C1A0

Artificial
Intelligence
Exposition

C1A0 EXPO

OUR SPONSORS

C1A0 EXPO 2019

MAIN SPONSOR

iren

C1A0 EXPO 2019

SPONSOR

GRUPPO FOS
soluzioni ad alta tecnologia

IFM Infomaster
indigo.ai

avanade

ESA
ECOLOGIA SOLUZIONE AMBIENTE

Daitek

firedesktop
burn document barriers
FINCANTIERI Si
evolving integration

RINA

ERG
EVOLVING ENERGIES

MAPS
SHARING KNOWLEDGE

UniCredit

FLAIRBIT
DATA CENTRIC SOLUTIONS

GIE
GIOVANI IMPRENDITORI

C1A0 EXPO 2019

SPONSOR TECNICO

upen SEARCH

C1A0 EXPO

State-of-the-art concepts in NLP

(... and their limits)

Cristiano De Nobili, Ph.D.

Harman - Samsung & AINDO

Who am I?

Theoretical
Particle Physicist

Language of Nature

Ph.D. Statistical Physics
(Entanglement!)

Natural (Human) Language

Artificial Intelligence

Cristiano-De-Nobili

@denocris

@denocris

Trieste: Sea, Coffee & Science

ma per seguir virtute e canoscenza

SISSA

MHPC

Master in High Performance Computing

ICTP

The Abdus Salam
International Centre
for Theoretical Physics

Outline

- Language & Computers: a bird's-eye view
- Recurrent Neural Network (the state-of-the-art till yesterday)
- Auto-encoders and Seq2Seq
- The Attention and Self-Attention Mechanism
- Transformers
- Cool NLP Applications
- Limits: energy and… (we will see later)

Language

Humans do **3** basic things with language that machines also can do, or at least attempt.

We can listen!

... computers can now easily translate text to voice: translate voice to text.
better at transforming
handwritten notes to typed text.

Language

But humans also perform a fourth function…

We can understand language(s)!

?

That's what NLP is trying to achieve

Language is a hard task for a machine

Dai diamanti non nasce niente, dal letame nascono i fiori.

Una lunga vacanza. Una lunga coda.

"Giulia, sei libera domani alle sei?"

Language is a hard task for a machine

Dai diamanti non

. Una lunga coda.

"Giulia,

So,
how a machine can understand language?

Words and Numbers

Words and numbers have always been thought to be at odds. You can be a man of letters or a man of science.

There are poets, philosophers & journalist on one side. Scientists & engineers on the other.

But, what is the language of a machine?

Therefore, thanks to NLP, words and number can become best friends after many centuries!

Dai diamanti non nasce niente, dal letame nascono i fiori.

,

WORDS <-> NUMBERS

Dai

fiori

79 24 34 14 172 5 127 674 42 9 183 221

,

Yes, but this is just the first step...

VECTORS

$$\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix}$$

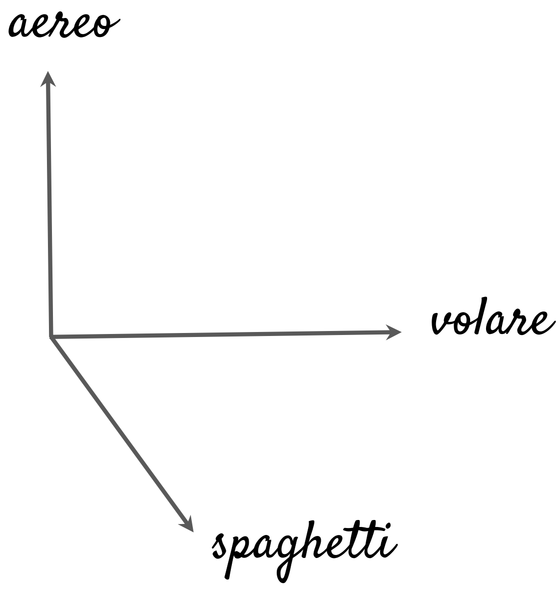
Dai diamanti non nasce niente, dal letame nascono i fiori.

$$\begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

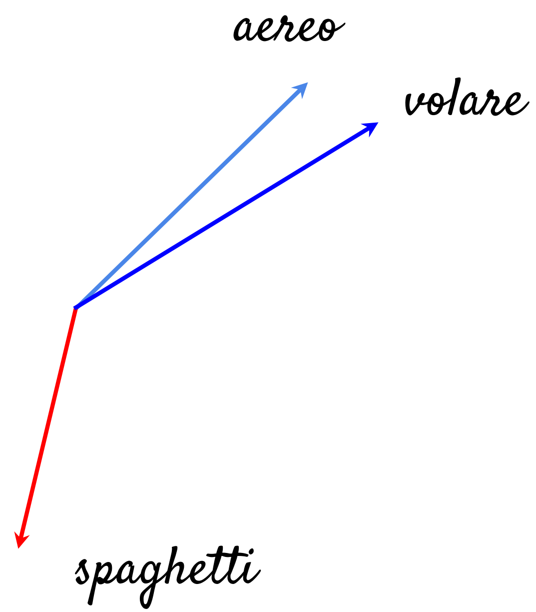
However...

- these are sparse vectors
- ... and orthogonal

DENSE VECTORS



all words are equally distant!



some words are more similar!

(`Word2Vec, Glove`) arXiv: 1301.3781

DENSE VECTORS

(from discrete to continuous variables)

$$\begin{bmatrix} 0.12 \\ \vdots \\ 0.23 \\ 0.78 \end{bmatrix}$$

$$\begin{bmatrix} 0.55 \\ \vdots \\ 0.17 \\ 0.61 \end{bmatrix}$$

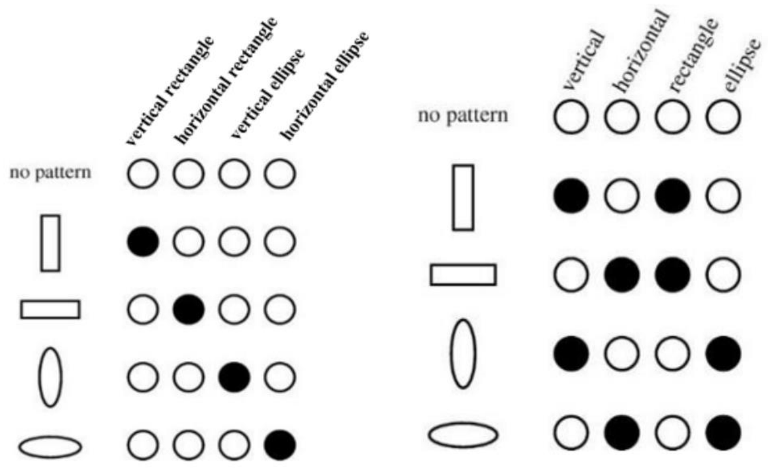
Dai diamanti non nasce niente, dal letame nascono i fiori.

$$\begin{bmatrix} 0.67 \\ \vdots \\ 0.42 \\ 0.28 \end{bmatrix}$$

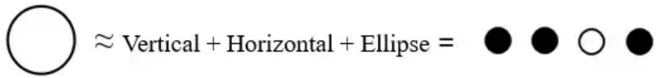
They can encode more complex structures/relations

- syntactics
- semantics and more

DENSE REPRESENTATIONS: forget for a while about words...



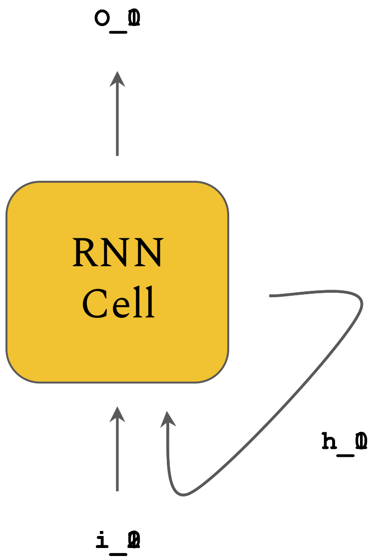
SPARSE **DENSE**



- One concept is represented by more than one dot
- One dot represents more than one concept

RECURRENT NEURAL NETWORKS

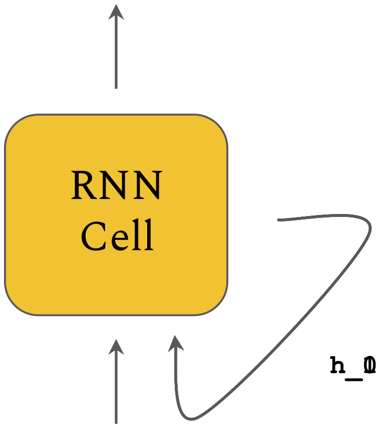
(state-of-the-art till yesterday…)



(state-of-the-art till yesterday...)

"Ciao come stai?"

Used for many NLP tasks!



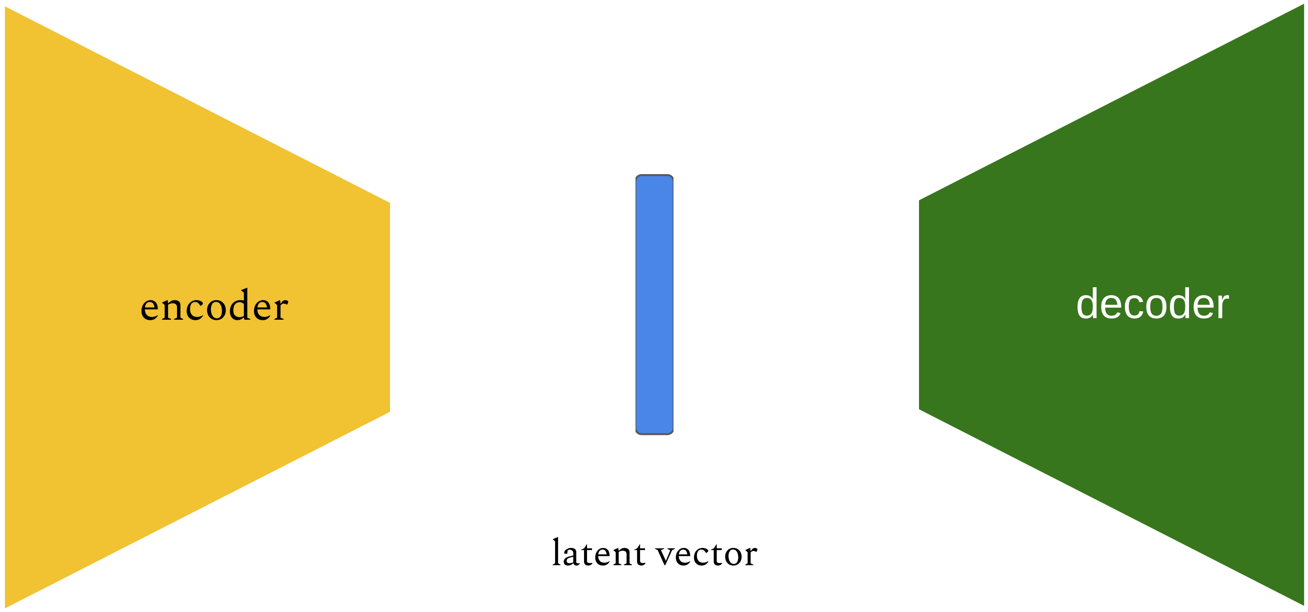
contains the memory of "ciao"come"

"Ciao come stai?"

"Hi, how are you?"

We are going to consider translation!

Autoencoders & Seq2Seq

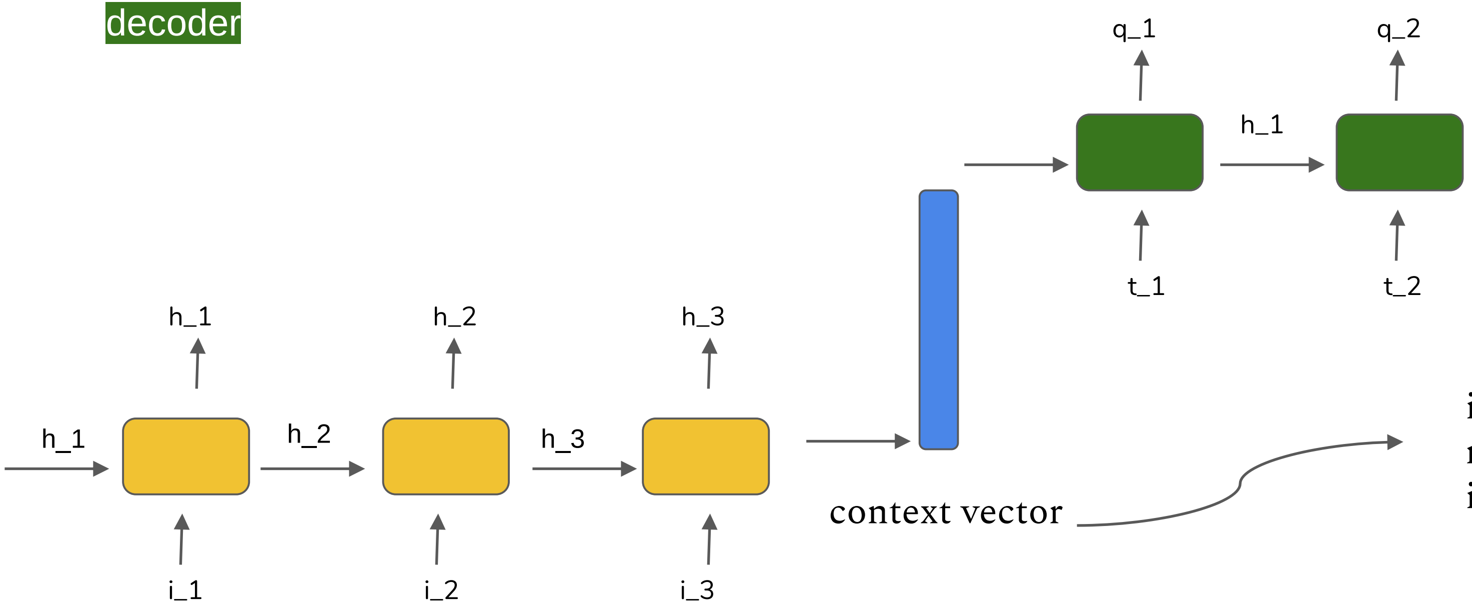


embedding dim < input/output dim

RNN Seq2Seq

encoder

decoder



it contains all the memory of the input sequence!

it loses memory for long sentences!

RNN Cons

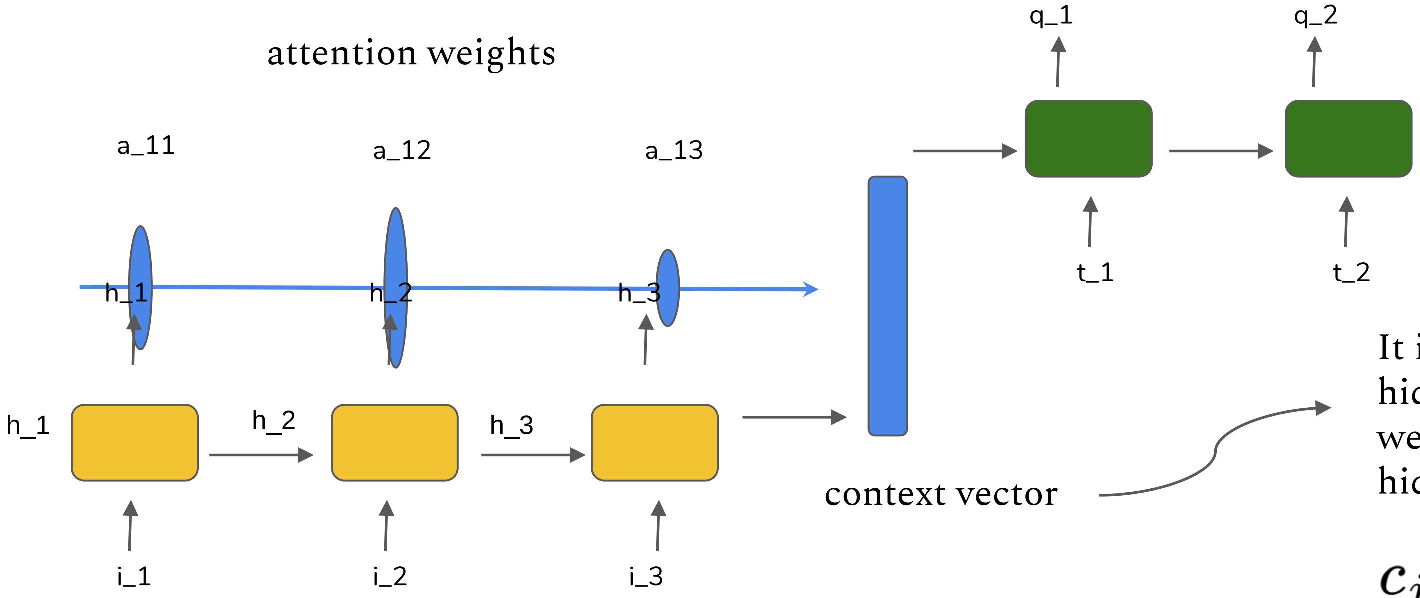
- they cannot remember/summarize long sequences.
    - they cannot learn long-term dependencies

- they are slow because sequential (not parallelizable)

CNNs solve the second bullet, but hardly the second.

Attention solves both!

Attention Mechanism



It is not just the last hidden state, but a weighted sum encoder's hidden states!

$$c_j = \sum_i \alpha_{i,j} h_i$$

It is computed at each time step **j** of the decoder

This is the encoder-decoder attention:

The context vectors enable the decoder to focus on certain parts of the input when predicting its output.

Attention Weights

y_1 y_2

h_1

$$c_j = \sum_i \alpha_{i,j} h_i$$

How are they computed?

How are they learned?

the last
but a
n encoder's
!

$\alpha_{i,j} h_i$

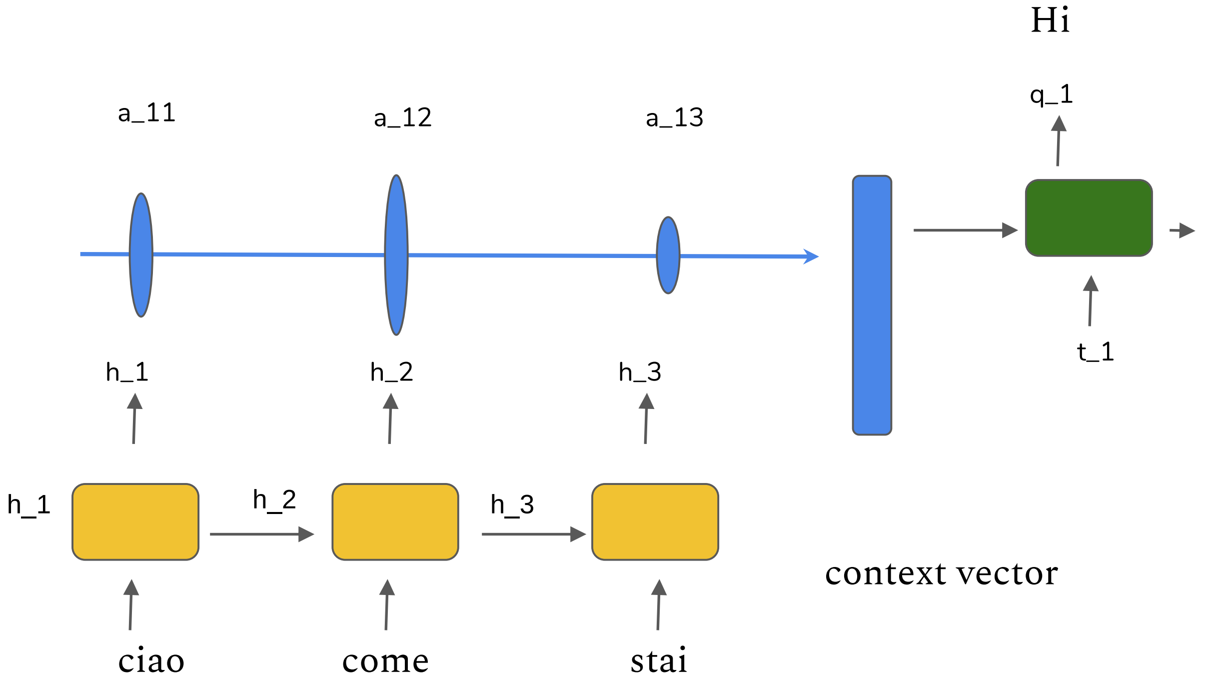
d at each
the decoder

Attention Weights: how are they computed?

$$\boxed{\alpha_{i,j} = h_i^T \cdot q_j}$$

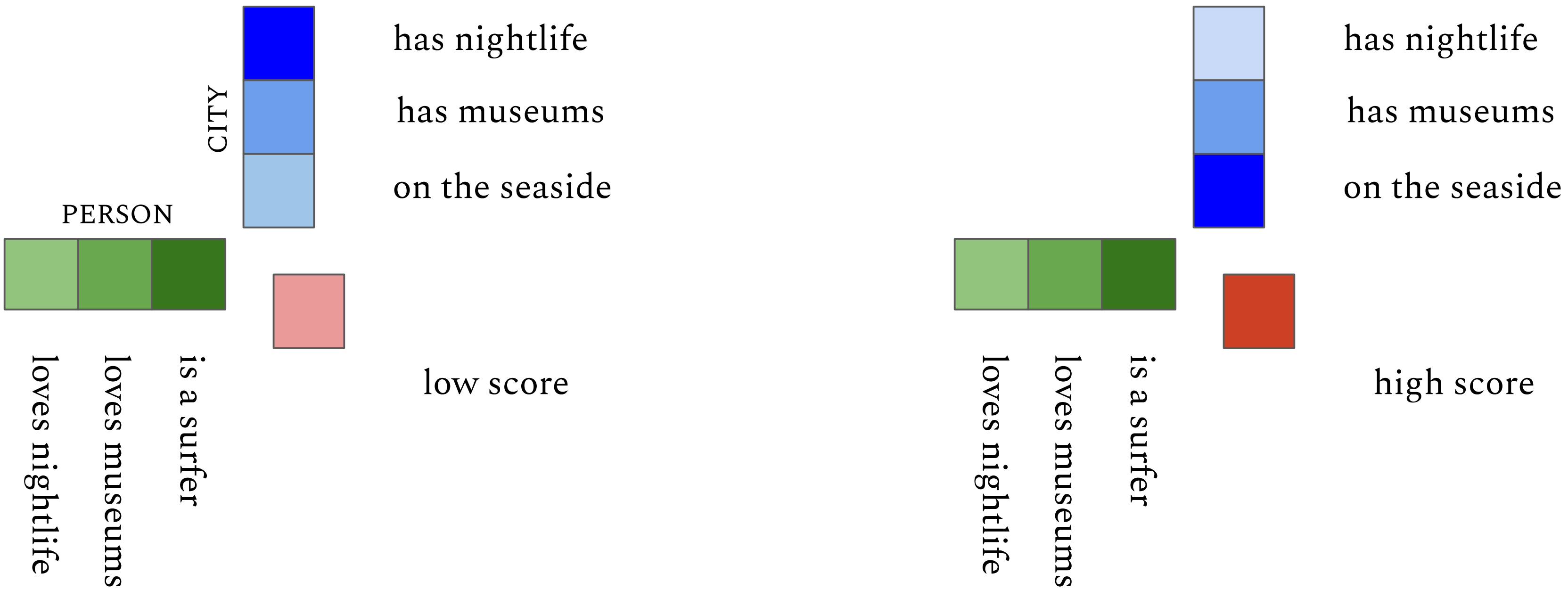
$$c_j = \sum_i \alpha_{i,j} h_i$$

$$c_1 = \alpha_{1,1} h_1 + \alpha_{1,2} h_2 + \alpha_{1,3} h_3$$



$$\alpha_{i,j} = h_i^T \cdot q_j$$

Attention Weights: how are they computed?



We then say that the person "attends" more to the city on the right!

Attention Weights: Alignment

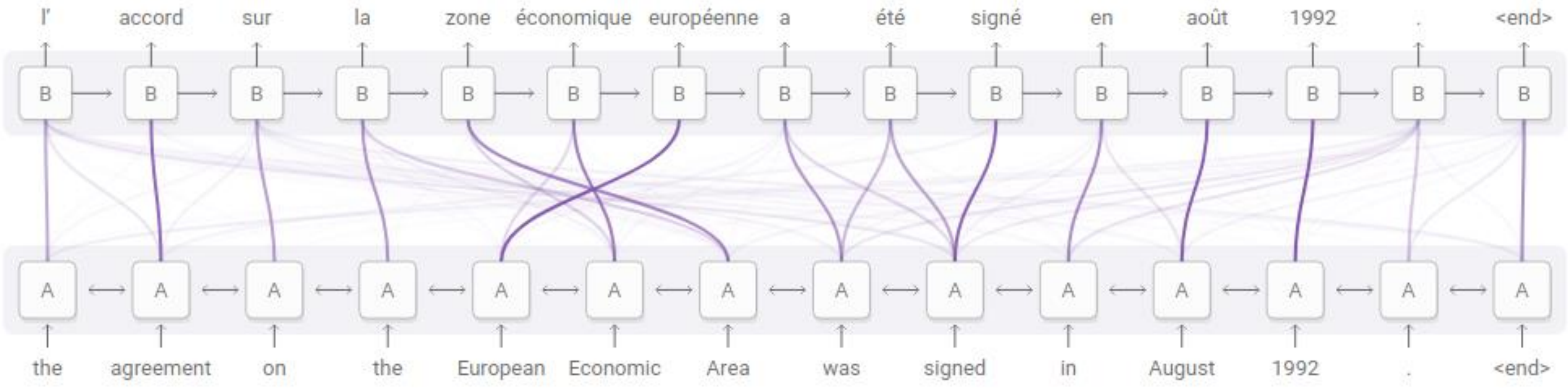
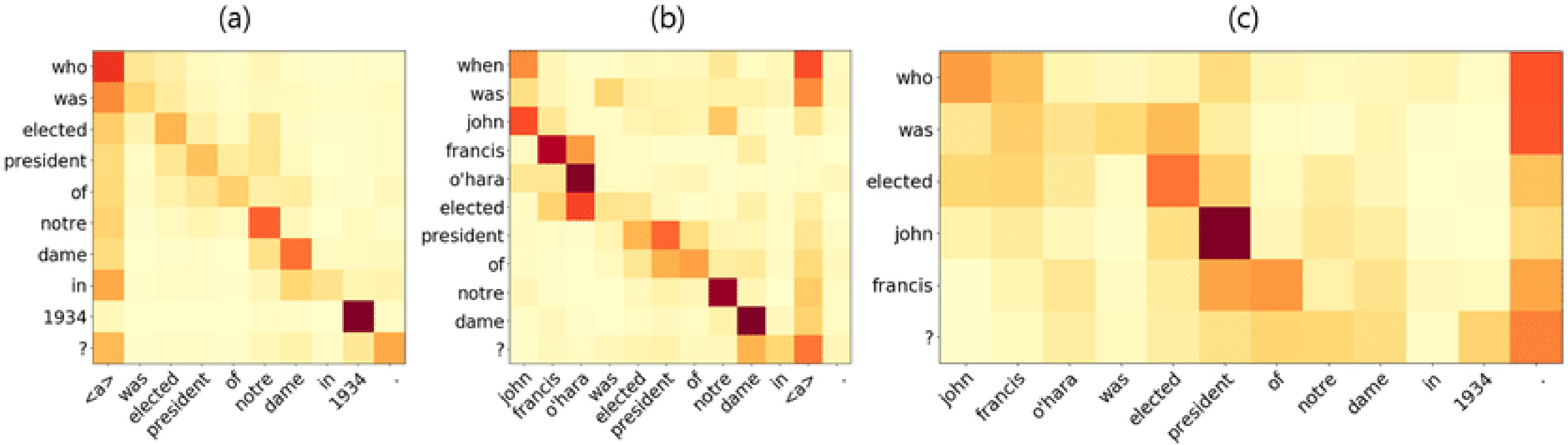


Diagram derived from Fig. 3 of Bahdanau, *et al.* 2014

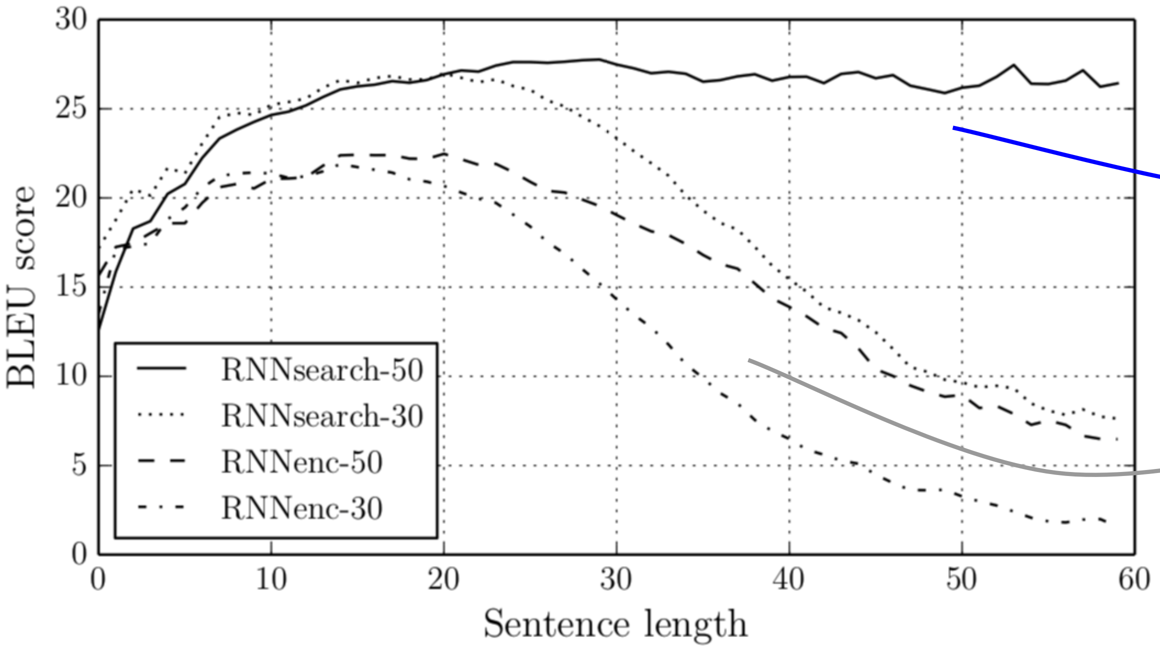
Attention weights measures the alignment between input and output sentences

Attention Matrix



Attention weights measures the alignment between input and output sentences

Attention Matrix



WITH ATTENTION!

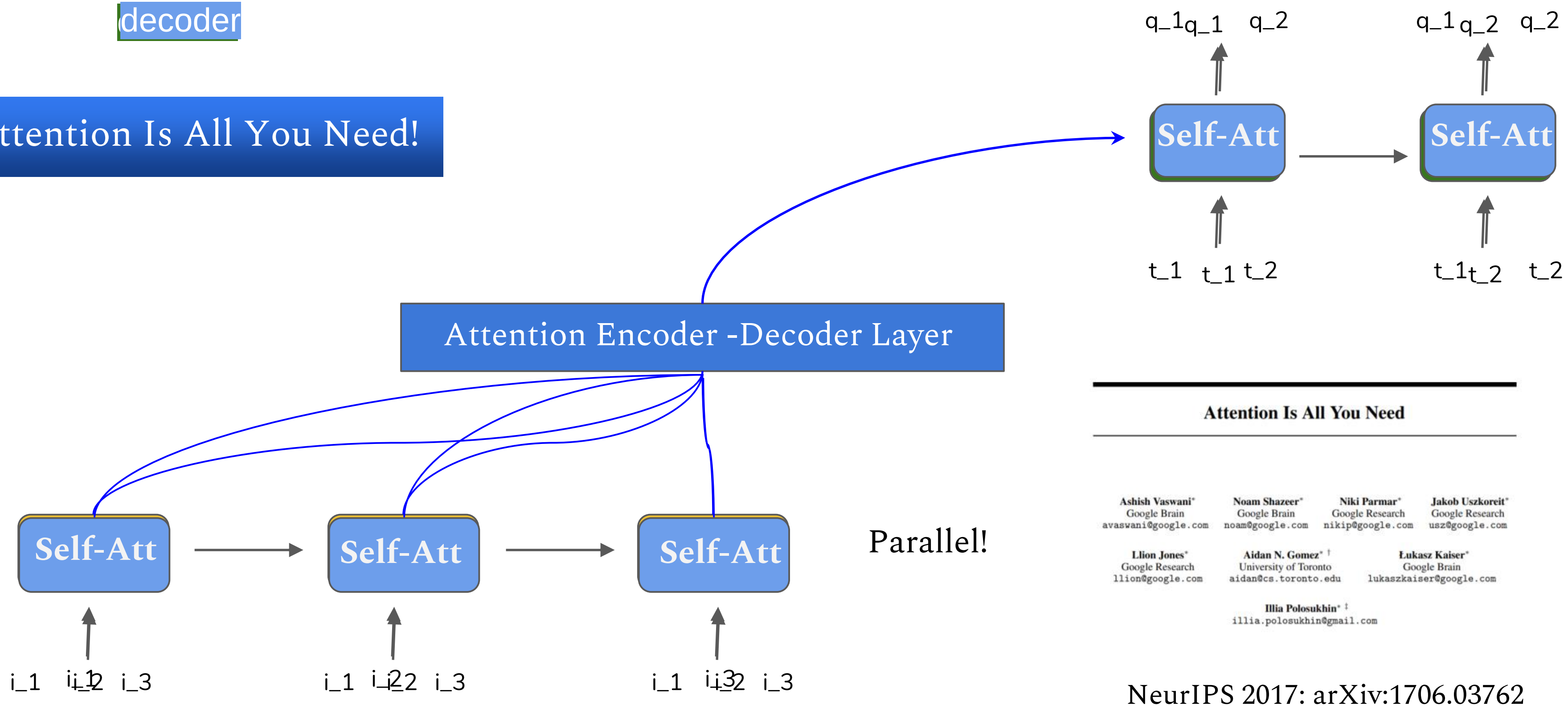
NO ATTENTION!

Transformers: Basics

encoder

decoder

Attention Is All You Need!



Parallel!

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

NeurIPS 2017: arXiv:1706.03762

SELF-ATTENTION?

So far we have seen the **encoder-decoder attention**. Together with it, the fundamental operation of any transformer architecture is **self-attention**.

Self-attention is an attention mechanism relating different positions of a single sequence in order to compute a representation of the same sequence.

y_1 y_2 y_3 y_4 y_5

dal letame nascono i fiori

x_1 x_2 x_3 x_4 x_5

SELF-ATTENTION?

$$y_i = \sum_j w_{ij} x_j, \quad w' = x_i^T x_j \quad w_{ij} = \text{softmax}(w'_{ij})$$

y_2 = (*letame* x *dal*) x *dal* + ...

dal *letame* *nascono* *i* *fiori*

x_1 x_2 x_3 x_4 x_5

SELF-ATTENTION?

$$y_i = \sum_j w_{ij} x_j, \quad w' = x_i^T x_j \quad w_{ij} = \text{softmax}(w'_{ij})$$

y_2 = (letame x dal) x dal + (letame x letame) x letame + ...

dal	letame	nascono	i	fiori
x_1	x_2	x_3	x_4	x_5

SELF-ATTENTION?

$$y_i = \sum_j w_{ij} x_j, \quad w' = x_i^T x_j \quad w_{ij} = \mathrm{softmax}(w'_{ij})$$

y_2 = (letame x dal) x dal + (letame x letame) x letame + (letame x nascono) x nascono +

dal	letame	nascono	i	fiori
x_1	x_2	x_3	x_4	x_5

SELF-ATTENTION?

$$y_i = \sum_j w_{ij} x_j, \quad w' = x_i^T x_j \quad w_{ij} = \text{softmax}(w'_{ij})$$

y_2 = (letame x dal) x dal + (letame x letame) x letame + (letame x nascono) x nascono + (letame x i) x i +

dal letame nascono i fiori

x_1 x_2 x_3 x_4 x_5

SELF-ATTENTION?

$$y_i = \sum_j w_{ij} x_j, \quad w' = x_i^T x_j \quad w_{ij} = \text{softmax}(w'_{ij})$$

y_2 = (letame x dal) x dal + (letame x letame) x letame + (letame x nascono) x nascono + (letame x i) x i + (letame x fiori) x fiori

dal	letame	nascono	i	fiori
x_1	x_2	x_3	x_4	x_5

SELF-ATTENTION LAYER

y_dal y_letame y_nascono y_i y_fiori

y_letame is a weighted sum over all the embedding vectors in the first sequence, weighted by their (normalized) dot-product with x_letame

x_dal x_letame x_nascono x_i x_fiori

SELF-ATTENTION LAYER: more detailed

$$y_i = \sum_j w_{ij} x_j, \quad w' = x_i^T x_j \quad w_{ij} = \text{softmax}(w'_{ij})$$

y_2 = (*letame* x *dal*) x *dal* + (*letame* x *letame*) x *letame* + (*letame* x *nascono*) x *nascono* + (*letame* x *i*) x *i* + (*letame* x *fiori*) x *fiori*

In self-attention, each input vector (let's say **x_2**) is used in three different ways in the self attention operation:

- **query:** it is compared to every other vector to establish the weights for its own output **y_2**
- **key:** it is compared to every other vector to establish the weights for the output of the j-th word **y_j**
- **value:** it is used as part of the weighted sum to compute each output vector once the weights have been established

SELF-ATTENTION LAYER: more detailed

$$\cancel{y_i = \sum_j w_{ij} x_j, \quad w' = x_i^T x_j \quad w_{ij} = \text{softmax}(w'_{ij})}$$

In the basic self-attention written above, each input vector **`x_i`** must play all three roles.

Its life can be made a bit easier by deriving new vectors for each role (query, key, value), by applying a linear transformation to the original input vector.

$$\mathbf{q}_i = \mathbf{W}_q \mathbf{x}_i \qquad \mathbf{k}_i = \mathbf{W}_k \mathbf{x}_i \qquad \mathbf{v}_i = \mathbf{W}_v \mathbf{x}_i$$

$$w'_{ij} = \mathbf{q}_i{}^T \mathbf{k}_j$$

$$w_{ij} = \text{softmax}(w'_{ij})$$

$$\mathbf{y}_i = \sum_j w_{ij} \mathbf{v}_j \,.$$

*Peter Bloem Blog

SELF-ATTENTION LAYER: more detailed

$$\cancel{y_i = \sum_j w_{ij} x_j, \quad w' = x_i^T x_j} \quad w_{ij} = \text{softmax}(w'_{ij})$$

In the basic self-attention written above, each input vector **`x_i`** must play all three roles.

Its life can be made a bit easier by deriving new vectors for each role (query, key, value), by applying a linear transformation to the original input vector.

$$\mathbf{q}_i = \mathbf{W}_q \mathbf{x}_i \qquad \mathbf{k}_i = \mathbf{W}_k \mathbf{x}_i \qquad \mathbf{v}_i = \mathbf{W}_v \mathbf{x}_i$$

$$w'_{ij} = \mathbf{q}_i^T \mathbf{k}_j$$
$$w_{ij} = \text{softmax}(w'_{ij})$$
$$\mathbf{y}_i = \sum_j w_{ij} \mathbf{v}_j \,.$$

- the vector **q** encodes the word that is paying attention ('it is querying the other words').
- **k** encodes the word to which attention is being paid.

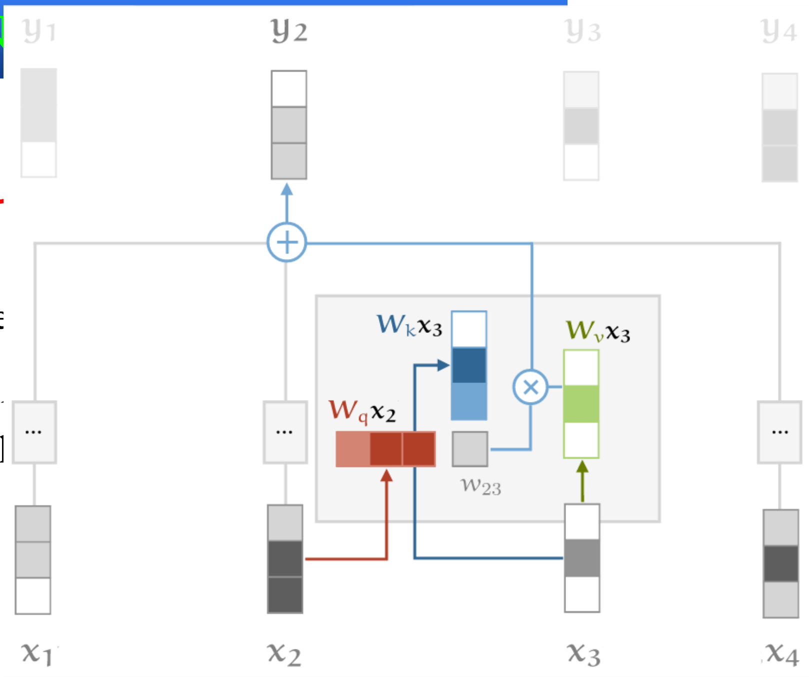
*Peter Bloem Blog

SELF-ATTENTION

$$y_i = \sum_j w_{ij} x_j, \quad w' =$$

In the basic self-attention written above, e

Its life can be made a bit easier by deriving new vec … linear transformation to tl



$$\mathbf{q}_i = \mathbf{W}_q \mathbf{x}_i \qquad \mathbf{k}_i = \mathbf{W}_k \mathbf{x}_i \qquad \mathbf{v}_i = \mathbf{W}_v \mathbf{x}_i$$

$$w'_{ij} = \mathbf{q}_i^\top \mathbf{k}_j$$

$$w_{ij} = \text{softmax}(w'_{ij})$$

$$\mathbf{y}_i = \sum_j w_{ij} \mathbf{v}_j \,.$$

- the vector **q** encodes the word that is paying attention ('it is querying the other words').
- **k** encodes the word to which attention is being paid.

*Peter Bloem Blog

Positional Encoding

y_dal y_letame y_nascono y_i y_fiori

Self - Attention

is PERMUTATIONAL INVARIANT

x_dal x_letame x_nascono x_i x_fiori

Positional Encoding

y_dal y_letame y_nascono y_i y_fiori

Self - Attention

is PERMUTATIONAL INVARIANT

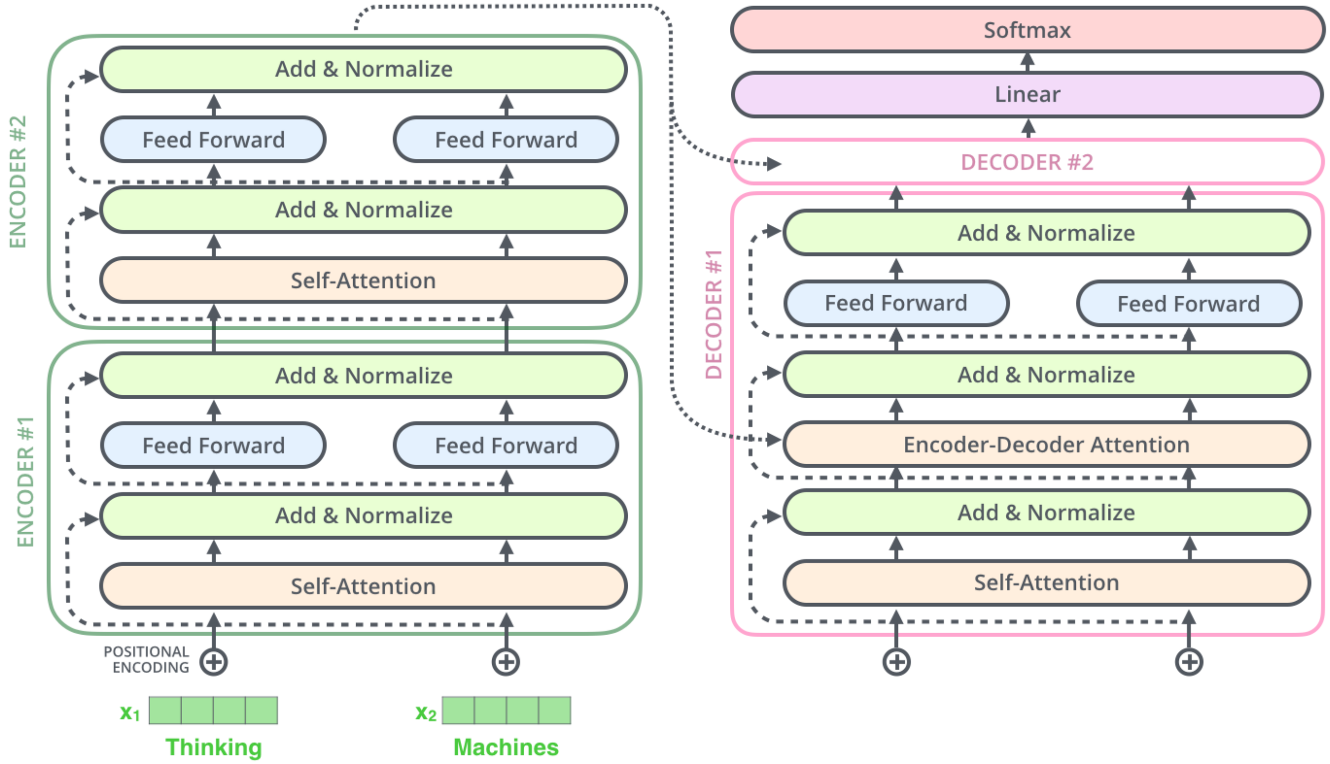
We mark each word with its absolute position

pos_1 pos_2 pos_3 pos_4 pos_5

+

x_dal x_letame x_nascono x_i x_fiori

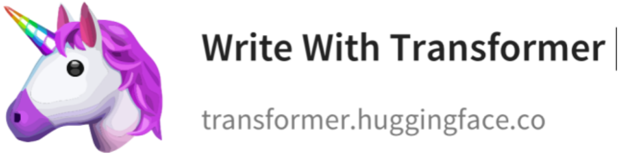
Transformers



NLP popular models

arXiv:1810.04805

"Language Models are Unsupervised Multitask Learners"



All of them are based on Transformers!

Applications

- spell checker
- auto-completion
- machine translation
- word sense disambiguation
- chat bots & virtual assistants
- sentiment analysis & social media marketing
- summarizing text
- text classification
- sentiment analysis
- ...

Machine translation is a huge application for NLP that allows us to overcome barriers to communicating

Non-Standard Applications

SIGNALL

Start-up based in Budapest. They developed a technology leveraging AI (computer vision + NLP) that is able to recognize and translate sign language.

Non-Standard Applications

AIRBUS

Aircraft Maintenance: NLP helps mechanics synthesize information from enormous aircraft manuals. It can also find meaning in the descriptions of problems reported verbally or handwritten from pilots.

Neurodegenerative Disease

Neurodegenerative diseases causing dementia are known to affect a person's speech and language. NLP is used to identify those defects.

Predicting probable Alzheimer's disease using linguistic deficits and biomarkers, Orimaye et al. (2017)
A new diagnostic approach for the identification of patients with neurodegenerative cognitive complaints, Al-Hameed et al. (2019)

Genomics

Transcription, the biological process through which DNA is transcribed into RNA, is heavily regulated by DNA-binding transcription factors. Transformers are used for the transcription factor binding site prediction task.

An Attention-Based Model for Transcription Factor Binding Site Prediction, Gunjan Baid (Berkeley, Thesis)

LIMIT #1: Efficiency

90×10^9 neurons
firing 10^3 time/s
each 10^4 connections

2×10^9 Mflops (ops/s)

Energy

20 watt

Serial + Massively Parallel

10^9 operations/s
5×10^9 transistors/cpu

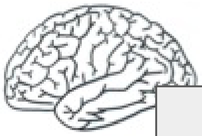
8×10^9 Mflops (ops/s)

Energy

2.5×10^7 watt

Mostly Serial

LIMIT #1: Efficiency



90 x 10^9 ne
firing 10^3
each 10^4 con

2 x 10^9 Mflops

Energ

20 watt

erations/s
ansistors/cpu

flops (ops/s)

nergy

2.5 x 10^7 watt

Serial + Massively Parallel

Mostly Serial

Green AI

Roy Schwartz*◇ Jesse Dodge*◇♣ Noah A. Smith◇♡ Oren Etzioni◇

◇Allen Institute for AI, Seattle, Washington, USA
♣ Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
♡ University of Washington, Seattle, Washington, USA

July 2019

aiXiv:1907.10597

LIMIT #2

I do not know how to define this limit…

…è di finezze che si distingue una persona di spessore.

Quell'uomo era così onesto che anche il caffè lo prendeva corretto.

Non contare sulle altre persone, la somma potrebbe essere zero.

Un'errore è corretto per coerenza.

LIMIT #2

Siamo fatti di sorrisi e silenzi,

sorrisi e silenzi.

I primi vincono,

i secondi passano.

Will AI be able to generate it?

Cristiano-De-Nobili

Thanks

@denocris

@denocris

C1A0 EXPO

OUR SPONSORS

MAIN SPONSOR

iren

SPONSOR

GRUPPO FOS soluzioni ad alta tecnologia

IFM Infomaster

indigo.ai

avanade

ESA ECOLOGIA SOLUZIONE AMBIENTE

aitek

firedesktop burn document barriers

FINCANTIERI SI

RINA

ERG EVOLVING ENERGIES

MAPS SHARING KNOWLEDGE

UniCredit

FLAIRBIT DATA CENTRIC SOLUTIONS

GIOVANI IMPRENDITORI CONFINDUSTRIA GENOVA

SPONSOR TECNICO

upen SEARCH

C1A0 EXPO 2019

C1A0
EXPO

15-16
Novembre 2019
Genova
#C1A0EXPO