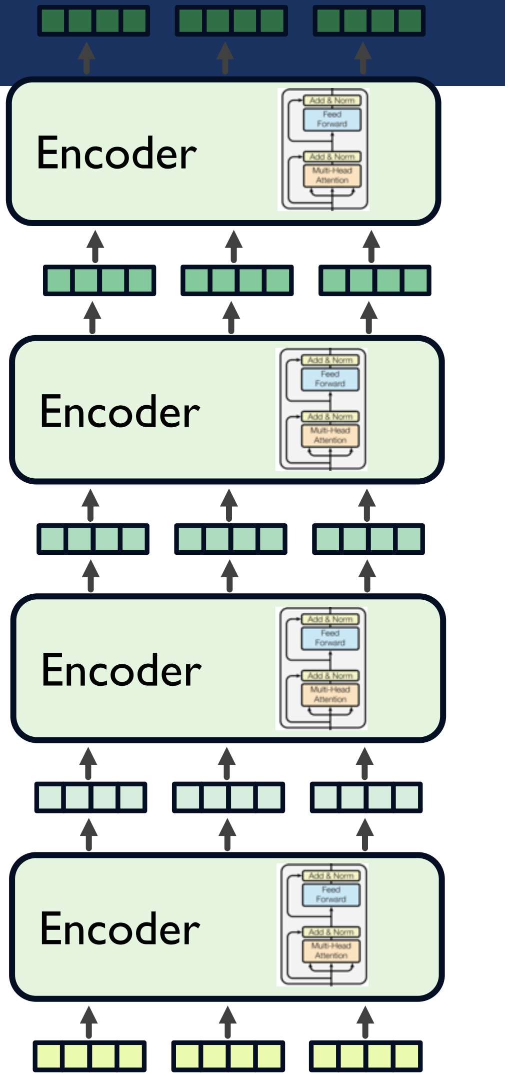
BERT (DEVLIN ET AL, 2018)

Why should it work?

- It is just a piece of the Transformer architecture (next in few slides) ago.

The GREAT IDEA: Pre-Training the encoder

- Pre-trained on a large corpus of text and then fine-tuned for specific tasks like question answering, sentiment analysis, etc.



Images from `https://jalammar.github.io/illustrated-bert/`

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.

NO PRE-TRAINING NO PARTY!
THE REVOLUTION OF PRE-TRAINING IN NLP

- **Simple idea:** train a (possibly large) model on a different task and re-use it on your task
  - circumventing the need for training from scratch
  - facilitating "quicker", more effective deployment of the model

- **Precedent in Computer Vision**:
  - This strategy mirrors developments in computer vision
  - Architectures pre-trained on classification tasks using datasets like ImageNet
  - When applied on related task, these "starting point" achieve very good results

- **Addressing Overfitting in Large Models**:
  - With **increasing model sizes** and parameter counts, the **risk of overfitting grows**
  - Pre-training on vast datasets mitigates this by providing a broad learning base.

TOWARDS FOUNDATION MODELS

- **Emergence of Foundation Models in NLP**:
  - Large-scale models trained on linguistic tasks, forming a versatile base that can be fine-tuned for various specific applications.
- **Everybody worked on customizing Foundation Models:**
  - Leverage the extensive knowledge encapsulated in Foundation Models by fine-tuning them for particular NLP tasks.

- If you are interested in foundation models
  - [Zhou et al, 2023] A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT
  - `https://arxiv.org/abs/2302.09419`

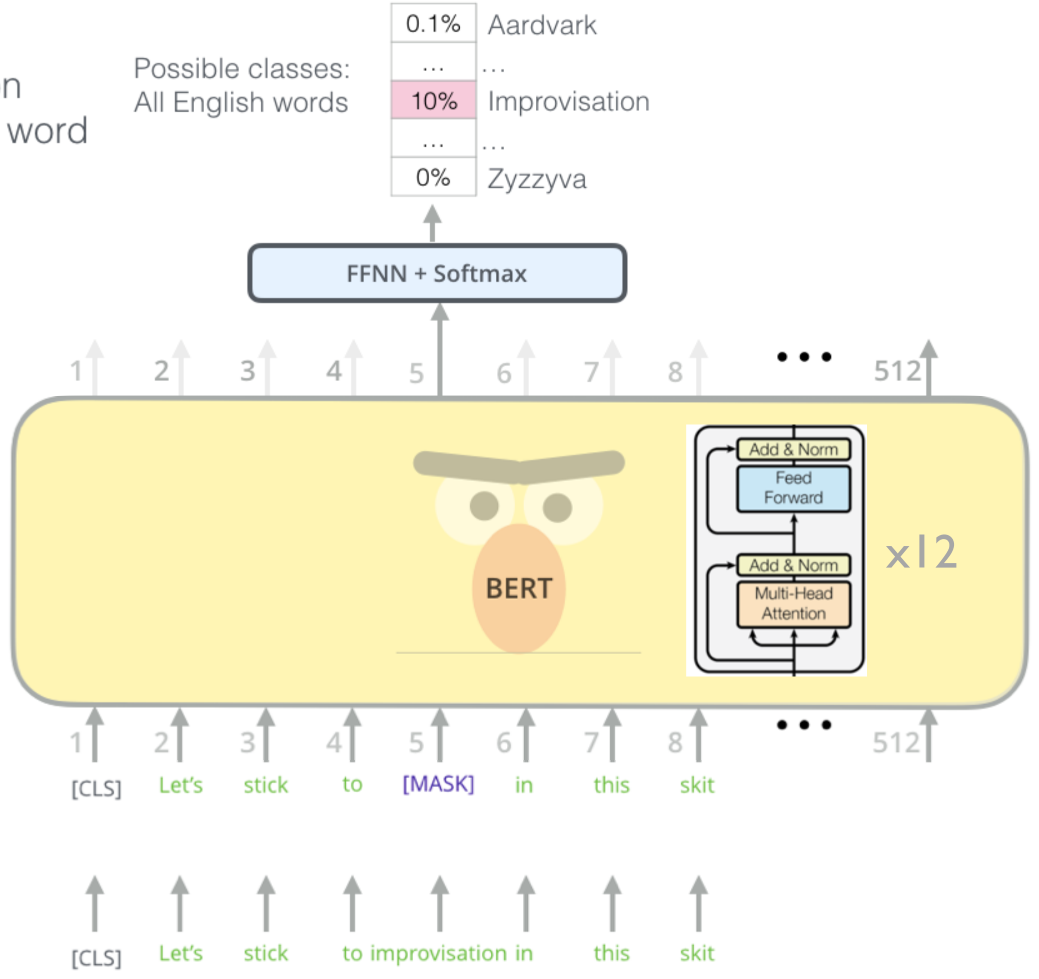
PRETRAINING BERT

- BERT takes a sequence of tokens as input
  - Utilizes **self-attention across layers** to **generate context-aware** representations of each token in the sequence.
  - In each layer, h=12 $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$ matrices
- **Pre-training tasks:**
  - **Masked-language modeling**

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

Randomly mask 15% of tokens

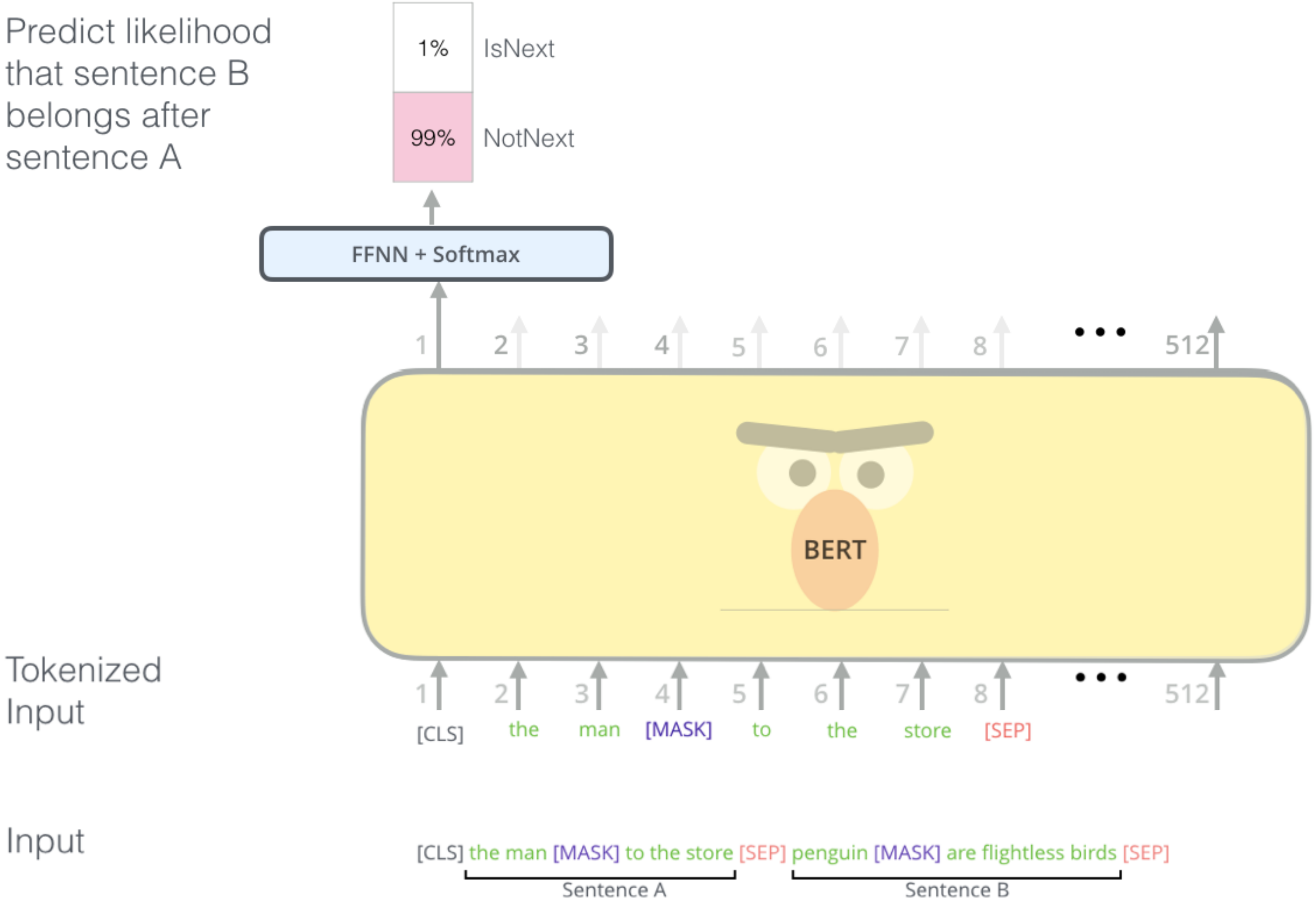
Input



PRETRAINING BERT (2)

- BERT takes a sequence of tokens as input
  - Utilizes **self-attention across layers** to **generate context-aware** representations of each token in the sequence.
  - In each layer, h=12 $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$ matrices
- **Pre-training tasks:**
  - Masked-language modeling
  - **Next sentence prediction**

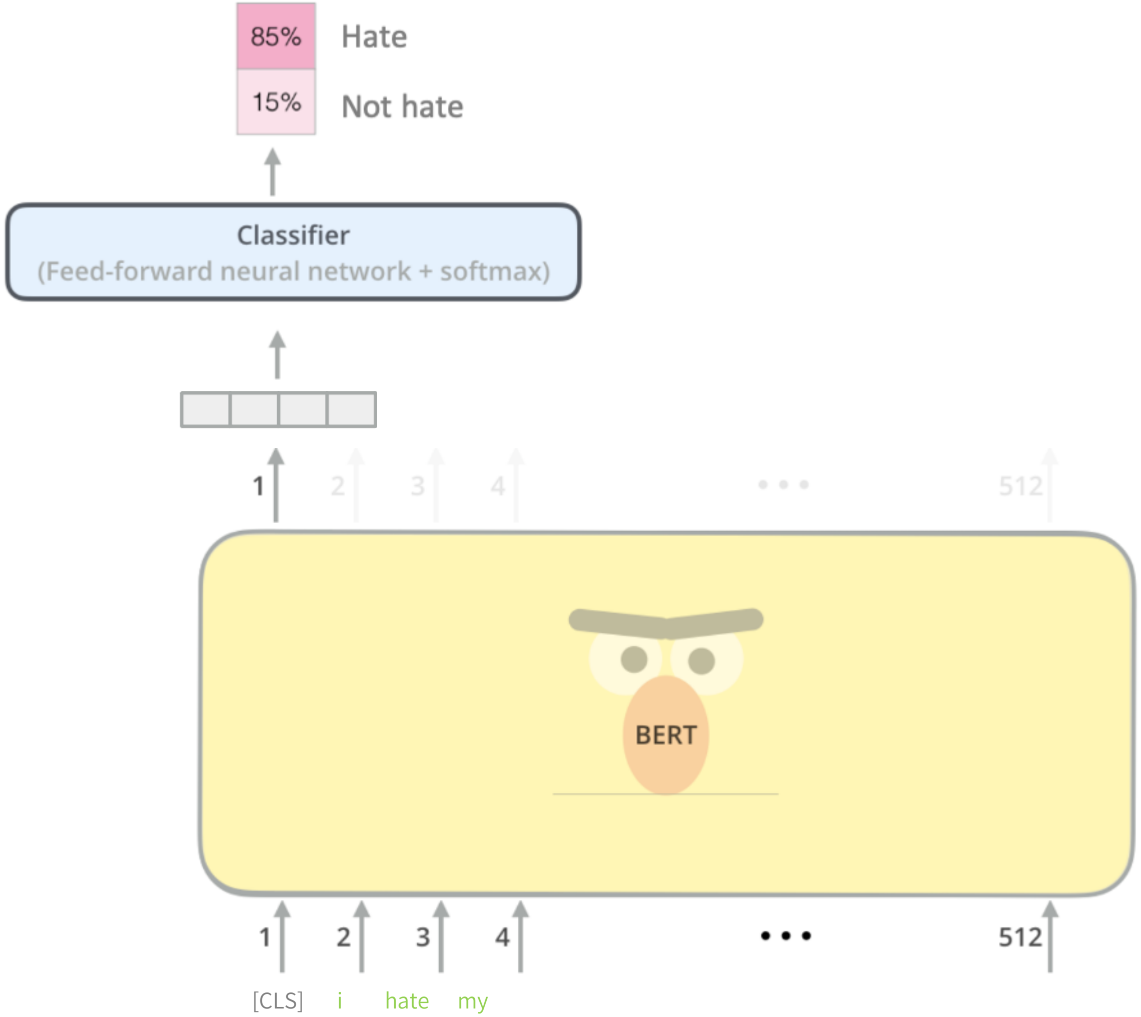
Pretrained using the Toronto BookCorpus (800M words) and English Wikipedia (2,500M words)



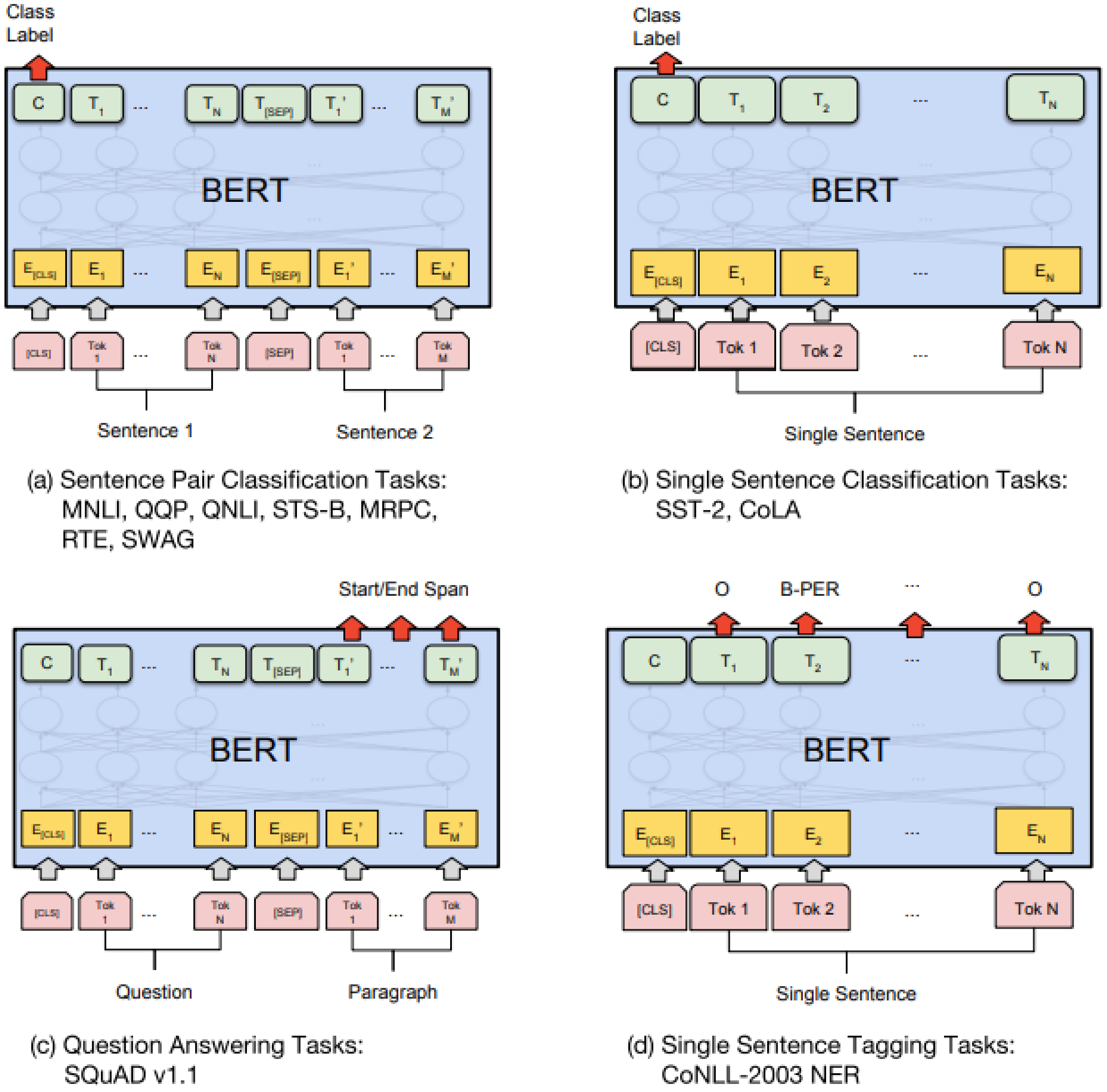
Predict likelihood that sentence B belongs after sentence A
1% IsNext
99% NotNext
FFNN + Softmax
1 2 3 4 5 6 7 8 ••• 512
BERT
Tokenized Input
1 2 3 4 5 6 7 8 ••• 512
[CLS] the man [MASK] to the store [SEP]
Input
[CLS] the man [MASK] to the store [SEP] penguin [MASK] are flightless birds [SEP]
Sentence A
Sentence B

BERT AND FINE-TUNING

- Once pretrained, we can **apply it to new sentences**
- BERT will **produce encoded representations** for each input symbol
- And it can be used in **different classification tasks**, just adding a new (linear) classifier…
- … through fine-tuning of the entire architecture
- not trivial to forget what learned during the pre-training



Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.



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

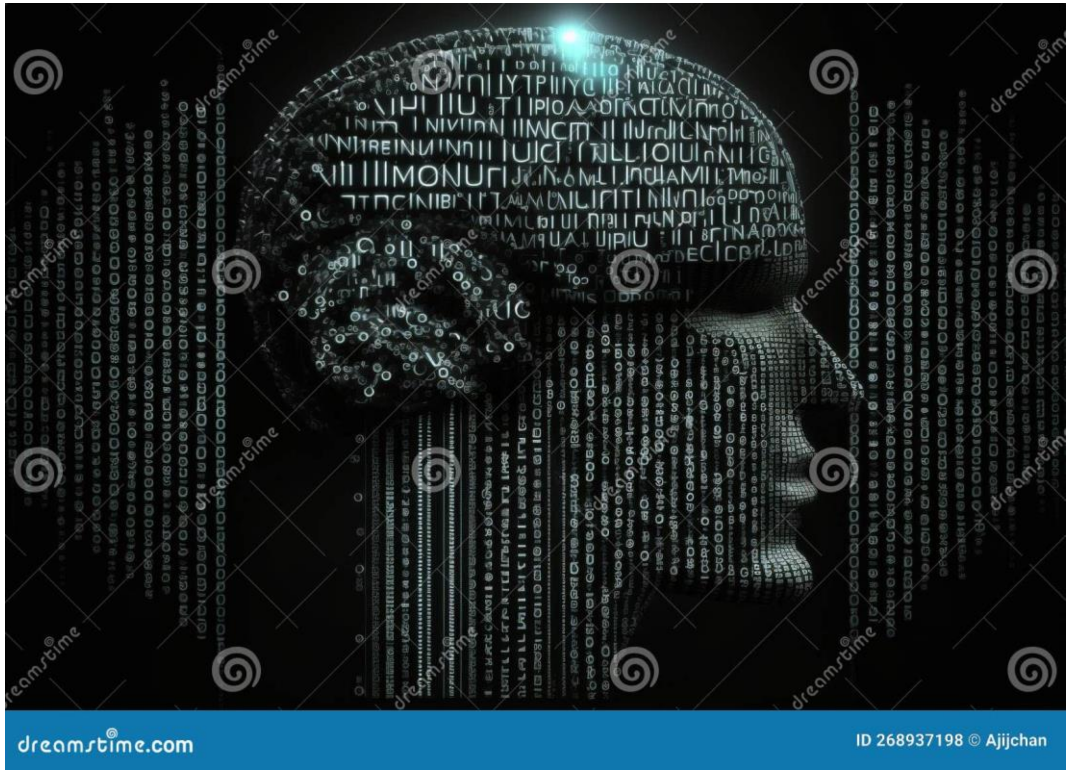
(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

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

LANGUAGE MODELING AND REASONING

- Logical Entailment: the axiomatic «logical» view
- Training Automatic Entailment systems
  - From formal logic to NL
  - Recognizing Textual Entailment as a learning modality without any training example
- Applied Textual Entailment
  - Classification of Sentence Pairs as a new task
  - New Task description as Prompting
- Prompting Applications



ENTAILMENT: THE «LOGICAL» VIEW

- Logical implication is used to express the **entailment relationship** between two subformulas

$$A \rightarrow B \qquad\qquad \forall x\, A(x) \rightarrow B(x)$$

- Logics helps in expressing **logical reasoning schemata** through normalized forms, e.g.,

$$A \rightarrow B \equiv \neg A \vee B \qquad \forall x\, A(x) \rightarrow B(x) \equiv \neg A(e) \vee B(e)$$ (after Skolemization)

- or equivalent variants

$$A \rightarrow B \equiv \neg(A \wedge \neg B) \qquad \forall x\, A(x) \rightarrow B(x) \equiv \forall x\, \neg(A(x) \wedge \neg B(x))$$

ENTAILMENT: SEMANTICS

- Logical implication is tightly related to **semantics,** as it is the basis for an efficent approach to logical reasoning.
- Infact $\{A\} \vDash B$ iff $\{\} \vDash (A \rightarrow B)$ (Worlds where A is true also make B true, i.e. $A \rightarrow B$ is a tautology)

- B is semantically implied by A (only) if $(A \rightarrow B)$ is a tautology. This is used for the algorithms based on **proof by contradiction**, i.e.,

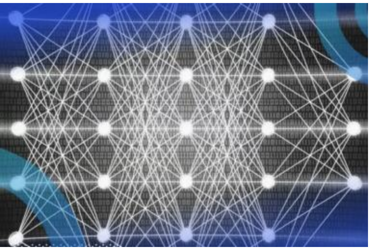
$\{A\} \vDash B$ iff $\{A, \neg B\} \vDash \bot$ or (with $\bot$ denoting the always false formula)

$\{\Delta, A\} \vDash B$ iff $\{\Delta, A, \neg B\} \vDash \bot$

HOW TO DECIDE ABOUT ENTAILMENT THROUGH TRANSFOMERS

- Logical implication (such as $\{A\} \vDash B$) is usually managed through **a chain of deductive steps** (as in logic programming) from the input query (i.e. a theorem to be demonstrated) to its fully resolved facts, or through contadictions
- Limitations: not formal treatment of uncertainty, poor coverage (the axiomatic system Δ is not fully known a priori), pervasive complexity within large knowledge bases.

- **Neural Networks can be adopted to limit the impact of incompleteness or noise in the reference rules and minimze the risk of mistakes in the entailment decision**.
  - **LANGUAGE KNOWLEDGE** allows to employ linguistic semantics for approximating logical deductions
  - The **deduction chain** can be successul or not: this implies that **the entire inference can be mapped into a BINARY CLASSIFICATION TASK**
  - The input correspond to a pair A and B of the sentenced corresponding respectively to the hypothesys (A) and the sentence corresponding to the thesis (B)



ENTAILMENT & TRANSFOMERS

A possible process is

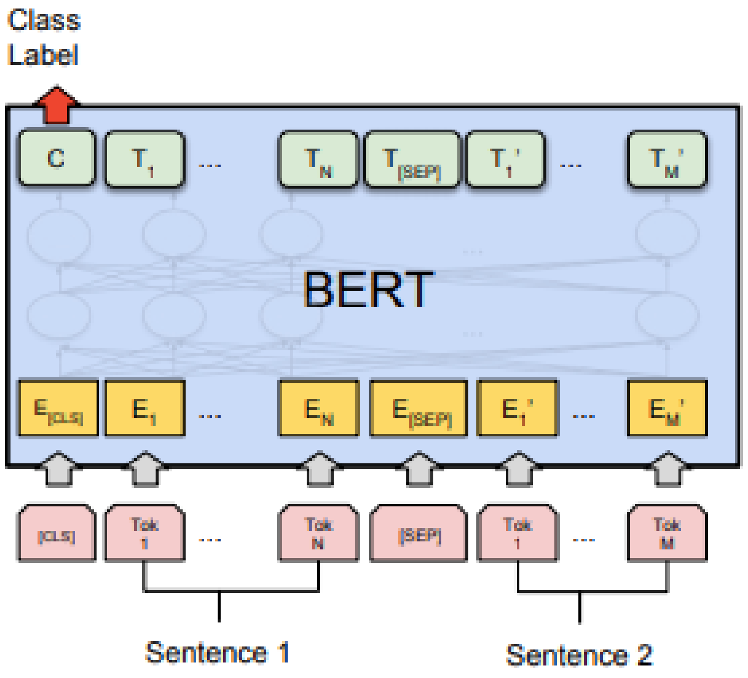
- Map the logical rules (as axioms) into a training dataset
- Map a new potential theorem into a natural language sentence
- Make the sentence the input of a NNs
- Solve the inference task of accepting/rejecting the entailment as a **binary classification task**

In other words, given a training set of axioms such as

- Δ: $\{A_1 \rightarrow B_1, \dots, A_n \rightarrow B_n\}$
- Induce a function RTE such that for every future pair (A_i, B_j)
  - $h(A_i, B_j) = true$ iff $\{\Delta, A_i\} \vDash B_j$ or alternatively $h(A_i \rightarrow B_j) = true$ iff $\{\Delta, A_i\} \vDash B_j$

THE ROLE OF TRASFORMERS

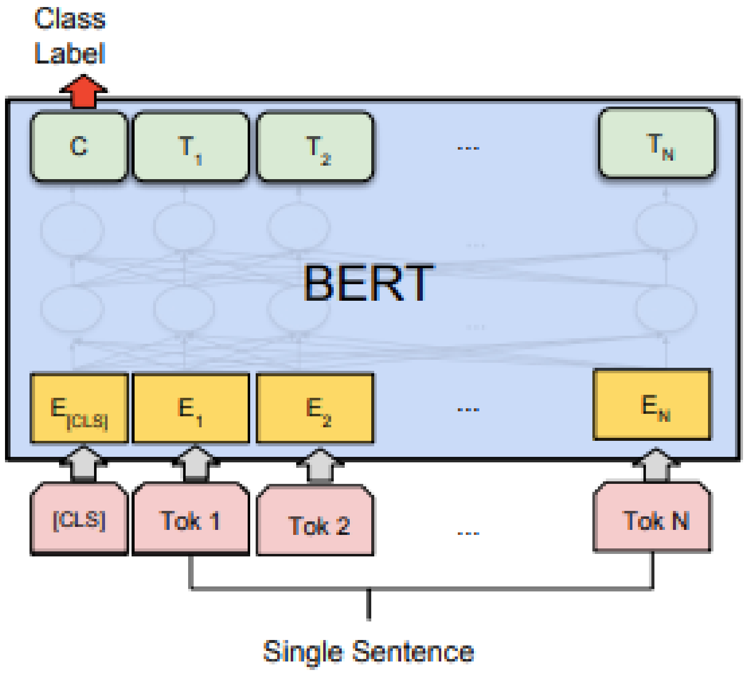
- First setting
  - $h(A_i, B_j) = true$ iff $\{\Delta, A_i\} \Vdash B_j$
  - Input given by 2 sentences
  - BERT used as the encoder
  - A stacked classifier is trained on labeled pairs
  - Type of Inference:
    - PARAPHRASING
    - TEXTUAL ENTAILMENT



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

THE ROLE OF TRASFORMERS (2)

- Second setting
  - $h(A_i \rightarrow B_j) = true$ iff $\{\Delta, A_i\} \Vdash B_j$
  - Input given 1 sentence expressing the task over A_i and B_j
  - BERT used as the encoder
  - A stacked classifier is trained on labeled pairs
  - Example (PARAPHRASING):
  - «The sentence B_j has the same meaning of sentence A_i»
  - «Sentence A_i means the same as B_j»

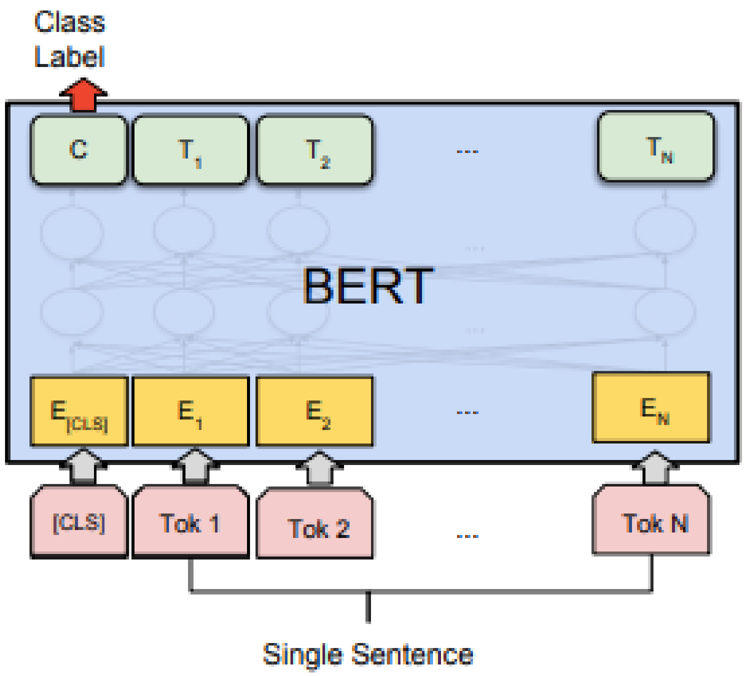


(b) Single Sentence Classification Tasks: SST-2, CoLA

THE ROLE OF TRASFORMERS (3)

- Second setting
  - $h(A_i \rightarrow B_j) = true$ iff $\{\Delta, A_i\} \Vdash B_j$
  - Input given 1 sentence expressing the task over A_i and B_j
  - BERT used as the encoder
  - A stacked classifier is trained on labeled pairs

  - Example (TEXTUAL ENTAILMENT):
  - «The sentence B_j is implied by sentence A_i»
  - «Sentence A_i guarantees the truth of B_j»



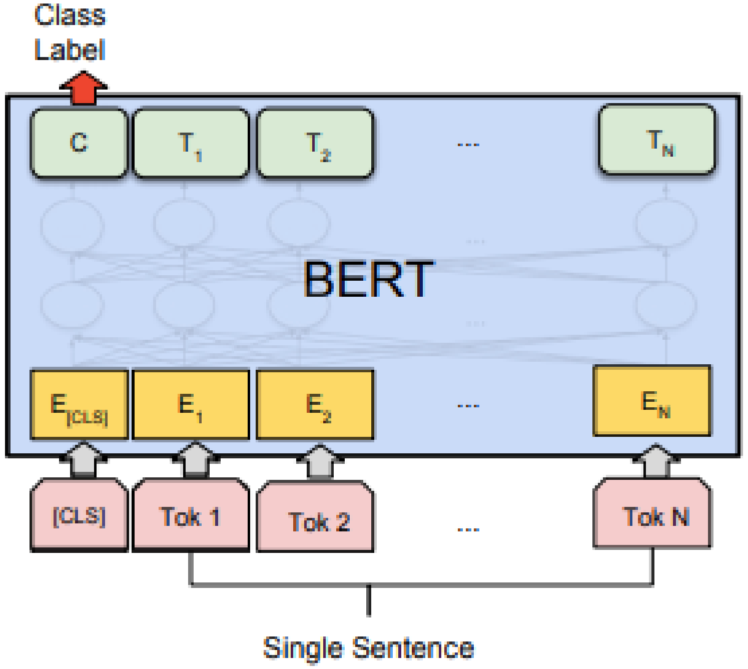
(b) Single Sentence Classification Tasks: SST-2, CoLA

NEURAL ENTAILMENT: APPLICATIONS

- The setting

$$h(A_i \rightarrow B_j) = true \text{ iff } \{\Delta, A_i\} \Vdash B_j$$

- correspond to sentences that depend on complex interactions between A_i and B_j mapped into an individual sentences
  - **BERT can be always used as the encoder**
  - The **stacked classifier is an automatic entailment recognition tool**
- Future TEXTUAL ENTAILMENT tasks, e.g., :
  - TOPICAL CLASSIFICATION
    - «The sentence B_j is classified by label A_i», «Label A_i corresponds to the topic of B_j»
  - SENTIMENT ANALYSIS:
    - «A_i implies the sentiment label B_j», «A_i expresses sentiment B_j»



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

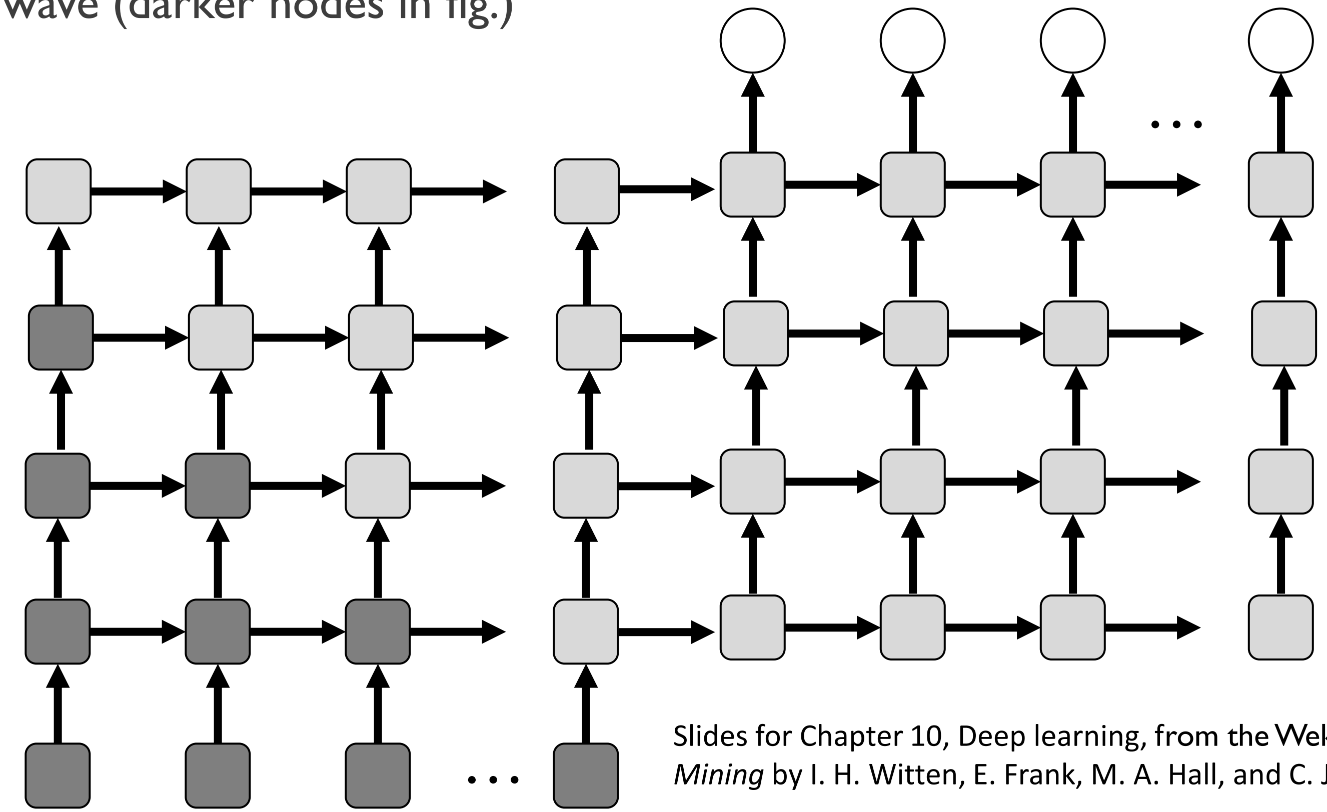
RETI NEURALI AVANZATE: INTERNALS

LA ATTENZIONE ED I TRANSFORMERS

METODI E ARCHITETTURE

ENCODER-DECODER DEEP ARCHITECTURES

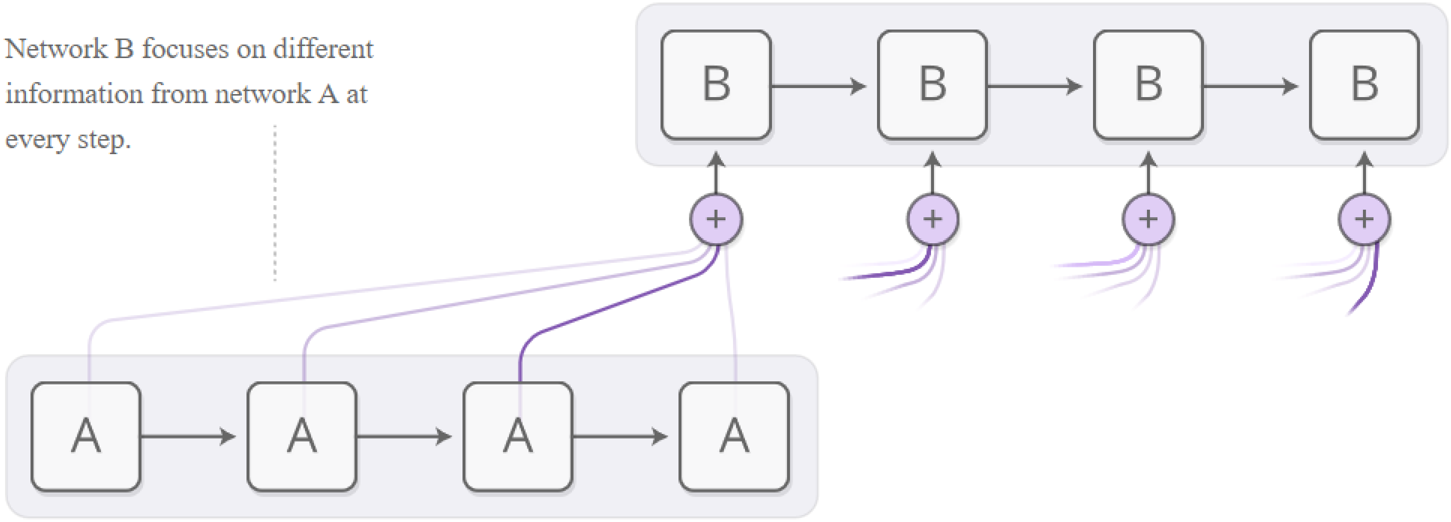
- Given enough data, a deep encoder-decoder architecture (see below) can yield results that compete with hand-engineered translation systems.
- The connectivity structure means that partial computations in the model can flow through the graph in a wave (darker nodes in fig.)



Slides for Chapter 10, Deep learning, from the Weka book, *Data Mining* by I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal

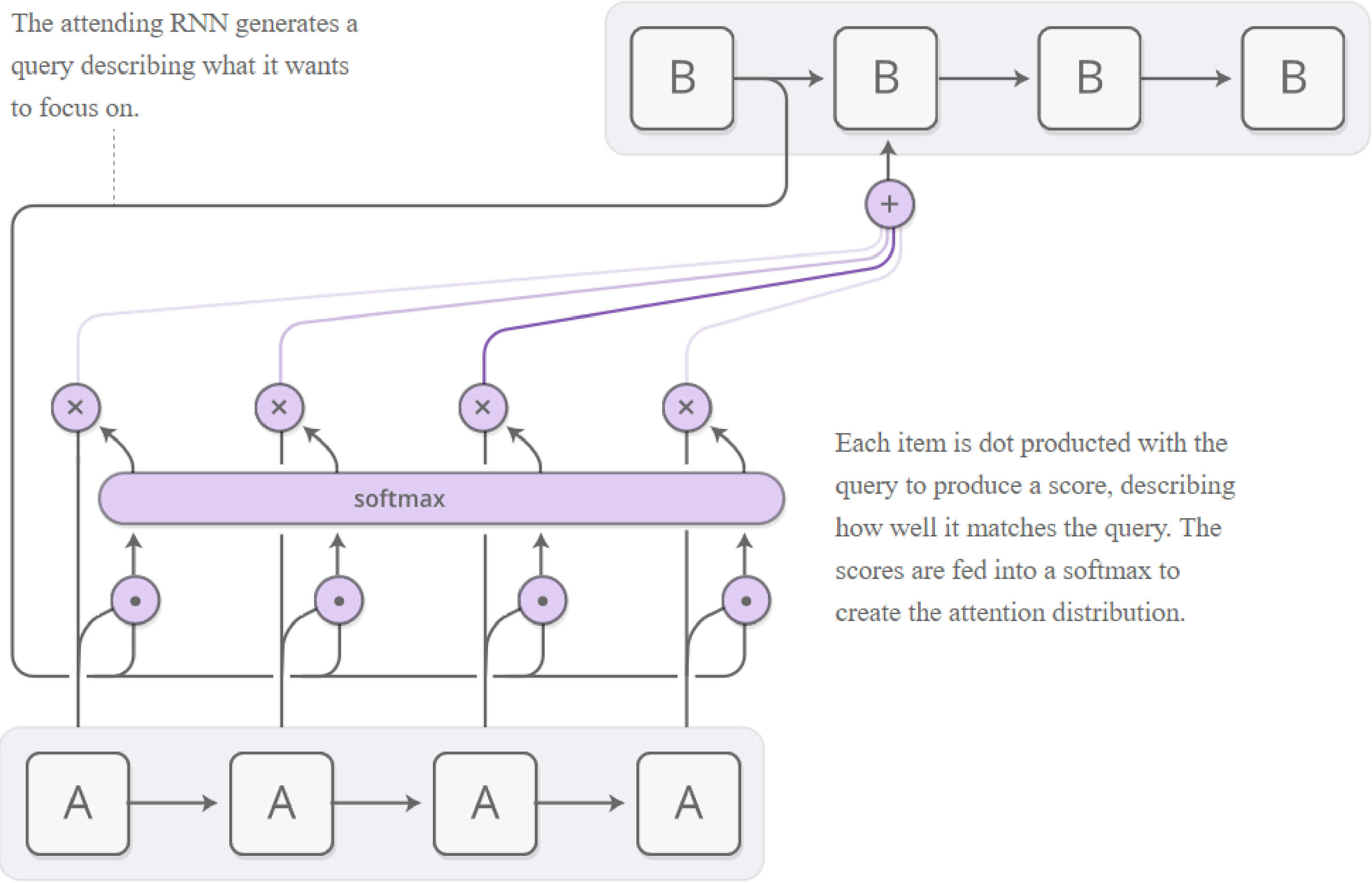
ATTENTION-BASED RNNS

- A NN (e.g. B) is used to attend the outcome of a second network A, e.g. (Vaswani et al., 2017)



Network B focuses on different information from network A at every step.

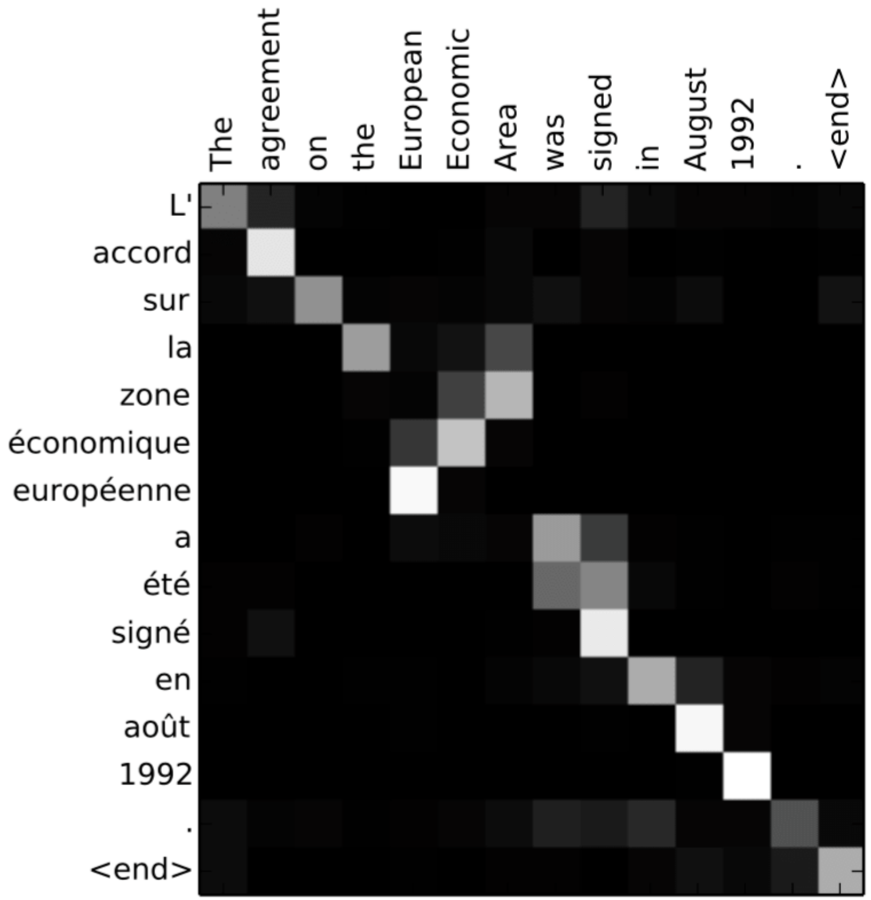
ATTENTION-BASED RNNS

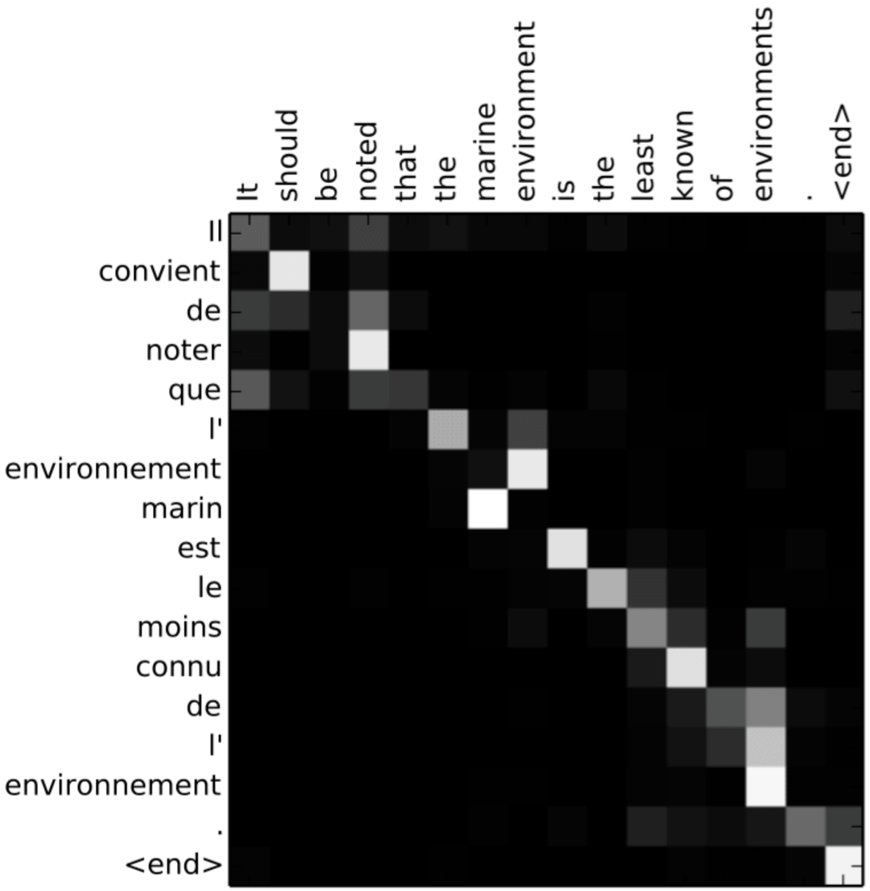


The attending RNN generates a query describing what it wants to focus on.

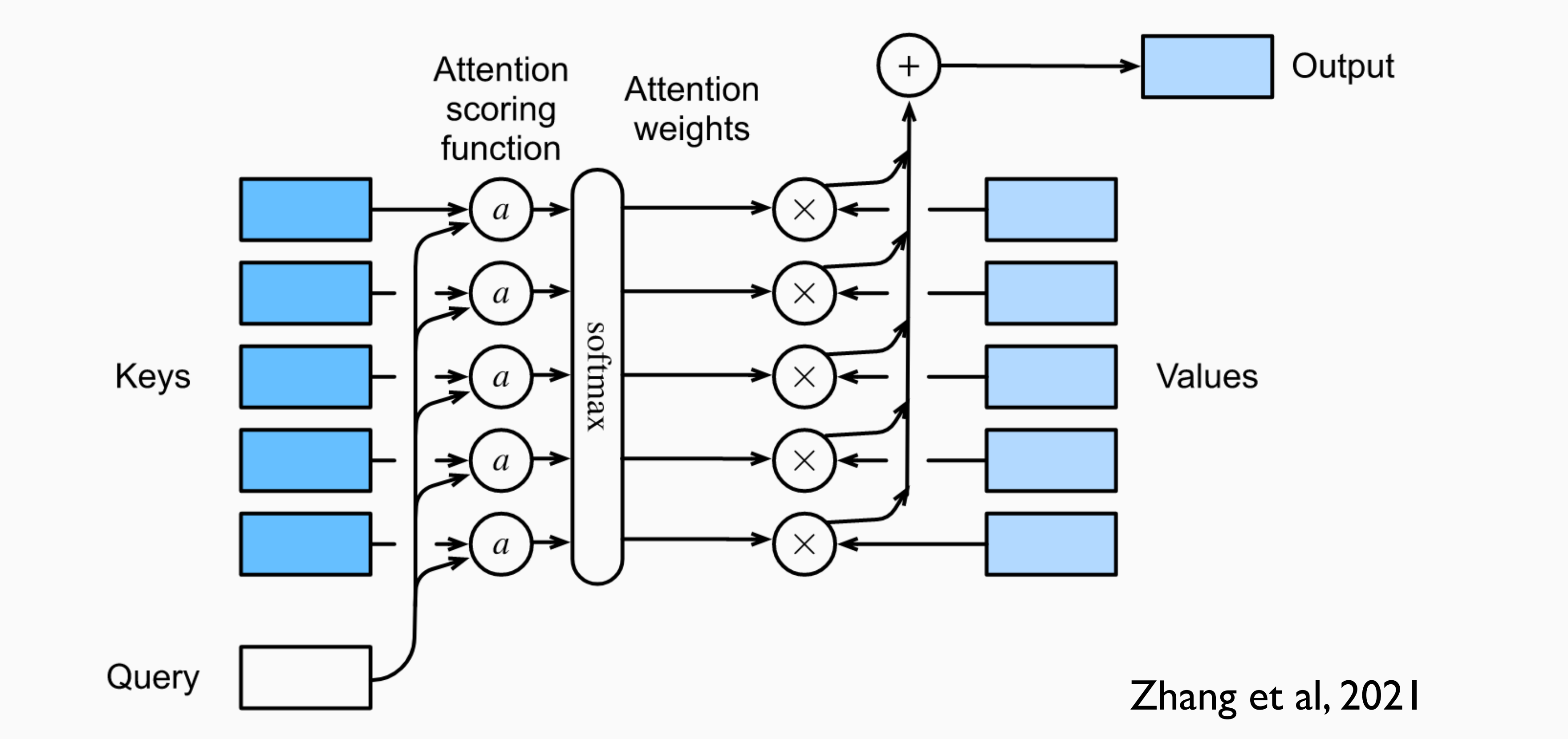
Each item is dot producted with the query to produce a score, describing how well it matches the query. The scores are fed into a softmax to create the attention distribution.

ATTENTION IN MACHINE TRANSLATION



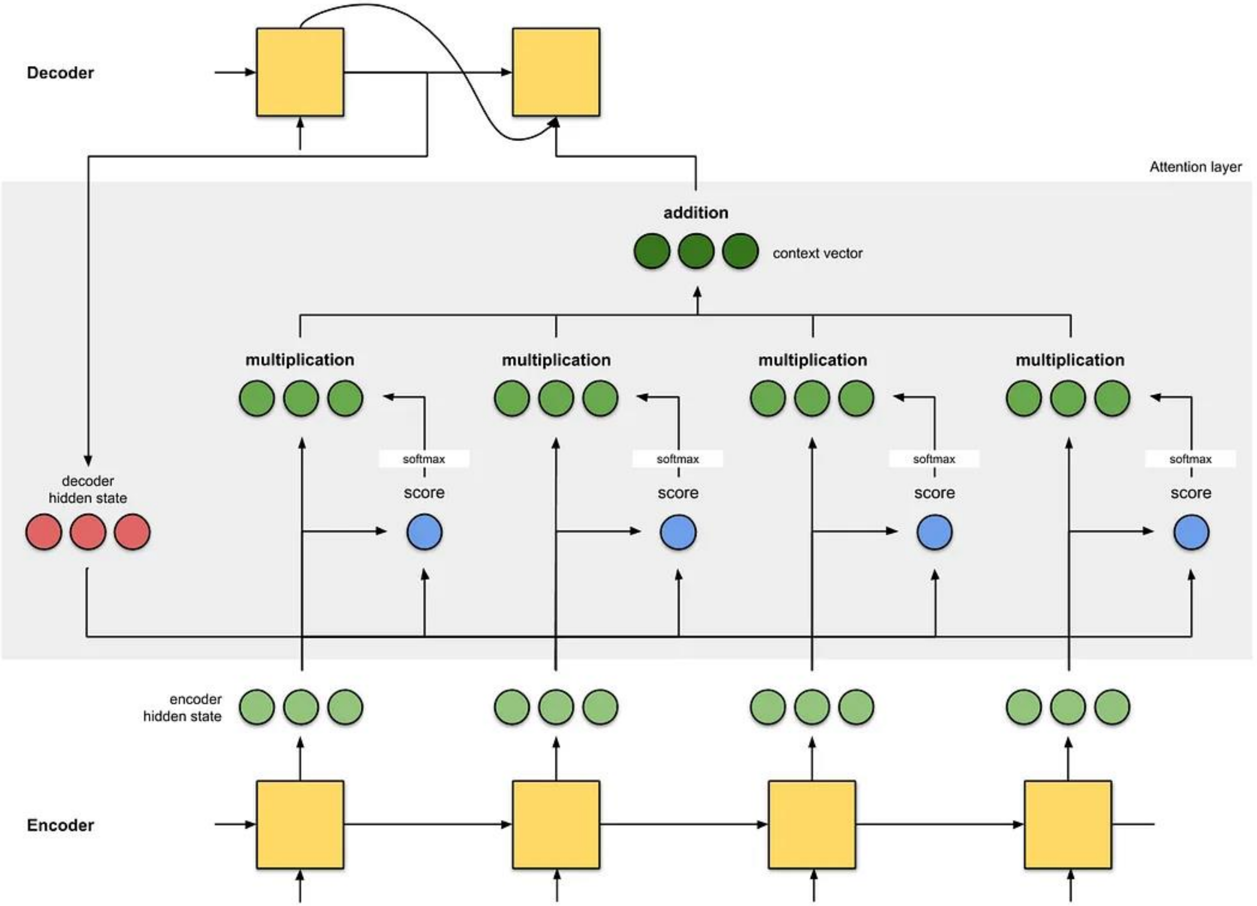


ATTENTION FUNCTIONS



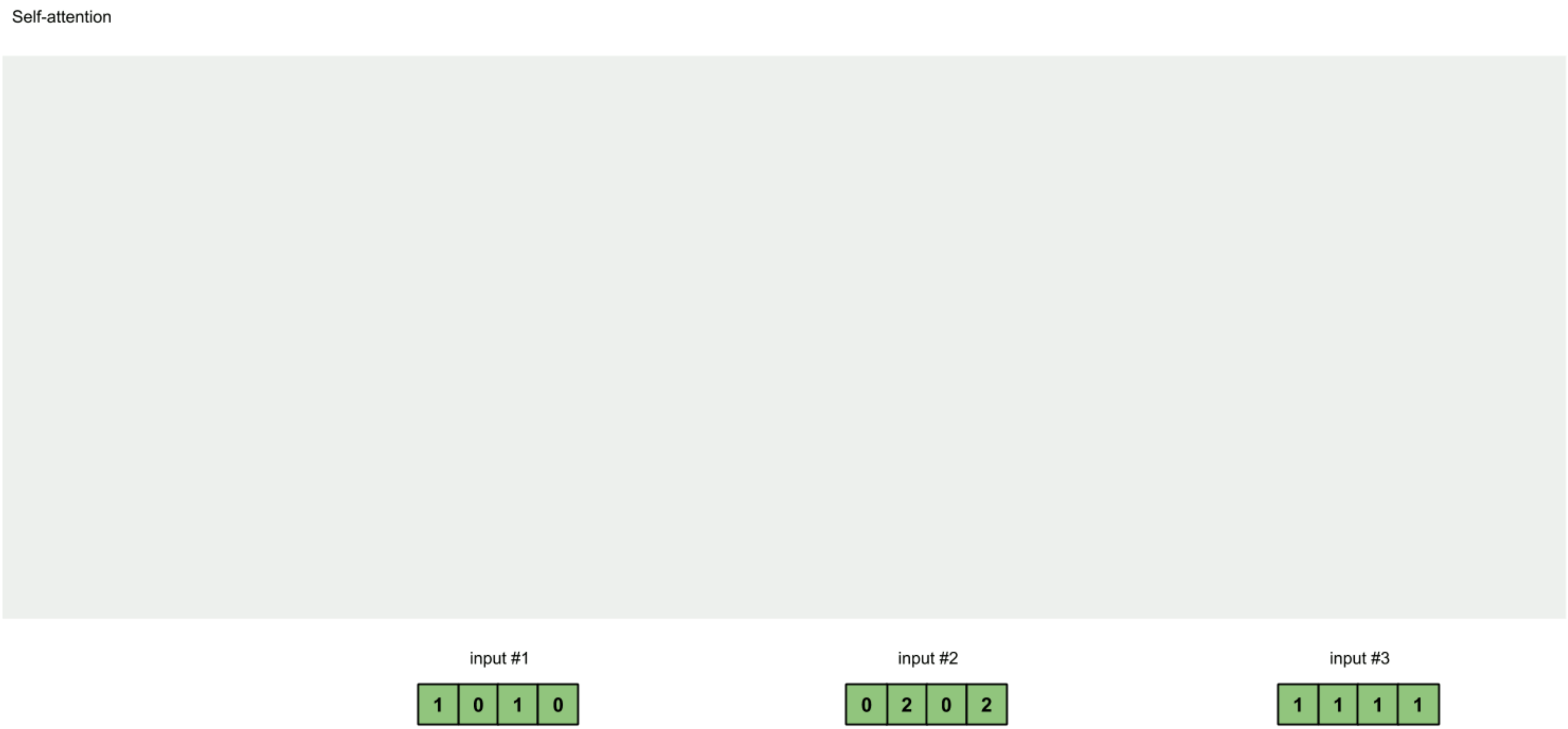
Zhang et al, 2021

ATTENTION IN SEQ2SEQ MODELS



by Manuel Romero: from *Attn: Illustrated Attention*, by Raimi Karim, Towards Data Science, Jan 20, 2019

SELF-ATTENTION



1	0	1	0

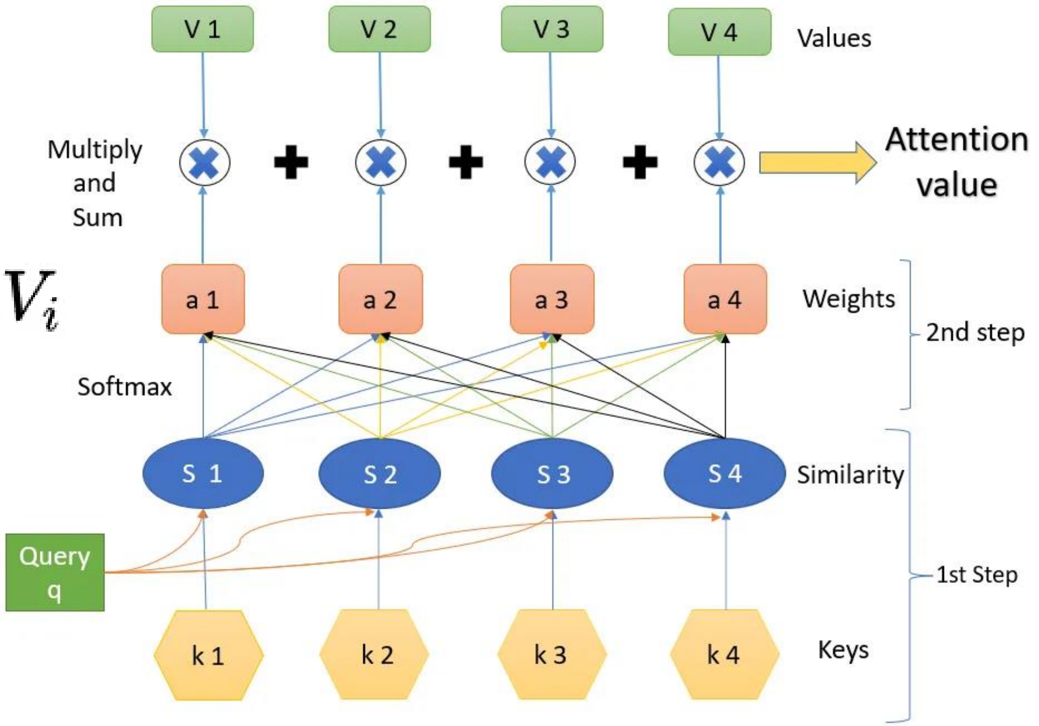
0	2	0	2

1	1	1	1

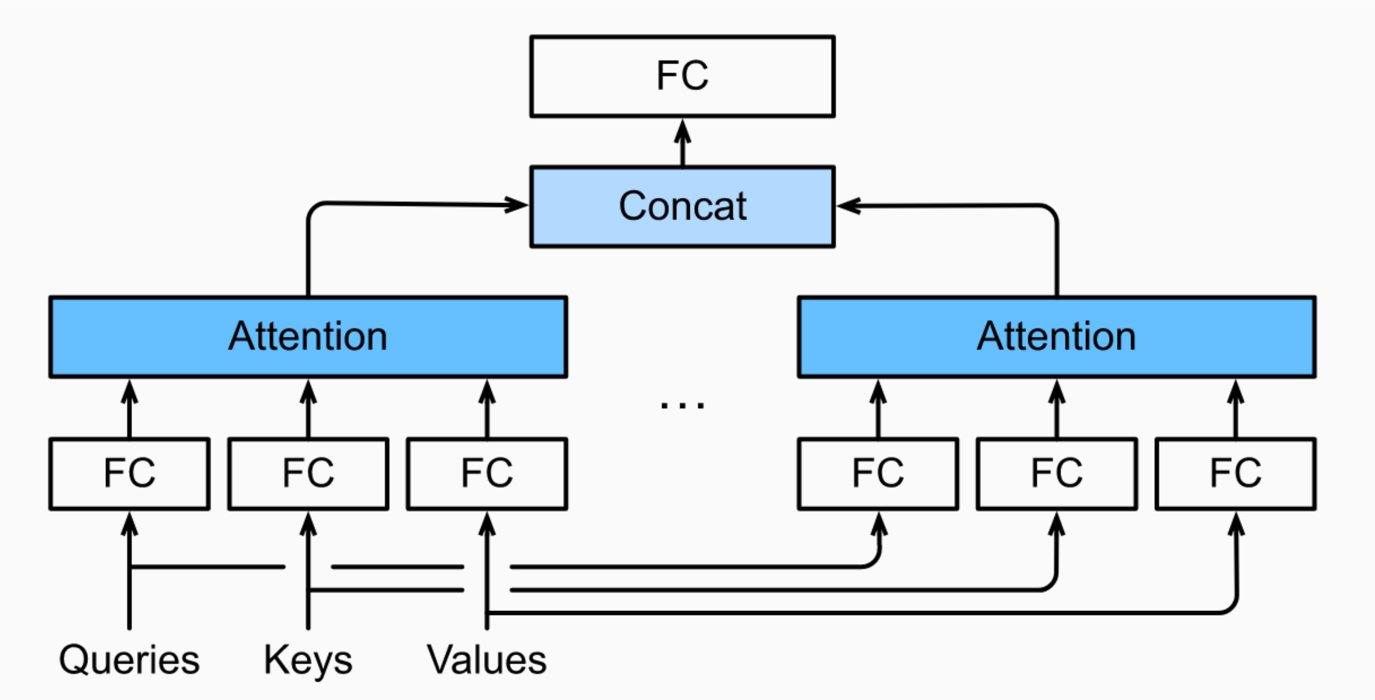
From https://colab.research.google.com/drive/1rPk3ohrmVclqhH7uQ7qys4oznDdAhpzF by Manuel Romero

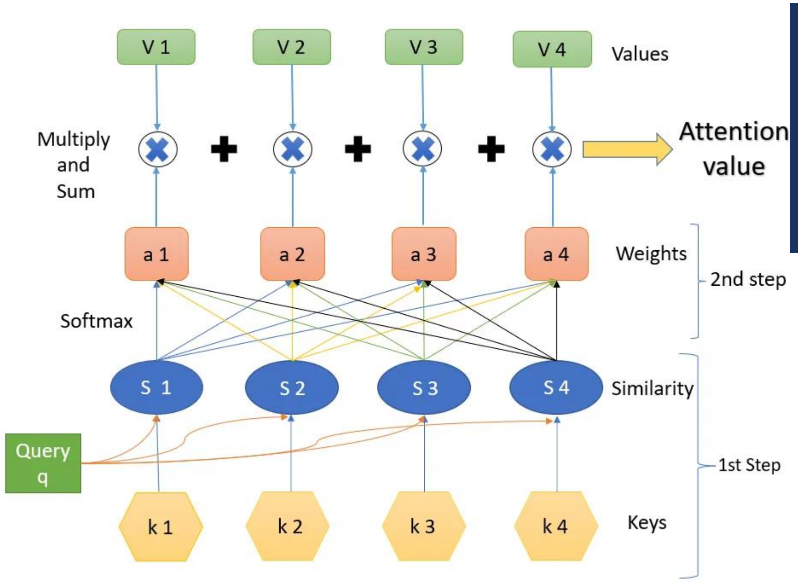
THE ATTENTION INFORMATION FLOW

$$attention\ value = \sum_i a_i V_i$$

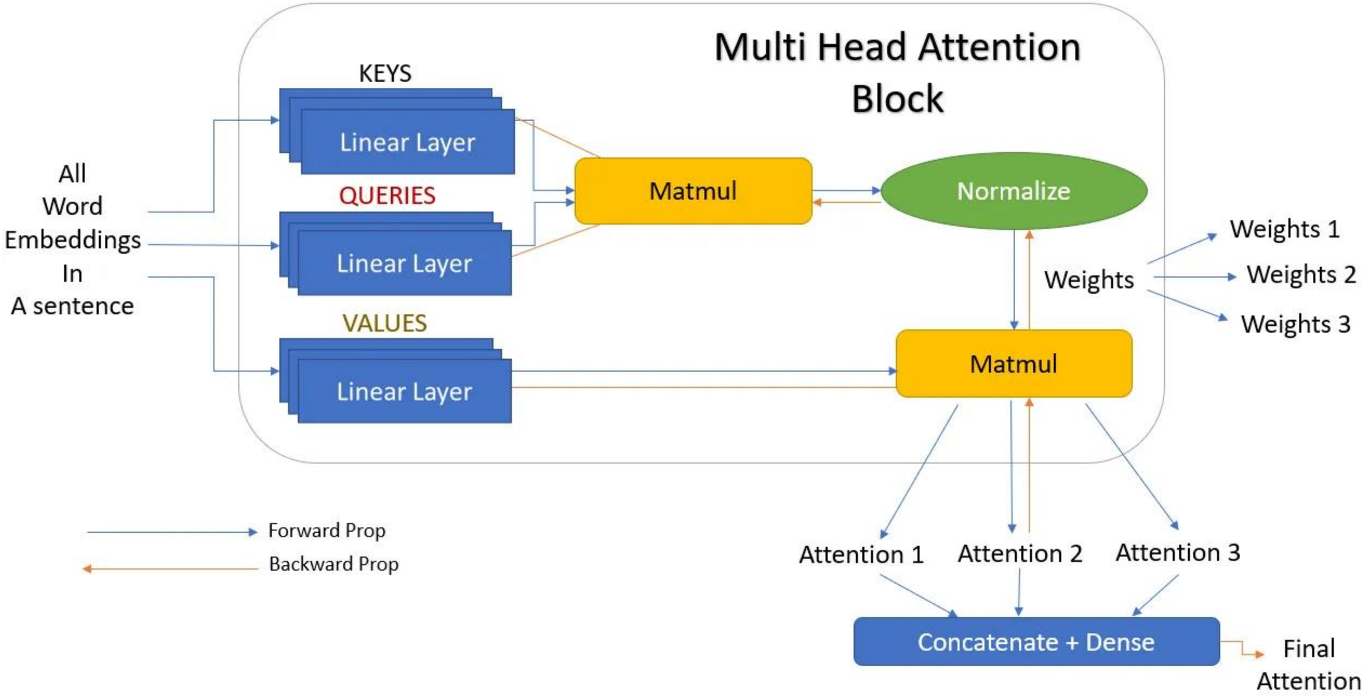


ATTENTION: MULTIHEAD





MULTIHEAD ATTENTION AND TRAINING



ATTENTION IN MACHINE TRANSLATION

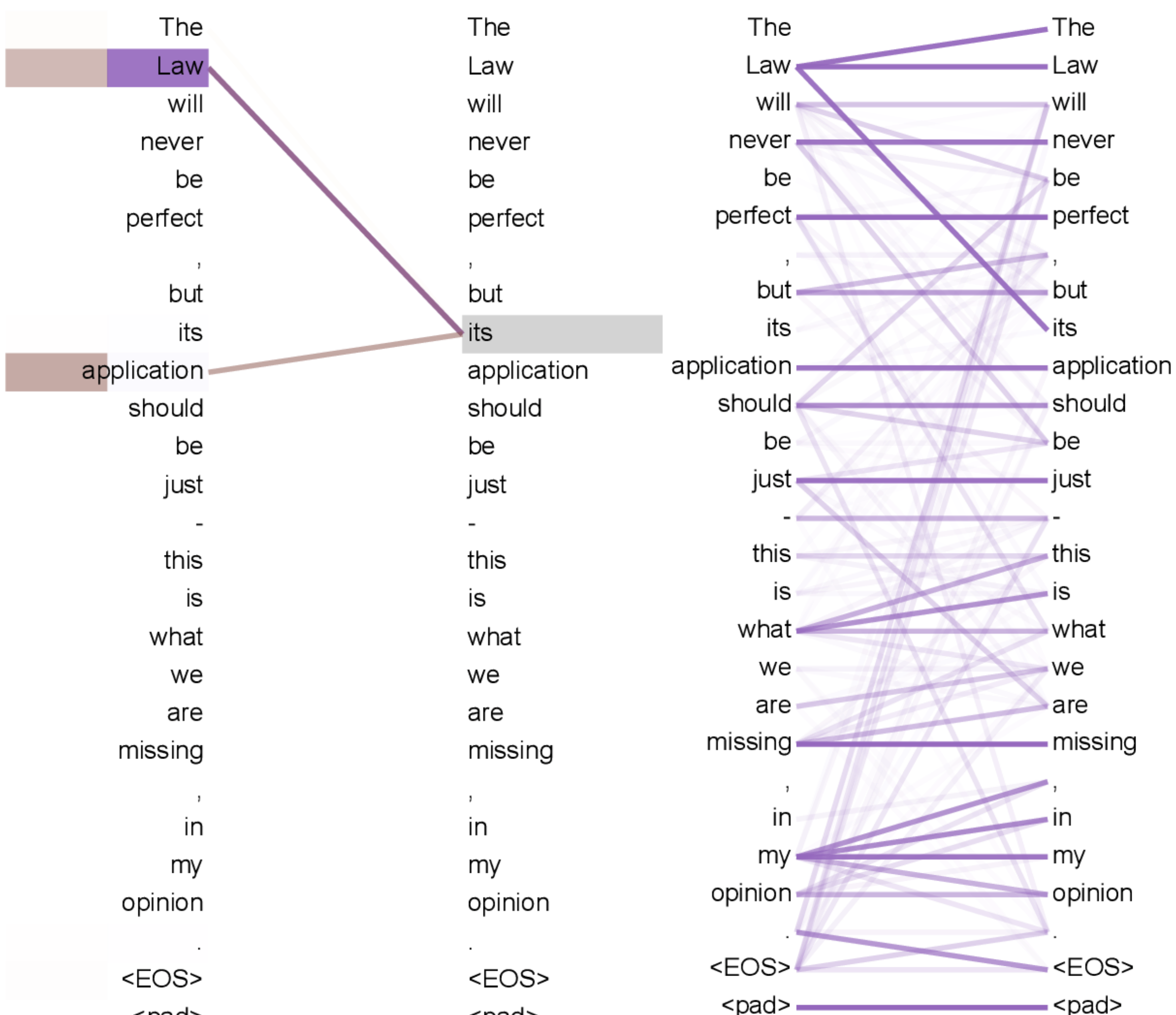
- Multihead attention is first captured at the encoding level between words in the input
- The different levels encode attention across multiple groups of word
- During Decoding the overall attention is used to condition individual emissions left to right
- As a results, emissions are made dependent on the entire input sequence and all dependencies are captured
- Queries are individual words embeddings, while keys are trained so that attention weights are learned from examples during training
- All attentions are thus targeted to minimize (decoding) errors

ATTENTION & ENCONDING

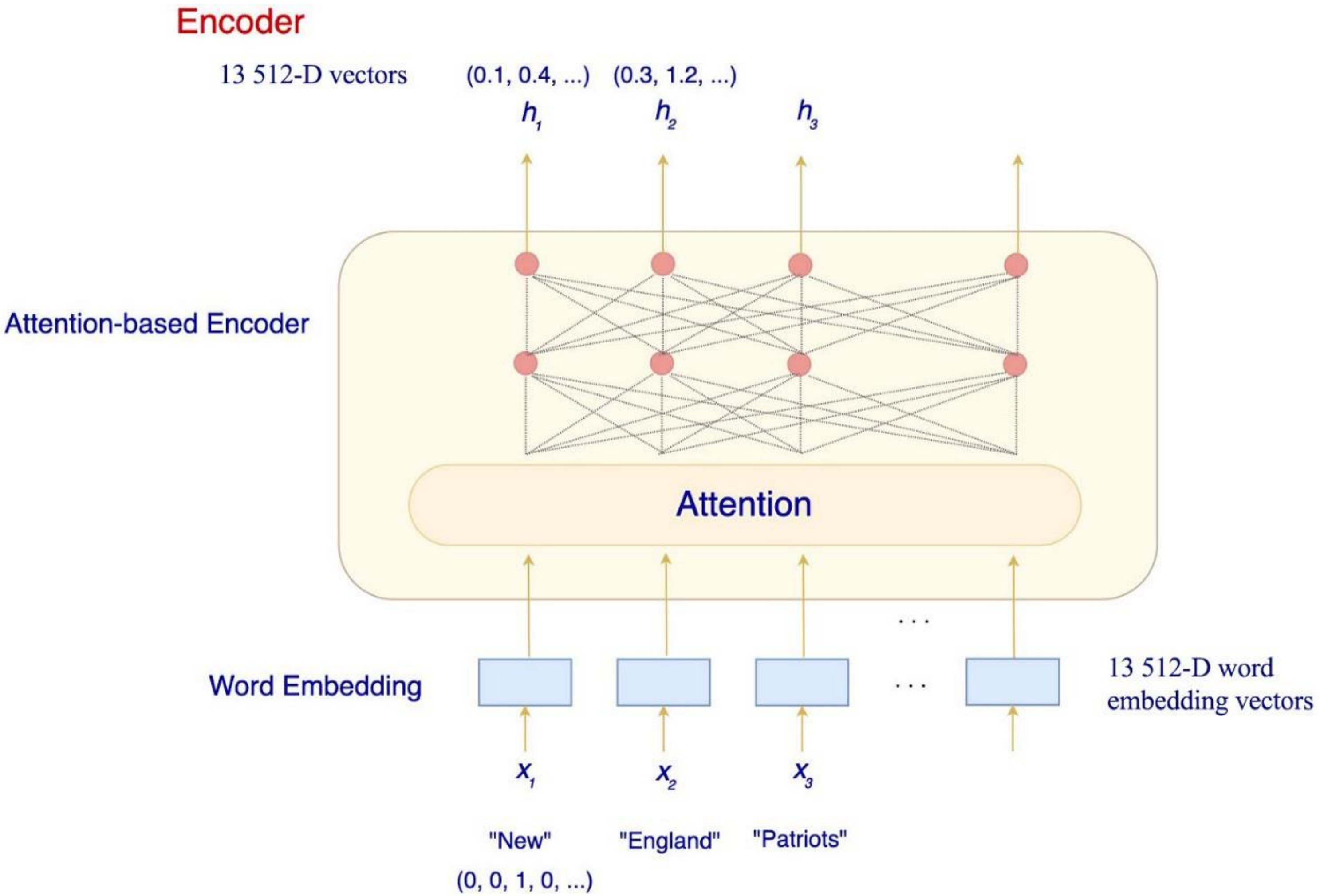
- In a decoding process (e.g. machine translation) there are **three** kinds of dependencies for neural architectures
- Dependencies are **independently** established between
    1. the ***input and output*** tokens
    2. the ***input tokens themselves***
    3. the ***output tokens themselves***

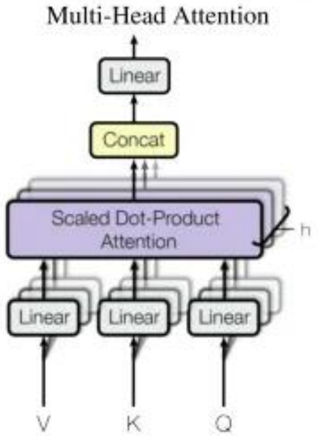
- Examples:
    - Machine Translation
    - QA where the query the answer paragraph is the input and the matched answer is the output

ATTENTION AND ANAPHORA



BERT: EXPLOITING ATTENTION FOR NLP





BERT & NLP: TRAINING THE ENCODER (ONLY)

- How to *train* (i.e. optimize) the encoding?
- **Two General** and **complex** tasks are proposed in (Devlin et al., 2018) are
  - Masked Language Modeling (15%)
    - Inpired by Distributional Hypothesis
    - Can be Simulated and does not require any labeling
  - Next Sentence Prediction
    - Inspired by Textual Inference tasks (e.g. Textual Entailment)
    - Can be Simulated and does not require any labeling
- Source Representations
  - Words? And why not subword? (in the BERT jargon) Word Pieces!!
    - Useful to deal with out-of-vocabulary phenomena

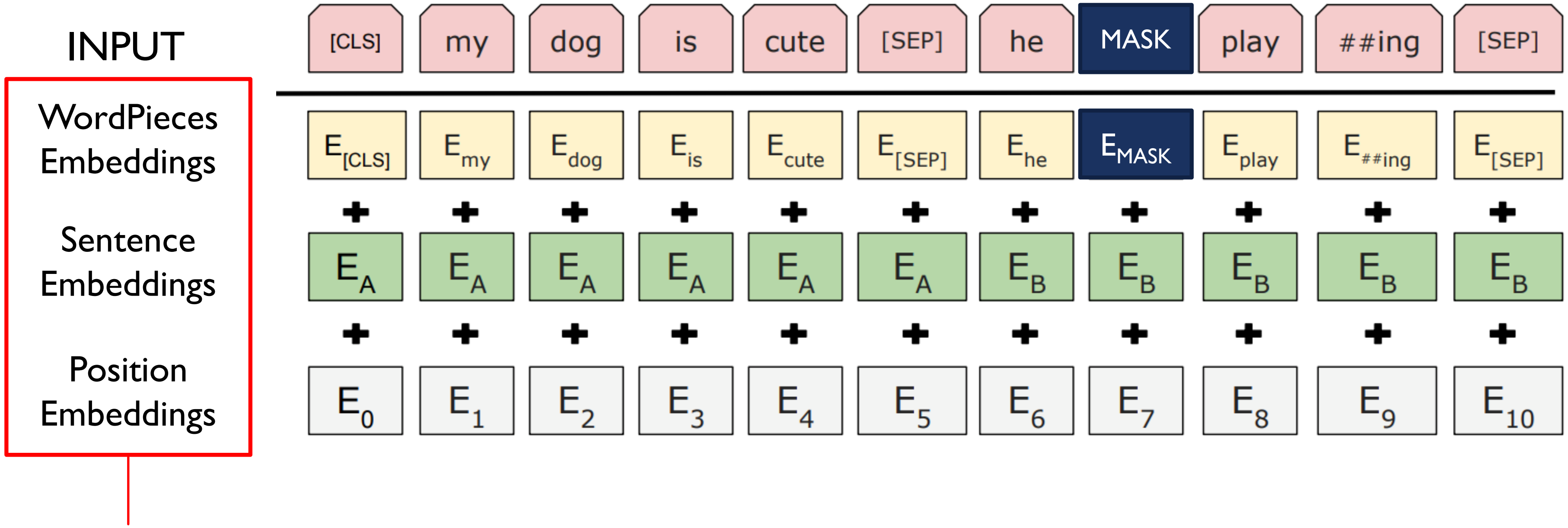
BERT (DEVLIN ET AL. ’18)

Pretraining on two unsupervised prediction tasks:

- **Masked Language Model**: given a sentence *s* with missing words, reconstruct *s*
  - Example: Amazon <MASK> amazing → Amazon is amazing
  - In BERT the language modeling is deeply Bidirectional, while in ELMo the forward and backward LMs were two independent branches of the NN

- **Next Sentence Prediction**: given two sentences s_1 and s_2, the task is to understand whether s_2 is the actual sentence that follows s_1
  - 50% of the training data are positive examples: s_1 and s_2 are actually consecutive sentences
  - 50% of the training data are negative examples: s_1 and s_2 are randomly chosen from the corpus

BERT PRETRAINING: INPUT REPRESENTATIONS

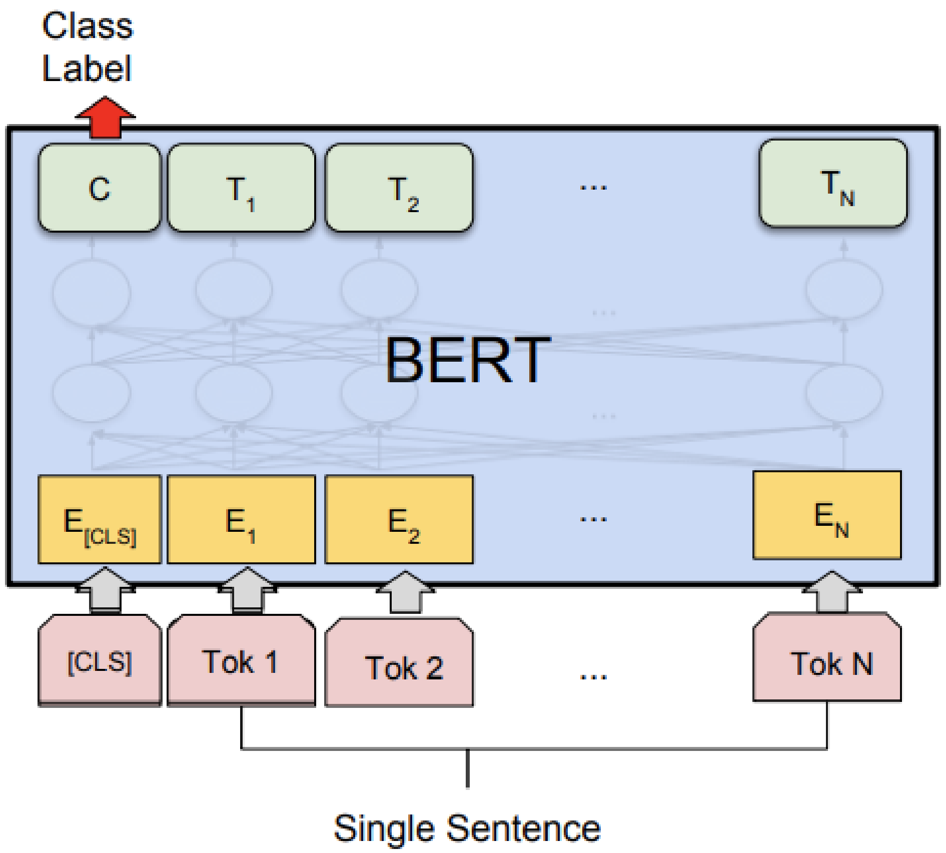


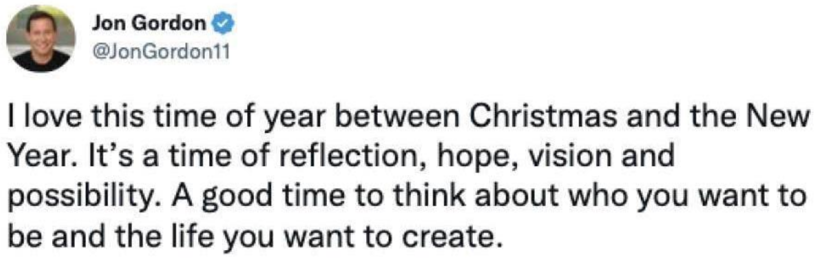
INPUT	[CLS]	my	dog	is	cute	[SEP]	he	MASK	play	##ing	[SEP]
WordPieces Embeddings	$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{MASK}	E_{play}	$E_{\#\#ing}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+
Sentence Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+	+	+	+	+	+	+	+	+	+
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}

All these embeddings are learned during the (pre)training process

In pre-training 15% of the input tokens are masked for the masked LM task

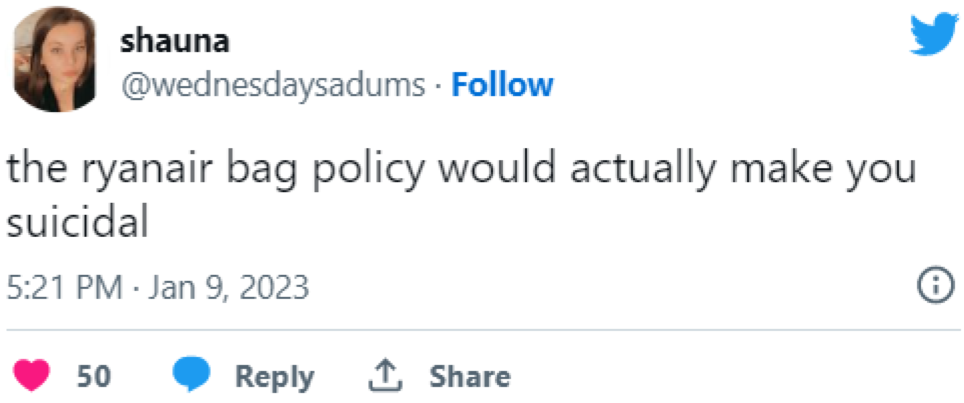
BERT (DEVLIN ET AL. '18) TASKS: INDIVIDUAL SENTENCE CLASSIFICATION




Jon Gordon
@JonGordon11

I love this time of year between Christmas and the New Year. It's a time of reflection, hope, vision and possibility. A good time to think about who you want to be and the life you want to create.

☺


shauna
@wednesdaysadums · Follow

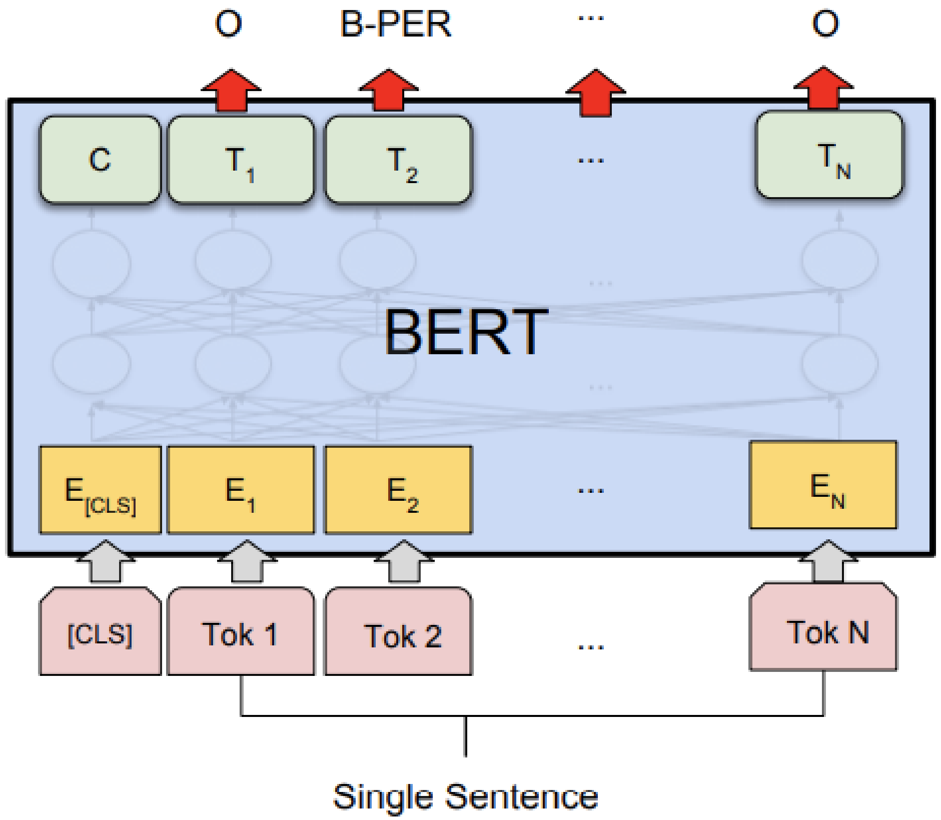
the ryanair bag policy would actually make you suicidal

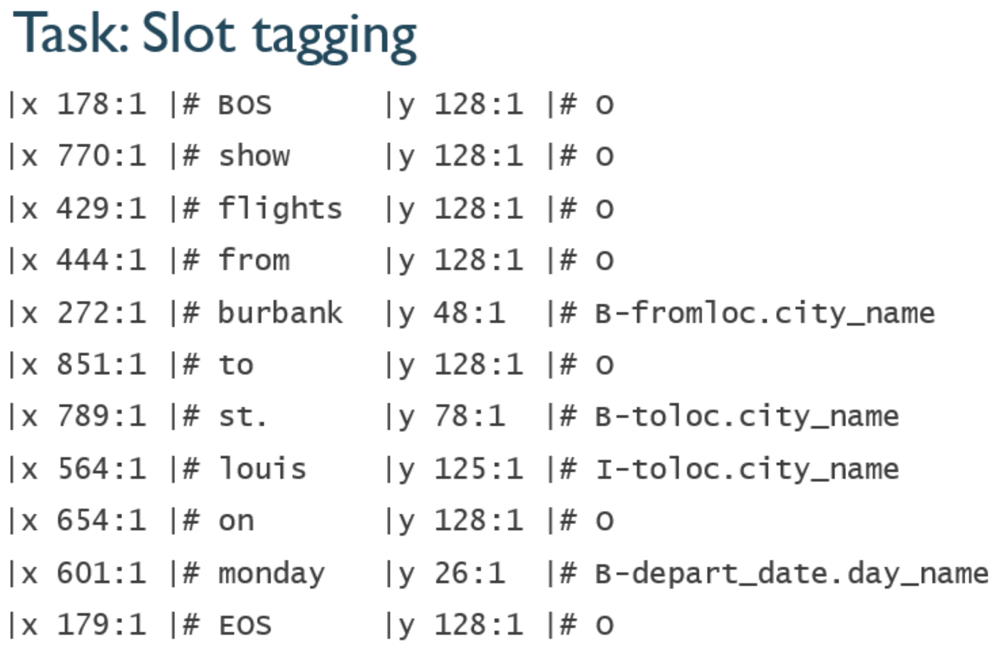
5:21 PM · Jan 9, 2023

☹

BERT for single sentence classification (Sentiment analysis, Intent Classification, etc.)

BERT (DEVLIN ET AL. '18) TASKS: SEQUENCE TAGGING



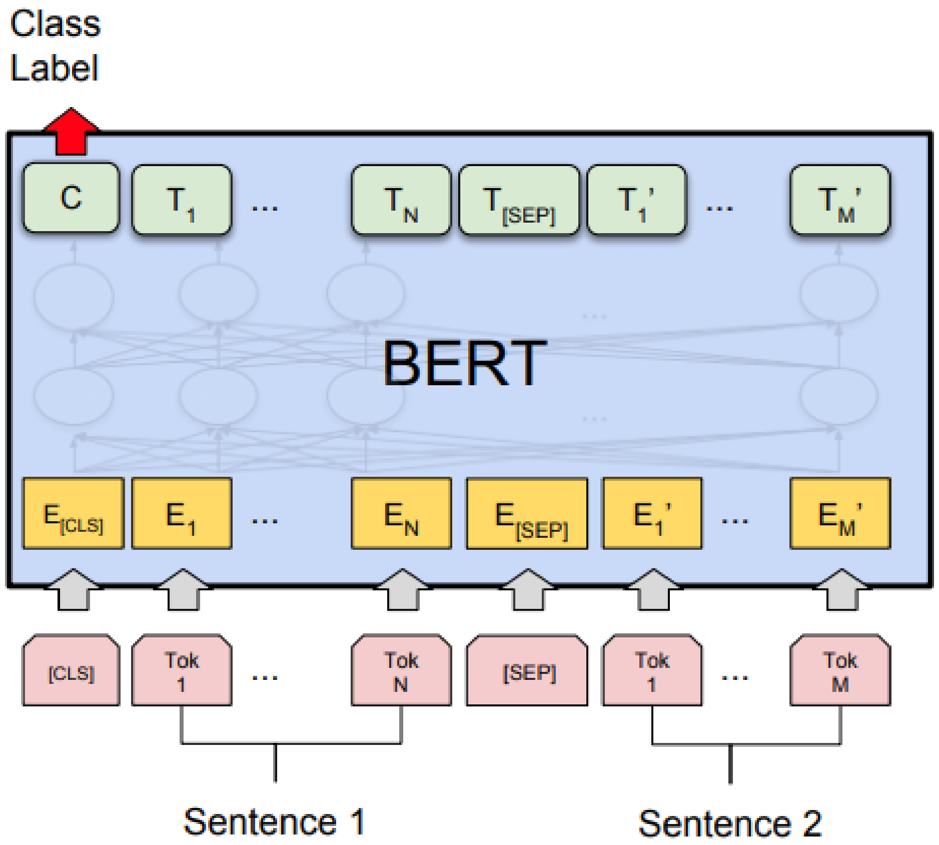


Task: Slot tagging

```
|x 178:1 |# BOS      |y 128:1 |# O
|x 770:1 |# show     |y 128:1 |# O
|x 429:1 |# flights  |y 128:1 |# O
|x 444:1 |# from     |y 128:1 |# O
|x 272:1 |# burbank  |y 48:1  |# B-fromloc.city_name
|x 851:1 |# to       |y 128:1 |# O
|x 789:1 |# st.      |y 78:1  |# B-toloc.city_name
|x 564:1 |# louis    |y 125:1 |# I-toloc.city_name
|x 654:1 |# on       |y 128:1 |# O
|x 601:1 |# monday   |y 26:1  |# B-depart_date.day_name
|x 179:1 |# EOS      |y 128:1 |# O
```

BERT for Sequence Tagging Tasks (e.g., POS tagging, Named Entity Recognition, etc.)

BERT (DEVLIN ET AL. '18) TASKS: SENTENCE PAIRS CLASSIFICATION



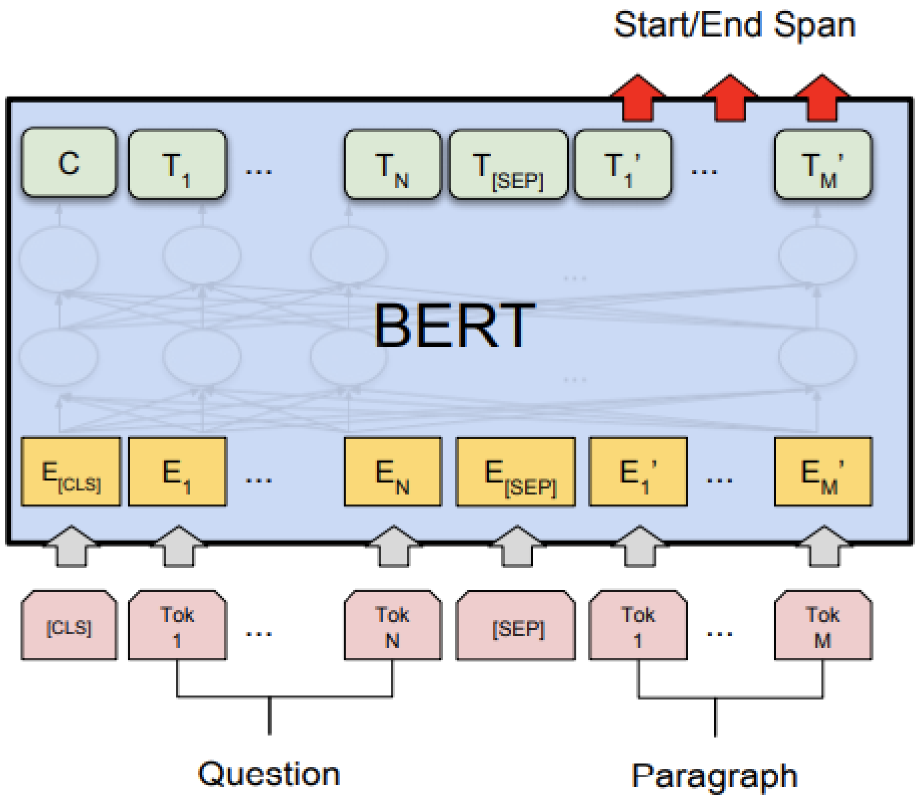
Answer selection in QA: Decide if A contains an answer to Q:
Q:"What is the Capital of Italy?"
A:"Rome, as the capital of Italy, is located …..."

RTE: Given T decide if H is true (or not)
T:"Rome is the Capital of Italy."
H:"Rome is in Italy."

PI: Given S1 and S2 decide if they are paraphrases (or not)
S:1 "Rome is the Capital of Italy."
S2:"Italy has Rome as its own Capital town."

BERT for sentence pairs classification (answer selection in QA, Recognizing Textual Entailment, Paraphrase Identification)

BERT (DEVLIN ET AL. '18) TASKS: QUESTION ANSWERING



Answer Span Selection in QA:

Decide which part of the text A corresponds to the answer to the query Q:

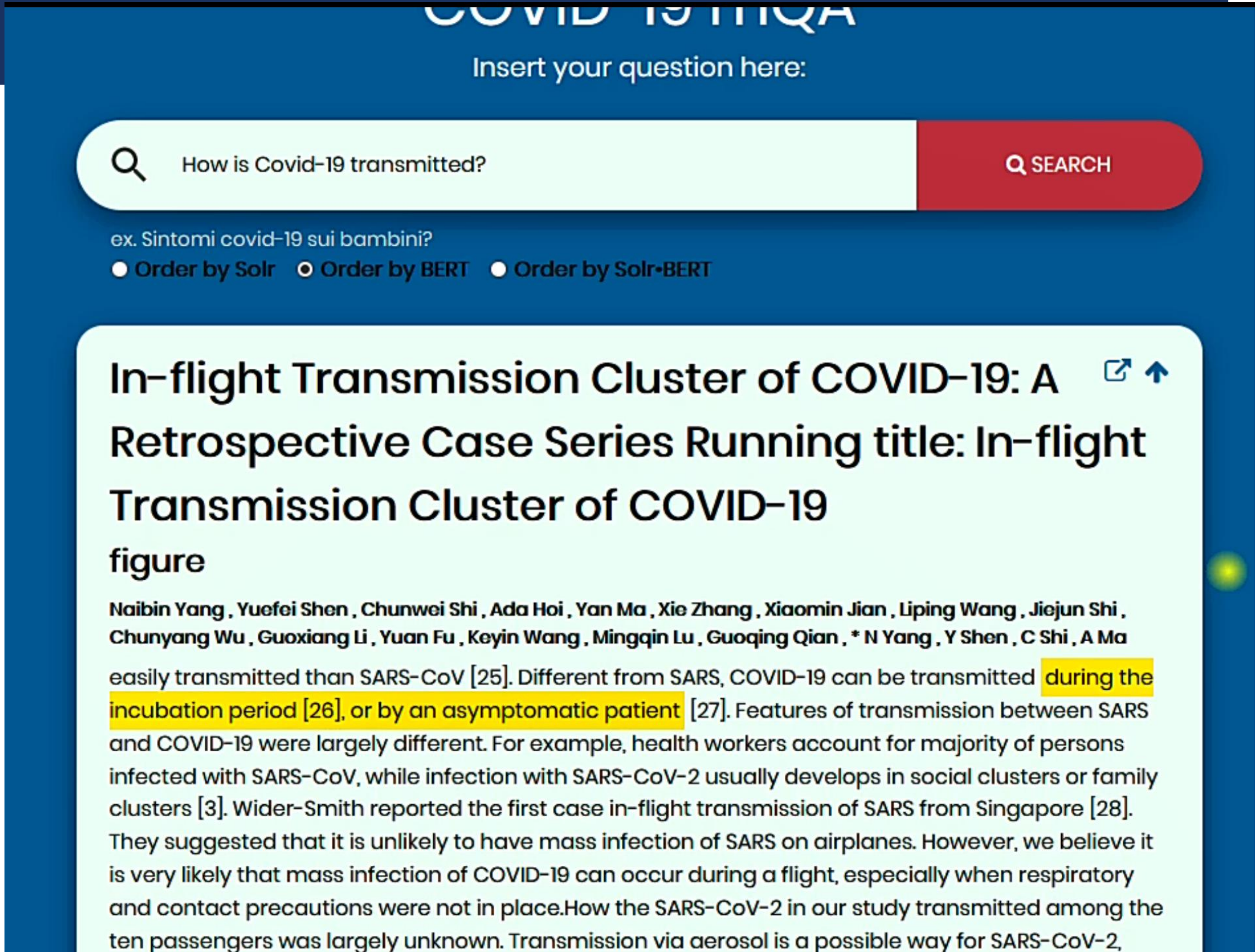
Q:"What is the Capital of Italy?"

A:"<Start>Rome<End>, as the capital of Italy, ….."

BERT for Answer Span Selection in Question Answering

A QA EXAMPLE ON SQUAD

- Question Answering even across languages
  - Query in Italian
  - Answer span over English Texts



Insert your question here:

How is Covid-19 transmitted?

ex. Sintomi covid-19 sui bambini?

○ Order by Solr ● Order by BERT ○ Order by Solr•BERT

In-flight Transmission Cluster of COVID-19: A Retrospective Case Series Running title: In-flight Transmission Cluster of COVID-19

figure

Naibin Yang , Yuefei Shen , Chunwei Shi , Ada Hoi , Yan Ma , Xie Zhang , Xiaomin Jian , Liping Wang , Jiejun Shi , Chunyang Wu , Guoxiang Li , Yuan Fu , Keyin Wang , Mingqin Lu , Guoqing Qian , * N Yang , Y Shen , C Shi , A Ma

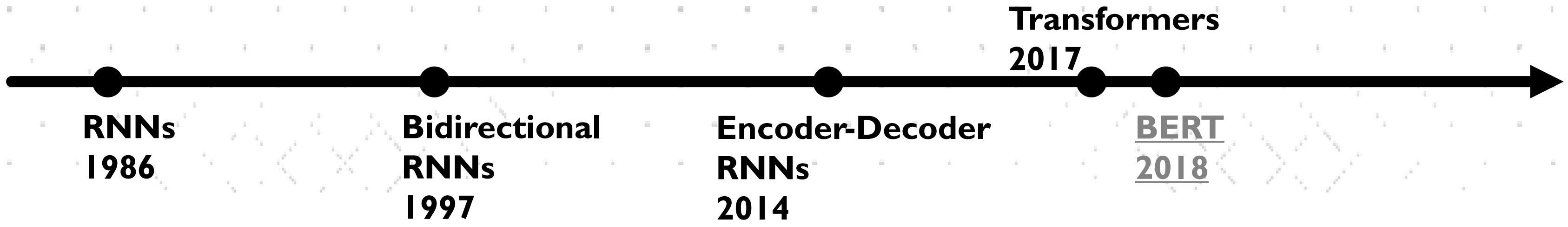
easily transmitted than SARS-CoV [25]. Different from SARS, COVID-19 can be transmitted during the incubation period [26], or by an asymptomatic patient [27]. Features of transmission between SARS and COVID-19 were largely different. For example, health workers account for majority of persons infected with SARS-CoV, while infection with SARS-CoV-2 usually develops in social clusters or family clusters [3]. Wider-Smith reported the first case in-flight transmission of SARS from Singapore [28]. They suggested that it is unlikely to have mass infection of SARS on airplanes. However, we believe it is very likely that mass infection of COVID-19 can occur during a flight, especially when respiratory and contact precautions were not in place.How the SARS-CoV-2 in our study transmitted among the ten passengers was largely unknown. Transmission via aerosol is a possible way for SARS-CoV-2,

RETI NEURALI AVANZATE:

DALL'AUTOENCODING ALLA IA GENERATIVA

METODI E ARCHITETTURE

Machine learning paradigms underlying ChatGPT



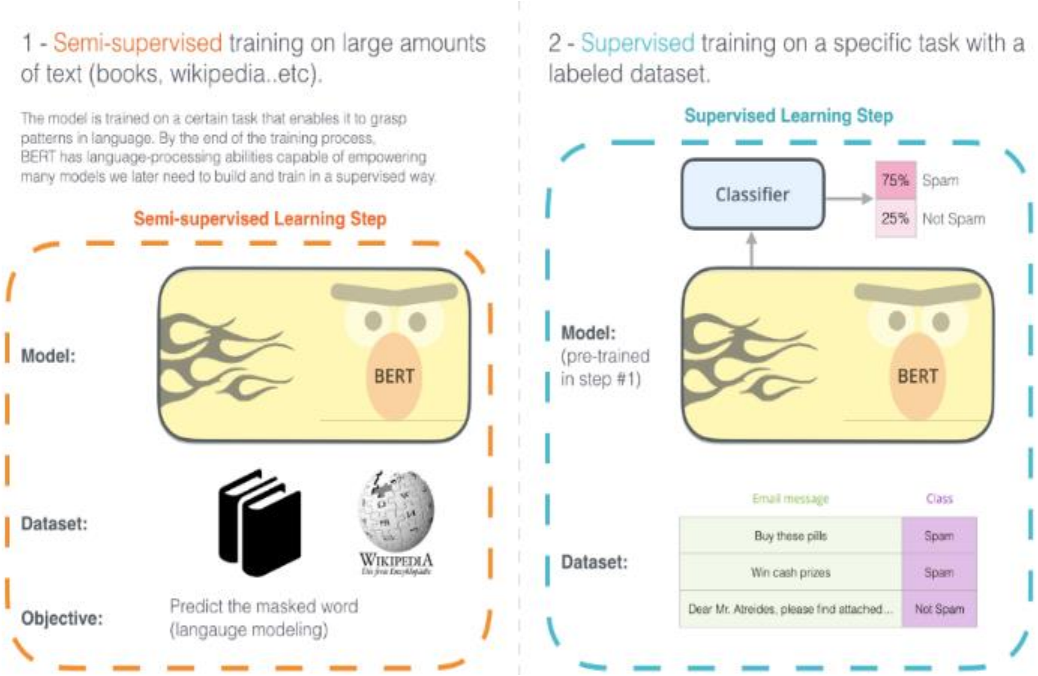
Transformers 2017

RNNs 1986

Bidirectional RNNs 1997

Encoder-Decoder RNNs 2014

BERT 2018



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:

Dataset:

Objective: Predict the masked word (langauge modeling)

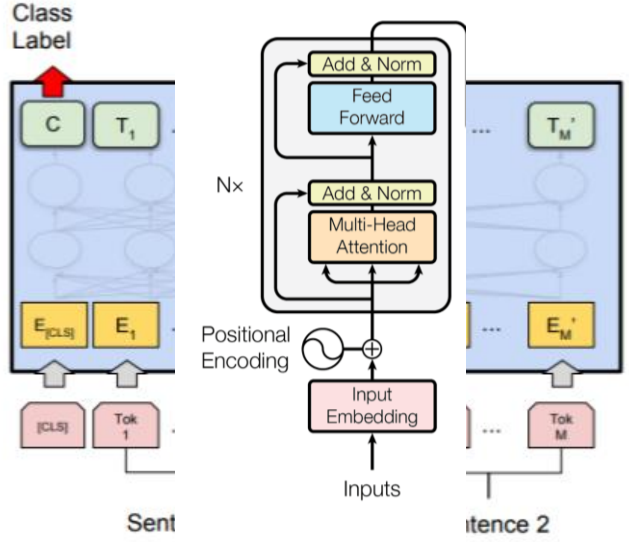
2 - Supervised training on a specific task with a labeled dataset.

Supervised Learning Step

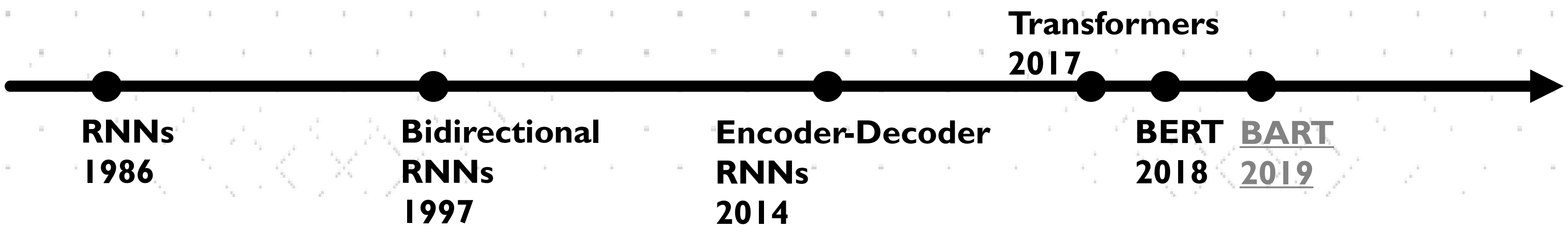
Model: (pre-trained in step #1)

Dataset:

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



Machine learning paradigms underlying ChatGPT: BART



Transformers 2017

RNNs 1986

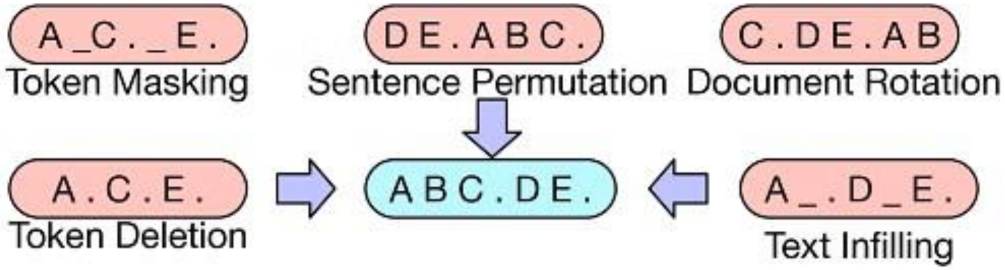
Bidirectional RNNs 1997

Encoder-Decoder RNNs 2014

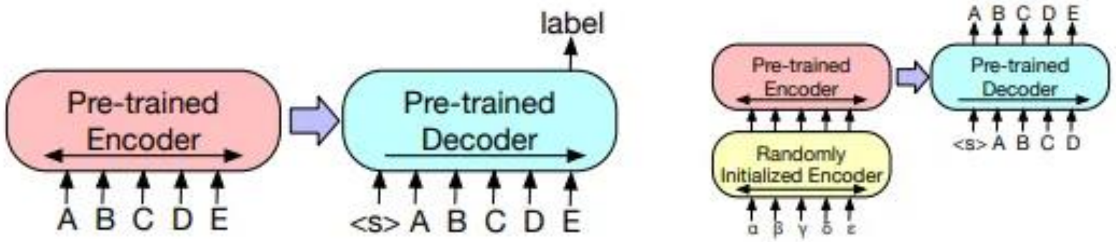
BERT 2018

BART 2019

BART pre-Training:



BART Fine-Tuning:



(a) To use BART for classification problems, the same input is fed into the encoder and decoder, and the representation from the final output is used.

(b) For machine translation, we learn a small additional encoder that replaces the word embeddings in BART. The new encoder can use a disjoint vocabulary.

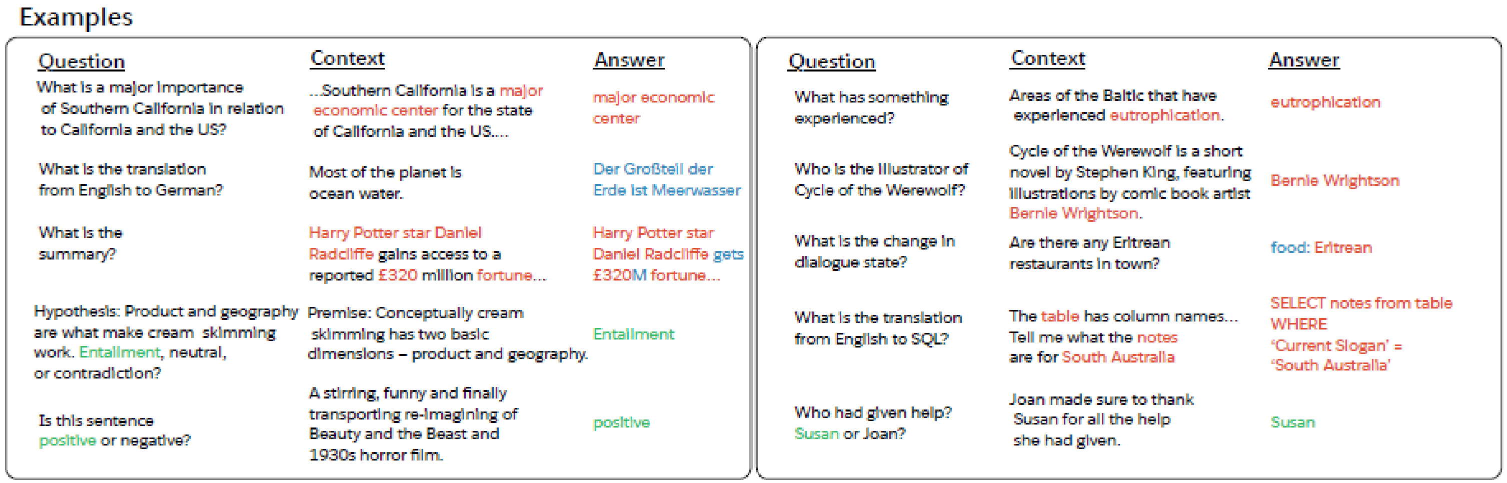
Figure 3: Fine tuning BART for classification and translation.

GPT-2: DECODER ONLY ARCHITECTURES (RADFORD ET AL., 2019)

- “We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText”
- GPT-2 is a large transformer-based language model with 1.5 billion parameters, trained on a dataset of 8 million web pages.
- GPT-2 is trained with a simple objective: predict the next word, given all of the previous words within some text.
- The diversity of the dataset causes this simple goal to contain naturally occurring demonstrations of many tasks across diverse domains.
- GPT-2 is a direct scale-up of GPT, with more than 10X the parameters and trained on more than 10X the amount of data

GPT-2: SOURCES OF INSIPIRATION

- Multitask QA Networks (MQAN) (McCann et al, 2018)



Examples

Question	Context	Answer
What is a major importance of Southern California in relation to California and the US?	...Southern California is a major economic center for the state of California and the US....	major economic center
What is the translation from English to German?	Most of the planet is ocean water.	Der Großteil der Erde ist Meerwasser
What is the summary?	Harry Potter star Daniel Radcliffe gains access to a reported £320 million fortune...	Harry Potter star Daniel Radcliffe gets £320M fortune...
Hypothesis: Product and geography are what make cream skimming work. Entailment, neutral, or contradiction?	Premise: Conceptually cream skimming has two basic dimensions – product and geography.	Entailment
Is this sentence positive or negative?	A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.	positive
What has something experienced?	Areas of the Baltic that have experienced eutrophication.	eutrophication
Who is the illustrator of Cycle of the Werewolf?	Cycle of the Werewolf is a short novel by Stephen King, featuring illustrations by comic book artist Bernie Wrightson.	Bernie Wrightson
What is the change in dialogue state?	Are there any Eritrean restaurants in town?	food: Eritrean
What is the translation from English to SQL?	The table has column names... Tell me what the notes are for South Australia	SELECT notes from table WHERE 'Current Slogan' = 'South Australia'
Who had given help? Susan or Joan?	Joan made sure to thank Susan for all the help she had given.	Susan

Figure 1: Overview of the decaNLP dataset with one example from each decaNLP task in the order presented in Section 2. They show how the datasets were pre-processed to become question answering problems. Answer words in red are generated by pointing to the context, in green from the question, and in blue if they are generated from a classifier over the output vocabulary.

- Our speculation is that a language model with sufficient capacity will begin to learn to infer and perform the tasks demonstrated in natural language sequences in order to better predict them, regardless of their method of procurement. If a language model is able to do this it will be, in effect, performing unsupervised multitask learning.

GPT-2: ARCHITECTURE AND TASKS

- From (Radford et al., 2017, GPT paper)

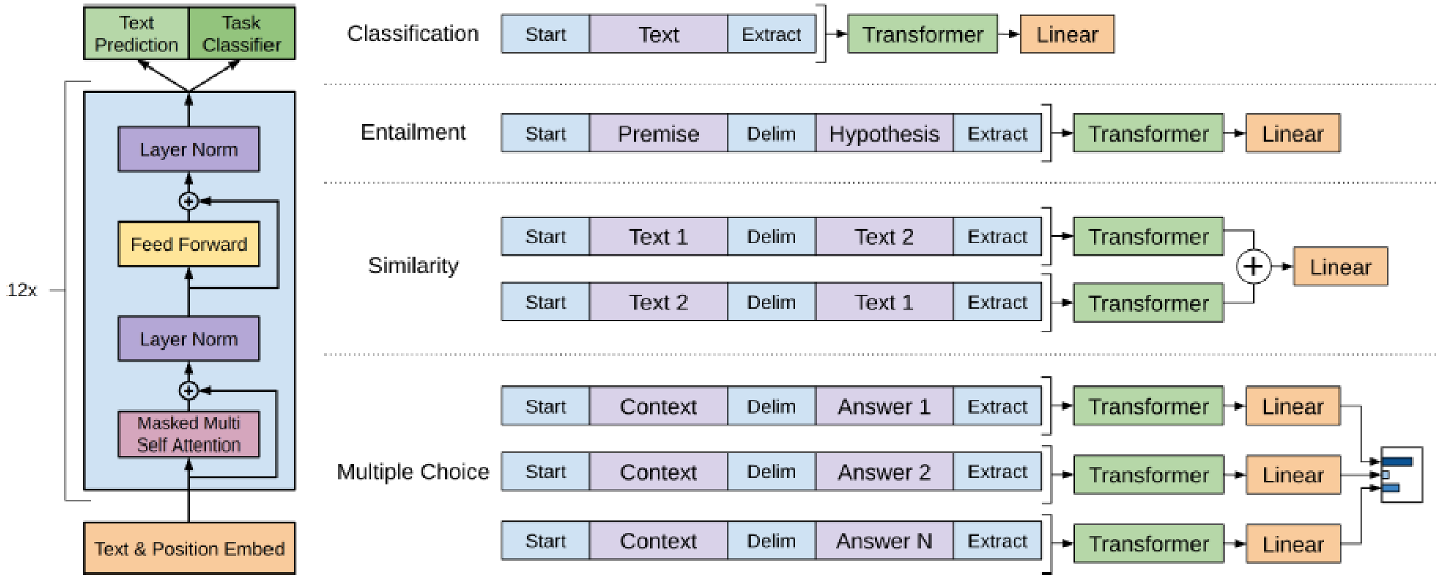
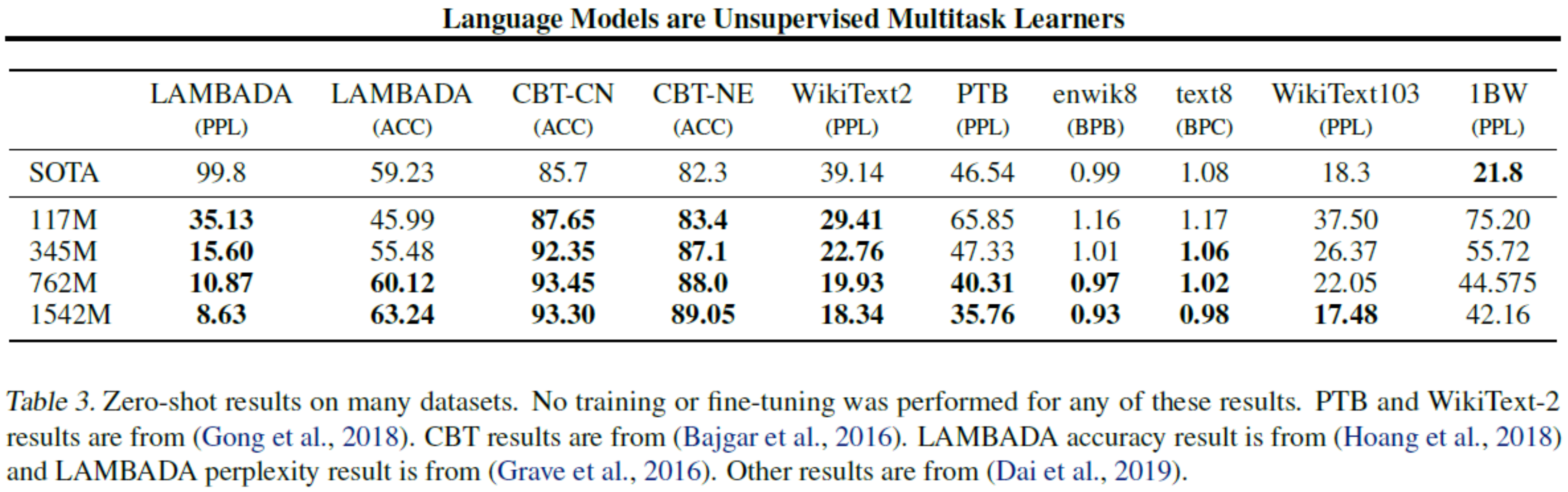


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

GPT-2: RESULTS OVER DIFFERENT TASKS



Language Models are Unsupervised Multitask Learners

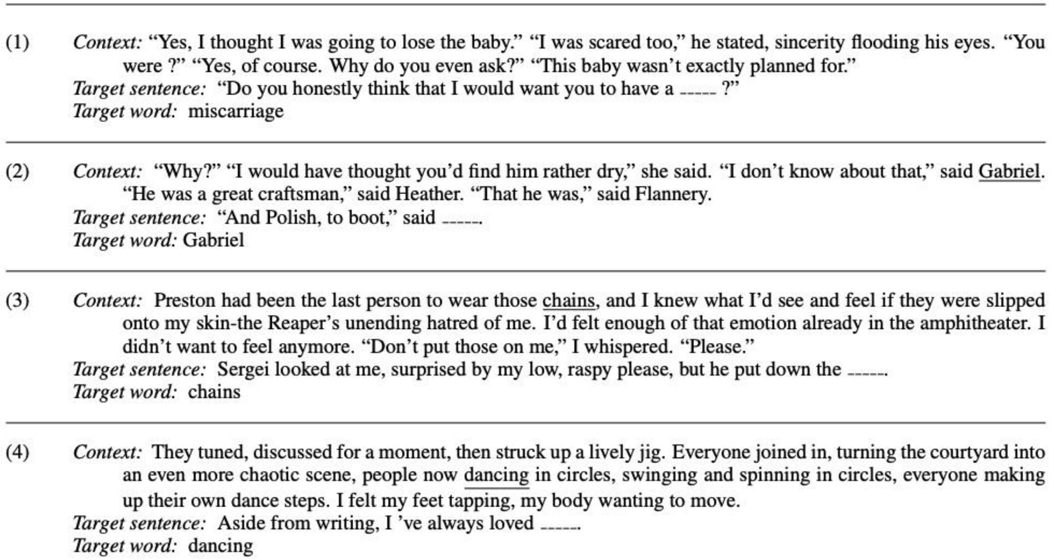
	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

- The LAMBADA dataset (Paperno et al., 2016)
  - It tests the ability of systems to model long-range dependencies in text.
  - The task is to predict the final word of sentences which require at least 50 tokens of context for a human to successfully predict.

GPT-2: RESULTS ON LAMBADA

- The LAMBADA dataset (Paperno et al., 2016)
  - It tests the ability of systems to model long-range dependencies in text.
  - The task is to predict the final word of sentences which require at least 50 tokens of context for a human to successfully predict.



(1) *Context:* "Yes, I thought I was going to lose the baby." "I was scared too," he stated, sincerity flooding his eyes. "You were ?" "Yes, of course. Why do you even ask?" "This baby wasn't exactly planned for."
Target sentence: "Do you honestly think that I would want you to have a ____ ?"
Target word: miscarriage

(2) *Context:* "Why?" "I would have thought you'd find him rather dry," she said. "I don't know about that," said Gabriel. "He was a great craftsman," said Heather. "That he was," said Flannery.
Target sentence: "And Polish, to boot," said ____.
Target word: Gabriel

(3) *Context:* Preston had been the last person to wear those chains, and I knew what I'd see and feel if they were slipped onto my skin-the Reaper's unending hatred of me. I'd felt enough of that emotion already in the amphitheater. I didn't want to feel anymore. "Don't put those on me," I whispered. "Please."
Target sentence: Sergei looked at me, surprised by my low, raspy please, but he put down the ____.
Target word: chains

(4) *Context:* They tuned, discussed for a moment, then struck up a lively jig. Everyone joined in, turning the courtyard into an even more chaotic scene, people now dancing in circles, swinging and spinning in circles, everyone making up their own dance steps. I felt my feet tapping, my body wanting to move.
Target sentence: Aside from writing, I 've always loved ____.
Target word: dancing

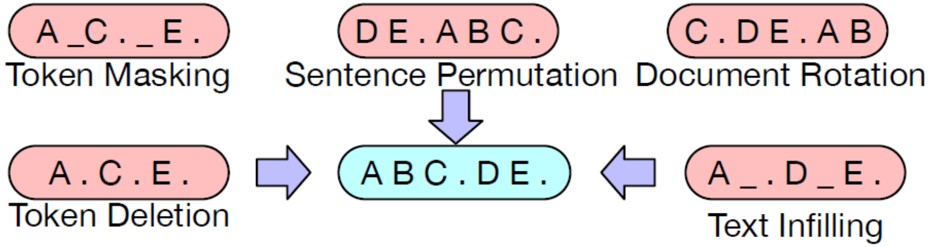
- GPT-2 improves the state of the art from 99.8 (Grave et al., 2016) to 8.6 perplexity and increases the accuracy of LMs on this test from 19% (Dehghani et al., 2018) to 52.66%. Adding a stop-word filter as an approximation to this further increases accuracy to 63.24%.
- Investigating GPT-2's errors showed most predictions are valid sentence continuations, but are not valid final words

BART (LEWIS ET AL., 2019) - FACEBOOK

- Enconding decoding architecture based on Pretraining and fine tuned towards different tasks such as:

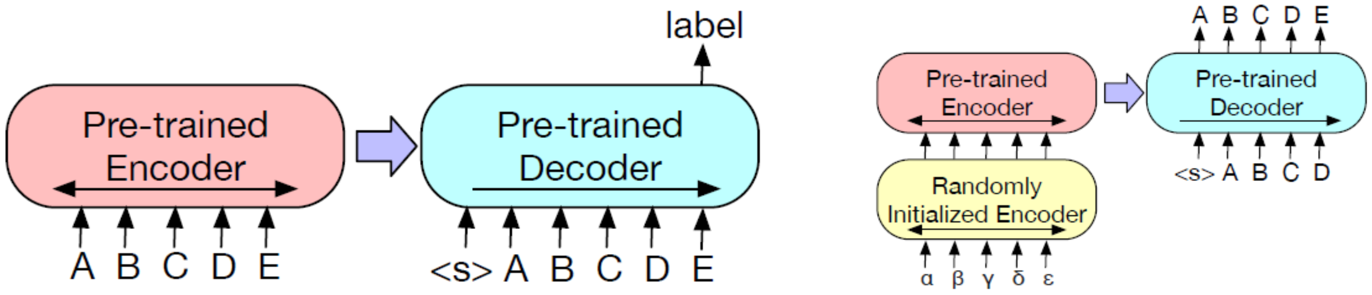
  RTE, SA, …

- Two stages of PRETRAINING
  - Text is first corrupted with an arbitrary noising function,
  - A sequence-to-sequence model is learned to reconstruct the original text.



- FINE TUNING:
  - **MNLI** (Williams et al., 2017), a bitext classification task to predict whether one sentence **entails** another. The fine-tuned model concatenates the two sentences with appended an EOS token, and passes them to both the BART encoder and decoder. In contrast to BERT, the representation of the EOS token is used to classify the sentences relations.
  - ELI5 (Fan et al., 2019), a long-form abstractive question answering dataset. Models generate answers conditioned on the concatenation of a question and supporting documents.

APPLYING BART



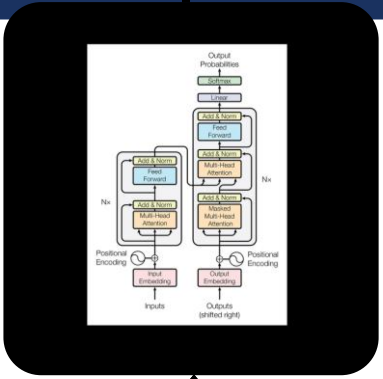
(a) To use BART for classification problems, the same input is fed into the encoder and decoder, and the representation from the final output is used.

(b) For machine translation, we learn a small additional encoder that replaces the word embeddings in BART. The new encoder can use a disjoint vocabulary.

Figure 3: Fine tuning BART for classification and translation.

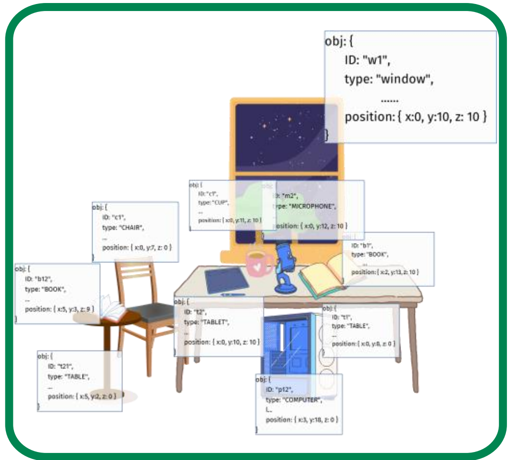
GRUT: THE OVERALL FLOW

Output: `TAKING(Theme(b1))`



GrUT-IT

Command: *"Prendi il volume sul tavolo vicino la finestra*"


obj: { ID: "w1", type: "window", ... position: { x:0, y:10, z: 10 } }
obj: { ID: "c1", type: "CUP", ... position: { x:0, y:11, z: 10 } }
obj: { ID: "m2", type: "MICROPHONE", ... position: { x:0, y:12, z: 10 } }
obj: { ID: "c1", type: "CHAIR", ... position: { x:0, y:7, z: 0 } }
obj: { ID: "b1", type: "BOOK", ... position: { x:2, y:13, z: 10 } }
obj: { ID: "b12", type: "BOOK", ... position: { x:5, y:3, z: 9 } }
obj: { ID: "t2", type: "TABLET", ... position: { x:0, y:10, z: 10 } }
obj: { ID: "t1", type: "TABLE", ... position: { x:0, y:8, z: 0 } }
obj: { ID: "t21", type: "TABLE", ... position: { x:5, y:2, z: 0 } }
obj: { ID: "p12", type: "COMPUTER", ... position: { x:3, y:18, z: 0 } }

Entities Retrieval

Linguistic Extraction

Input: *Command + MD*

MD: *b1, conosciuto anche come libro o volume, è un'istanza della classe BOOK, t1, conosciuto anche come tavolo o scrivania, è un'istanza della classe TABLE # b1 è vicino t1*

Hromei et al, 2022, "Embedding Contextual Information in Seq2seq Models for Grounded Semantic Role Labeling"

EXPERIMENTAL EVALUATION

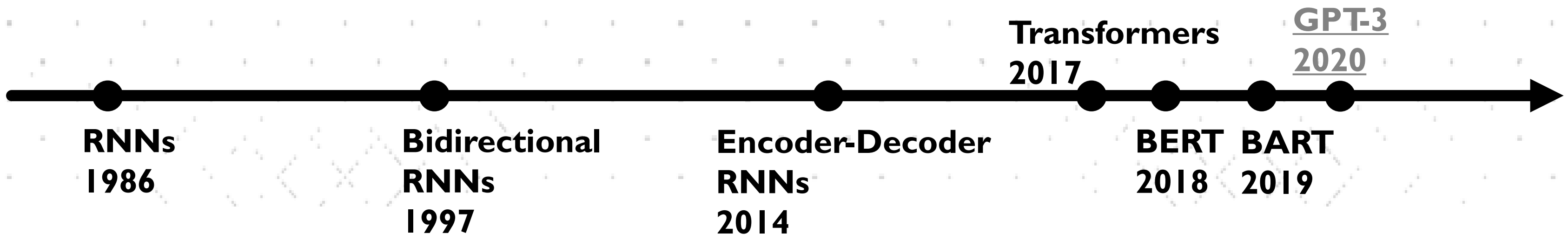
FP = Frame Prediction
AIC = Argument Identification and Classification
EM = Exact Match
HM = Head Match

Model	Learning Rate	FP	AIC-Exact Match	AIC-Head Match
LU4R	-	*95.32%*	*77.67%*	*86.35%*
GrUT-IT	$5 \cdot 10^{-5}$	96.86%	82.30%	85.19%

LU4R: `TAKING(Theme("libro"))`
GrUT-IT: `TAKING(Theme(b1))`

Results here are reported as F1 values on 10-fold cross-validation schema with 80/10/10 data split.
Performance for LU4R is reported in *italic* as it is not entirely comparable with.

Machine learning paradigms underlying ChatGPT



RNNs 1986

Bidirectional RNNs 1997

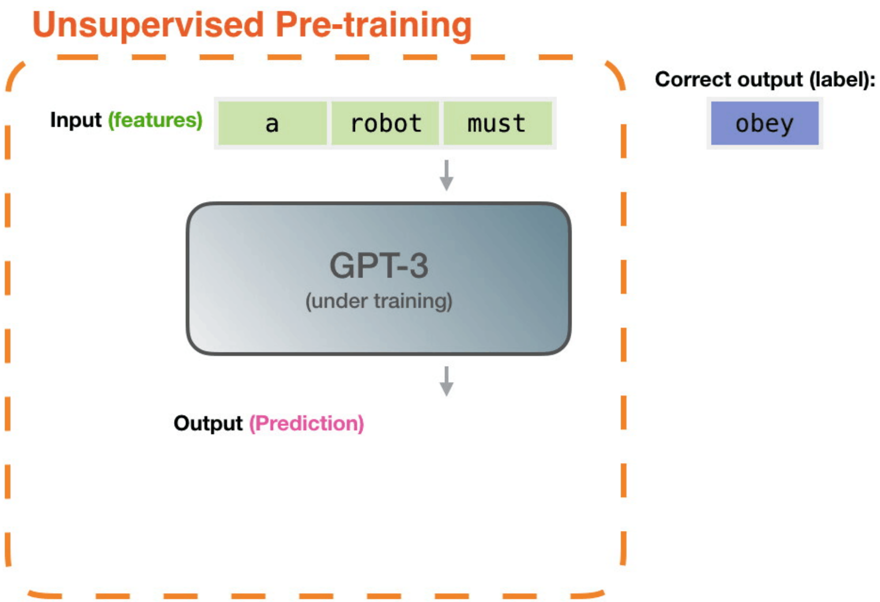
Encoder-Decoder RNNs 2014

Transformers 2017

BERT 2018

BART 2019

GPT-3 2020



Unsupervised Pre-training

Input (features) a robot must

GPT-3 (under training)

Output (Prediction)

Correct output (label): obey

GPT3: NOVELTY

- «Language Models are Few-Shot Learners" (Brown et al., 2020)

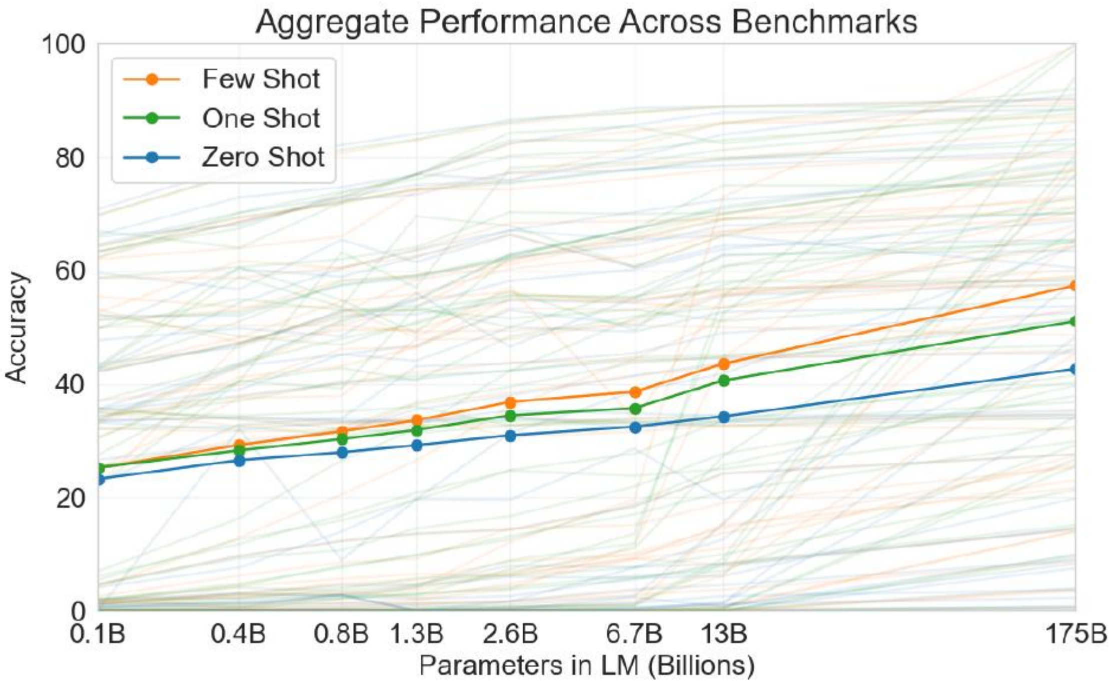
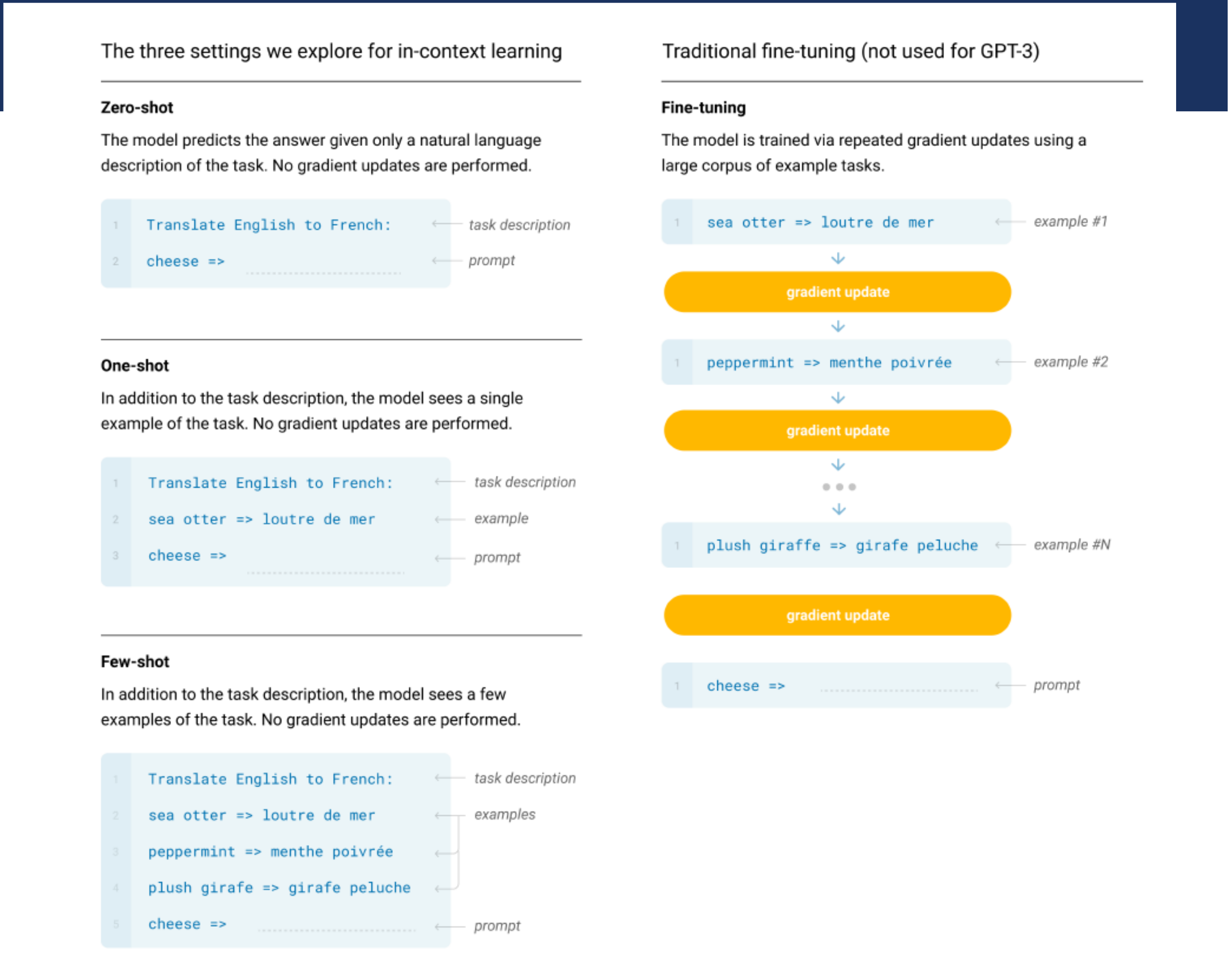


Figure 1.3: Aggregate performance for all 42 accuracy-denominated benchmarks While zero-shot performance improves steadily with model size, few-shot performance increases more rapidly, demonstrating that larger models are more proficient at in-context learning. See Figure 3.8 for a more detailed analysis on SuperGLUE, a standard NLP benchmark suite.

PROMPTING VS. LEARNING



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.

```
Translate English to French:
cheese =>
```

One-shot

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

```
Translate English to French:
sea otter => loutre de mer
cheese =>
```

Few-shot

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

```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

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.

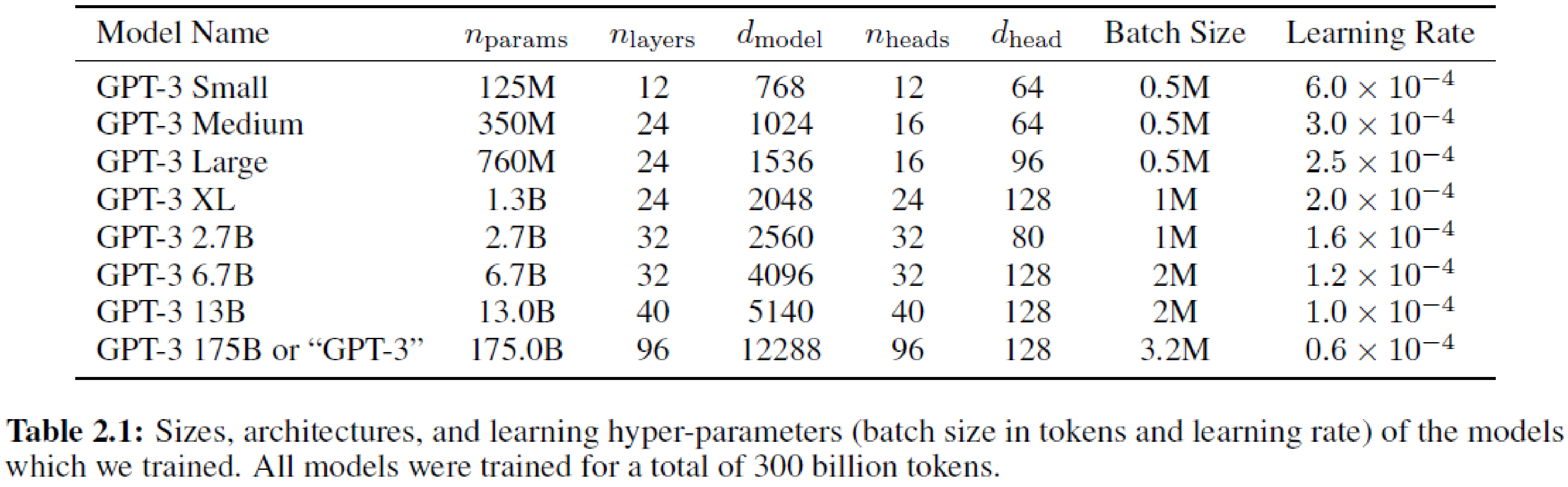
```
sea otter => loutre de mer
```

```
peppermint => menthe poivrée
```

```
plush giraffe => girafe peluche
```

```
cheese =>
```

GPT-3: SIZE

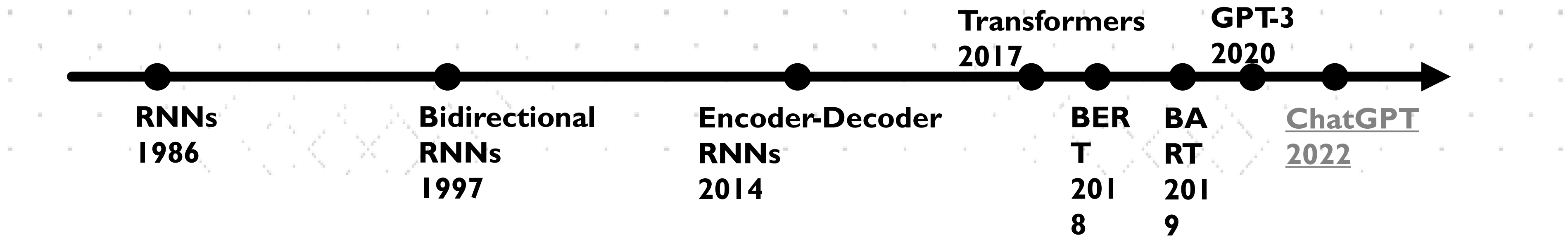


Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or “GPT-3”	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

- Here n_{params} is the total number of trainable parameters, n_{layers} is the total number of layers, d_{model} is the number of units in each bottleneck layer (we always have the feedforward layer four times the size of the bottleneck layer, d_{ff}=4xd_{model}), and d_{head} is the dimension of each attention head.
- All models use a context window of n_{ctx} = 2048 tokens

Machine learning paradigms underlying ChatGPT



RNNs 1986

Bidirectional RNNs 1997

Encoder-Decoder RNNs 2014

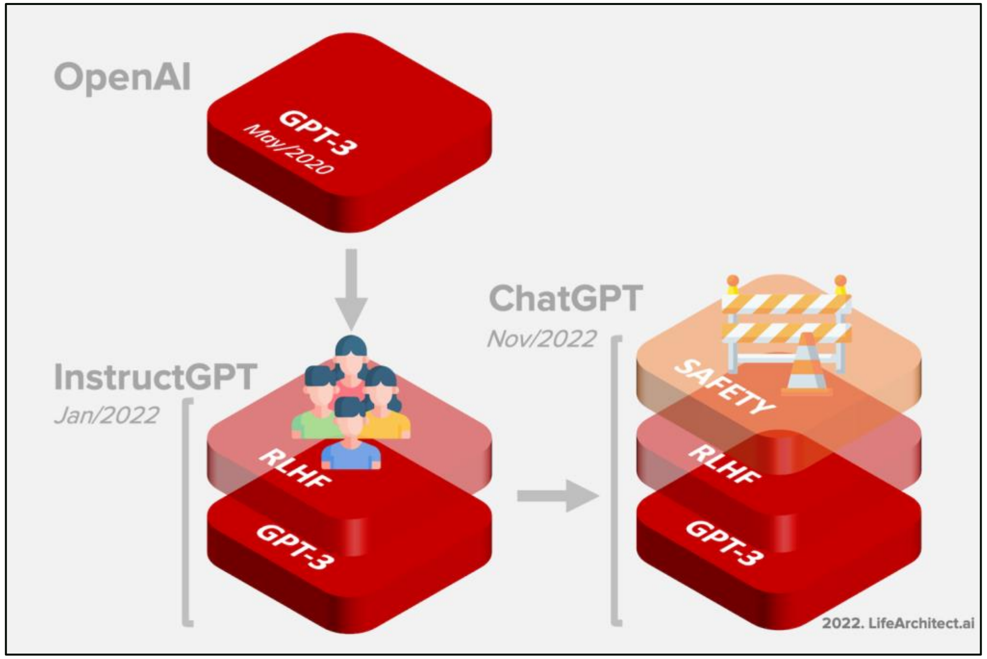
Transformers 2017

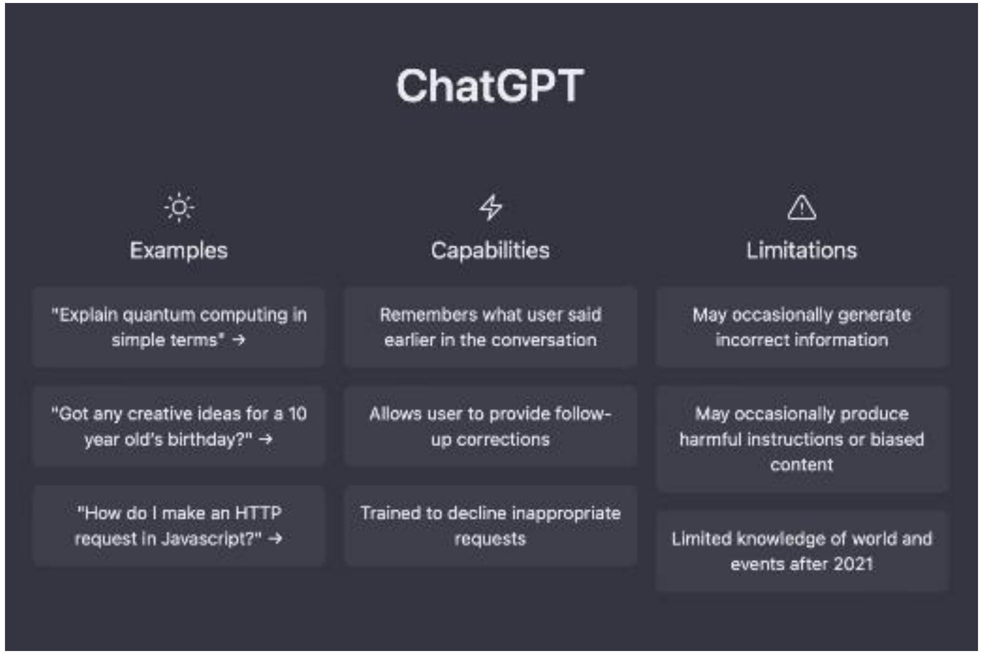
BERT 2018

BART 2019

GPT-3 2020

ChatGPT 2022





ChatGPT

Examples

"Explain quantum computing in simple terms" →

"Got any creative ideas for a 10 year old's birthday?" →

"How do I make an HTTP request in Javascript?" →

Capabilities

Remembers what user said earlier in the conversation

Allows user to provide follow-up corrections

Trained to decline inappropriate requests

Limitations

May occasionally generate incorrect information

May occasionally produce harmful instructions or biased content

Limited knowledge of world and events after 2021

LIMITATIONS OF GPT-3

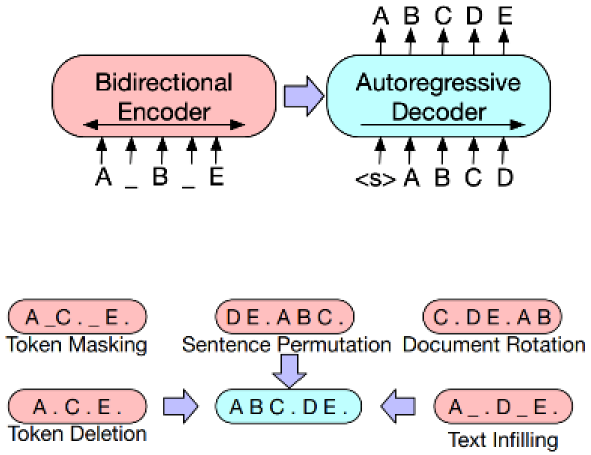
- Large language models often express unintended behaviors such as making up facts, generating biased or toxic text, or simply not following user instructions. This is because the language modeling objective is misaligned.
- The idea: aligning language models by **training them to act in accordance with the user's intention** (Leike et al., 2018).
  - explicit intentions such as following instructions
  - implicit intentions such as staying truthful, and not being biased, toxic, or otherwise harmful.
- Overall Objective: language models should be helpful (they should help the user solve their task), honest (they shouldn't fabricate information or mislead the user), and harmless (they should not cause physical, psychological, or social harm to people or the environment).

INSTRUCT GPT

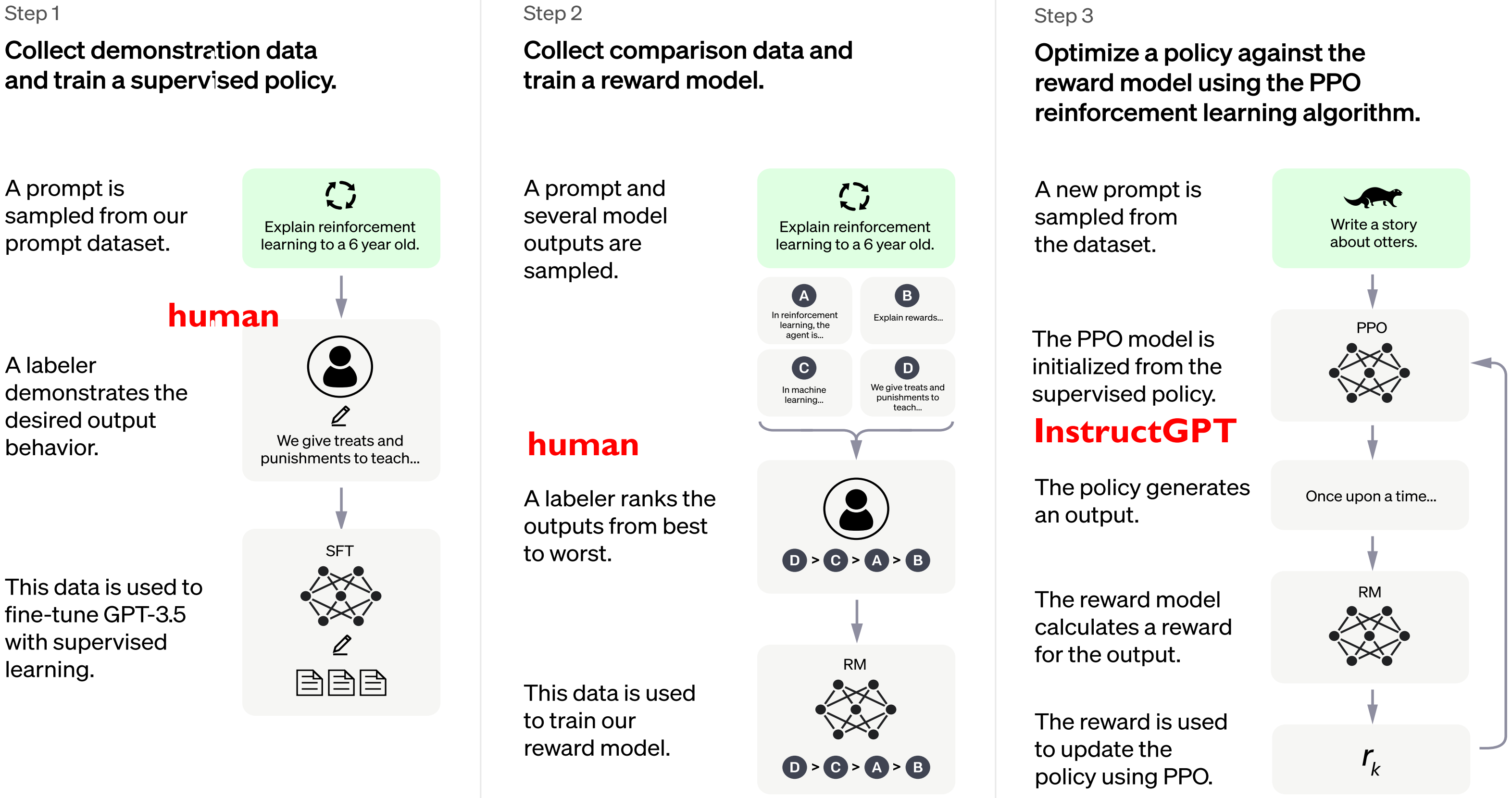
- **Step 1**: Collect demonstration data, and train a supervised policy. Labelers provide demonstrations of the desired behavior on the input prompt distribution. Then, fine-tuning of a pretrained GPT-3 model on this data using supervised learning is carried out.
- **Step 2**: Collect comparison data, and train a reward model. A dataset of comparisons between model outputs is collected: labelers indicate which output they prefer for a given input. A reward model to predict the human-preferred output is then trained.
- **Step 3:** Optimize a policy against the reward model using PPO. We use the output of the RM as a scalar reward. We fine-tune the supervised policy to optimize this reward using the proximal policy optimization (PPO) algorithm (Schulman et al., 2017).

At the heart of ChatGPT (from BART to ChatGPT)

BART Training-steps



ChatGPT Training-steps



Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.

Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

Explain reinforcement learning to a 6 year old.
We give treats and punishments to teach...
SFT
Explain reinforcement learning to a 6 year old.
A
In reinforcement learning, the agent is...
B
Explain rewards...
C
In machine learning...
D
We give treats and punishments to teach...
D > C > A > B
RM
D > C > A > B
Write a story about otters.
PPO
Once upon a time...
RM
r_k

human

human

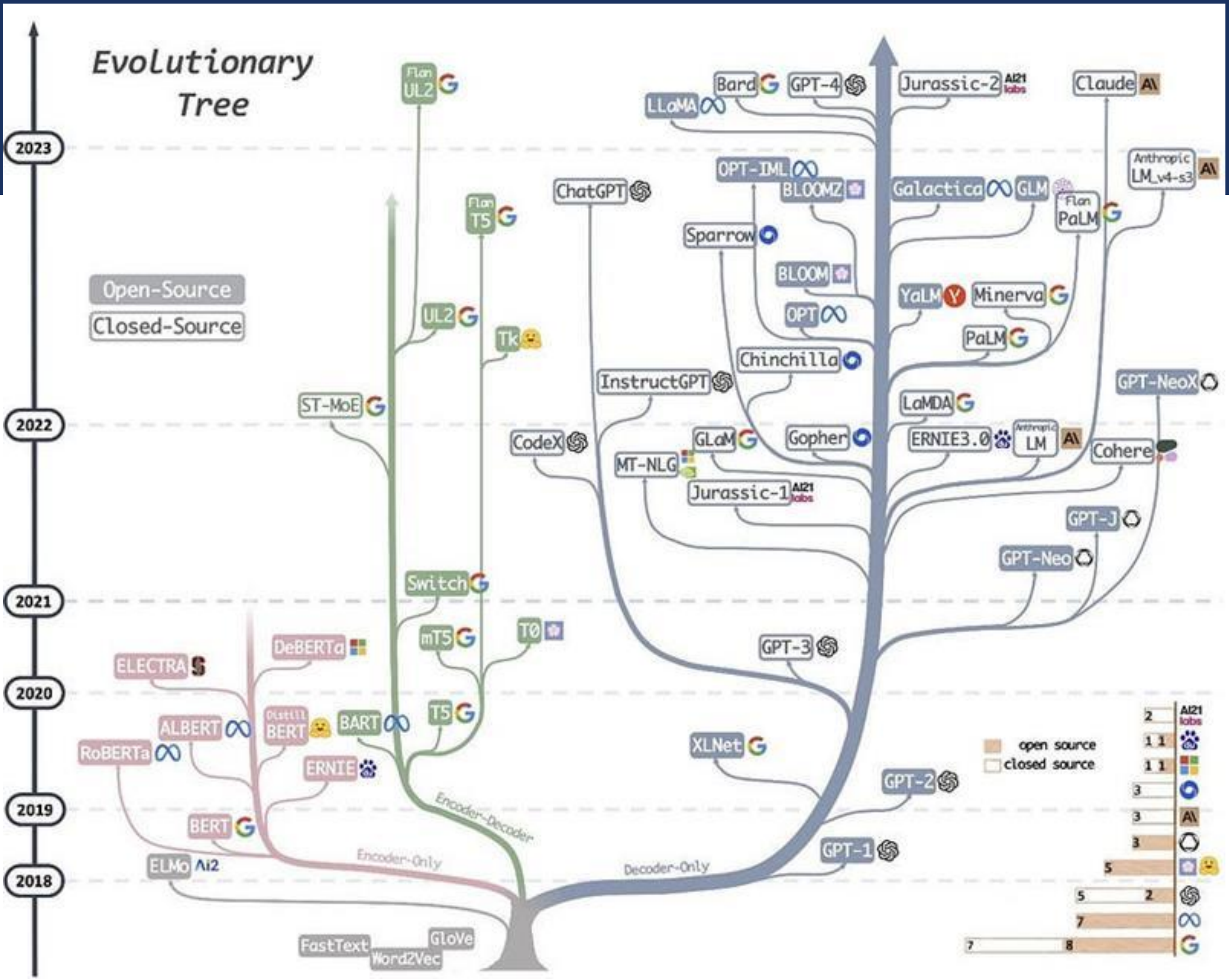
InstructGPT

Fine tune text-davinci-003 to get InstructGPT

The Environment

from Ouyang, L., Wu, J., Jiang, et al. (2022). *Training language models to follow instructions with human feedback*

FOUNDATIONAL MODELS



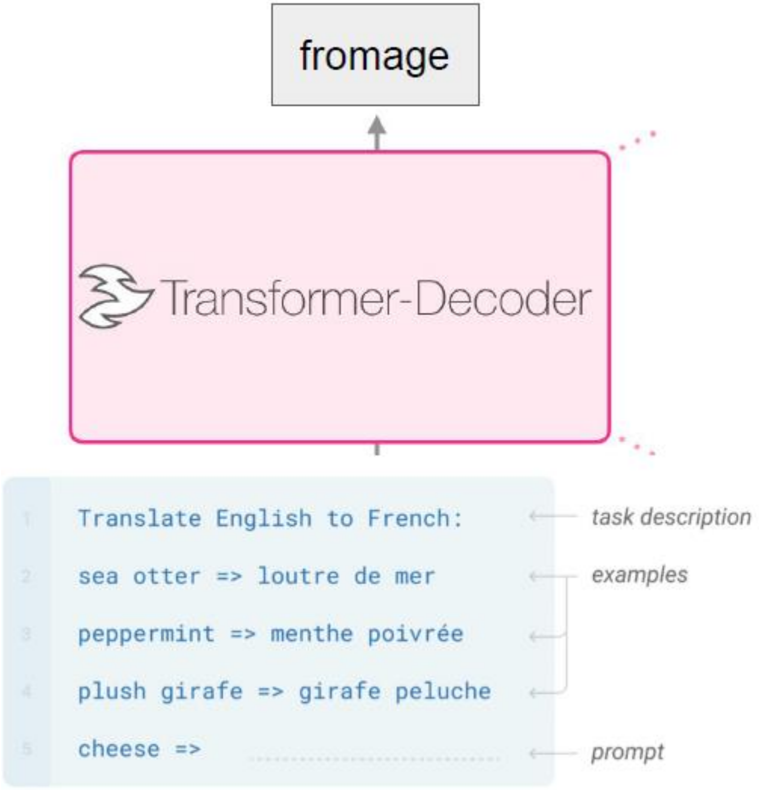
MORE ON PROMPTING

LEARNING MODALITIES

- Fine Tuning (see BERT/BART)
- In-context learning
- Prompting

IN-CONTEXT LEARNING

- Pretrain a large language model on a task
- Manually design a «prompt» that shows how to define a novel taks as a generation task
- There is no need to train further the model, i.e. update model weights



```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

Brown et al. 2020

PROMPTING

- “A good prompt is one that is specific and provides enough context for the model to be able to generate a response that is relevant to the task.” (GPT-3)

- Earliest work in prompts traces back to GPT-1/2 (Radford et al., 2018,2019)
- If LMs are given good prompts they can achieve significant zero-shot performance on NLP tasks ranging from sentiment classification to reading comprehension

PROMPT BASED FINE TUNING

FINE TUNING: more paremeters for the stacked classifier, more examples (even in few-shot scenarios)

PROMPT-BASED FINE TUNING: need for good prompts, no further parameters to tune

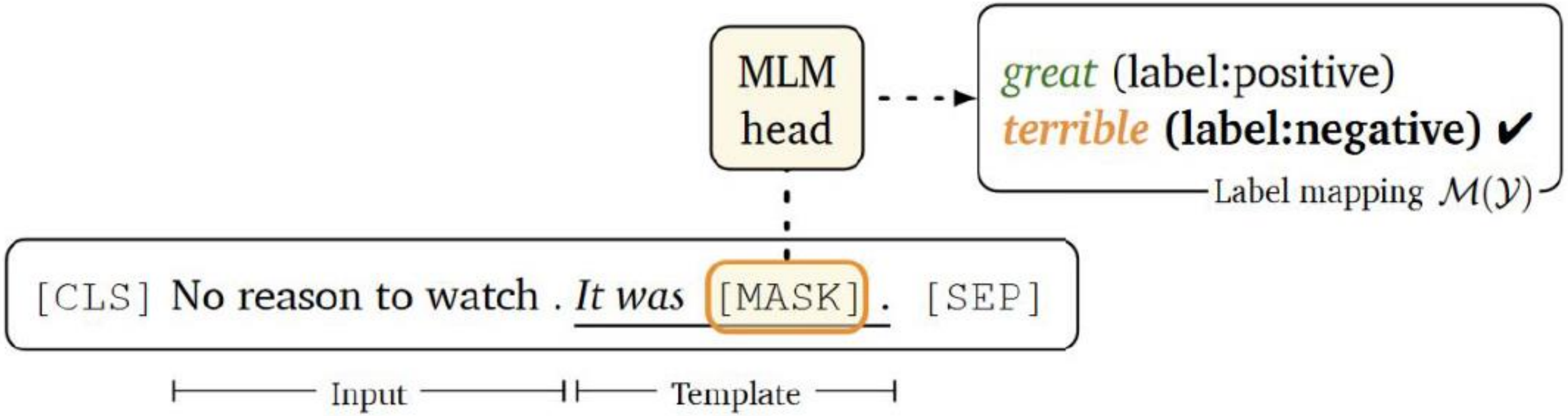
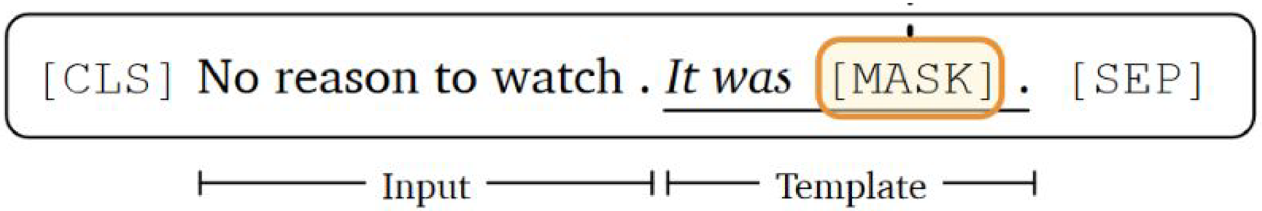

MLM head
great (label:positive)
terrible (label:negative) ✔
Label mapping $\mathcal{M}(\mathcal{Y})$
[CLS] No reason to watch . It was [MASK] . [SEP]
Input
Template

Image Source: Making Pre-trained Language Models Better Few-shot Learners, Gao, et al. 2021

PROMPT-BASED FINE TUNING: THE PROCESS

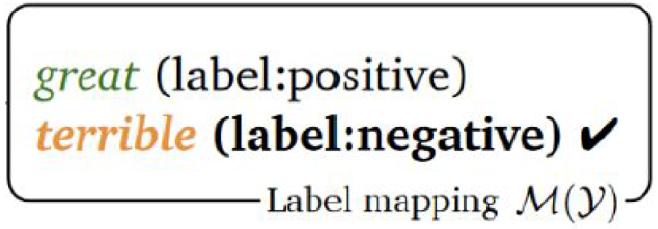
Input: x_1 = No reason to watch.

Step 1. Formulate the downstream task into a (Masked) LM problem using a *template:*



[CLS] No reason to watch . *It was* [MASK] . [SEP]

Step 2. Choose a *label word mapping* $\mathcal{M}$, which maps task labels to individual words.



great (label:positive)
terrible **(label:negative)** ✓

Image Source: Making Pre-trained Language Models Better Few-shot Learners, Gao, et al. 2021

PROMPT-BASED FINE TUNING: THE PROCESS

Step 3. Fine-tune the LM to fill in the correct label word.

$$p(y \mid x_{\text{in}}) = p\left(\texttt{[MASK]} = \mathcal{M}(y) \mid x_{\text{prompt}}\right)$$
$$= \frac{\exp\left(\mathbf{w}_{\mathcal{M}(y)} \cdot \mathbf{h}_{\texttt{[MASK]}}\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(\mathbf{w}_{\mathcal{M}(y')} \cdot \mathbf{h}_{\texttt{[MASK]}}\right)},$$

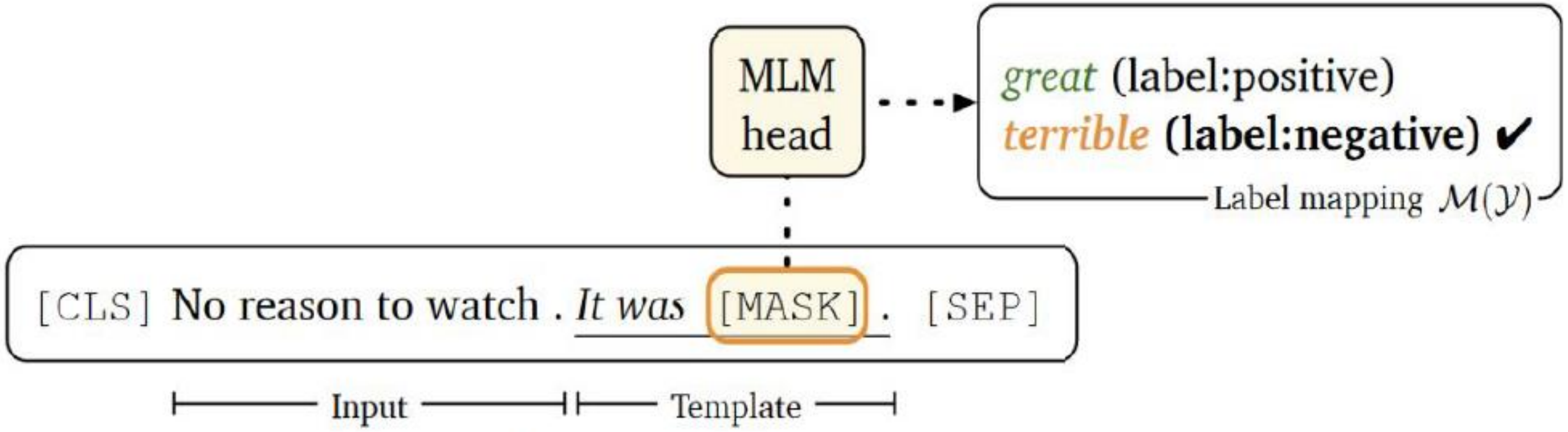
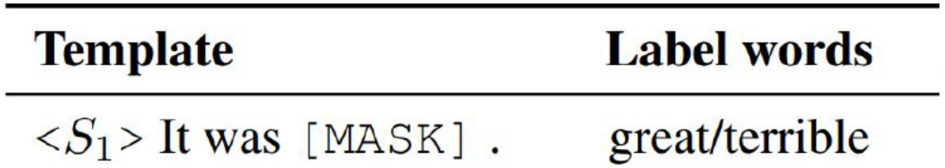

MLM head
great (label:positive)
terrible (label:negative) ✔
Label mapping $\mathcal{M}(\mathcal{Y})$
[CLS] No reason to watch . It was [MASK] . [SEP]
Input
Template

Image Source: Making Pre-trained Language Models Better Few-shot Learners, Gao, et al. 2021

PROMPT BASED FINE TUNING: TASKS

SST-2: sentiment analysis.

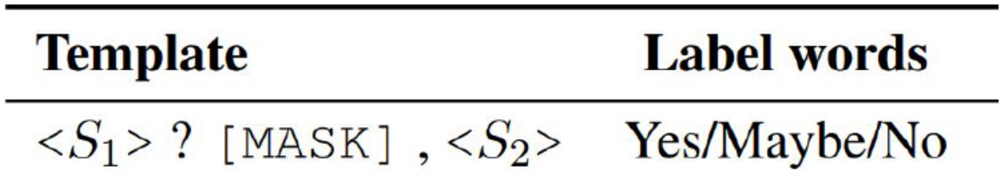
- E.g. **S1** = “The movie is ridiculous”. **Label**: negative.
- Manual prompt:



Template	**Label words**
$<S_1>$ It was `[MASK]` .	great/terrible

SNLI: Natural Language Inference

- **S1** = “A soccer game with multiple males playing”. **S2** = “Some men are playing sport”. **Label**: Entailment.
- Manual prompt:



Template	**Label words**
$<S_1>$? `[MASK]` , $<S_2>$	Yes/Maybe/No

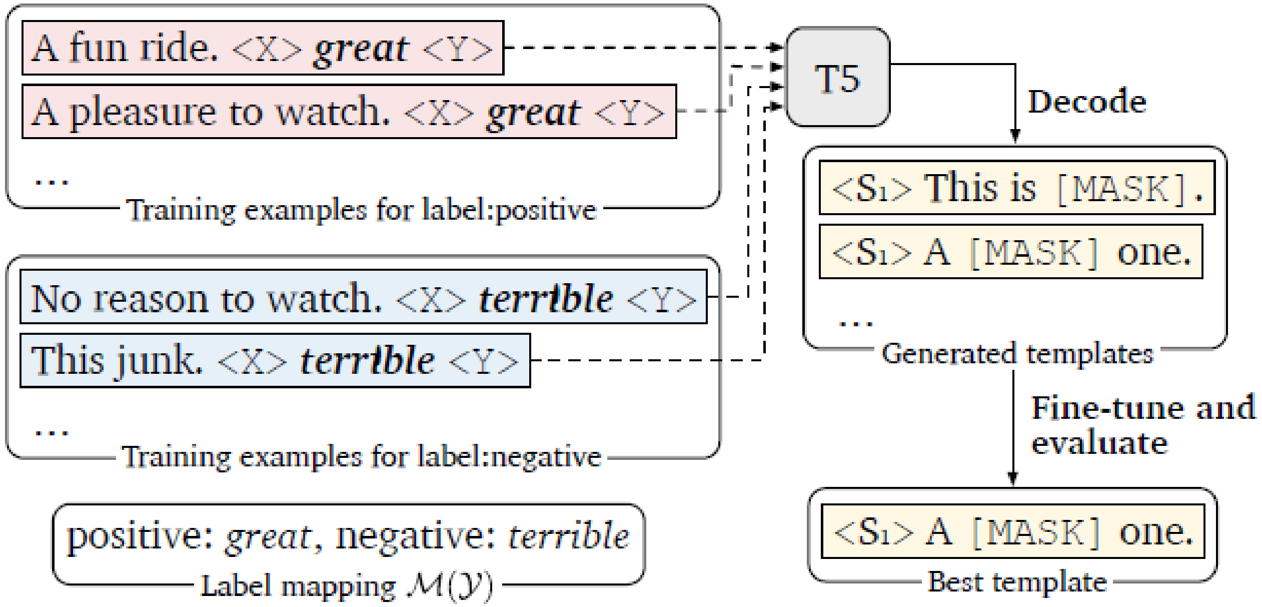
PROMPTING

GPT-3 🤖

Very very large language
Unchanged Model Parameters
Usually human-designed **prompts** and **demonstrations**

PET 🐱

LM-BFF 👯



A fun ride. <X> *great* <Y>

A pleasure to watch. <X> *great* <Y>

…

Training examples for label:positive

No reason to watch. <X> *terrible* <Y>

This junk. <X> *terrible* <Y>

…

Training examples for label:negative

positive: *great*, negative: *terrible*

Label mapping $\mathcal{M}(\mathcal{Y})$

T5

Decode

<S₁> This is [MASK].

<S₁> A [MASK] one.

…

Generated templates

Fine-tune and evaluate

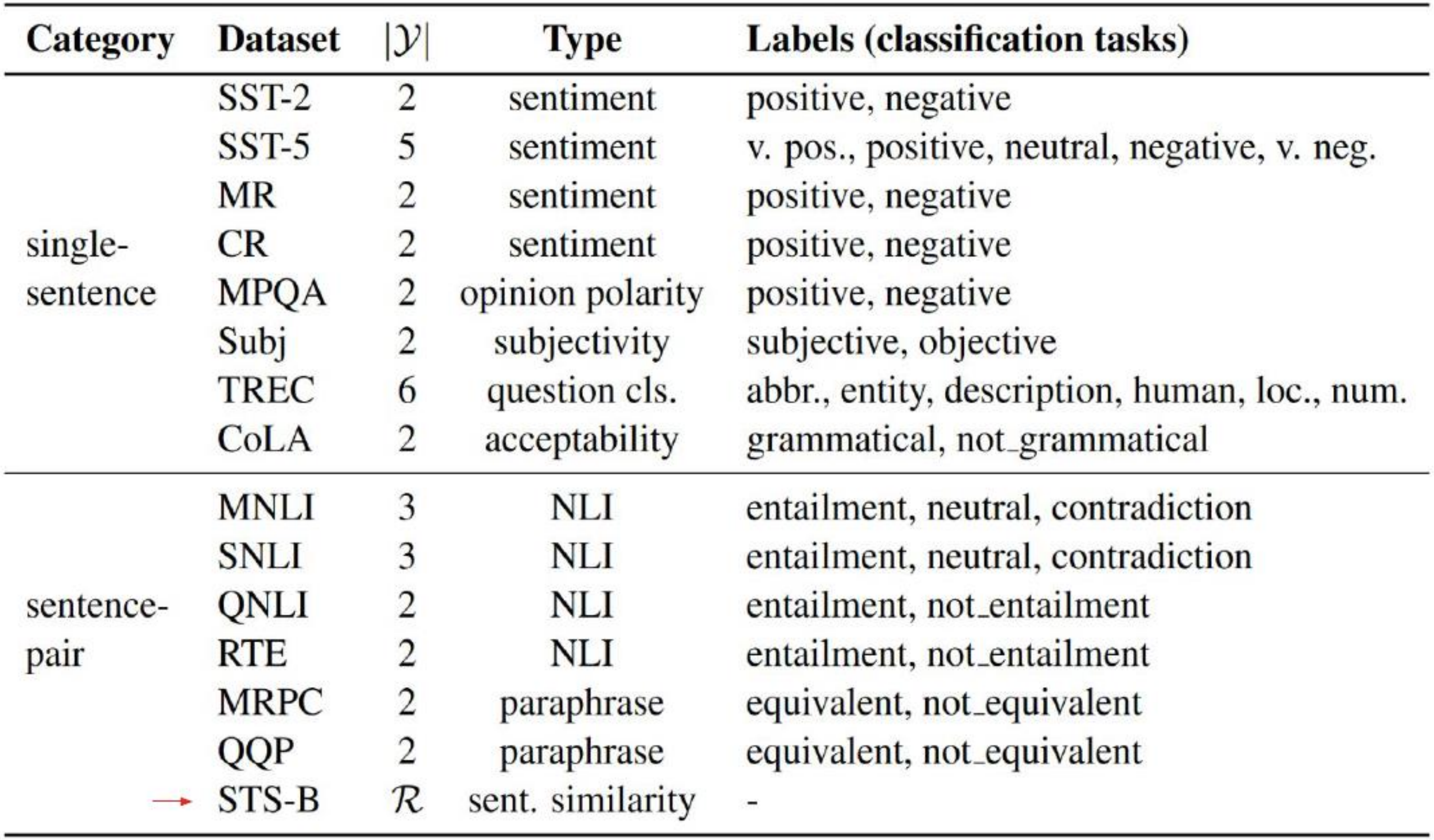
<S₁> A [MASK] one.

Best template

demonstrations

Figure 2: Our approach for template generation.

DATASETS

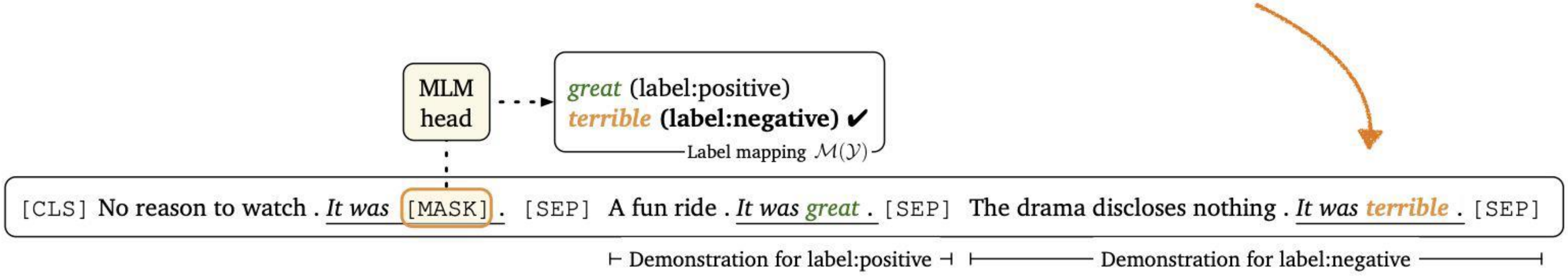


Category	Dataset	$\|\mathcal{Y}\|$	Type	Labels (classification tasks)
single-sentence	SST-2	2	sentiment	positive, negative
	SST-5	5	sentiment	v. pos., positive, neutral, negative, v. neg.
	MR	2	sentiment	positive, negative
	CR	2	sentiment	positive, negative
	MPQA	2	opinion polarity	positive, negative
	Subj	2	subjectivity	subjective, objective
	TREC	6	question cls.	abbr., entity, description, human, loc., num.
	CoLA	2	acceptability	grammatical, not_grammatical
sentence-pair	MNLI	3	NLI	entailment, neutral, contradiction
	SNLI	3	NLI	entailment, neutral, contradiction
	QNLI	2	NLI	entailment, not_entailment
	RTE	2	NLI	entailment, not_entailment
	MRPC	2	paraphrase	equivalent, not_equivalent
	QQP	2	paraphrase	equivalent, not_equivalent
→	STS-B	$\mathcal{R}$	sent. similarity	-

Source: Making Pre-trained Language Models Better Few-shot Learners, Gao, et al. 2021

PROMPT BASED ON DEMONSTRATION

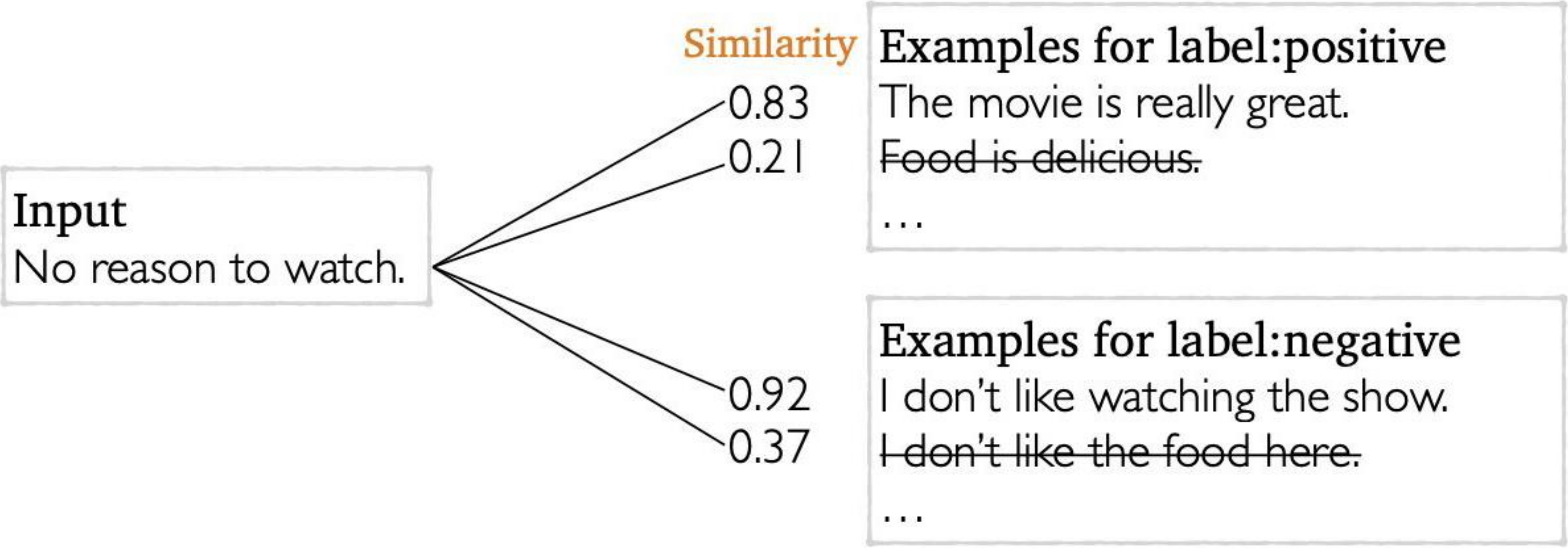
- Demonstration is based on the idea that in few-shot learning you can exemplify a task by using instances from the training set that demonstrate how to solve a task


MLM head
great (label:positive)
terrible (label:negative) ✔
Label mapping $\mathcal{M}(\mathcal{Y})$
[CLS] No reason to watch . It was [MASK] . [SEP] A fun ride . It was great . [SEP] The drama discloses nothing . It was terrible . [SEP]
Demonstration for label:positive
Demonstration for label:negative

Prompt-based fine-tuning with demonstrations

- Selective demonstration (INTUITION): Apply **demonstrations** that are **semantically close** to the input for optimal results

EXAMPLES OF DEMONSTRATIONS



Input

No reason to watch.

Similarity

0.83

0.21

Examples for label:positive

The movie is really great.

~~Food is delicious.~~

…

0.92

0.37

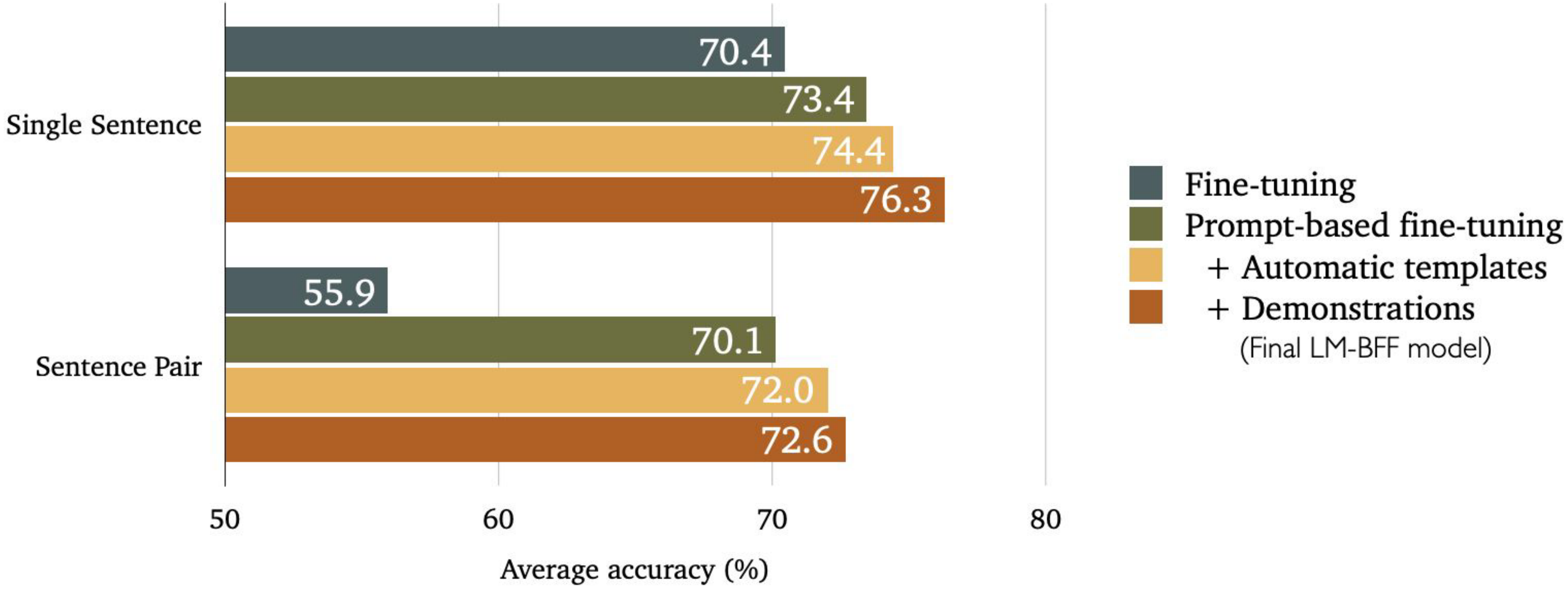
Examples for label:negative

I don't like watching the show.

~~I don't like the food here.~~

…

PROMPTING WITH DEMOSTRATIONS



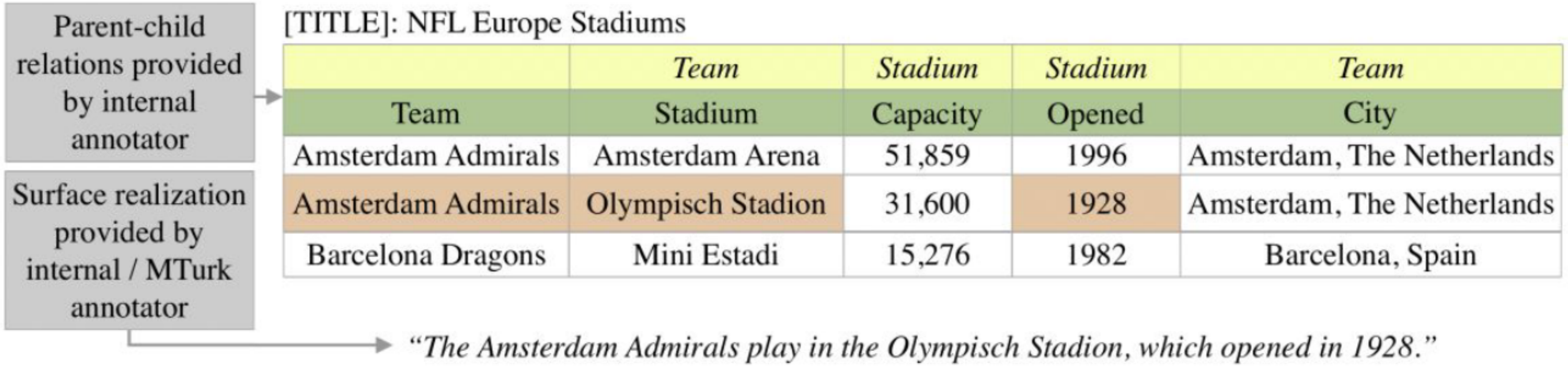
APPLICAZIONI DELLE RETI NEURALI: USE CASES

IMMAGINI: OBJECT DETECTION, ENCODING, MAP COLOURING, RAGS

TASK: TABLES-TO-TEXT

3 Datasets

- DART: Triplets similar to WebNLG, but bigger and on all Wikipedia tables. **(Open domain)**
- (Nan et al. 2021)



Parent-child relations provided by internal annotator

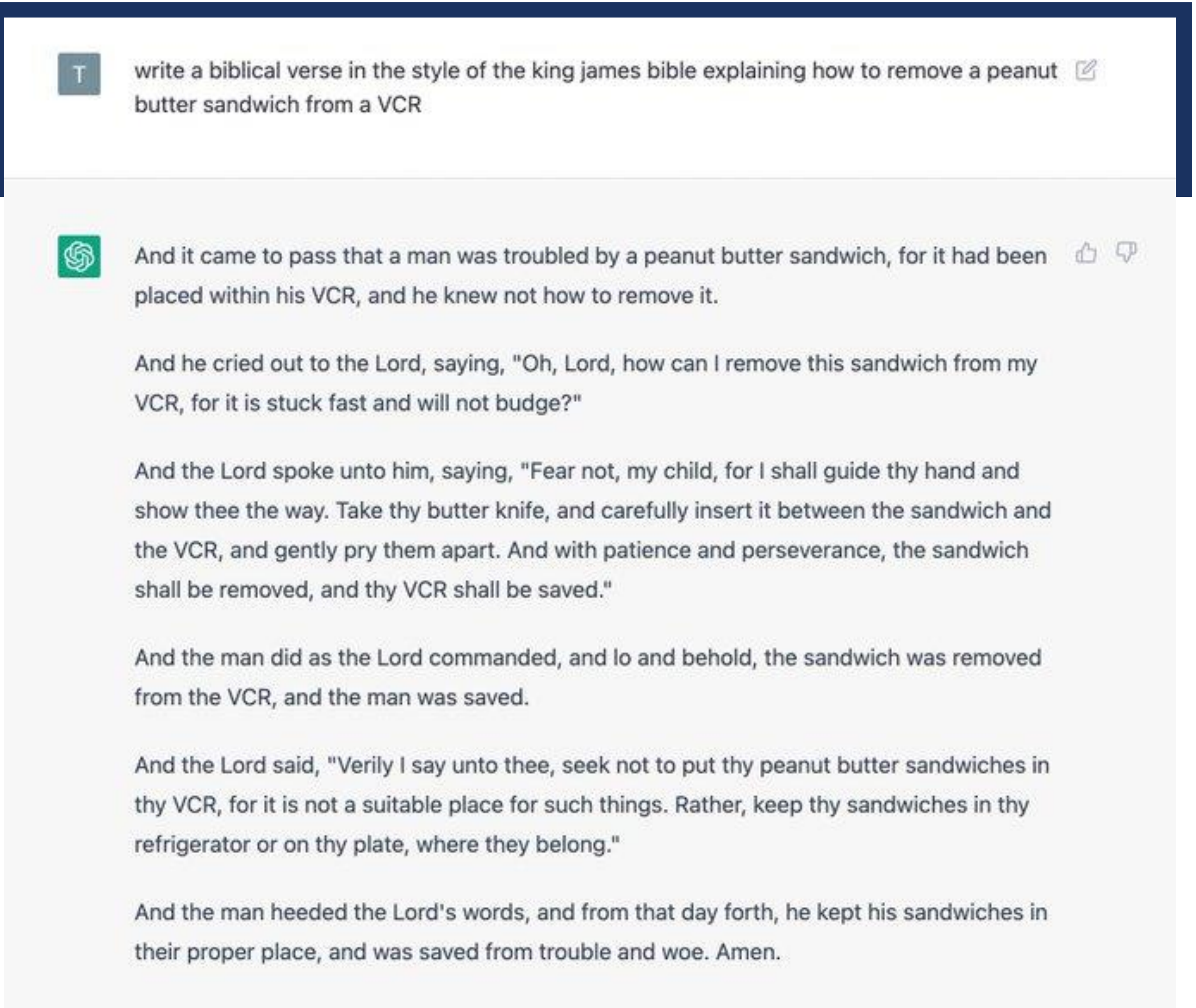
Surface realization provided by internal / MTurk annotator

[TITLE]: NFL Europe Stadiums

	Team	*Stadium*	*Stadium*	*Team*
Team	Stadium	Capacity	Opened	City
Amsterdam Admirals	Amsterdam Arena	51,859	1996	Amsterdam, The Netherlands
Amsterdam Admirals	Olympisch Stadion	31,600	1928	Amsterdam, The Netherlands
Barcelona Dragons	Mini Estadi	15,276	1982	Barcelona, Spain

"The Amsterdam Admirals play in the Olympisch Stadion, which opened in 1928."

TASK: NEURAL AUTHORING



T

write a biblical verse in the style of the king james bible explaining how to remove a peanut butter sandwich from a VCR

And it came to pass that a man was troubled by a peanut butter sandwich, for it had been placed within his VCR, and he knew not how to remove it.

And he cried out to the Lord, saying, "Oh, Lord, how can I remove this sandwich from my VCR, for it is stuck fast and will not budge?"

And the Lord spoke unto him, saying, "Fear not, my child, for I shall guide thy hand and show thee the way. Take thy butter knife, and carefully insert it between the sandwich and the VCR, and gently pry them apart. And with patience and perseverance, the sandwich shall be removed, and thy VCR shall be saved."

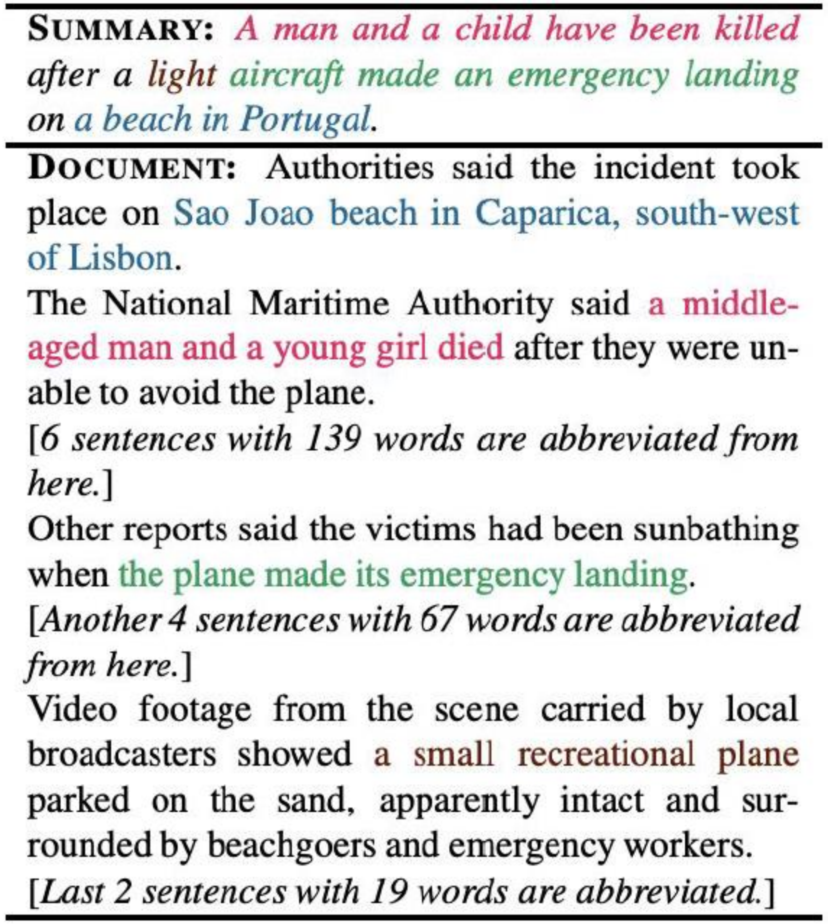
And the man did as the Lord commanded, and lo and behold, the sandwich was removed from the VCR, and the man was saved.

And the Lord said, "Verily I say unto thee, seek not to put thy peanut butter sandwiches in thy VCR, for it is not a suitable place for such things. Rather, keep thy sandwiches in thy refrigerator or on thy plate, where they belong."

And the man heeded the Lord's words, and from that day forth, he kept his sandwiches in their proper place, and was saved from trouble and woe. Amen.

SUMMARIZATION

- Given a long (fragment of) document generate few summary sentences
- XSUM data set
  - BBC News articles
  - Depends on several parts of the text, i.e., not just few sentence selection
  - Designed to amplify the abstractive power of the trasformer decoder
  - (Narayan et al, 2018)



SUMMARY: *A man and a child have been killed after a light aircraft made an emergency landing on a beach in Portugal.*

DOCUMENT: Authorities said the incident took place on Sao Joao beach in Caparica, south-west of Lisbon.

The National Maritime Authority said a middle-aged man and a young girl died after they were unable to avoid the plane.

[*6 sentences with 139 words are abbreviated from here.*]

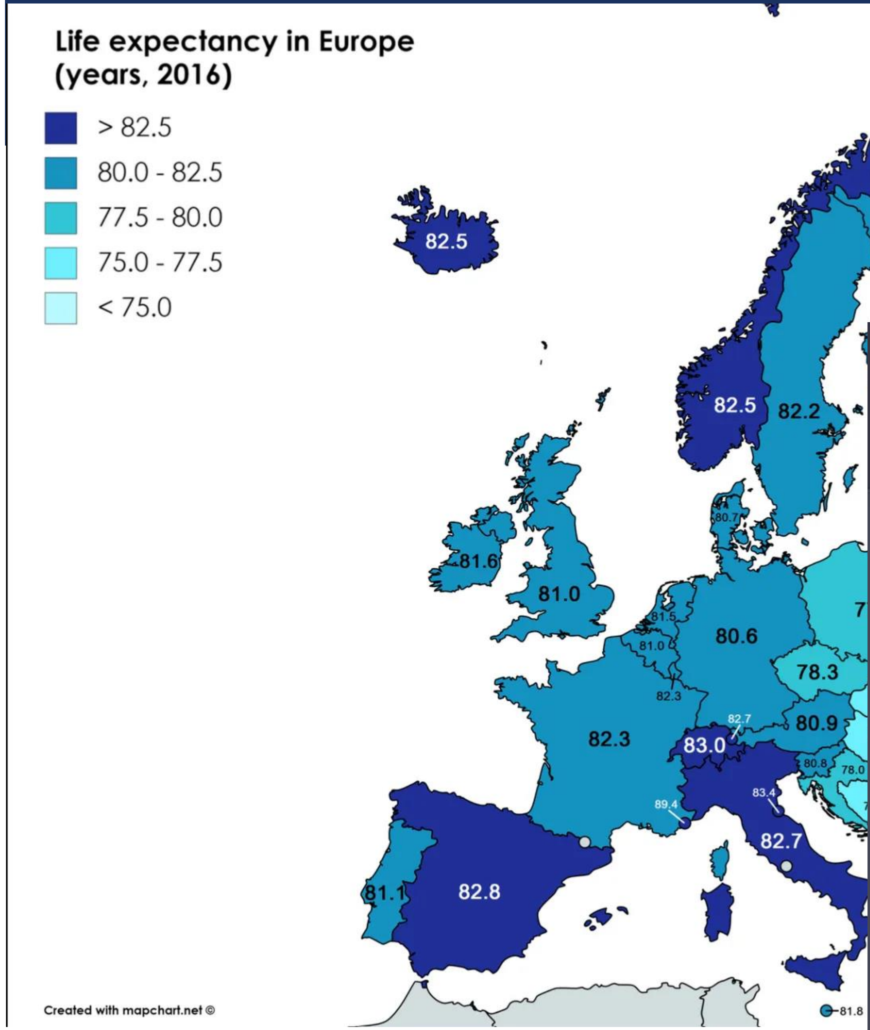
Other reports said the victims had been sunbathing when the plane made its emergency landing.

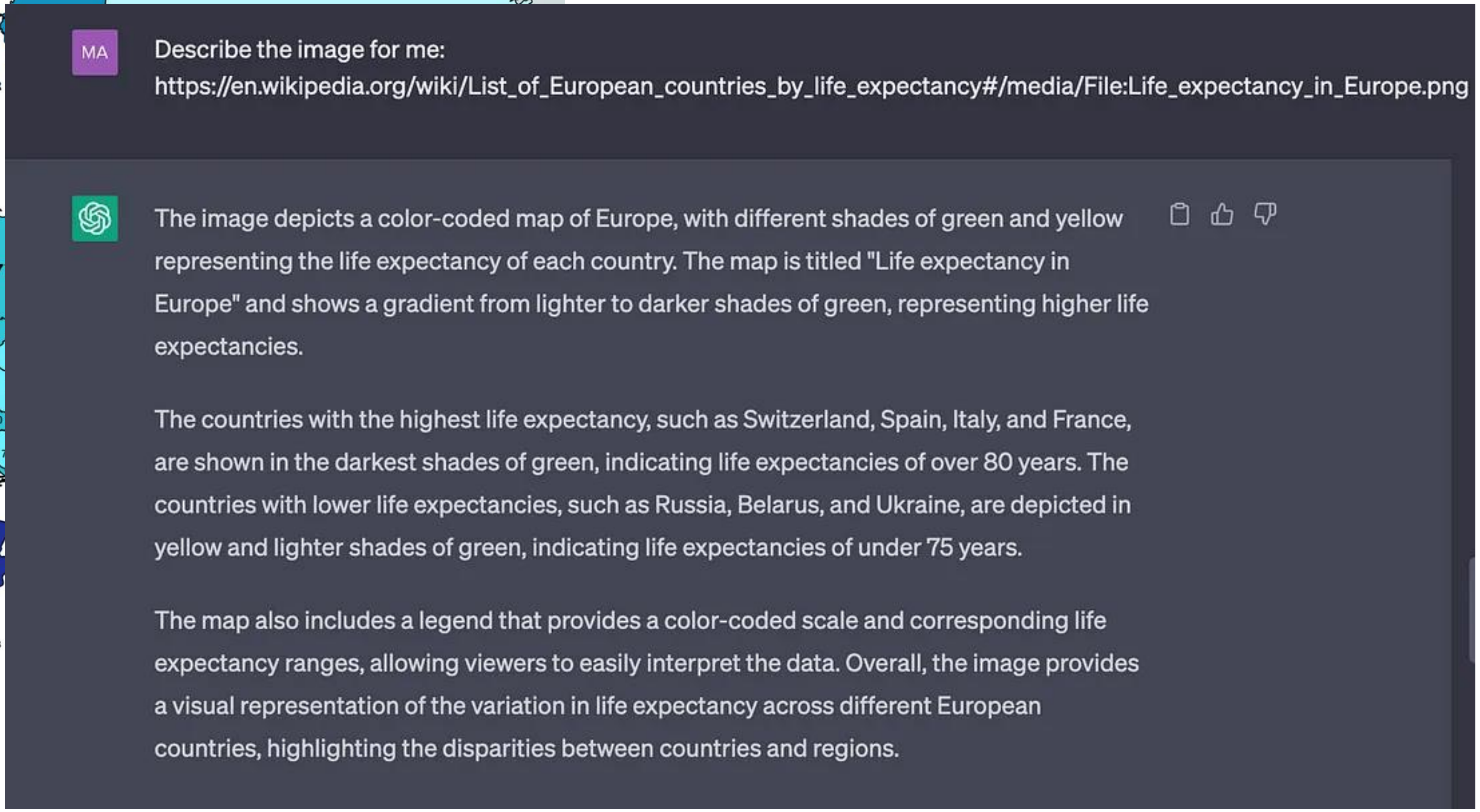
[*Another 4 sentences with 67 words are abbreviated from here.*]

Video footage from the scene carried by local broadcasters showed a small recreational plane parked on the sand, apparently intact and surrounded by beachgoers and emergency workers.

[*Last 2 sentences with 19 words are abbreviated.*]

CHAT GPT4





Describe the image for me:
https://en.wikipedia.org/wiki/List_of_European_countries_by_life_expectancy#/media/File:Life_expectancy_in_Europe.png

The image depicts a color-coded map of Europe, with different shades of green and yellow representing the life expectancy of each country. The map is titled "Life expectancy in Europe" and shows a gradient from lighter to darker shades of green, representing higher life expectancies.

The countries with the highest life expectancy, such as Switzerland, Spain, Italy, and France, are shown in the darkest shades of green, indicating life expectancies of over 80 years. The countries with lower life expectancies, such as Russia, Belarus, and Ukraine, are depicted in yellow and lighter shades of green, indicating life expectancies of under 75 years.

The map also includes a legend that provides a color-coded scale and corresponding life expectancy ranges, allowing viewers to easily interpret the data. Overall, the image provides a visual representation of the variation in life expectancy across different European countries, highlighting the disparities between countries and regions.

OBJECT DETECTION WITH CNNS

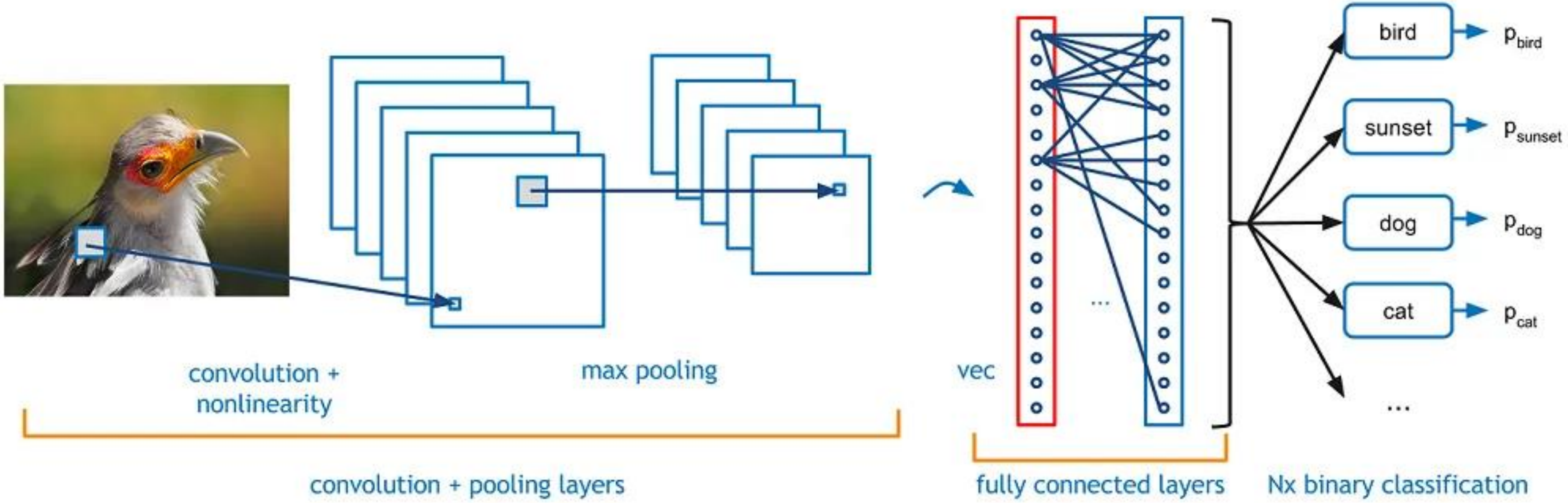
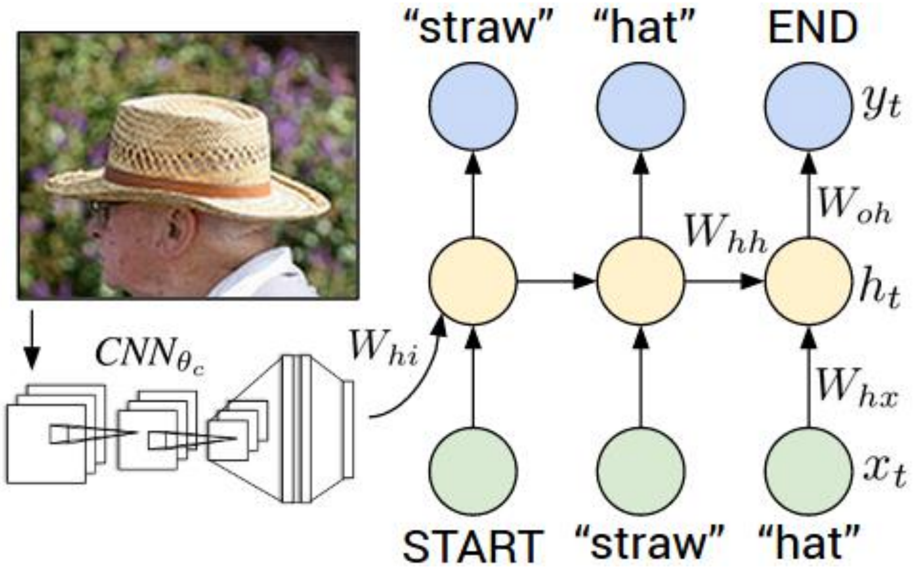


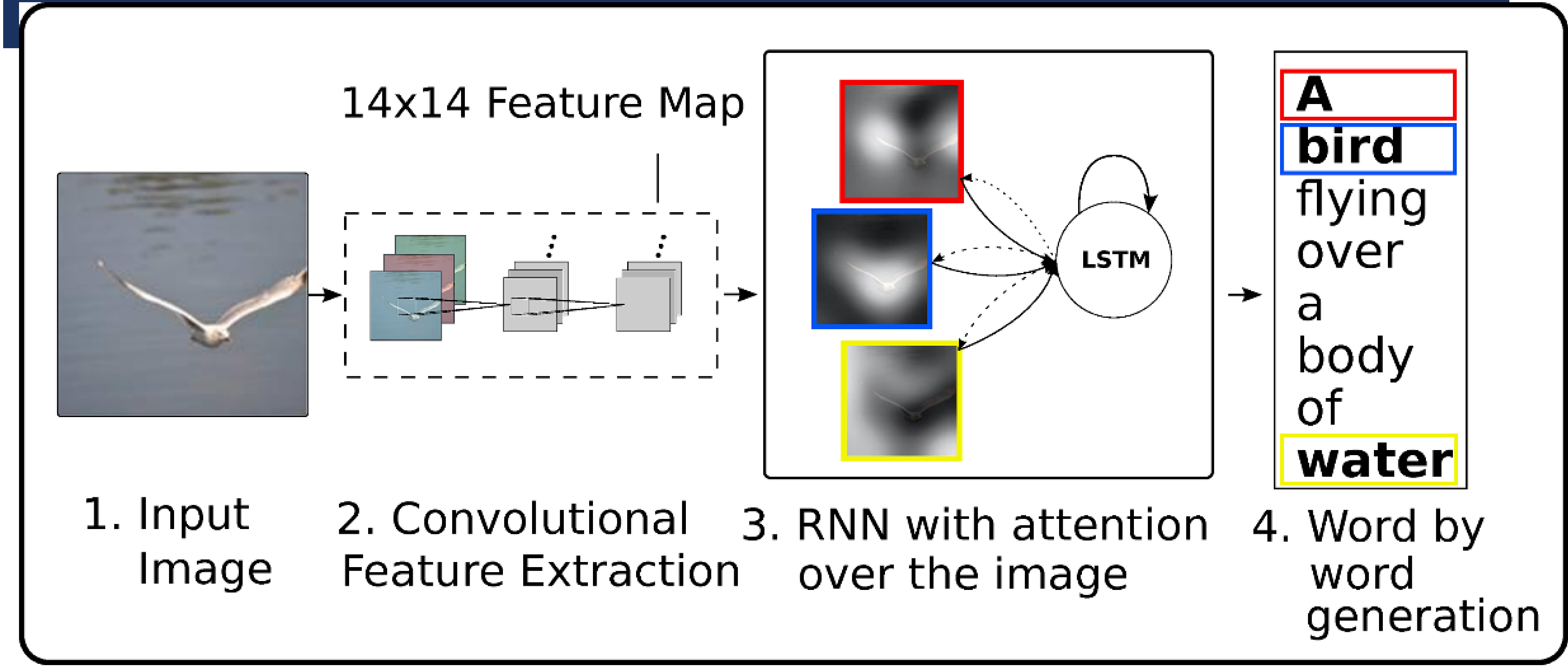
IMAGE CAPTIONING: ADVANCED ARCHITECTURES

- Image to captions
  - Convolutional Neural Network to learn a representation of the image
  - (Bi-directional) Recurrent Neural Network to generate a caption describing the image
    - its input is the representation computed from the CNN
    - its output is a sequence of words, i.e. the caption

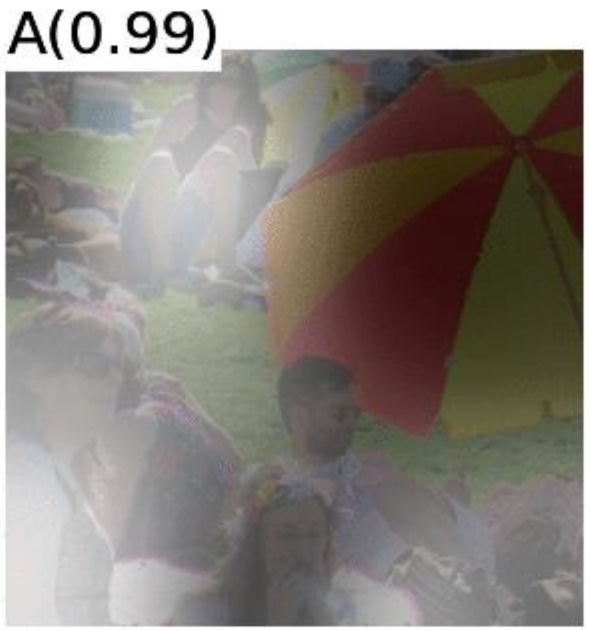


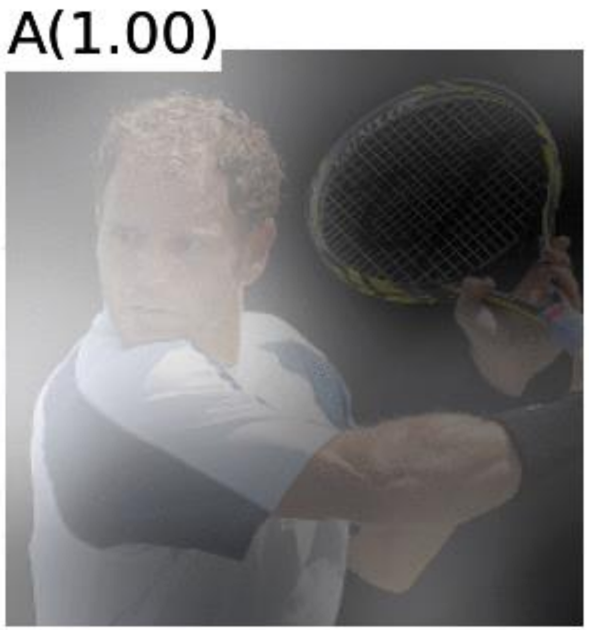
"baseball player is throwing ball in game."

THE ARCHITECTURE

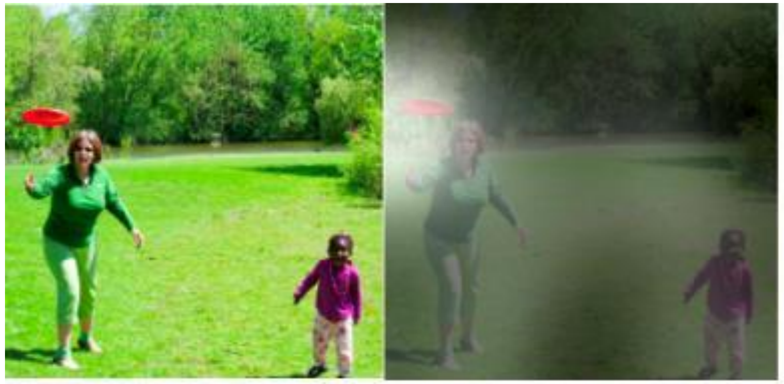

14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

ATTENTION: THE BRIDGE BETWEEN VISION AND LANGUAGE

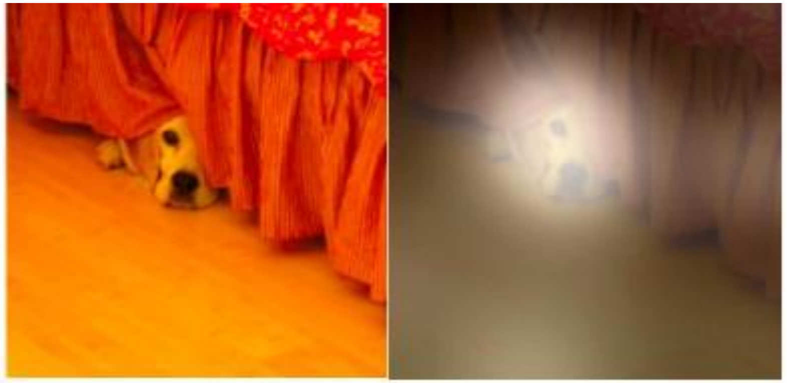




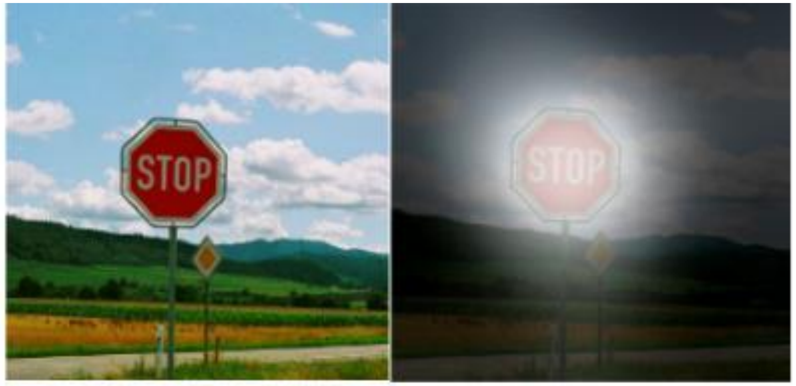
INTEGRATED VISION AND LANGUAGE PROCESSING: IMAGE CAPTIONING AND ATTENTION



A woman is throwing a frisbee in a park.

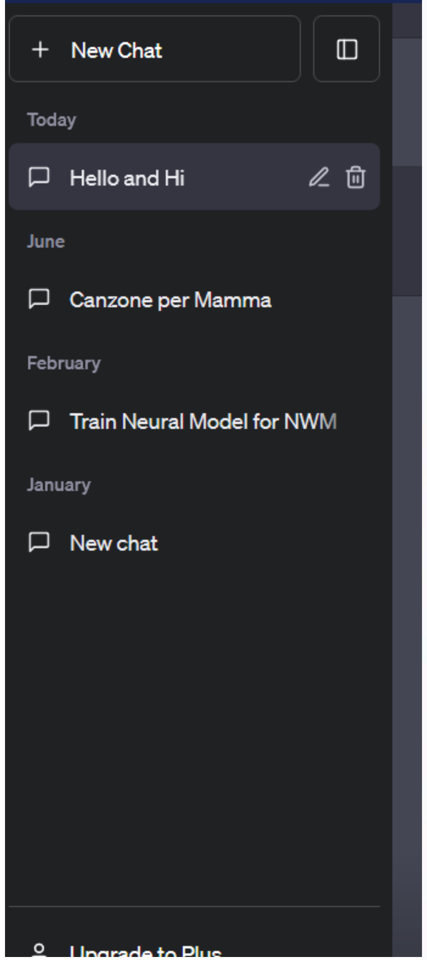


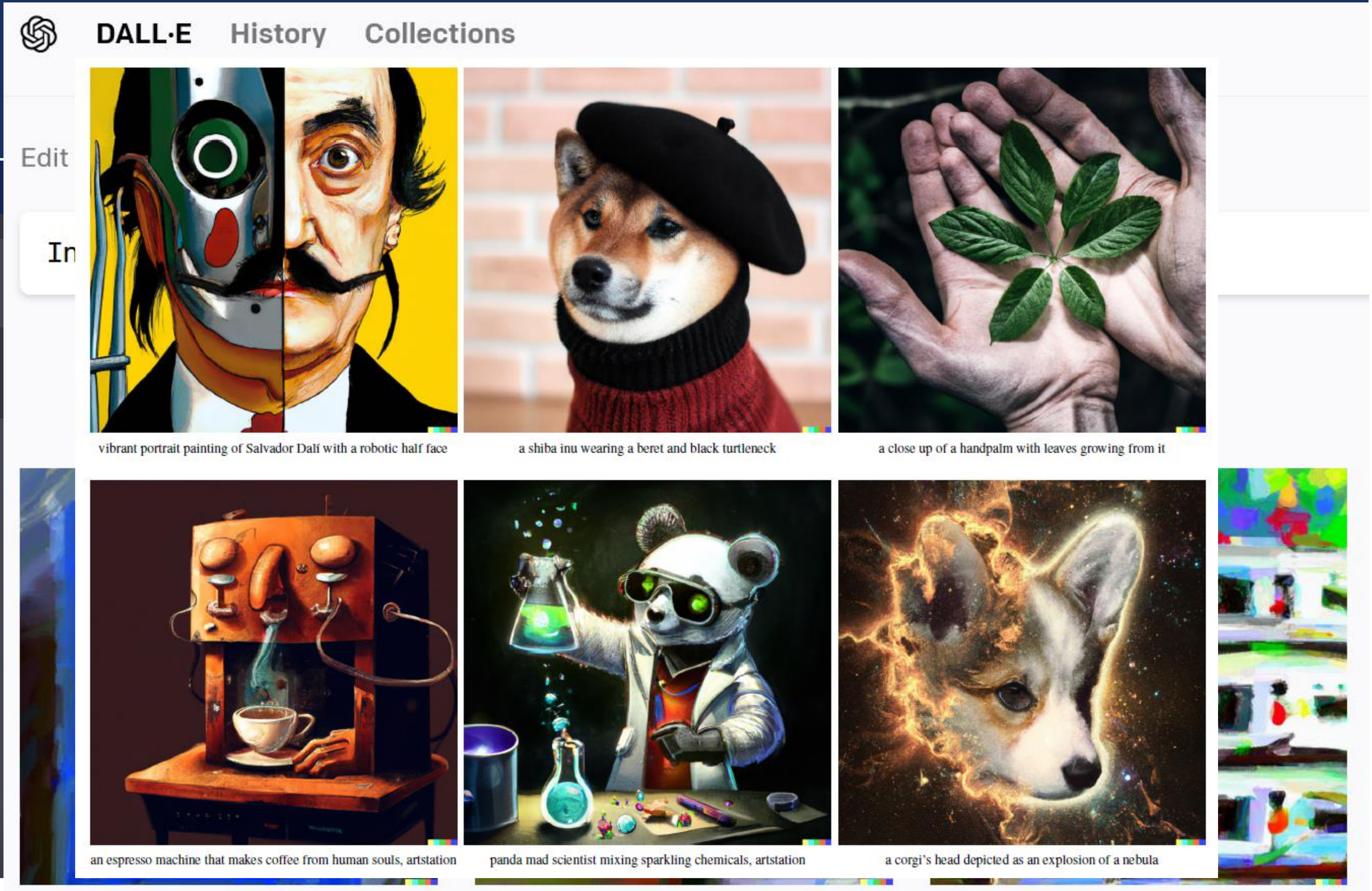
A dog is standing on a hardwood floor.



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

ESEMPI





vibrant portrait painting of Salvador Dalí with a robotic half face

a shiba inu wearing a beret and black turtleneck

a close up of a handpalm with leaves growing from it

an espresso machine that makes coffee from human souls, artstation

panda mad scientist mixing sparkling chemicals, artstation

a corgi's head depicted as an explosion of a nebula

NEURAL ENCODING-DECODING FOR DALL-E

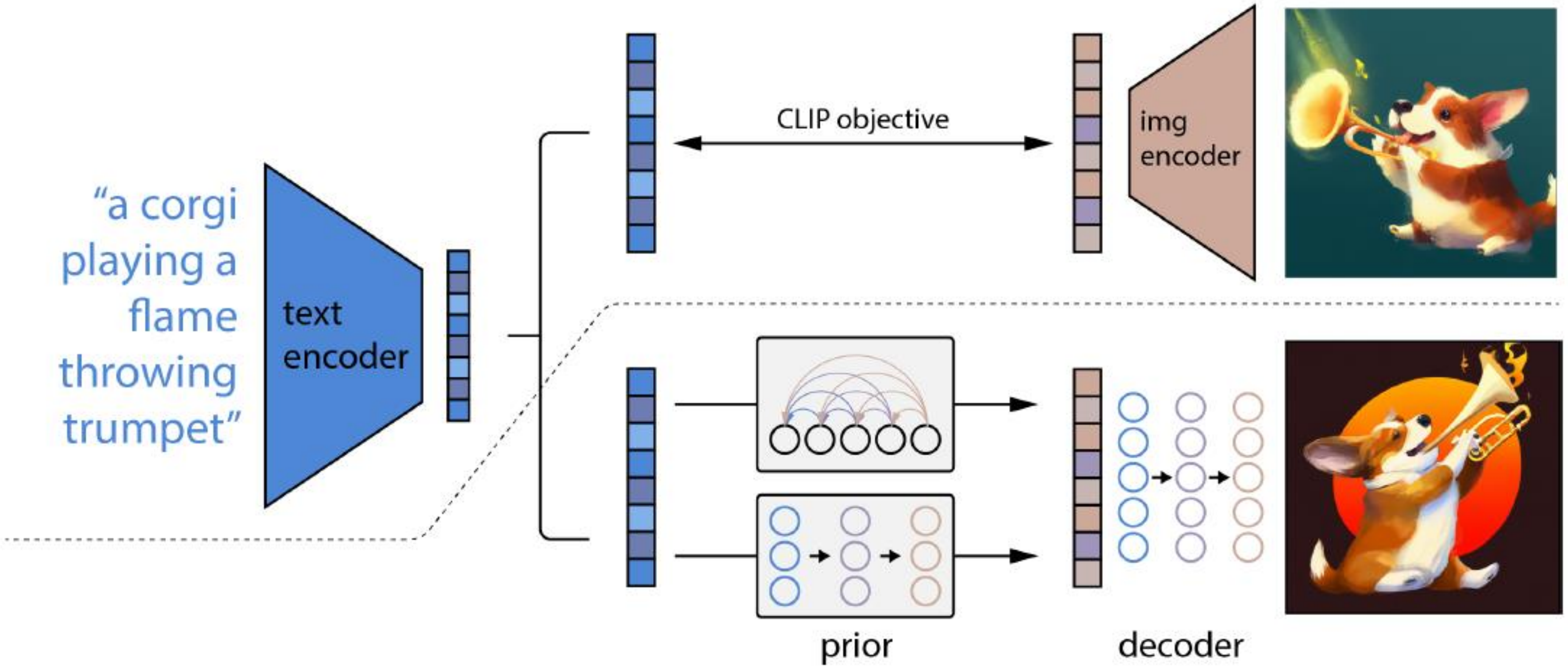


Figure 2: A high-level overview of unCLIP. Above the dotted line, we depict the CLIP training process, through which we learn a joint representation space for text and images. Below the dotted line, we depict our text-to-image generation process: a CLIP text embedding is first fed to an autoregressive or diffusion prior to produce an image embedding, and then this embedding is used to condition a diffusion decoder which produces a final image. Note that the CLIP model is frozen during training of the prior and decoder.

MULTIMODAL NNS: INTEGRATING IMAGE AND TEXTS IN CLIP

- Object Recognition usually employs ad hoc training data sets implying ad hoc CNN models
- The paper (*) demonstrates that the simple pre-training task of predicting **which caption goes with which image** is an efficient and scalable way to learn SOTA image representations from scratch on a dataset of 400 million (image, text) pairs collected from the internet.
- **After pre-training**, natural language is used to reference learned visual concepts (or describe new ones) enabling zero-shot transfer of the model to downstream tasks.
- **Zero-shot learning**: solving an object recognition task without ANY training example
- The IDEA: Optimizing the behaviours of image classifiers trained with natural language supervision at large scale.

(*) Learning Transferable Visual Models From Natural Language Supervision, Redford et al, 2021, https://arxiv.org/abs/2103.00020v1

CLIP
(CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING)

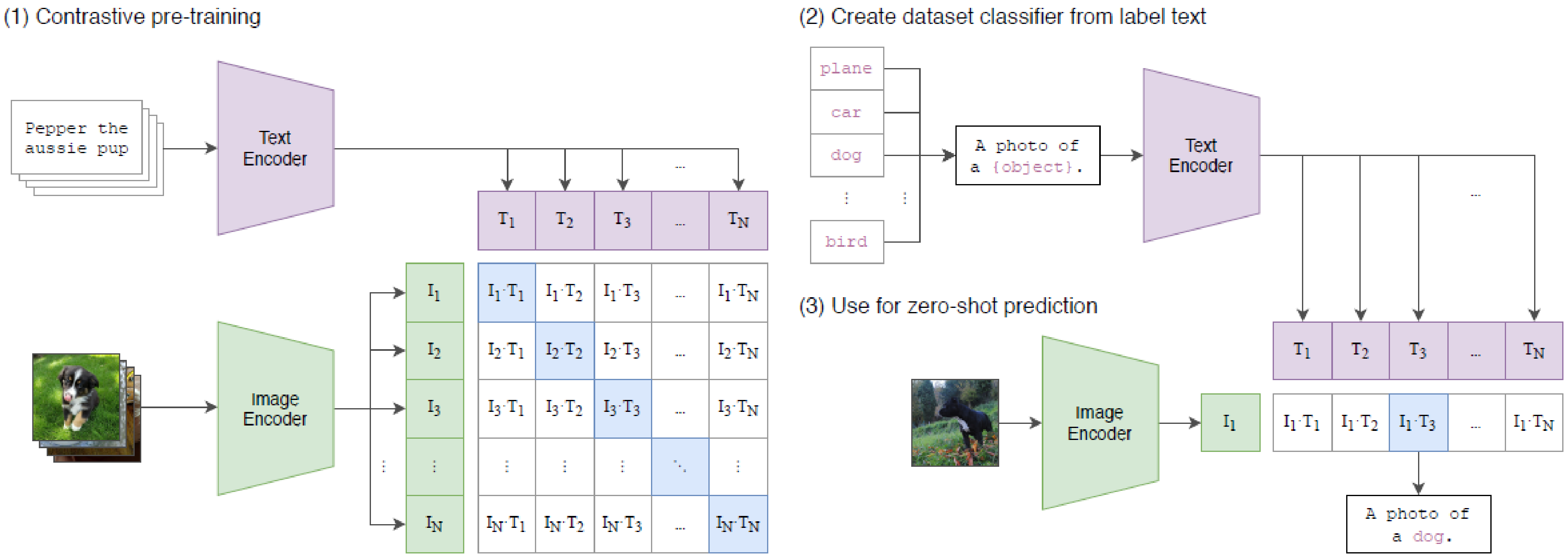
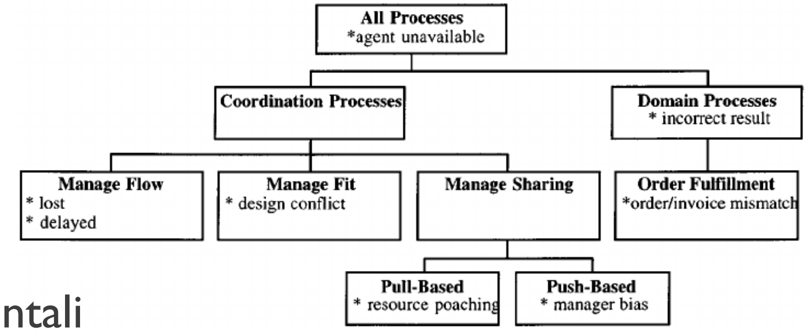
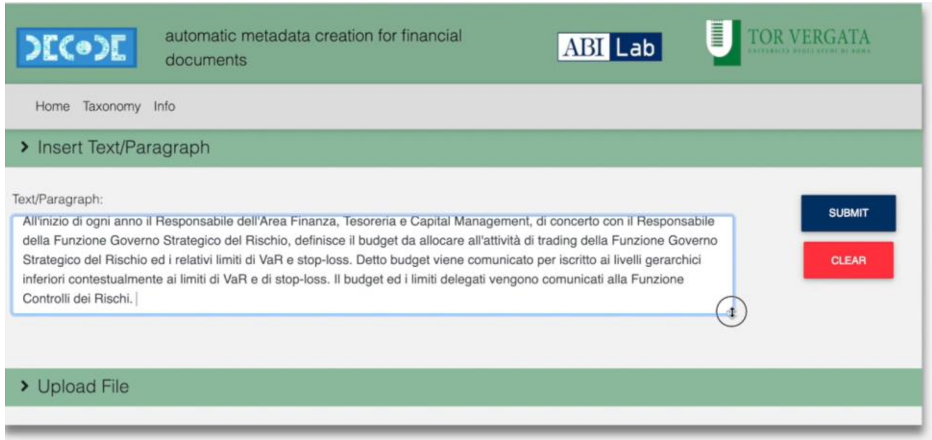

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T1 T2 T3 ... TN
I1 I2 I3 ... IN
I1·T1 I1·T2 I1·T3 ... I1·TN
I2·T1 I2·T2 I2·T3 ... I2·TN
I3·T1 I3·T2 I3·T3 ... I3·TN
IN·T1 IN·T2 IN·T3 ... IN·TN
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I1
I1·T1 I1·T2 I1·T3 ... I1·TN
A photo of a dog.

Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

BANKING: ABILABERT IN DECODE



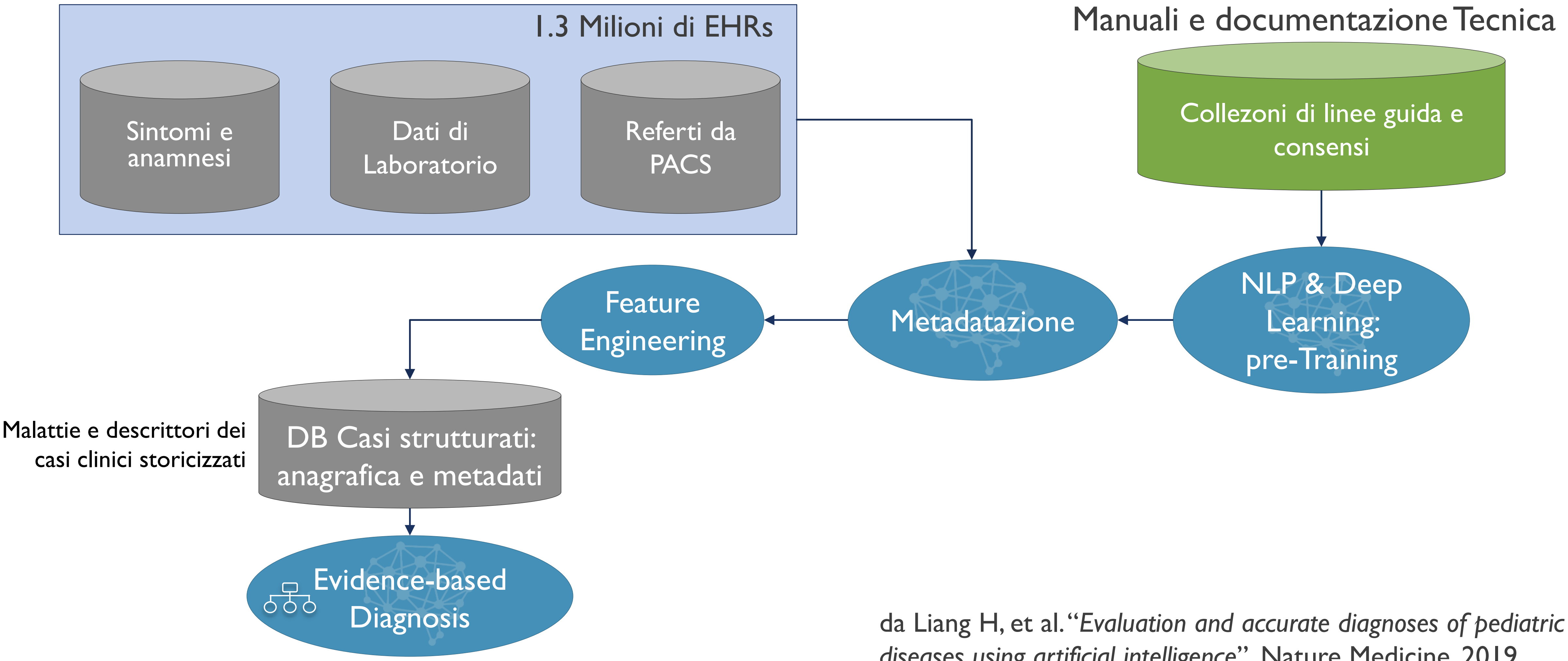
- 5 banche coordinate da ABILAB
- Una Process Taxonomy condivisa e differenti Basi di Dati Documentali
- Automatic Text-driven Process Mapping basato su reti neurali Trasformers



Text/Paragraph:

All'inizio di ogni anno il Responsabile dell'Area Finanza, Tesoreria e Capital Management, di concerto con il Responsabile della Funzione Governo Strategico del Rischio, definisce il budget da allocare all'attività di trading della Funzione Governo Strategico del Rischio ed i relativi limiti di VaR e stop-loss. Detto budget viene comunicato per iscritto ai livelli gerarchici inferiori contestualmente ai limiti di VaR e di stop-loss. Il budget ed i limiti delegati vengono comunicati alla Funzione Controlli dei Rischi.

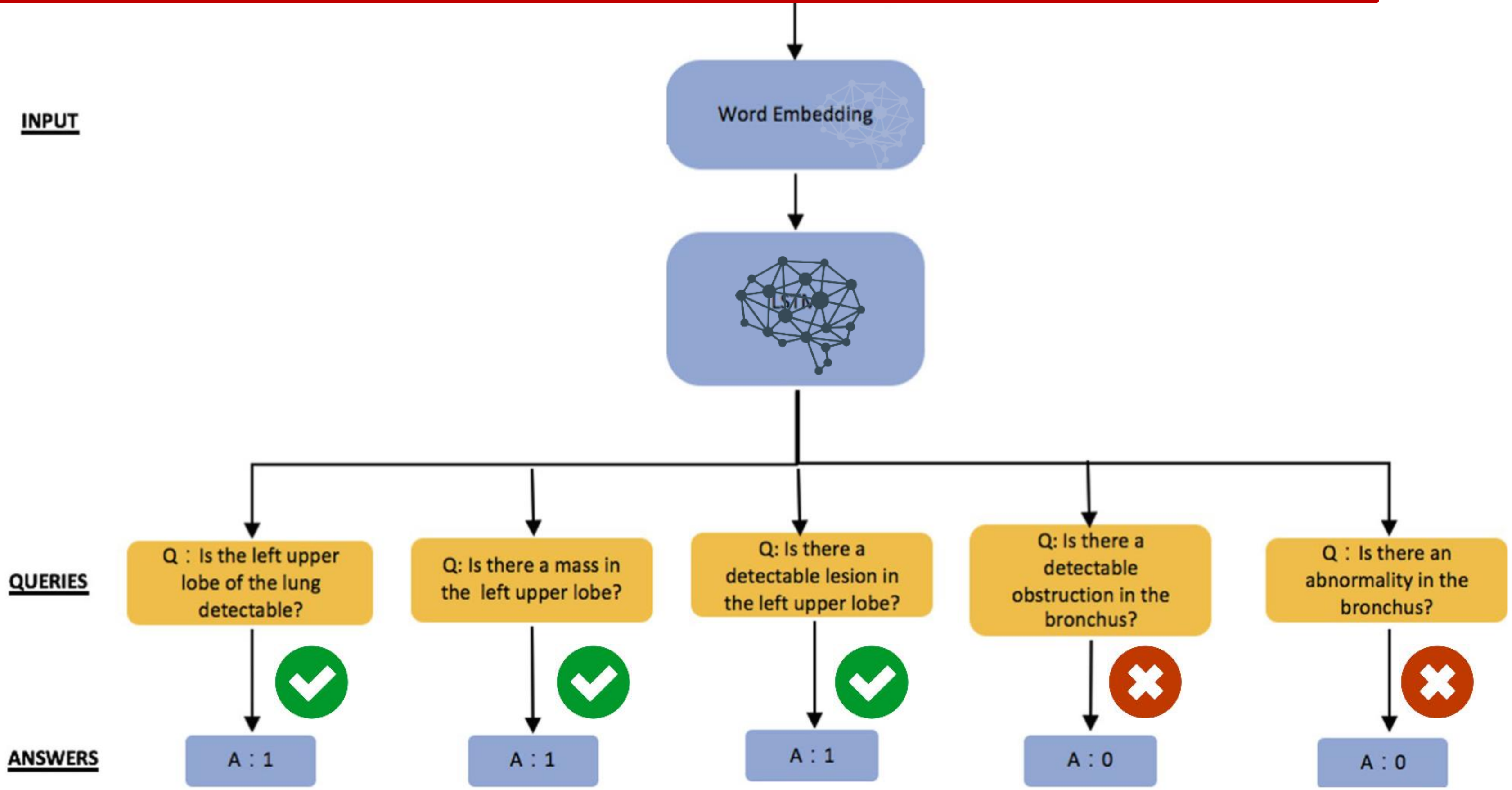
DIAGNOSI MALATTIE PEDIATRICHE: UN WORKFLOW ORIENTATO AL ML


1.3 Milioni di EHRs
Sintomi e anamnesi
Dati di Laboratorio
Referti da PACS
Manuali e documentazione Tecnica
Collezoni di linee guida e consensi
NLP & Deep Learning: pre-Training
Metadatazione
Feature Engineering
DB Casi strutturati: anagrafica e metadati
Malattie e descrittori dei casi clinici storicizzati
Evidence-based Diagnosis

da Liang H, et al. "*Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence*", Nature Medicine, 2019

MEDICAL INFORMATION EXTRACTION

INPUT: “Si **osserva *una lesione*** nel ***lobo superiore sinistro*** del ***polmone del paziente*** . ...”



INPUT

Word Embedding

LSTM

QUERIES

Q : Is the left upper lobe of the lung detectable?

Q: Is there a mass in the left upper lobe?

Q: Is there a detectable lesion in the left upper lobe?

Q: Is there a detectable obstruction in the bronchus?

Q : Is there an abnormality in the bronchus?

ANSWERS

A : 1

A : 1

A : 1

A : 0

A : 0

EVIDENCE BASED DIAGNOSIS: RISULTATI (11,926 PAZIENTI)

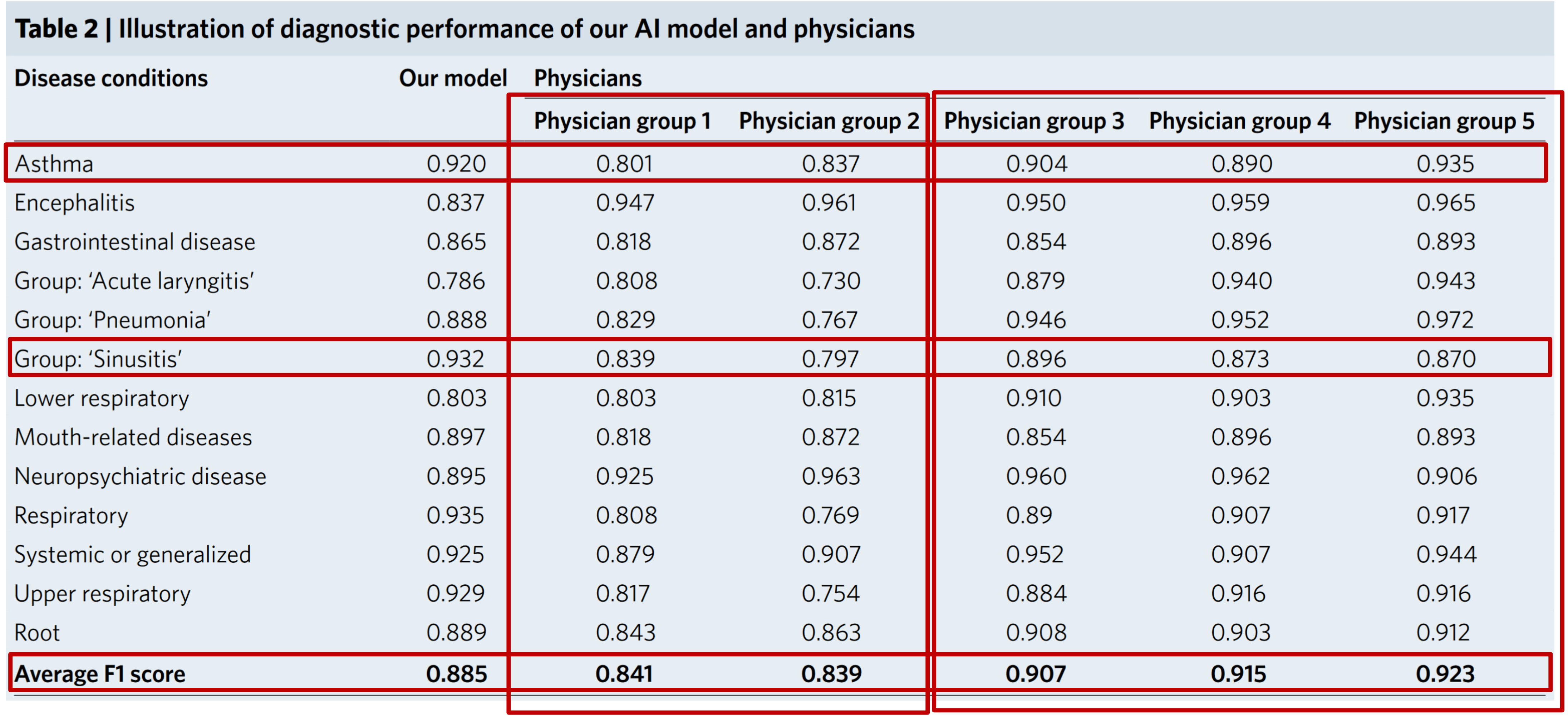


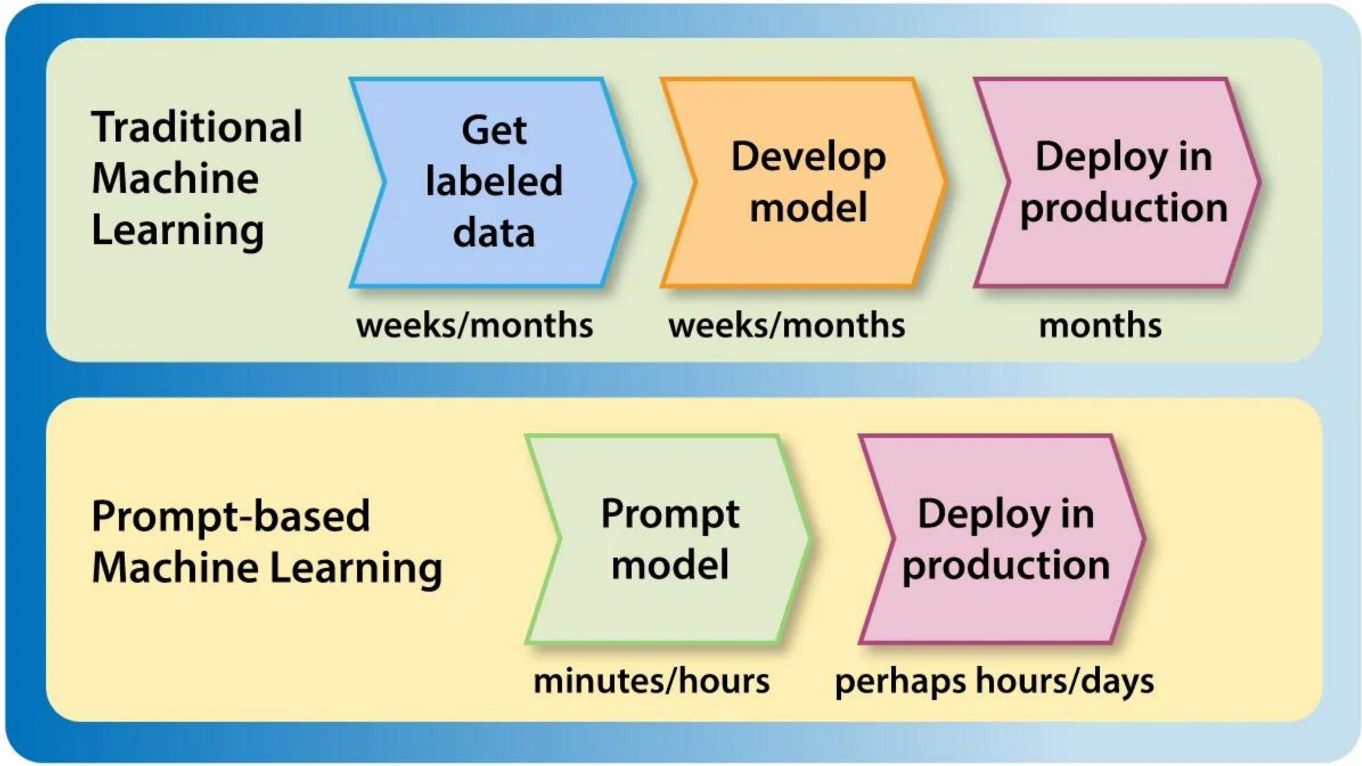
Table 2 | Illustration of diagnostic performance of our AI model and physicians

Disease conditions	**Our model**	**Physicians**				
		Physician group 1	**Physician group 2**	**Physician group 3**	**Physician group 4**	**Physician group 5**
Asthma	0.920	0.801	0.837	0.904	0.890	0.935
Encephalitis	0.837	0.947	0.961	0.950	0.959	0.965
Gastrointestinal disease	0.865	0.818	0.872	0.854	0.896	0.893
Group: 'Acute laryngitis'	0.786	0.808	0.730	0.879	0.940	0.943
Group: 'Pneumonia'	0.888	0.829	0.767	0.946	0.952	0.972
Group: 'Sinusitis'	0.932	0.839	0.797	0.896	0.873	0.870
Lower respiratory	0.803	0.803	0.815	0.910	0.903	0.935
Mouth-related diseases	0.897	0.818	0.872	0.854	0.896	0.893
Neuropsychiatric disease	0.895	0.925	0.963	0.960	0.962	0.906
Respiratory	0.935	0.808	0.769	0.89	0.907	0.917
Systemic or generalized	0.925	0.879	0.907	0.952	0.907	0.944
Upper respiratory	0.929	0.817	0.754	0.884	0.916	0.916
Root	0.889	0.843	0.863	0.908	0.903	0.912
Average F1 score	**0.885**	**0.841**	**0.839**	**0.907**	**0.915**	**0.923**

LARGE LANGUAGE MODELS

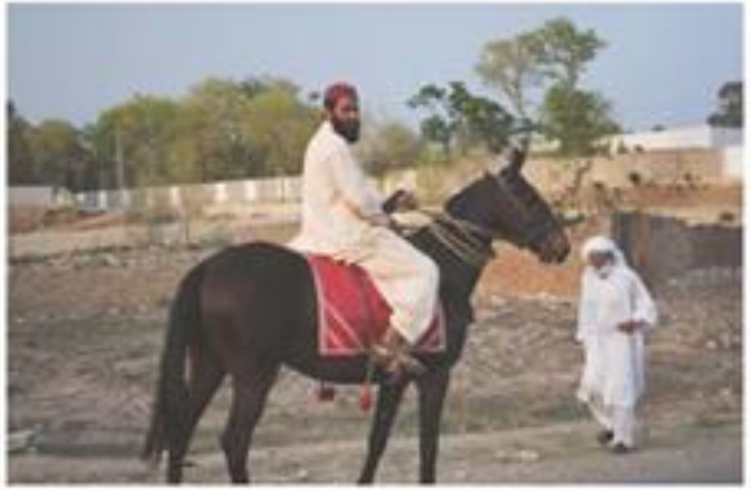
TRENDS

TRENDS …

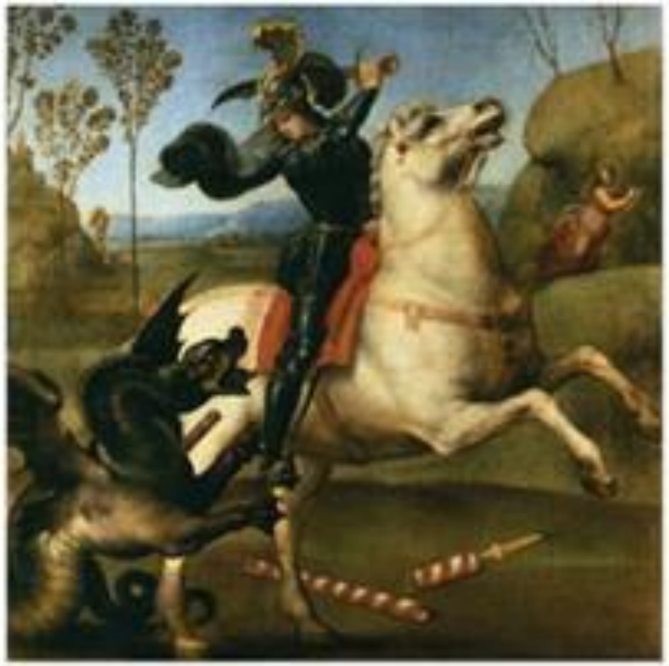

Traditional Machine Learning
Get labeled data
weeks/months
Develop model
weeks/months
Deploy in production
months
Prompt-based Machine Learning
Prompt model
minutes/hours
Deploy in production
perhaps hours/days

RIFLESSIONI

- Competenza, Razionalità ed Onniscenza
  - Un sistema di AI generativa ha una SIGNIFICATIVA COMPETENZA LINGUISTICA in analogia con i parlanti delle diverse lingue in cui esso è stato addestrato
  - E' RAZIONALE in senso linguistico poiché conosce le regole della comunicazione e le usa in modo *utile*
  - NON è ONNISCENTE
    - Errori di *senso comune*
    - Mostra talvolta incompetenza
    - Non è esperto dei diversi domini
  - NON è sempre completamente coerente
    - Allucinazioni



A person on a horse



A person on a horse ?

Raphael - Saint George Fighting the Dragon

Raphael, Public domain, via Wikimedia Commons

LLMS: POTENZIALITÀ E RISCHI

- Enorme flessibilità nella comprensione e generazione linguistica
- Capacità di affrontare nuovi task attraverso il prompting
- Forte capacità di specializzazione verso fenomeni semantici specifici (domini, enciclopedie, dati in tempo reale)
- Facile integrazione con competenze in altri ambiti cognitivi (machine vision)
- Forti limiti nella capacità di certificare i comportamenti linguistici
- Bulimia computazionale
- Limitata analogia con i processi cognitivi

INTEGRAZIONE DI CONOSCENZA E LLMS: RAG MODELS

- Retrieval Augmented Generation
  - A *generation time* si rende disponibile una informazione di contesto che qualifica la risposta
  - Essenziale per task knowledge intensive
  - Si applica sia al *pre-training* che al *fine-tuning* ed al *prompting*
  - Ha mostrato di mitigare le allucinazioni

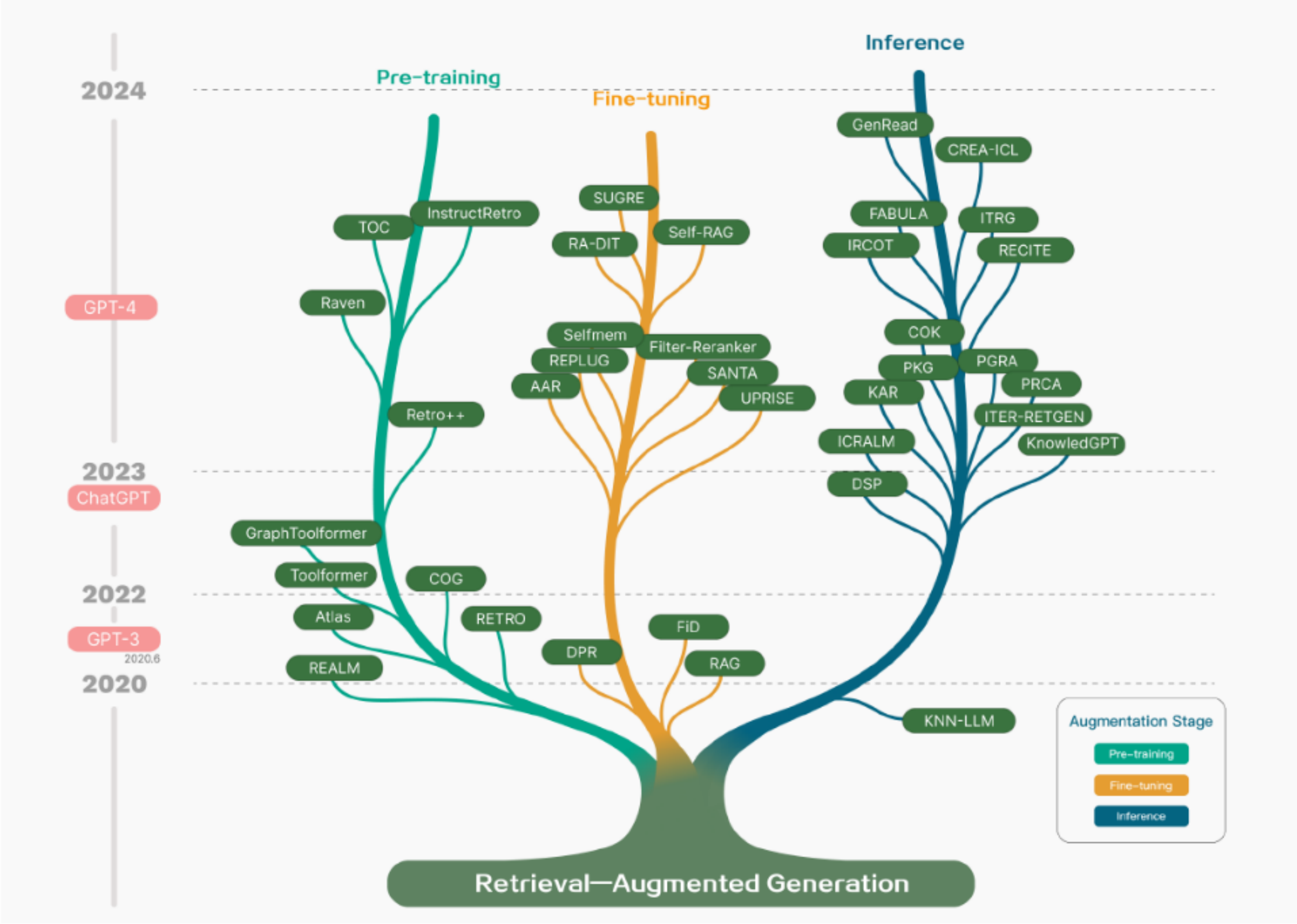


Figure 1: Technology tree of RAG research development featuring representative works

(Lewis et al, 2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. Proceedings of NIPS, Advances in Neural Information Processing Systems, 2020.

RAG MODELS: IL PROBLEMA E L'IDEA

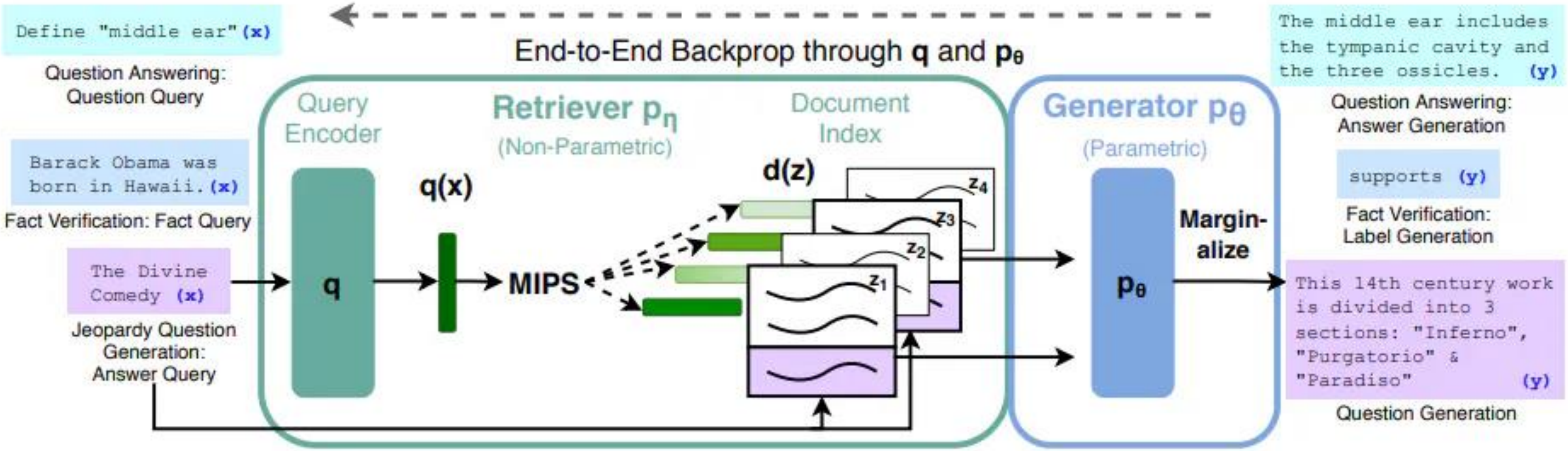
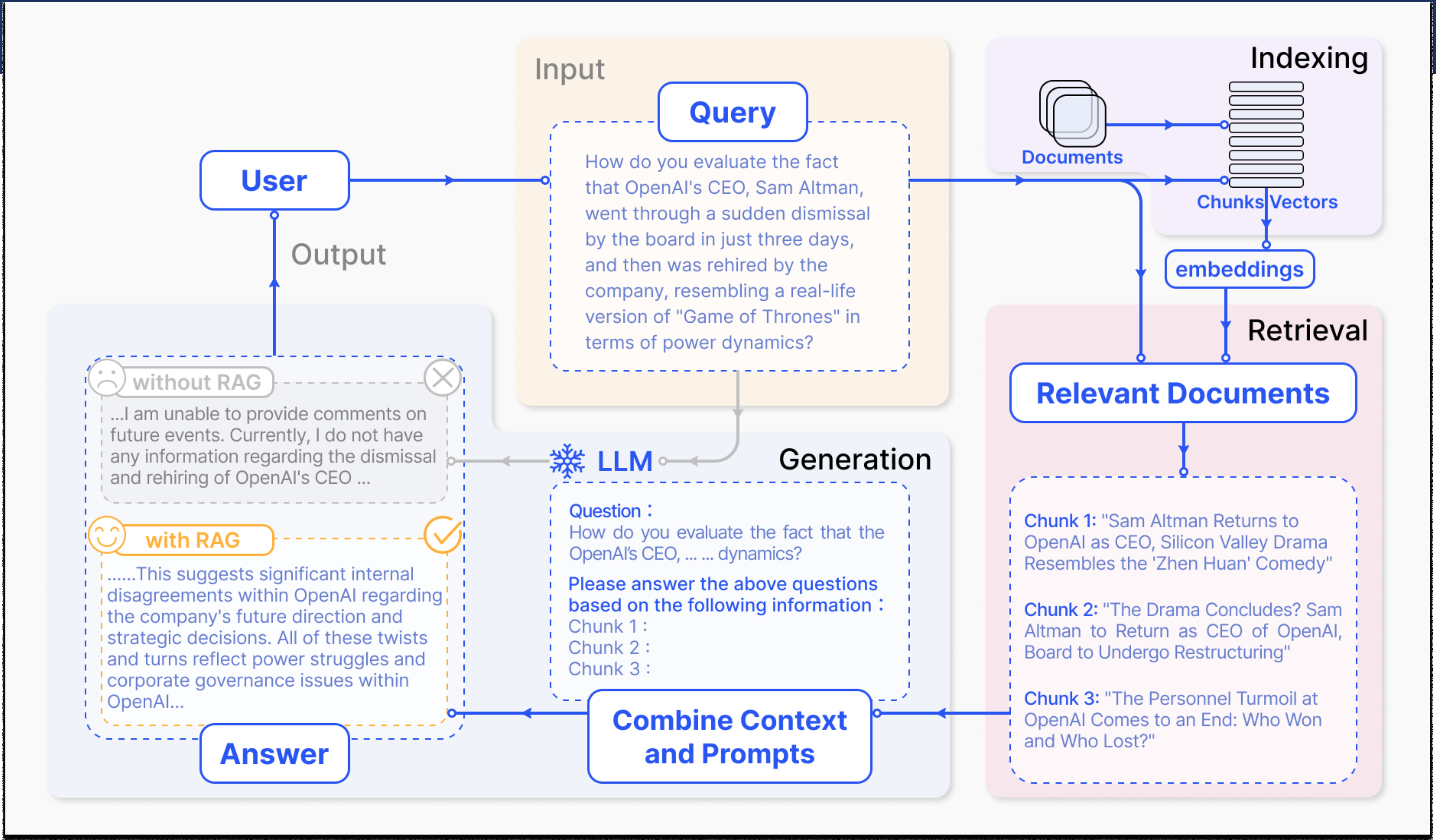

Define "middle ear" (x)
Question Answering: Question Query
Barack Obama was born in Hawaii. (x)
Fact Verification: Fact Query
The Divine Comedy (x)
Jeopardy Question Generation: Answer Query
End-to-End Backprop through q and pθ
Query Encoder
q
q(x)
Retriever pη
(Non-Parametric)
MIPS
Document Index
d(z)
z4
z3
z2
z1
Generator pθ
(Parametric)
pθ
Margin-alize
The middle ear includes the tympanic cavity and the three ossicles. (y)
Question Answering: Answer Generation
supports (y)
Fact Verification: Label Generation
This 14th century work is divided into 3 sections: "Inferno", "Purgatorio" & "Paradiso" (y)
Question Generation

Figure 1: Overview of our approach. We combine a pre-trained retriever (*Query Encoder* + *Document Index*) with a pre-trained seq2seq model (*Generator*) and fine-tune end-to-end. For query x, we use Maximum Inner Product Search (MIPS) to find the top-K documents z_i. For final prediction y, we treat z as a latent variable and marginalize over seq2seq predictions given different documents.

RAG MODELS: IL FLUSSO INFORMATIVO



How do you evaluate the fact that OpenAI's CEO, Sam Altman, went through a sudden dismissal by the board in just three days, and then was rehired by the company, resembling a real-life version of "Game of Thrones" in terms of power dynamics?

...I am unable to provide comments on future events. Currently, I do not have any information regarding the dismissal and rehiring of OpenAI's CEO ...

......This suggests significant internal disagreements within OpenAI regarding the company's future direction and strategic decisions. All of these twists and turns reflect power struggles and corporate governance issues within OpenAI...

Question：
How do you evaluate the fact that the OpenAI's CEO, ... dynamics?

Please answer the above questions based on the following information：
Chunk 1：
Chunk 2：
Chunk 3：

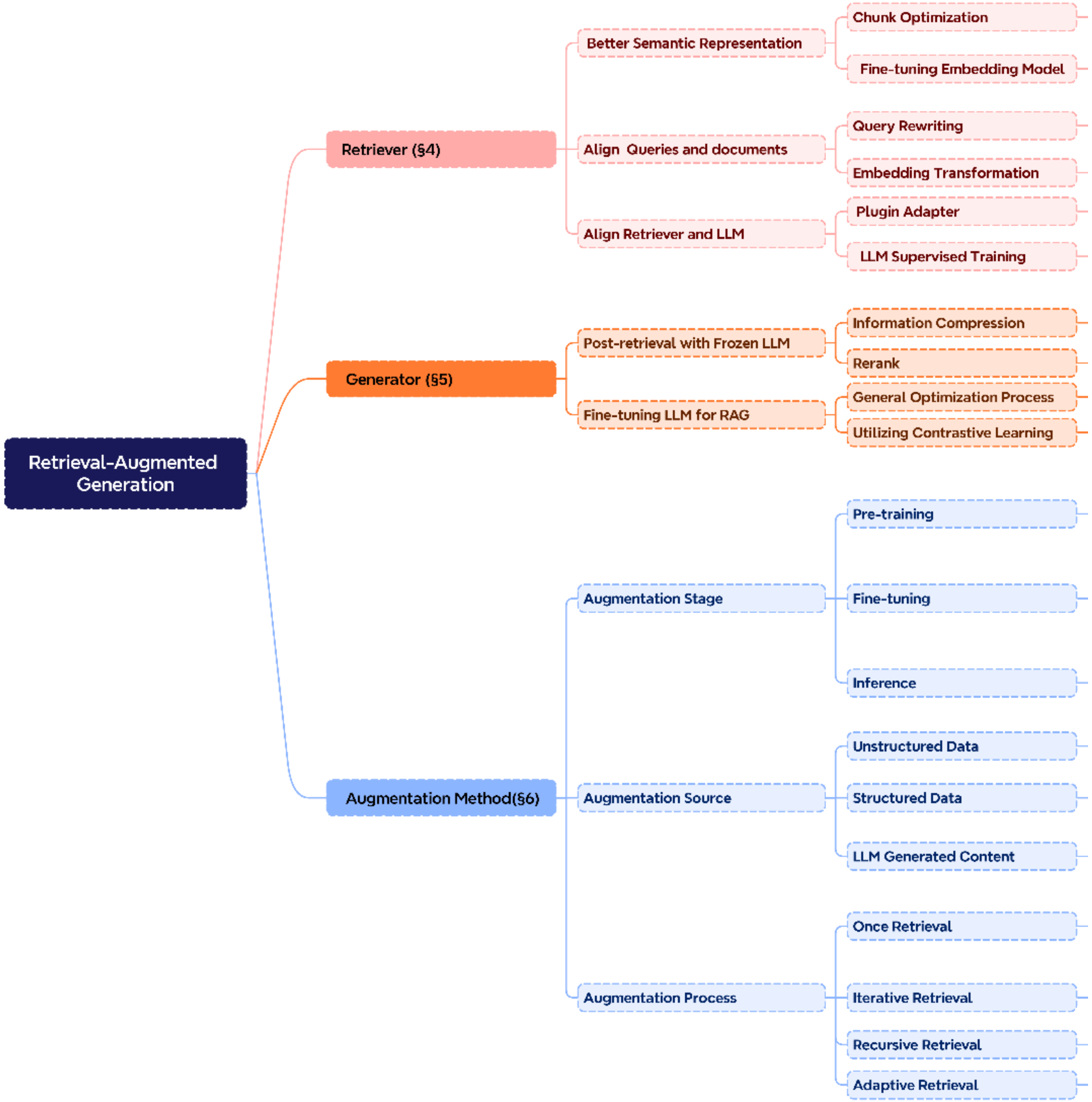
Chunk 1: "Sam Altman Returns to OpenAI as CEO, Silicon Valley Drama Resembles the 'Zhen Huan' Comedy"

Chunk 2: "The Drama Concludes? Sam Altman to Return as CEO of OpenAI, Board to Undergo Restructuring"

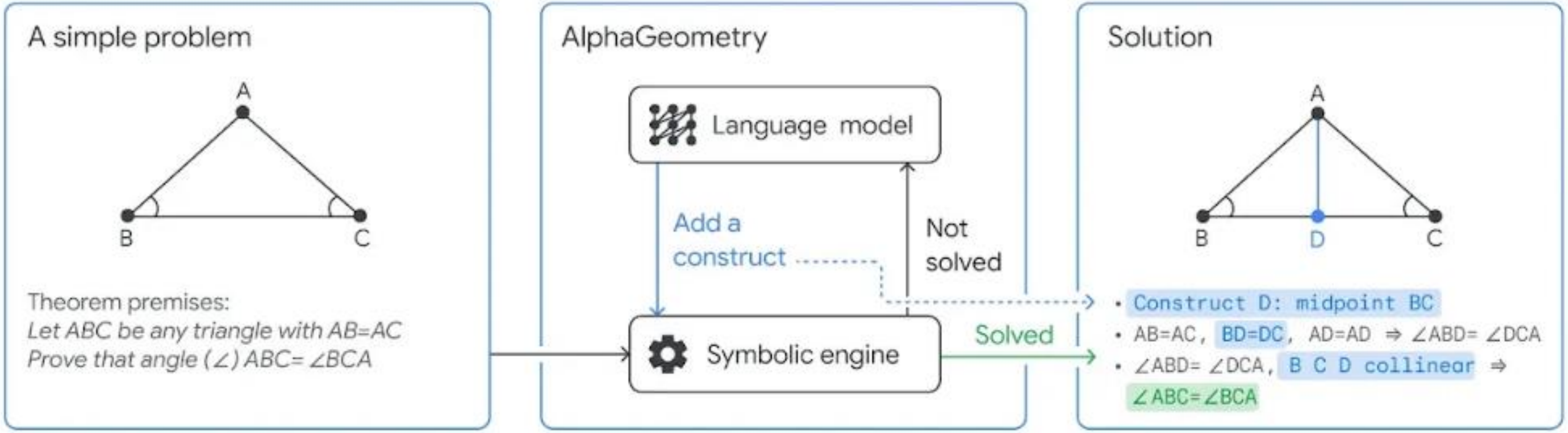
Chunk 3: "The Personnel Turmoil at OpenAI Comes to an End: Who Won and Who Lost?"

A RAG TAXONOMY

- Research is active in different directions
  - Retrieval
  - Generation
  - Textual, Logical and Procedural Augmentation
- DBs or KG are often explored as information sources

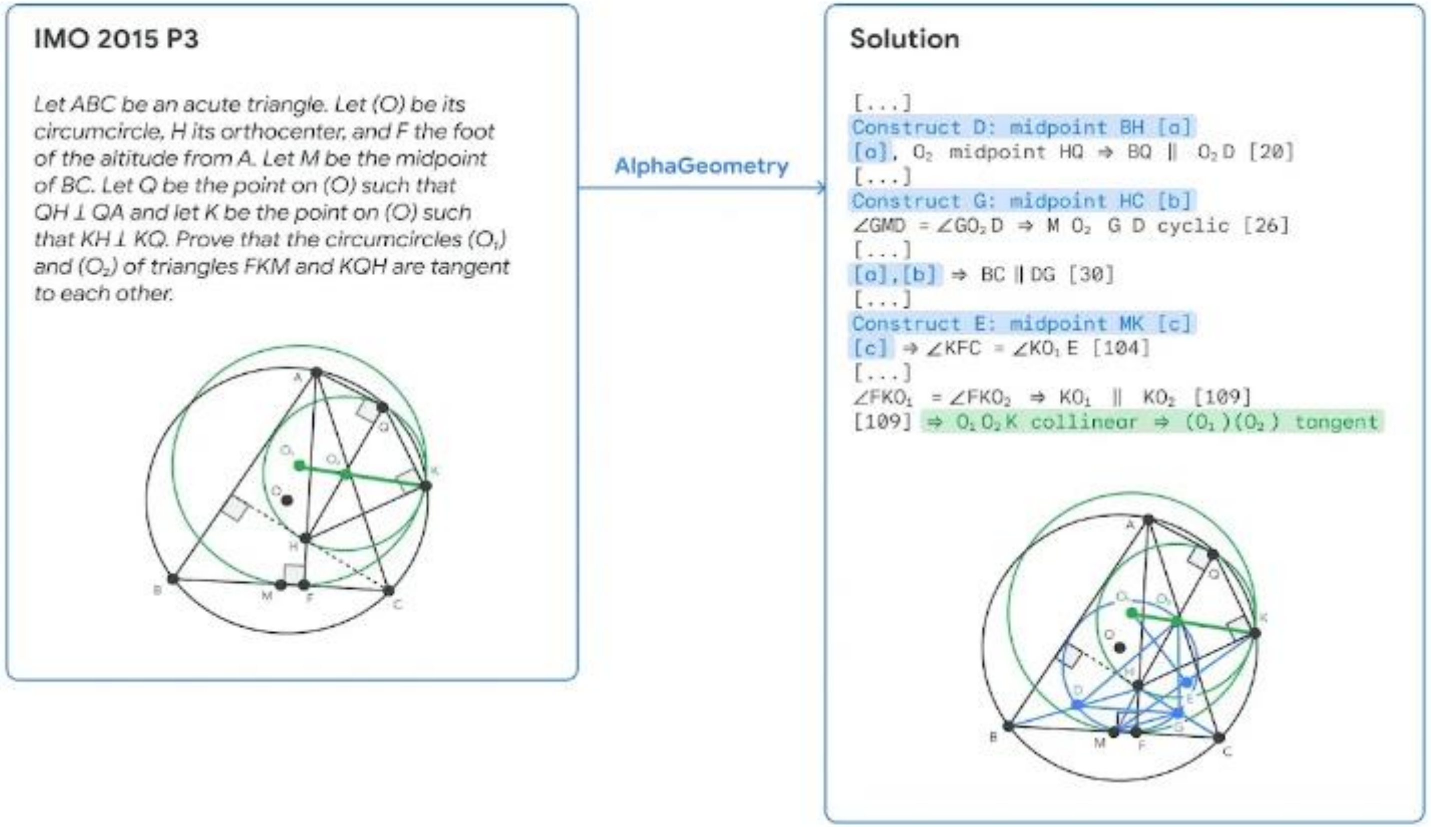


ALPHAGEOMETRY (GOOGLE DEEPMIND, JAN 2024)


A simple problem
A
B
C
Theorem premises:
Let ABC be any triangle with AB=AC
Prove that angle (∠) ABC= ∠BCA
AlphaGeometry
Language model
Add a construct
Not solved
Symbolic engine
Solved
Solution
A
B
D
C
• Construct D: midpoint BC
• AB=AC, BD=DC, AD=AD ⇒ ∠ABD= ∠DCA
• ∠ABD= ∠DCA, B C D collinear ⇒ ∠ABC=∠BCA

Trinh, Trieu H., Wu Yuhuai, Le Quoc V., He He, Luong Thang, Solving olympiad geometry without human demonstrations, Nature, 625, 2024.

ALPHAGEOMETRY (GOOGLE DEEPMIND, JAN 2024)



IMO 2015 P3

Let ABC be an acute triangle. Let (O) be its circumcircle, H its orthocenter, and F the foot of the altitude from A. Let M be the midpoint of BC. Let Q be the point on (O) such that QH ⊥ QA and let K be the point on (O) such that KH ⊥ KQ. Prove that the circumcircles (O_1) and (O_2) of triangles FKM and KQH are tangent to each other.

AlphaGeometry

Solution

[...]

Construct D: midpoint BH [a]

[a], O_2 midpoint HQ ⇒ BQ ∥ O_2D [20]

[...]

Construct G: midpoint HC [b]

∠GMD = ∠GO_2D ⇒ M O_2 G D cyclic [26]

[...]

[a],[b] ⇒ BC ∥ DG [30]

[...]

Construct E: midpoint MK [c]

[c] ⇒ ∠KFC = ∠KO_1E [104]

[...]

∠FKO_1 = ∠FKO_2 ⇒ KO_1 ∥ KO_2 [109]

[109] ⇒ $O_1$$O_2$K collinear ⇒ ($O_1$)($O_2$) tangent

Problem 3 of the 2015 International Mathematics Olympiad (left) and a condensed version of AlphaGeometry's solution (right). The blue elements are added constructs. AlphaGeometry's solution has 109 logical steps.

Trinh, Trieu H., Wu Yuhuai, Le Quoc V., He He, Luong Thang, Solving olympiad geometry without human demonstrations, Nature, 625, 2024.

BIBLIOGRAFIA: TRANSFORMERS

- (Vaswani 2017), Attention is all you need, https://arxiv.org/abs/1706.03762
- (Devlin et al 2018), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, https://arxiv.org/abs/1810.04805
- An interesting introduction to the attention mechanism:
  - **All you need to know about 'Attention' and 'Transformers' — In-depth Understanding — Part 1,** A. Sarkar, URL: https://towardsdatascience.com/all-you-need-to-know-about-attention-and-transformers-in-depth-understanding-part-1-552f0b41d021#4c16

- Other Task specific works:
  - Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473, 2014.
  - Effective Approaches to Attention-based Neural Machine Translation, Minh-Thang Luong Hieu Pham Christopher D. Manning, 2015, https://arxiv.org/abs/1508.04025v5
  - Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. Structured attention networks. In International Conference on Learning Representations, 2017.

BIBLIOGRAFIA: *BEYOND TRANSFORMER*

- (Vaswani 2017), Attention is all you need, https://arxiv.org/abs/1706.03762
- (Devlin et al 2018), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, https://arxiv.org/abs/1810.04805
- Rocktaschel et al., "Reasoning About Entailment With Neural Attention" (ICLR 2016)
- T5: (Wolf et al, 2019) Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. ArXiv, abs/1910.03771.
- BART Encoding-Decoding: Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.https://arxiv.org/abs/1910.13461
- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, "Improving Language Understanding by Generative Pre-Training", 2019
- 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: Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901. https://arxiv.org/abs/2005.14165, NeurIPS 2020.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, Paul F. Christiano: Learning to summarize with human feedback. NeurIPS 2022

GRAZIE DELL'ATTENZIONE

BASILI@INFO.UNIROMA2.IT