Learning General Language Processing Agents

Dani Yogatama

DeepMind

Language and Intelligence

A **uniquely human** ability that is a **core component** of our **intelligence**, **independent of** the **surface forms** it manifests in (Hockett, 1960).

Language and Intelligence

A **primary medium** through which we **acquire** new skills and knowledge (+visual perception).

Language and Intelligence

The **most effective** form of communication to **transmit** information and knowledge to others.
(Language for communication; Wittgenstein, 1953; Austin, 1975)

Language and Intelligence

A mechanism with which we **formulate** our **thought process**. (Language for thinking; Spelke, 2003)

Language and Intelligence

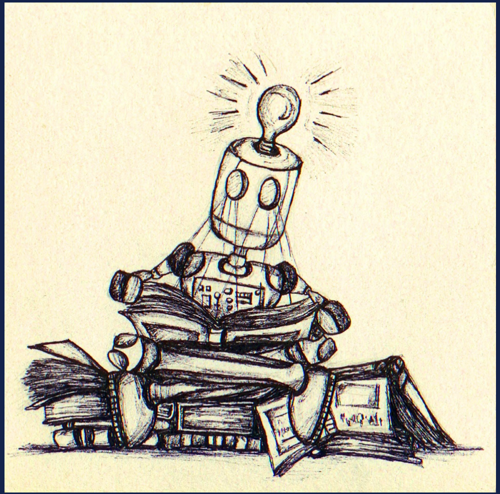
Language is key to **human intelligence** and is important for **artificial intelligence**.

General Linguistic Intelligence

The ability to **acquire**, **store**, and **reuse** knowledge (about a language's lexicon, syntax, semantics, and pragmatic conventions) to **adapt** to new tasks **quickly without forgetting** old ones.

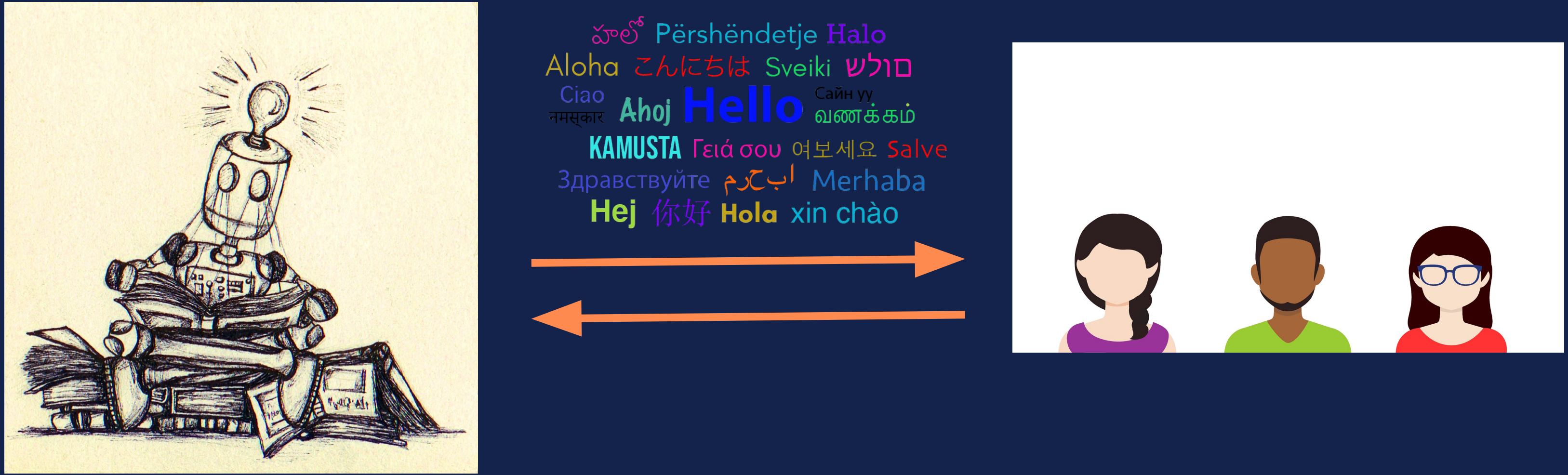
General Linguistic Intelligence

The ability to **acquire**, **store**, and **reuse** knowledge (about a language's lexicon, syntax, semantics, and pragmatic conventions) to **adapt** to new tasks **quickly without forgetting** old ones.



General Linguistic Intelligence

The ability to **acquire**, **store**, and **reuse** knowledge (about a language's lexicon, syntax, semantics, and pragmatic conventions) to **adapt** to new tasks **quickly without forgetting** old ones.



The State of Natural Language Processing

State-of-the-art models are based on increasingly larger transformers.

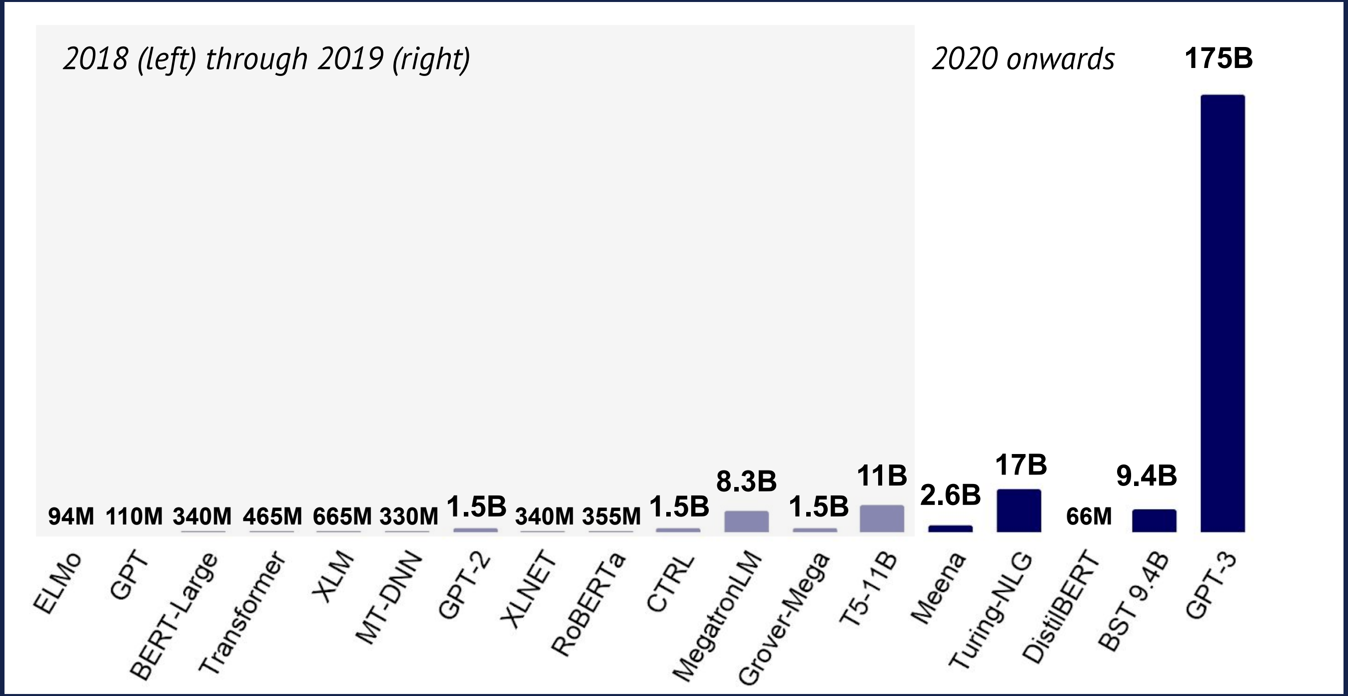


Figure taken from State of AI Report 2020.

Challenges: Human Learning vs. Machine Learning

Human	
``Large'' datasets	**Acquisition**

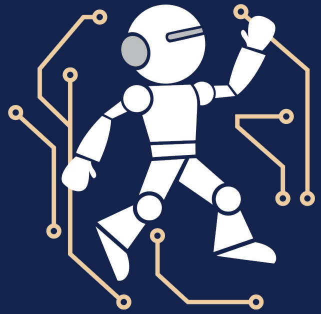
Challenges: Human Learning vs. Machine Learning

Human	
``Large'' datasets	**Acquisition**
Few examples	**Task Training**

Challenges: Human Learning vs. Machine Learning

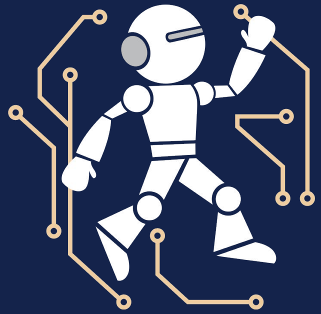
Human	
``Large'' datasets	**Acquisition**
Few examples	**Task Training**
Dataset agnostic	**Linguistic knowledge**
Generalizable to new tasks	**Generalization**

Challenges: Human Learning vs. Machine Learning



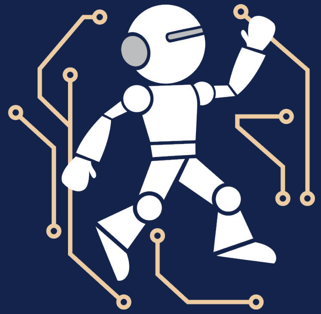
Human		Machine
``Large” datasets	**Acquisition**	Large datasets (representation learning)
Few examples	**Task Training**	
Dataset agnostic	**Linguistic knowledge**	
Generalizable to new tasks	**Generalization**	

Challenges: Human Learning vs. Machine Learning



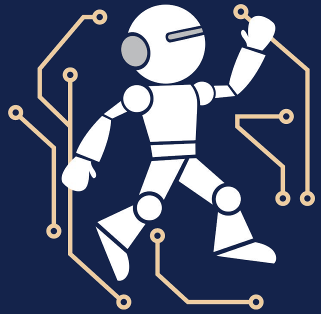
Human		Machine
``Large'' datasets	**Acquisition**	Large datasets (representation learning)
Few examples	**Task Training**	Large datasets (supervised fine tuning)
Dataset agnostic	**Linguistic knowledge**	
Generalizable to new tasks	**Generalization**	

Challenges: Human Learning vs. Machine Learning



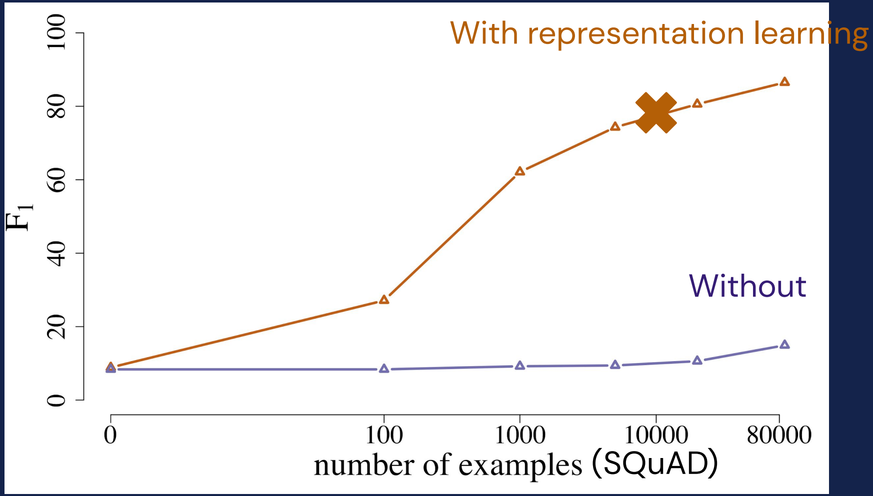
Human		Machine
``Large'' datasets	**Acquisition**	Large datasets (representation learning)
Few examples	**Task Training**	Large datasets (supervised fine tuning)
Dataset agnostic	**Linguistic knowledge**	Dataset specific
Generalizable to new tasks	**Generalization**	

Challenges: Human Learning vs. Machine Learning



Human		Machine
``Large" datasets	**Acquisition**	Large datasets (representation learning)
Few examples	**Task Training**	Large datasets (supervised fine tuning)
Dataset agnostic	**Linguistic knowledge**	Dataset specific
Generalizable to new tasks	**Generalization**	Forget previous tasks given a new task

The State of Natural Language Processing

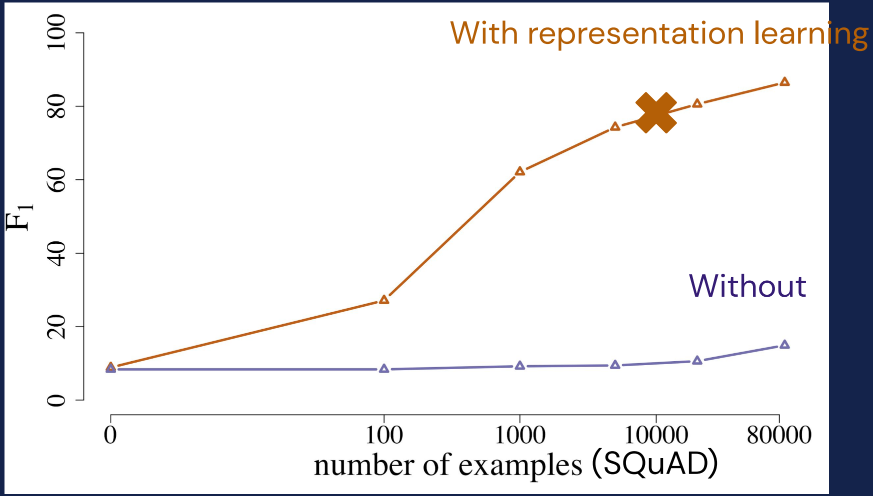


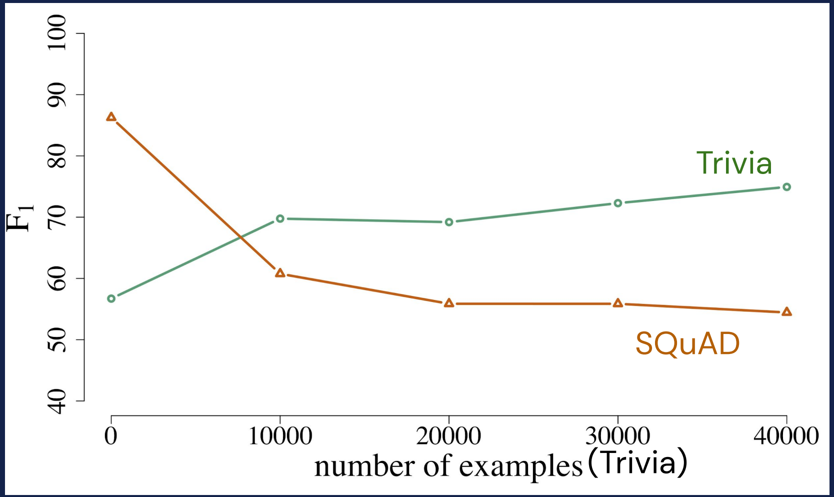
Yogatama et al., arXiv 2019

Model: BERT, Devlin et al. 2019

QA dataset: SQuAD, Rajpurkar et al., 2016

The State of Natural Language Processing





Yogatama et al., arXiv 2019

Model: BERT, Devlin et al. 2019

QA dataset: SQuAD, Rajpurkar et al., 2016

QA dataset 2: Trivia, Joshi et al., 2017

Research Areas

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

Research Areas

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

Training Paradigms

Model Architectures

Research Areas

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

Training Paradigms

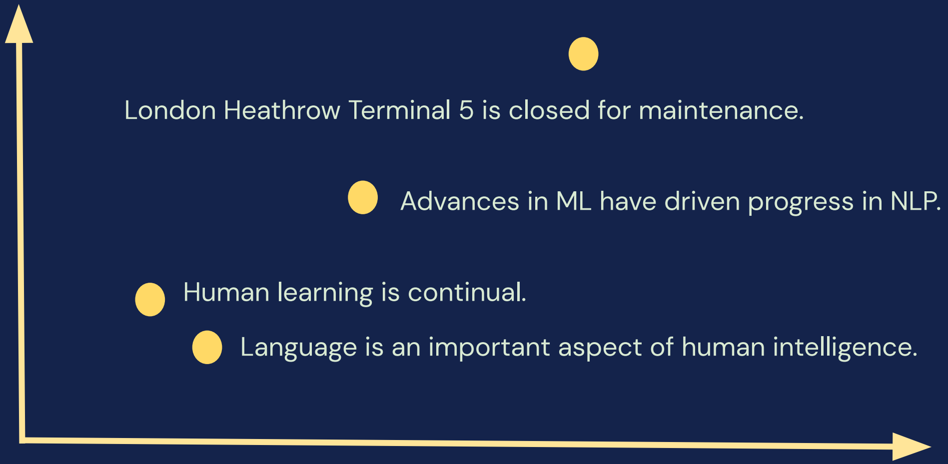
Representation Learning

Yogatama and Smith, ACL 2014

Yogatama et al., ACL 2015

Yogatama and Smith; ICML 2015

Kong, de Masson d'Autume, Ling, Yu, Dai, Yogatama; ICLR 2020



Research Areas

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

Training Paradigms

Generative Training

$$\mathcal{L} = \log p(\boldsymbol{x}, y)$$

Yogatama et al., TACL 2014

Yogatama et al., arXiv 2017

Kong, Melis, Ling, Yu, and Yogatama, ICLR 2018

Cao and Yogatama, arXiv 2020

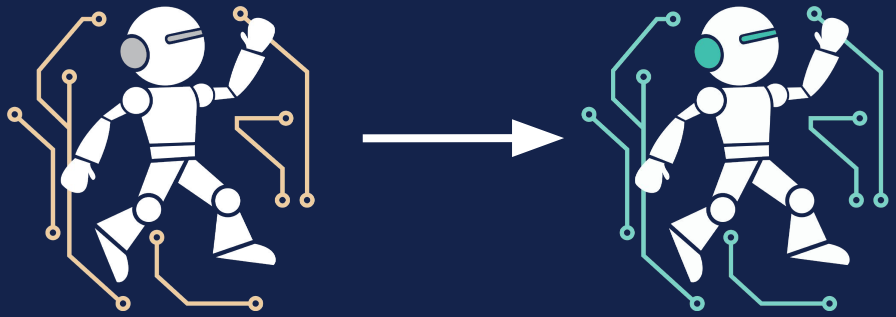
Research Areas

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

Training Paradigms

Few-shot and Transfer Learning

Yogatama and Mann, AISTATS 2014
Yogatama et al., EMNLP 2015
Artetxe, Ruder, Yogatama, ACL 2020



Research Areas

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

Model Architectures

Memory Networks

Yogatama et al., ICLR 2017

Yogatama et al., ICLR 2018

de Masson d'Autume, Ruder, Kong, Yogatama, NeurIPS 2019

Yogatama et al., TACL 2021

Research Areas

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

Model Architectures

Memory Networks

Yogatama et al., ICLR 2017

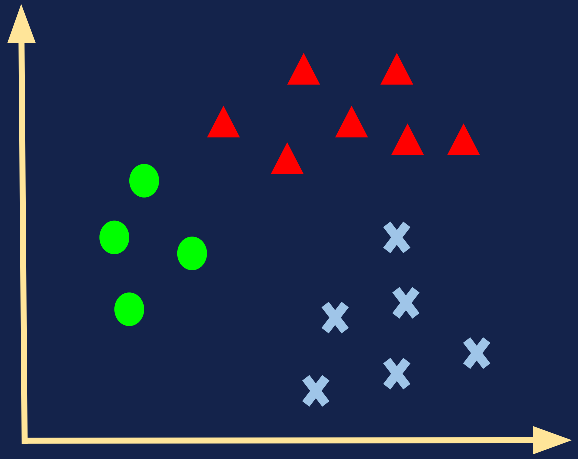
Yogatama et al., ICLR 2018

de Masson d'Autume, Ruder, Kong, Yogatama, NeurIPS 2019

Yogatama et al., TACL 2021

Research Areas

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.



Model Architectures

Memory Networks

Yogatama et al., ICLR 2017
Yogatama et al., ICLR 2018
de Masson d'Autume, Ruder, Kong, Yogatama, NeurIPS 2019
Yogatama et al., TACL 2021

Research Areas

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

Training Paradigms

Model Architectures

This Talk

- A framework for self-supervised language representation learning methods.

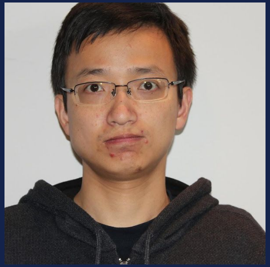
  Kong et al., ICLR 2020

- Semiparametric (memory-augmented) language models.

  Yogatama et al., TACL 2021

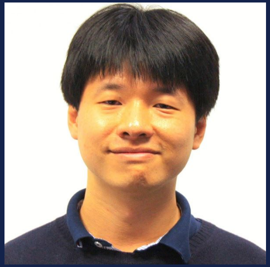
A Mutual Information Maximization Perspective of Language Representation Learning

Kong et al., ICLR 2020

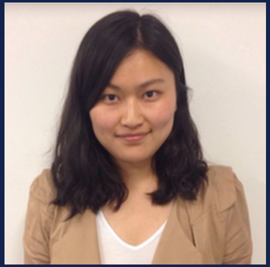


Lingpeng

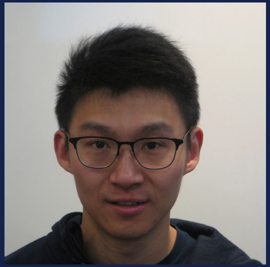
Cyprien



Wang



Lei

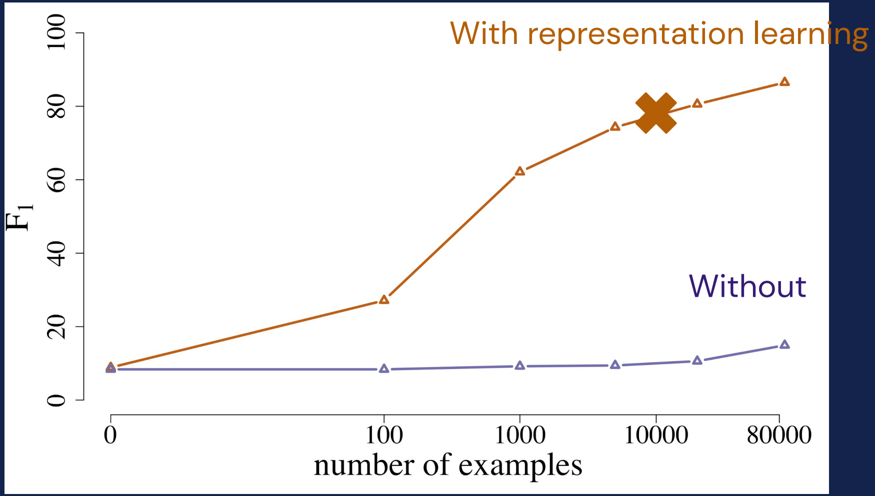


Zihang

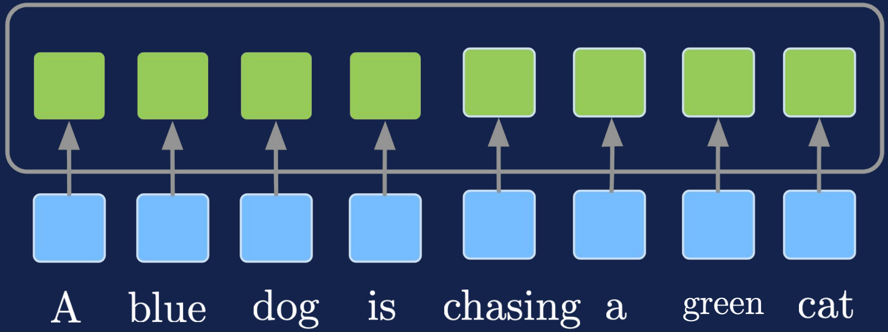
Dani

Text Representations

Good representations facilitate more efficient transfer.



Text Representations

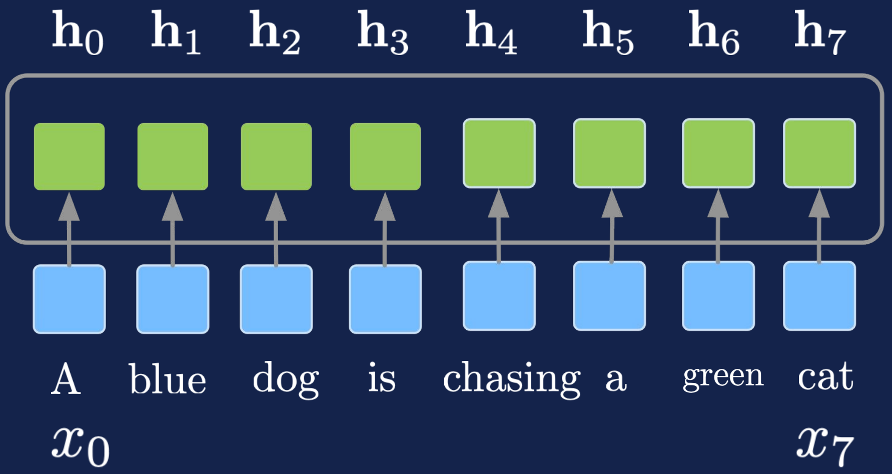


a sequence of vectors

(word representations)

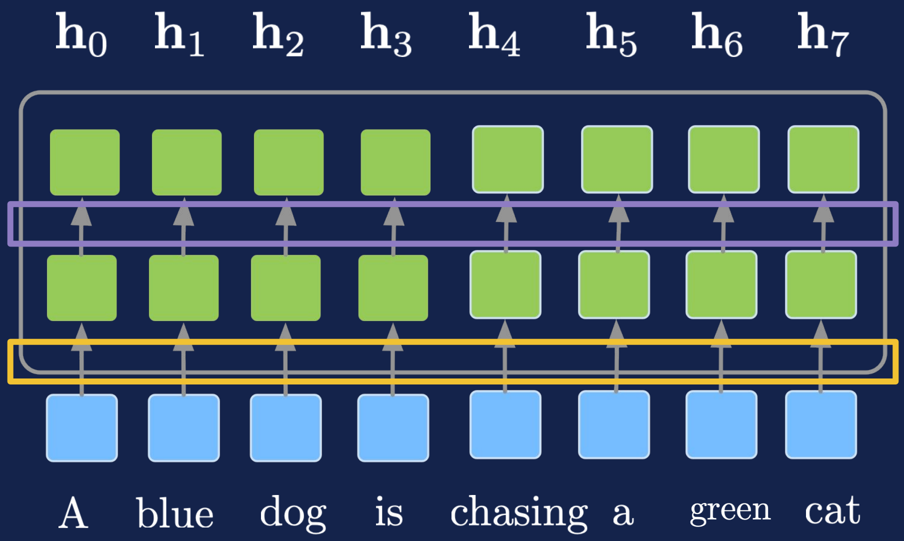
a sequence of words

Text Representations



$$\boxed{g_\omega(x_0, x_1, \ldots, x_7)} \rightarrow \mathbf{h}_0, \mathbf{h}_1, \ldots, \mathbf{h}_7$$

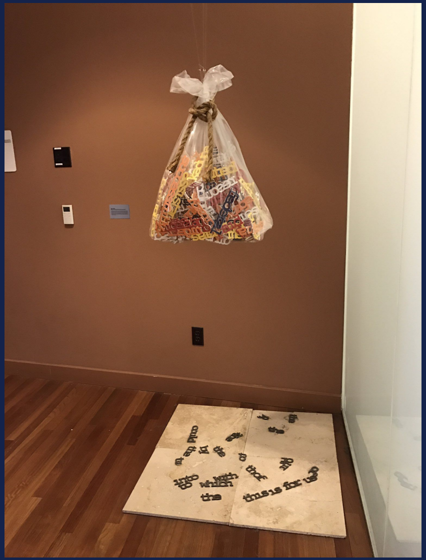
Text Representations



$$\boxed{g_\tau(\mathbf{i}_0, \mathbf{i}_1, \ldots, \mathbf{i}_7)} \rightarrow \mathbf{h}_0, \mathbf{h}_1, \ldots, \mathbf{h}_7$$

$$\boxed{g_\omega(x_0, x_1, \ldots, x_7)} \rightarrow \mathbf{i}_0, \mathbf{i}_1, \ldots, \mathbf{i}_7$$

Text Representations



https://twitter.com/SmithaMilli/status/837153616116985856/

Bag of words → **Word embeddings** → **Contextual word embeddings**

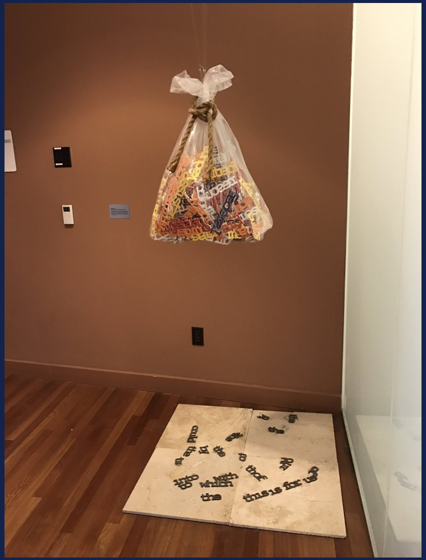
Skip gram, Mikolov et al., 2013.

GloVe, Pennington et al., 2014.

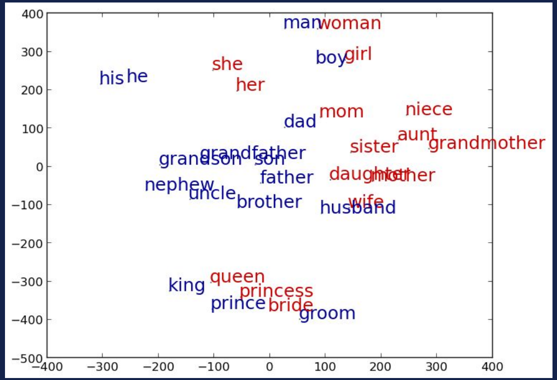
ELMo, Peters et al., 2018.

BERT, Devlin et al., 2019.

Text Representations



https://twitter.com/SmithaMilli/status/837153616116985856/



Bag of words

Word embeddings

Contextual word embeddings

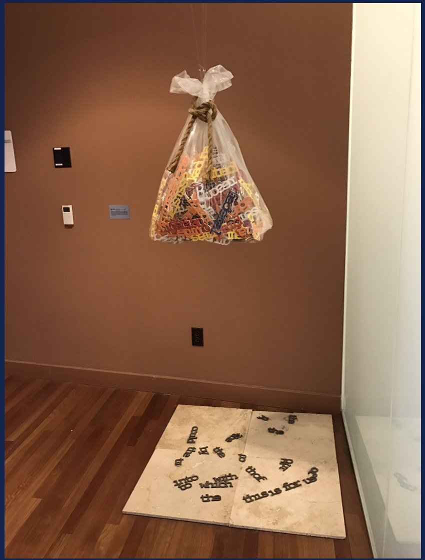
Skip gram, Mikolov et al., 2013.

GloVe, Pennington et al., 2014.

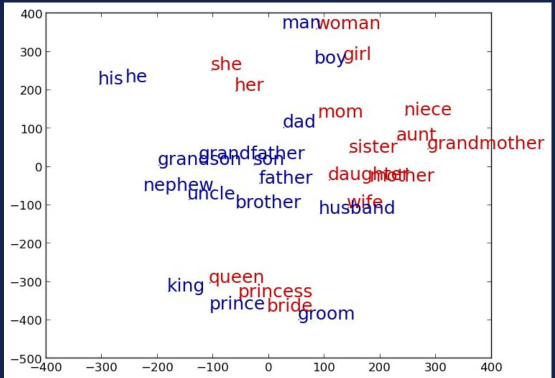
ELMo, Peters et al., 2018.

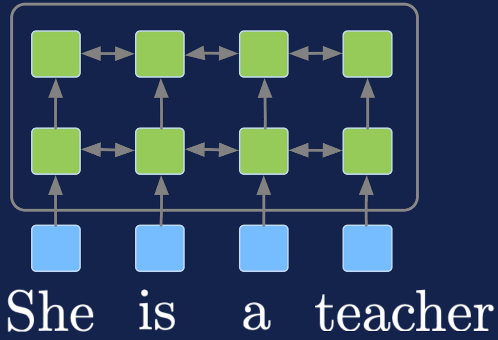
BERT, Devlin et al., 2019.

Text Representations



https://twitter.com/SmithaMilli/status/837153616116985856/





Bag of words

Word embeddings

Skip gram, Mikolov et al., 2013.

GloVe, Pennington et al., 2014.

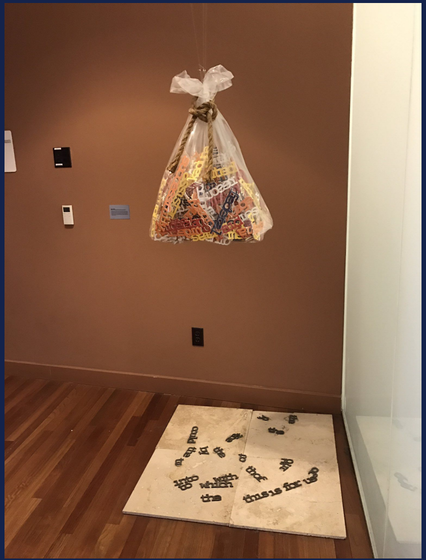
Contextual word embeddings

ELMo, Peters et al., 2018.

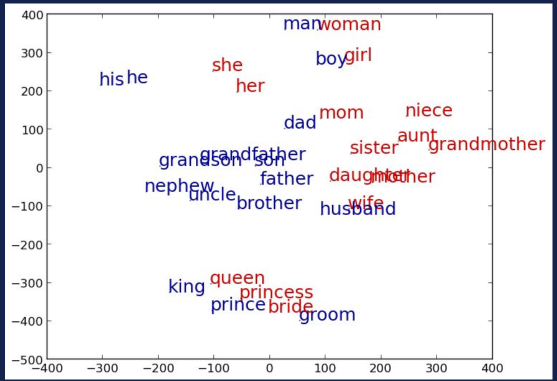
BERT, Devlin et al., 2019.

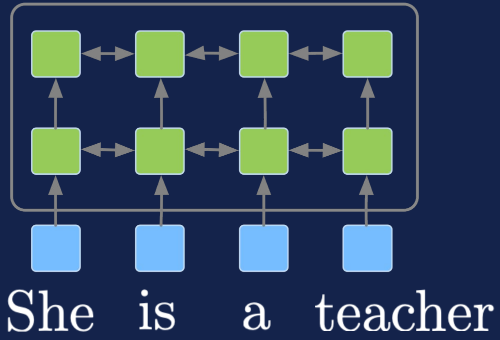
Text Representations

What has been the main driver of progress so far?



https://twitter.com/SmithaMilli/status/837153616116985856/





Bag of words

Word embeddings

Contextual word embeddings

Skip gram, Mikolov et al., 2013.

GloVe, Pennington et al., 2014.

ELMo, Peters et al., 2018.

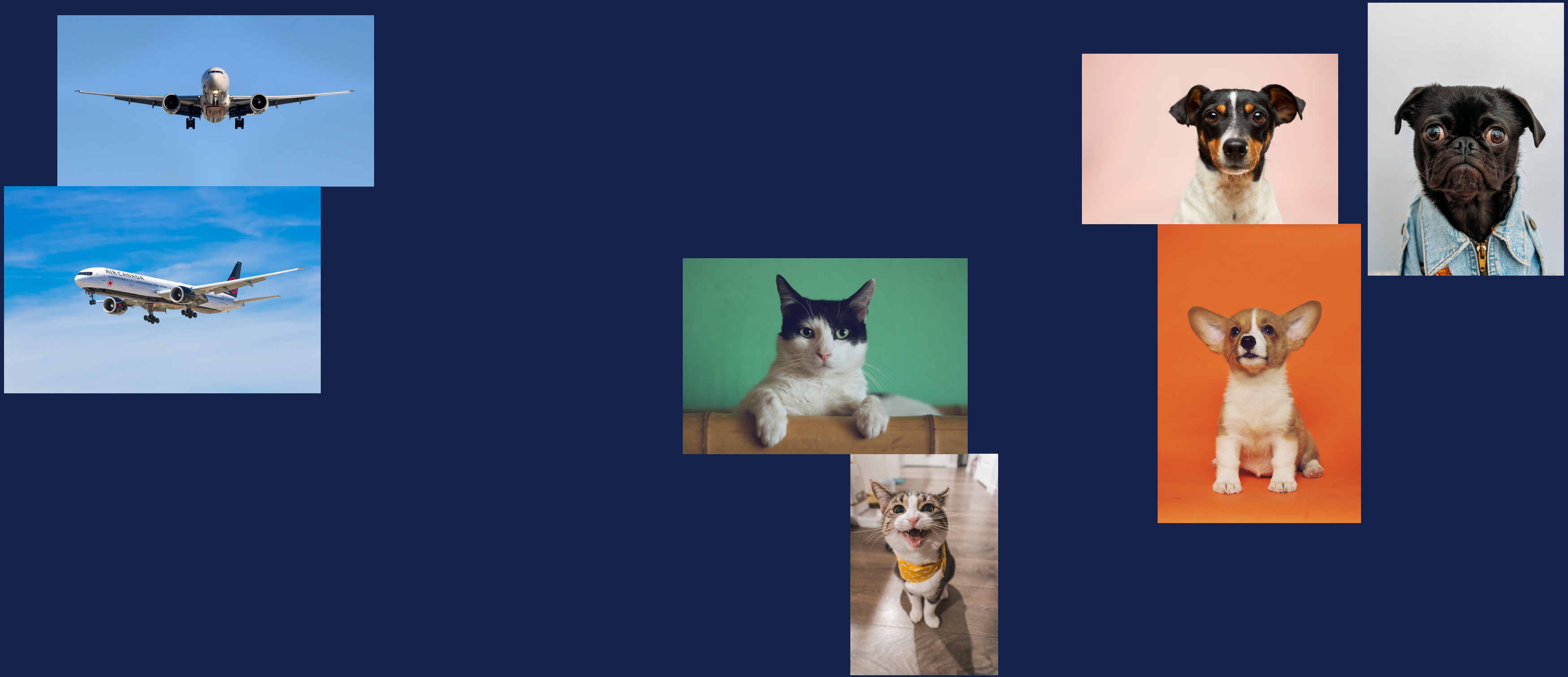
BERT, Devlin et al., 2019.

Contrastive Learning

Main assumption: representations should capture similarity (Arora et al., 2019).

Contrastive Learning

Main assumption: representations should capture similarity (Arora et al., 2019).



Contrastive Learning

Main assumption: representations should capture similarity (Arora et al., 2019).

Human learning is continual.

Advances in ML have driven progress in NLP.

Logistic regression can be used for classification.

Transformer uses self attention.

There are many direct flights between London and Tokyo.

London Heathrow Terminal 5 is closed for maintenance.

Contrastive Learning with InfoNCE

Main assumption: representations should capture similarity (Arora et al., 2019).

$$I(A, B) \geq \mathbb{E}_{p(A,B)}\left[\mathbb{E}_{p(C)}\left[\log \frac{\exp f_{\boldsymbol{\theta}}(a, b)}{\exp f_{\boldsymbol{\theta}}(a, b) + \sum_{c \neq b} \exp f_{\boldsymbol{\theta}}(a, c)}\right]\right]$$

InfoNCE objective

Logeswaran and Lee, 2018

van den Oord, et al., 2019

Contrastive Learning with InfoNCE

Main assumption: representations should capture similarity (Arora et al., 2019).

$$\boxed{I(A, B)} \geq \mathbb{E}_{p(A,B)}\left[\mathbb{E}_{p(C)}\left[\log \frac{\exp f_{\boldsymbol{\theta}}(a, b)}{\exp f_{\boldsymbol{\theta}}(a, b) + \sum_{c \neq b} \exp f_{\boldsymbol{\theta}}(a, c)}\right]\right]$$

InfoNCE objective

Logeswaran and Lee, 2018

van den Oord, et al., 2019

Contrastive Learning with InfoNCE

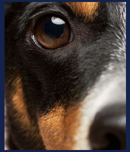
Main assumption: representations should capture similarity (Arora et al., 2019).

$$I(A,B) \geq \mathbb{E}_{p(A,B)}\left[\mathbb{E}_{p(C)}\left[\log \frac{\exp \boxed{f_{\boldsymbol{\theta}}(a,b)}}{\exp f_{\boldsymbol{\theta}}(a,b) + \sum_{c \neq b} \exp f_{\boldsymbol{\theta}}(a,c)}\right]\right]$$

InfoNCE objective

Logeswaran and Lee, 2018

van den Oord, et al., 2019



High when **a** and **b** go together

Contrastive Learning with InfoNCE

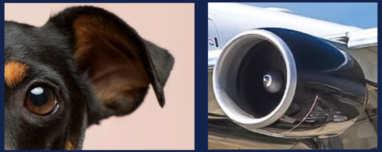
Main assumption: representations should capture similarity (Arora et al., 2019).

$$I(A,B) \geq \mathbb{E}_{p(A,B)}\left[\mathbb{E}_{p(C)}\left[\log \frac{\exp f_{\boldsymbol{\theta}}(a,b)}{\exp f_{\boldsymbol{\theta}}(a,b) + \sum_{c \neq b} \exp \boxed{f_{\boldsymbol{\theta}}(a,c)}}\right]\right]$$

InfoNCE objective

Logeswaran and Lee, 2018

van den Oord, et al., 2019



Low when **a** and **c** do not go together

Contrastive Learning with InfoNCE

$$\mathbb{E}_{p(A,B)}\left[\mathbb{E}_{p(C)}\left[\log \frac{\exp f_{\boldsymbol{\theta}}(a,b)}{\exp f_{\boldsymbol{\theta}}(a,b) + \sum_{c\neq b} \exp f_{\boldsymbol{\theta}}(a,c)}\right]\right]$$

The University of Waterloo is located in Canada

Contrastive Learning with InfoNCE

$$\mathbb{E}_{p(A,B)}\left[\mathbb{E}_{p(C)}\left[\log \frac{\exp \boxed{f_{\boldsymbol{\theta}}(a,b)}}{\exp f_{\boldsymbol{\theta}}(a,b) + \sum_{c \neq b} \exp f_{\boldsymbol{\theta}}(a,c)}\right]\right]$$

The $\overset{a}{\boxed{\text{University}}}$ of $\overset{b}{\boxed{\text{Waterloo}}}$ is located in Canada

Contrastive Learning with InfoNCE

$$\mathbb{E}_{p(A,B)}\left[\mathbb{E}_{p(C)}\left[\log \frac{\exp \boxed{f_{\boldsymbol{\theta}}(a,b)}}{\exp f_{\boldsymbol{\theta}}(a,b) + \sum_{c \neq b} \exp f_{\boldsymbol{\theta}}(a,c)}\right]\right]$$

$$\overset{a}{\boxed{\text{The University of}}}\ \overset{b}{\boxed{\text{Waterloo}}}\ \overset{a}{\boxed{\text{is located in Canada}}}$$

Contrastive Learning with InfoNCE

$$\mathbb{E}_{p(A,B)}\left[\mathbb{E}_{p(C)}\left[\log \frac{\exp \boxed{f_{\boldsymbol{\theta}}(a,b)}}{\exp f_{\boldsymbol{\theta}}(a,b) + \sum_{c \neq b} \exp f_{\boldsymbol{\theta}}(a,c)}\right]\right]$$

$$\underbrace{\boxed{\text{The University of Waterloo}}}_{a}\ \text{is located in}\ \underbrace{\boxed{\text{Canada}}}_{b}$$

Contrastive Learning with InfoNCE

$$\mathbb{E}_{p(A,B)}\left[\mathbb{E}_{p(C)}\left[\log \frac{\exp \boxed{f_{\boldsymbol{\theta}}(a,b)}}{\exp \boxed{f_{\boldsymbol{\theta}}(a,b)} + \sum_{c \neq b} \exp f_{\boldsymbol{\theta}}(a,c)}\right]\right]$$

a b

$\boxed{\text{The University of Waterloo}}$ is located in $\boxed{\text{Canada}}$

$$\boxed{f_{\boldsymbol{\theta}}(a,b)} = g_{\boldsymbol{\psi}}(b)^{\top} g_{\boldsymbol{\omega}}(a)$$

Contrastive Learning with InfoNCE

$$\mathbb{E}_{p(A,B)}\left[\mathbb{E}_{p(C)}\left[\log \frac{\exp f_{\boldsymbol{\theta}}(a,b)}{\exp f_{\boldsymbol{\theta}}(a,b) + \boxed{\sum_{c \neq b} \exp f_{\boldsymbol{\theta}}(a,c)}}\right]\right]$$

Tokyo
London
dog
cat

$\overset{a}{\boxed{\text{The University of Waterloo}}}$ is located in $\overset{b}{\boxed{\text{Canada}}}$

$$f_{\boldsymbol{\theta}}(a,b) = g_{\boldsymbol{\psi}}(b)^{\top} g_{\boldsymbol{\omega}}(a)$$

Contrastive Learning with InfoNCE

$$\mathbb{E}_{p(A,B)}\left[\boxed{\mathbb{E}_{p(C)}}\left[\log \frac{\exp f_{\boldsymbol{\theta}}(a,b)}{\exp f_{\boldsymbol{\theta}}(a,b) + \sum_{c \neq b} \exp f_{\boldsymbol{\theta}}(a,c)}\right]\right]$$

$\underbrace{\boxed{\text{The University of Waterloo}}}_{}$ is located in $\boxed{\text{Canada}}$

a — The University of Waterloo; b — Canada

$$f_{\boldsymbol{\theta}}(a,b) = g_{\boldsymbol{\psi}}(b)^{\top} g_{\boldsymbol{\omega}}(a)$$

Contrastive Learning with InfoNCE

$$\boxed{\mathbb{E}_{p(A,B)}}\left[\mathbb{E}_{p(C)}\left[\log \frac{\exp f_{\boldsymbol{\theta}}(a,b)}{\exp f_{\boldsymbol{\theta}}(a,b) + \sum_{c \neq b} \exp f_{\boldsymbol{\theta}}(a,c)}\right]\right]$$

$$\overset{a}{\boxed{\text{The University of Waterloo}}}\ \text{is located in}\ \overset{b}{\boxed{\text{Canada}}}$$

$$f_{\boldsymbol{\theta}}(a,b) = g_{\boldsymbol{\psi}}(b)^{\top} g_{\boldsymbol{\omega}}(a)$$

Contrastive Learning with InfoNCE

$$\mathbb{E}_{p(A,B)}\left[\mathbb{E}_{p(C)}\left[\log \frac{\exp f_{\boldsymbol{\theta}}(a,b)}{\exp f_{\boldsymbol{\theta}}(a,b) + \sum_{c\neq b} \exp f_{\boldsymbol{\theta}}(a,c)}\right]\right]$$

$\underbrace{\text{The University of Waterloo}}_{a}$ is located in $\underbrace{\text{Canada}}_{b}$

$$f_{\boldsymbol{\theta}}(a,b) = g_{\boldsymbol{\psi}}(b)^\top \boxed{g_{\boldsymbol{\omega}}}(a)$$

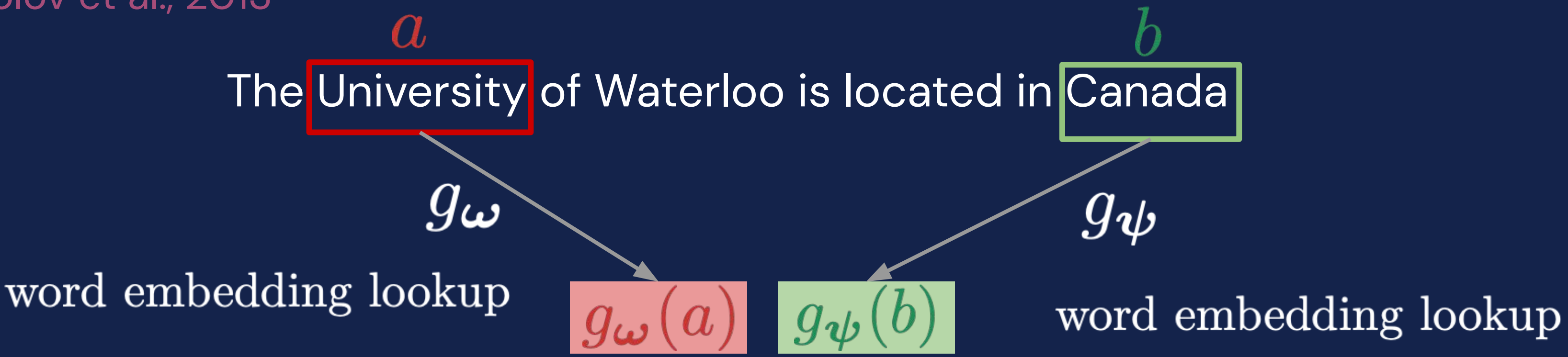
Skip-gram

Mikolov et al., 2013

The University of Waterloo is located in Canada

Skip-gram

Mikolov et al., 2013



BERT

Devlin et al., 2019

The University of Waterloo is located in Canada

BERT

Devlin et al., 2019

The University of Waterloo is located in $\overset{b}{\boxed{\text{Canada}}}$

g_ψ

$\boxed{g_\psi(b)}$ word embedding lookup

BERT

Devlin et al., 2019

a

$\boxed{\text{The University of Waterloo is located in Canada}}$

g_ω

transformer

Vaswani et al., 2017

$g_\omega(a)$ $g_\psi(b)$

Why is this interesting?

- A framework that unifies classical and modern word embedding methods.

		a	b	g_ω	g_ψ
Mikolov et al., 2013	**Skip-gram**	word	word	lookup	lookup
Devlin et al., 2019	**BERT**	context	word	transformer	lookup
Yang et al., 2019	**XLNet**	context	word	TXL++	lookup

Why is this interesting?

- A framework that unifies classical and modern word embedding methods.

		a	b	g_ω	g_ψ
Mikolov et al., 2013	**Skip-gram**	word	word	lookup	lookup
Devlin et al., 2019	**BERT**	context	word	transformer	lookup
Yang et al., 2019	**XLNet**	context	word	TXL++	lookup

- Provides connections to methods used in other domains (vision, speech).

Why is this interesting?

- A framework that unifies classical and modern word embedding methods.

		a	b	g_ω	g_ψ
Mikolov et al., 2013	**Skip-gram**	word	word	lookup	lookup
Devlin et al., 2019	**BERT**	context	word	transformer	lookup
Yang et al., 2019	**XLNet**	context	word	TXL++	lookup

- Provides connections to methods used in other domains (vision, speech).

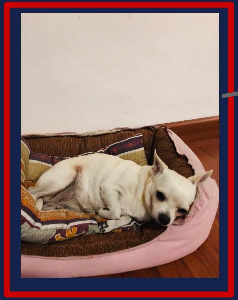
- Facilitates exchanges of ideas on how to improve representation learning models.

Model

Deep InfoMax (DIM; Hjelm et al., 2019)

Model

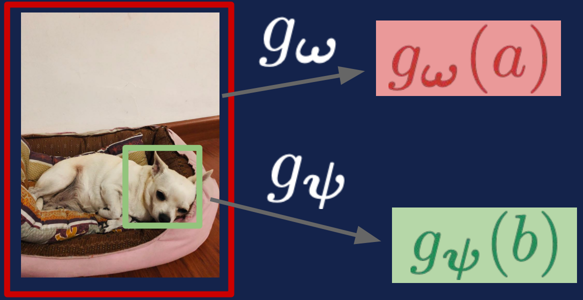
Deep InfoMax (DIM; Hjelm et al., 2019)



$\xrightarrow{g_\omega} \boxed{g_\omega(a)}$

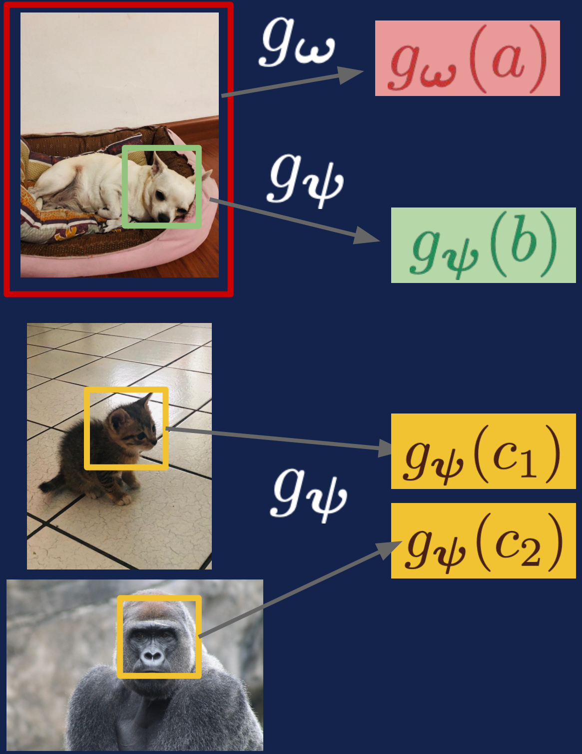
Model

Deep InfoMax (DIM; Hjelm et al., 2019)



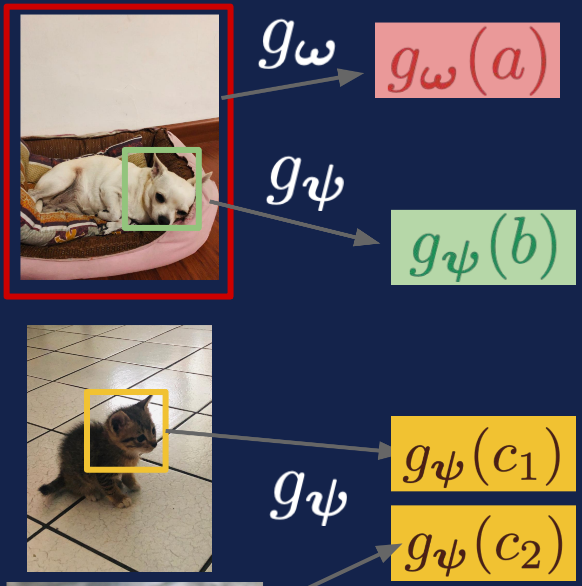
Model

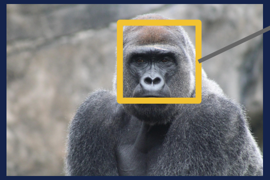
Deep InfoMax (DIM; Hjelm et al., 2019)



Model

Deep InfoMax (DIM; Hjelm et al., 2019)

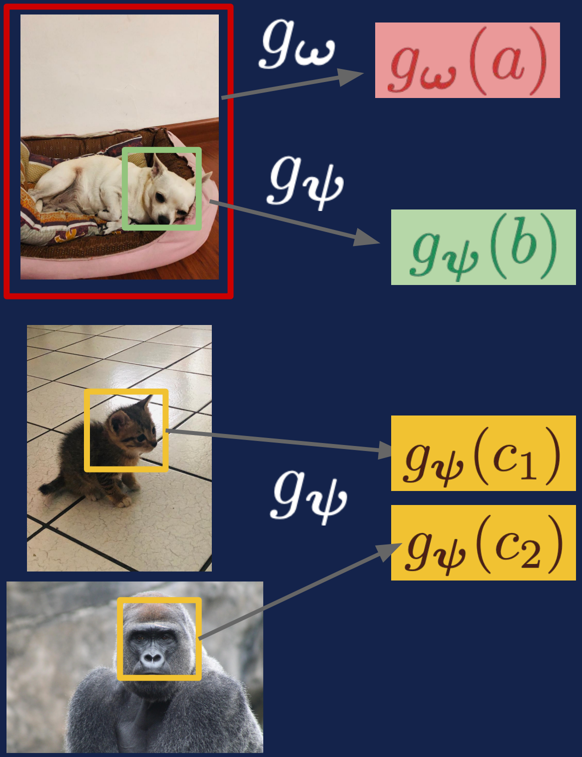




$$\mathcal{I}_{\mathrm{DIM}} = \mathbb{E}_{p(A,B)}\left[\mathbb{E}_{p(C)}\left[\log \frac{\exp \boxed{g_\omega(a)}^\top \boxed{g_\psi(b)}}{\exp \boxed{g_\omega(a)}^\top \boxed{g_\psi(b)} + \sum_{c\neq b} \exp \boxed{g_\omega(a)}^\top \boxed{g_\psi(c)}}\right]\right]$$

Model

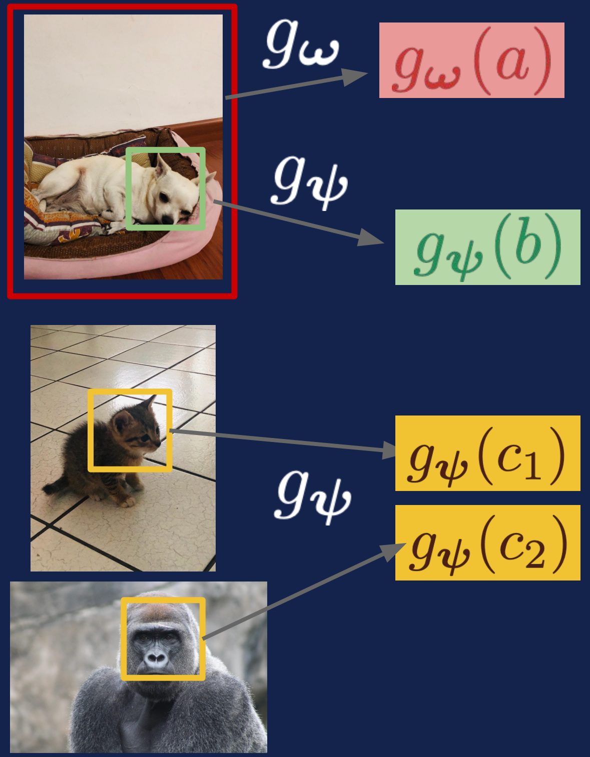
Deep InfoMax (DIM; Hjelm et al., 2019)

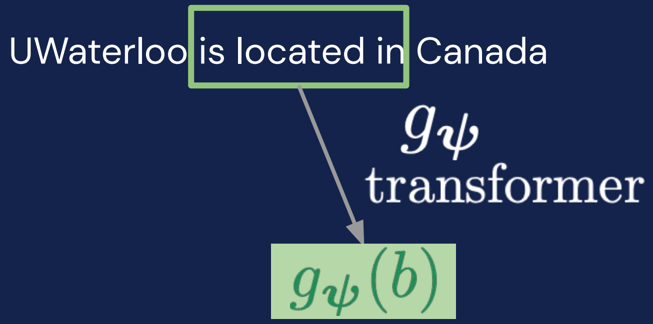


UWaterloo is located in Canada

Model

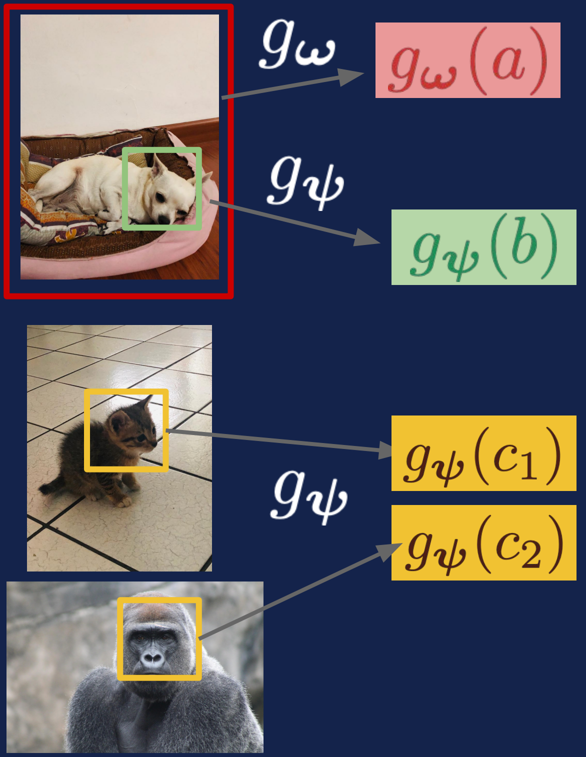
Deep InfoMax (DIM; Hjelm et al., 2019)

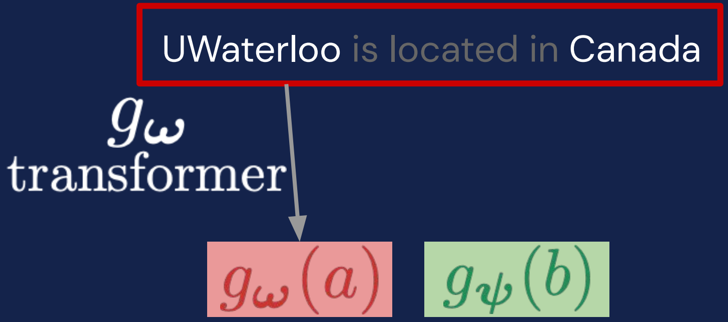




Model

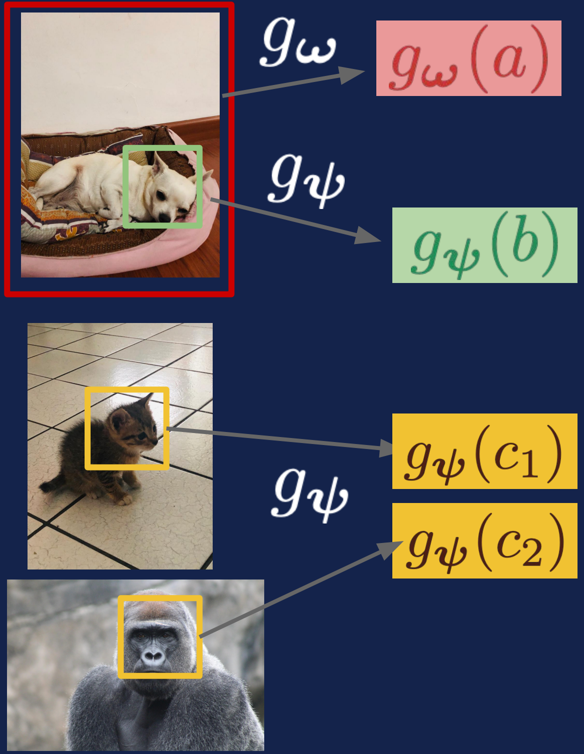
Deep InfoMax (DIM; Hjelm et al., 2019)





Model

Deep InfoMax (DIM; Hjelm et al., 2019)



UWaterloo is located in Canada

$g_\omega(a)$ $g_\psi(b)$

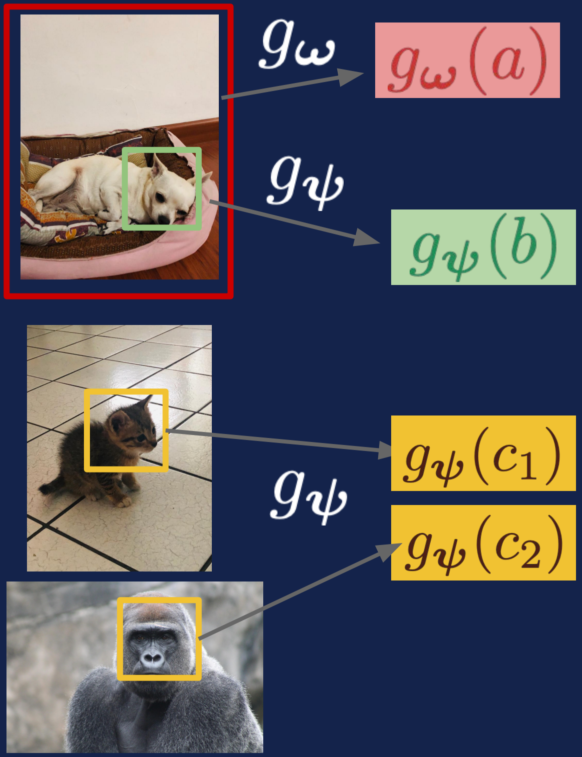
Starcraft II is a fun game

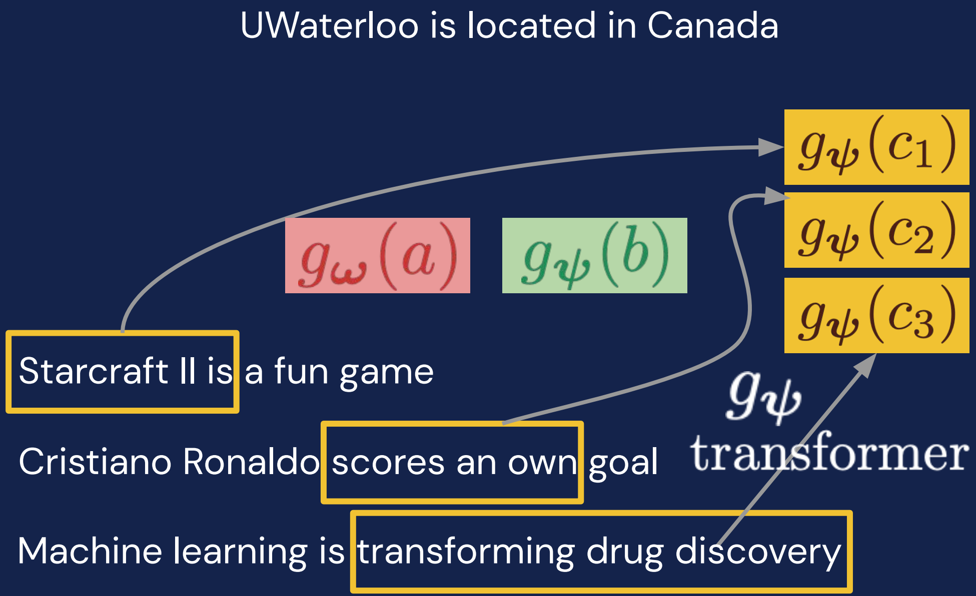
Cristiano Ronaldo scores an own goal

Machine learning is transforming drug discovery

Model

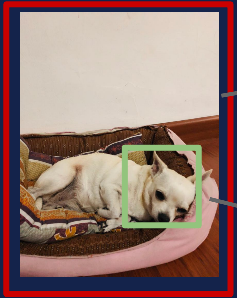
Deep InfoMax (DIM; Hjelm et al., 2019)





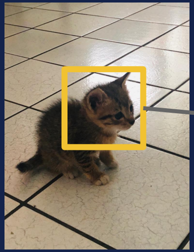
Model

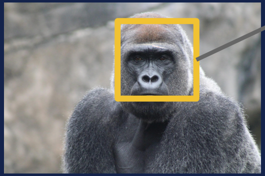
Deep InfoMax (DIM; Hjelm et al., 2019)



g_ω → $g_\omega(a)$

g_ψ → $g_\psi(b)$





g_ψ → $g_\psi(c_1)$, $g_\psi(c_2)$

UWaterloo is located in Canada

$g_\omega(a)$ $g_\psi(b)$

$g_\psi(c_1)$

$g_\psi(c_2)$

$g_\psi(c_3)$

Starcraft II is a fun game

Cristiano Ronaldo scores an own goal

Machine learning is transforming drug discovery

$$\mathcal{I}_{\text{DIM}} = \mathbb{E}_{p(A,B)}\left[\mathbb{E}_{p(C)}\left[\log \frac{\exp g_\omega(a)^\top g_\psi(b)}{\exp g_\omega(a)^\top g_\psi(b) + \sum_{c\neq b}\exp g_\omega(a)^\top g_\psi(c)}\right]\right]$$

Experiments

Question answering on SQuAD (Rajpurkar et al., 2016).

		F1
Small Model	BERT	90.9
	Ours	**91.4**
Large Model	BERT	92.7
	Ours	**93.1**

F1 scores (0–100), higher is better.

BERT: Devlin et al., 2019.

Takeaways

- It is possible to transfer ideas across domains when designing representation learning methods.

Takeaways

- It is possible to transfer ideas across domains when designing representation learning methods.

- Progress in language representation learning has largely been driven by advances in model architectures.

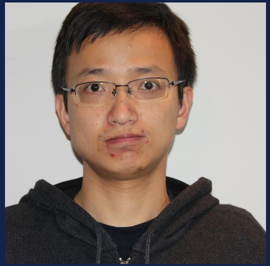
	a	b	$\boxed{g_\omega}$	g_ψ
Skip-gram	word	word	lookup	lookup
BERT	context	word	transformer	lookup
XLNet	context	word	TXL++	lookup

Adaptive Semiparametric Language Models

Yogatama et al., TACL 2021

Dani

Cyprien



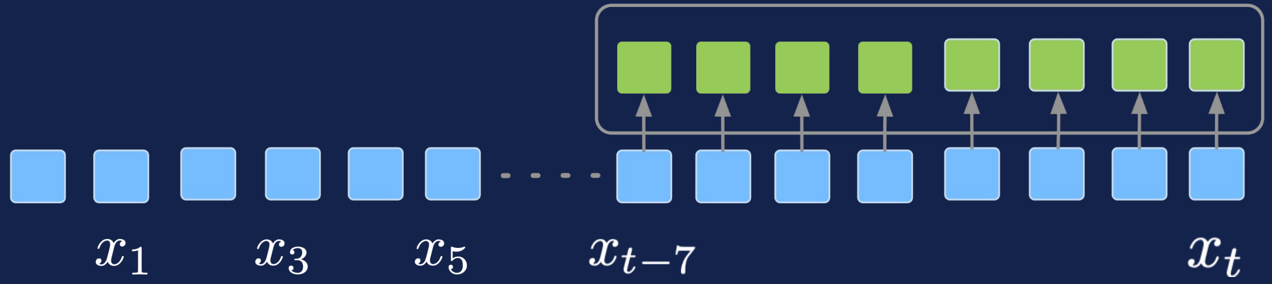
Lingpeng

Background

What are core limitations of of existing architectures?

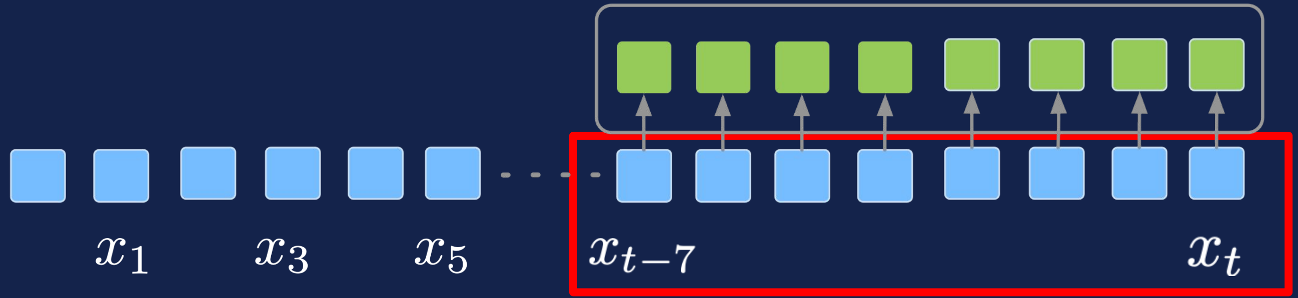
Background

State of the art architectures (transformers) are limited by the input sequence length.



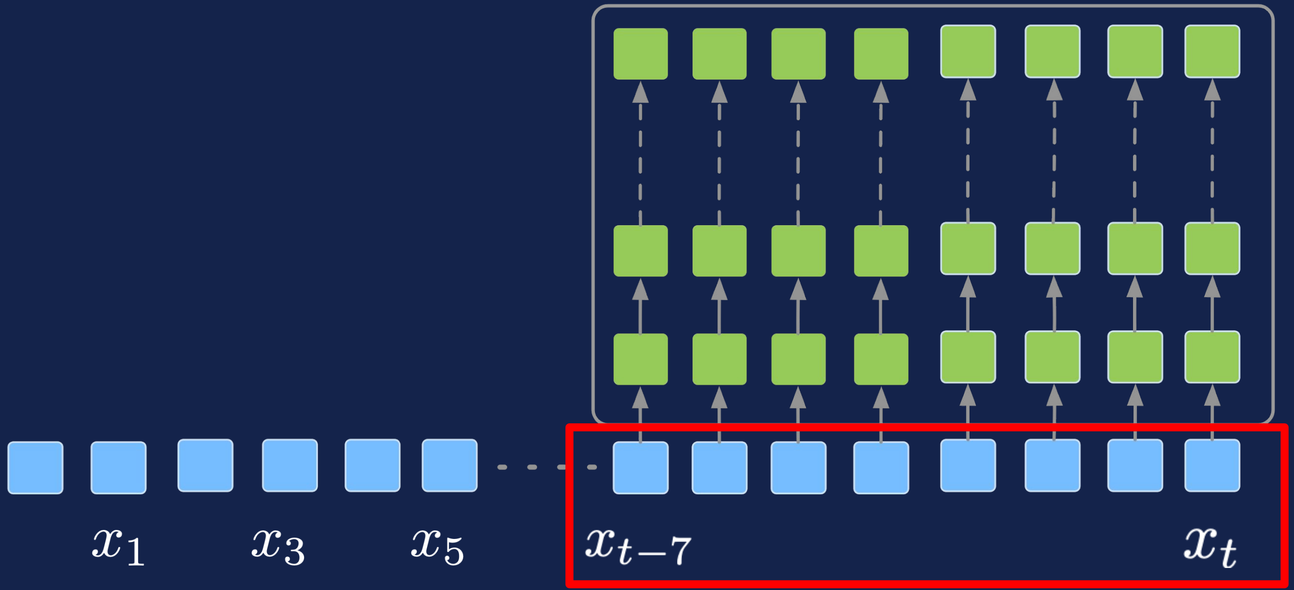
Background

State of the art architectures (transformers) are limited by the input sequence length.



Background

State of the art architectures (transformers) are limited by the input sequence length.

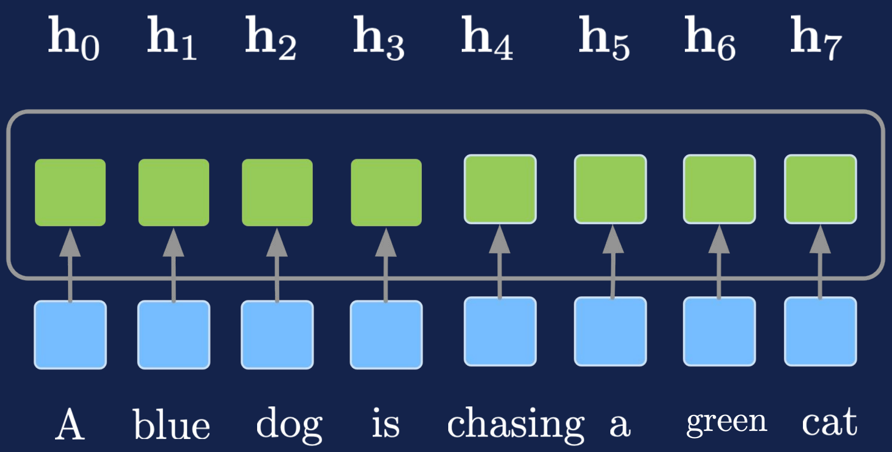


Background

Knowledge is encoded in the weights of a parametric neural network.

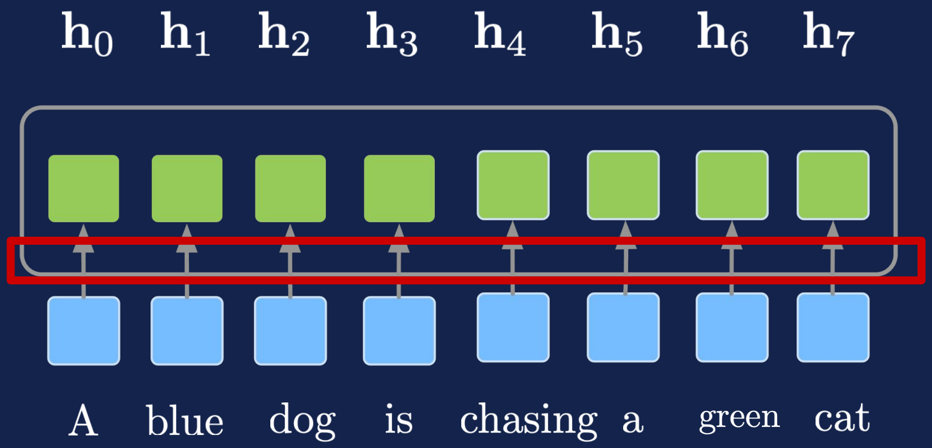
Background

Knowledge is encoded in the weights of a parametric neural network.



Background

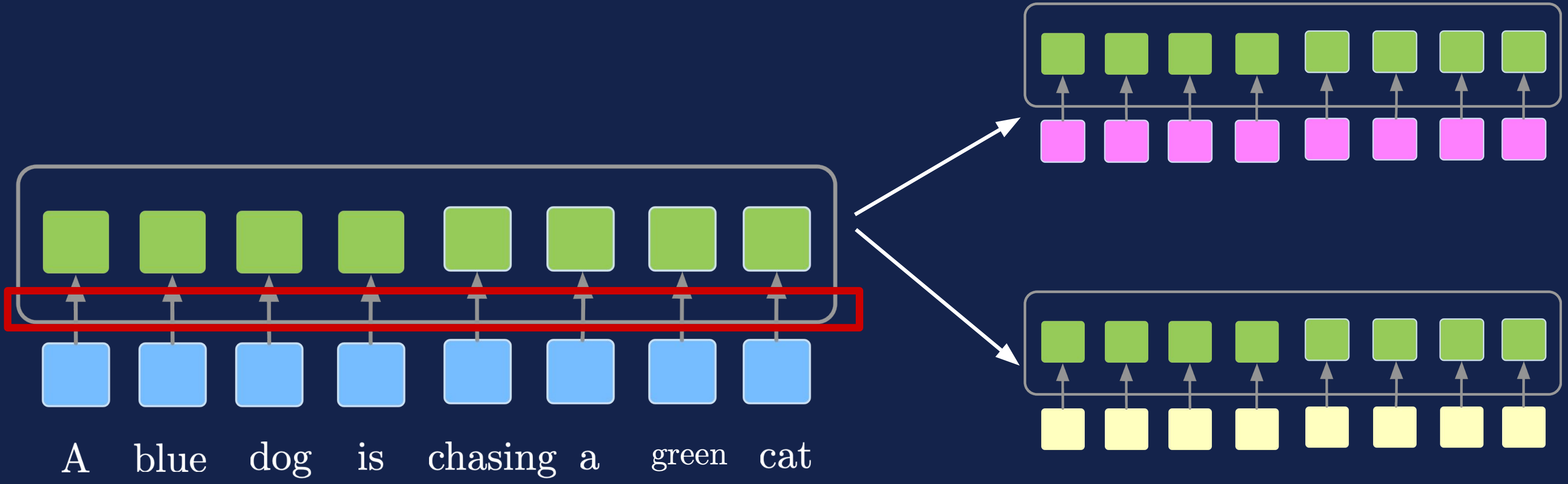
Knowledge is encoded in the weights of a parametric neural network.



$$g_\omega(x_0, x_1, \ldots, x_7) \rightarrow \mathbf{h}_0, \mathbf{h}_1, \ldots, \mathbf{h}_7$$

Background

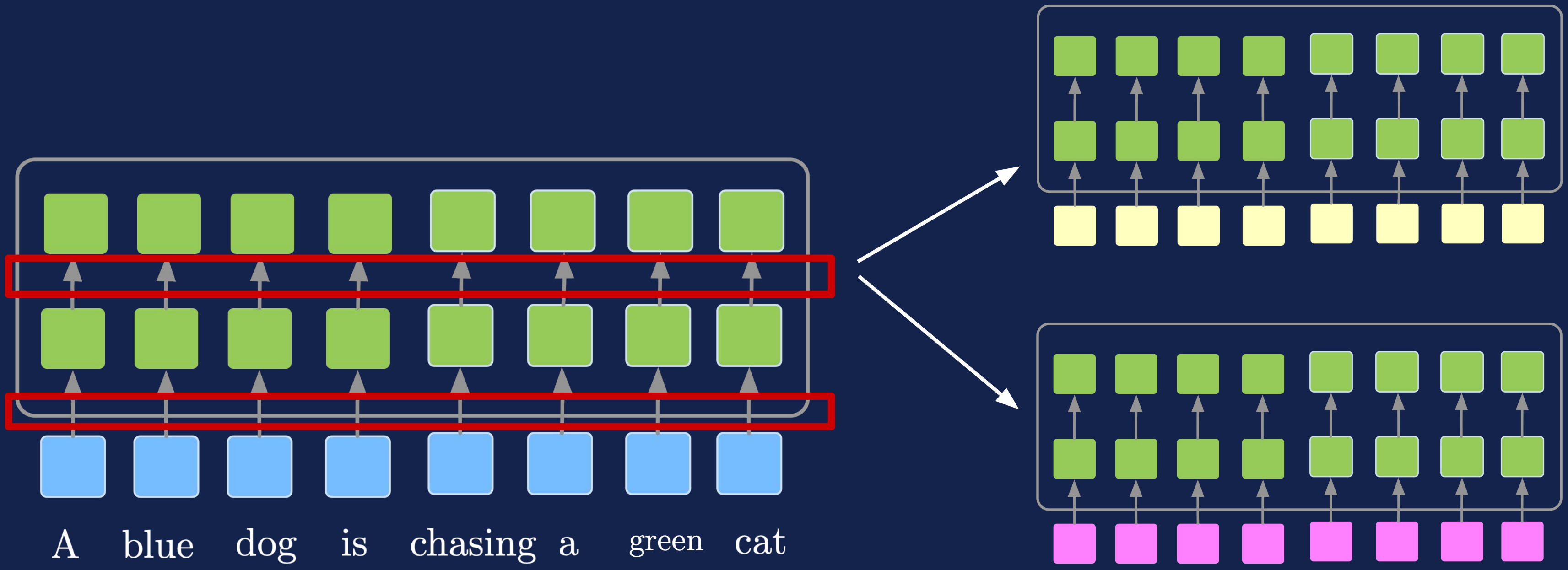
Knowledge is encoded in the weights of a parametric neural network.



Update weights with new knowledge ⟶ changes affect all examples (sequences).

Background

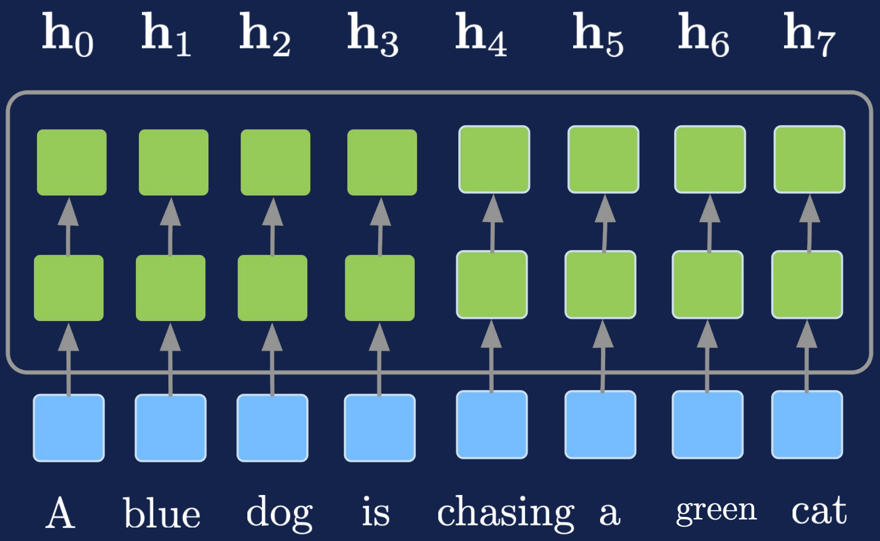
Knowledge is encoded in the weights of a parametric neural network.



Update weights with new knowledge → changes affect all examples (sequences).

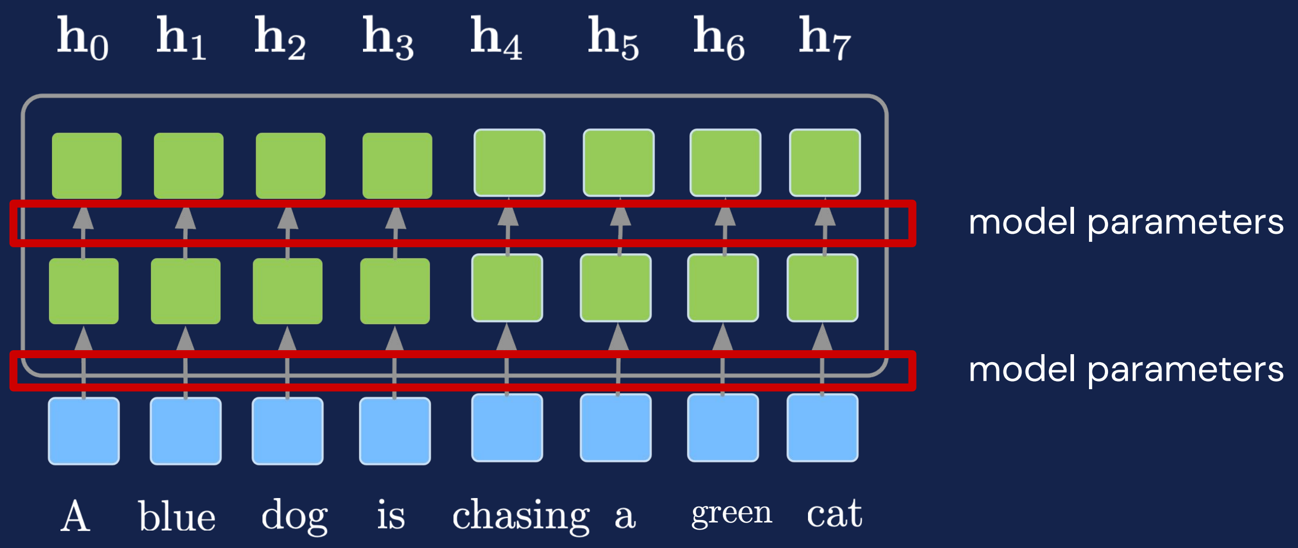
Background

Knowledge is encoded in the weights of a parametric neural network.



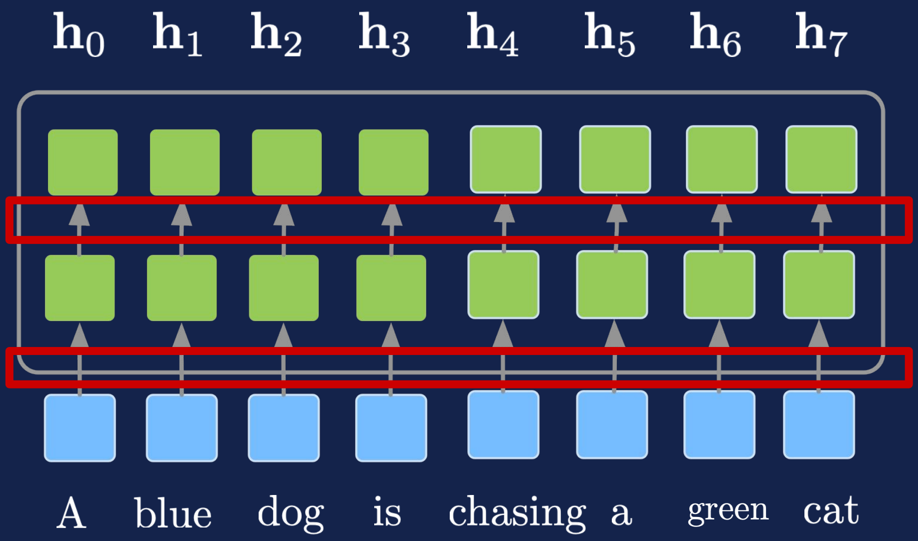
Background

Knowledge is encoded in the weights of a parametric neural network.



Background

Knowledge is encoded in the weights of a parametric neural network.

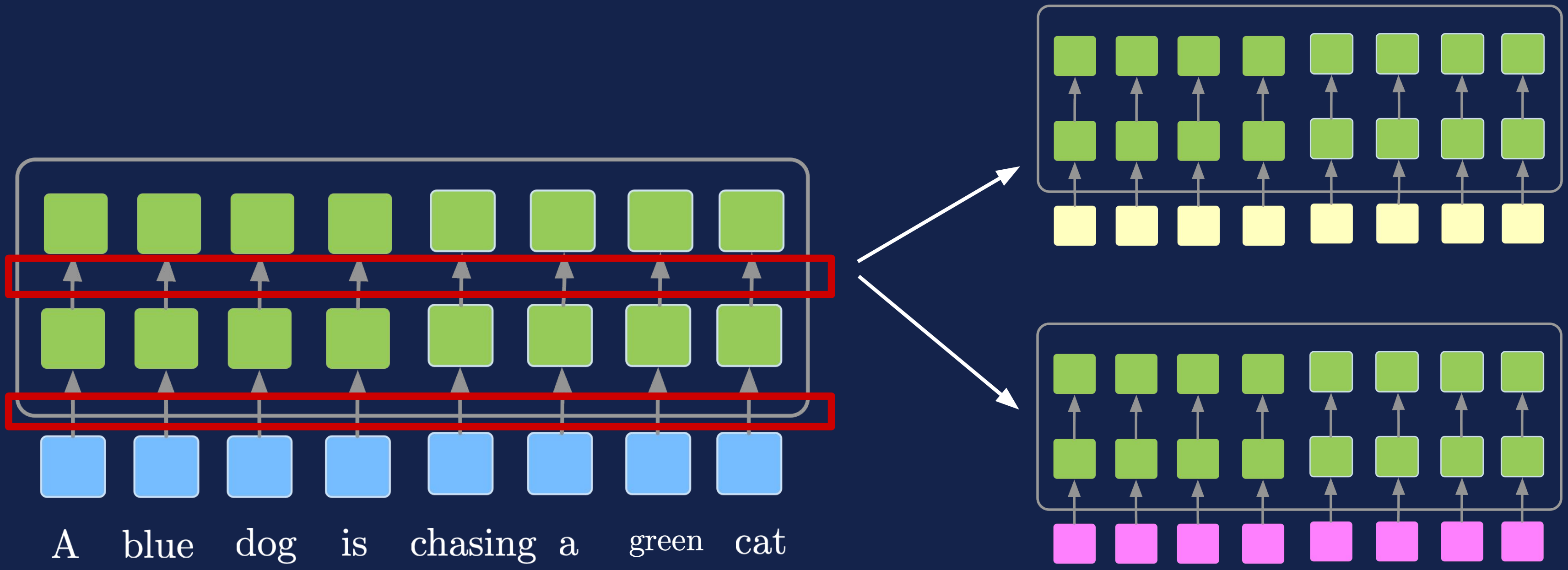


$$f_\alpha(\mathbf{i}_0, \mathbf{i}_1 \ldots, \mathbf{i}_7) \rightarrow \mathbf{h}_0, \mathbf{h}_1, \ldots, \mathbf{h}_7$$

$$f_\beta(x_0, x_1, \ldots, x_7) \rightarrow \mathbf{i}_0, \mathbf{i}_1, \ldots, \mathbf{i}_7$$

Background

Knowledge is encoded in the weights of a parametric neural network.



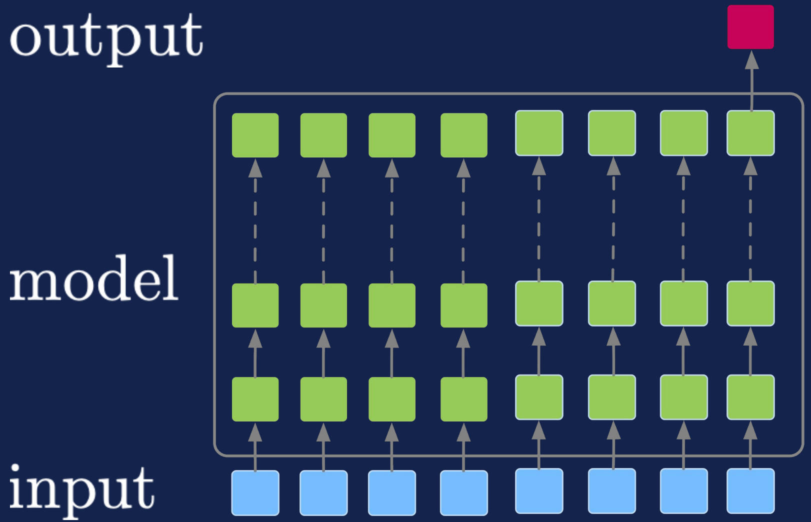
Update weights with new knowledge ⟶ changes affect all examples (sequences).

Semiparametric Models

Separation of computation and storage as an architectural bias.

Semiparametric Models

Separation of computation and storage as an architectural bias.

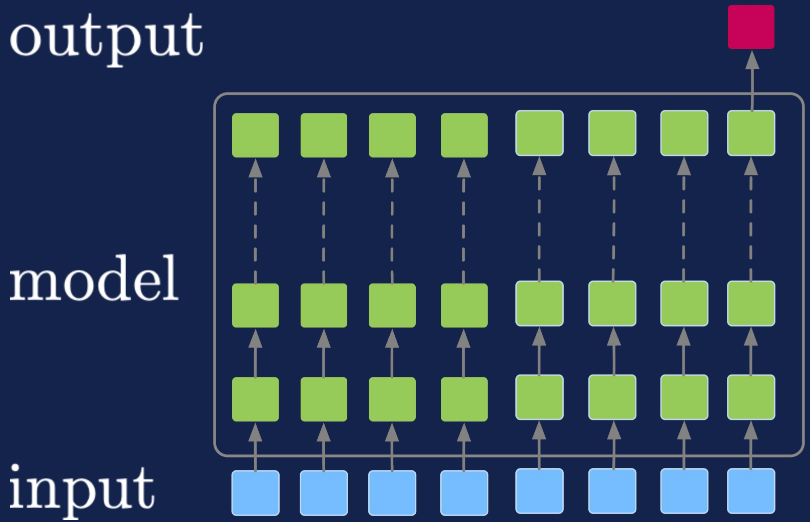


Computation: a neural network

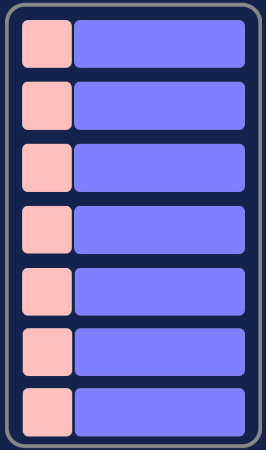
Nematzadeh, Ruder, Yogatama, ICLR Bridging AI and Cognitive Science Workshop 2020

Semiparametric Models

Separation of computation and storage as an architectural bias.



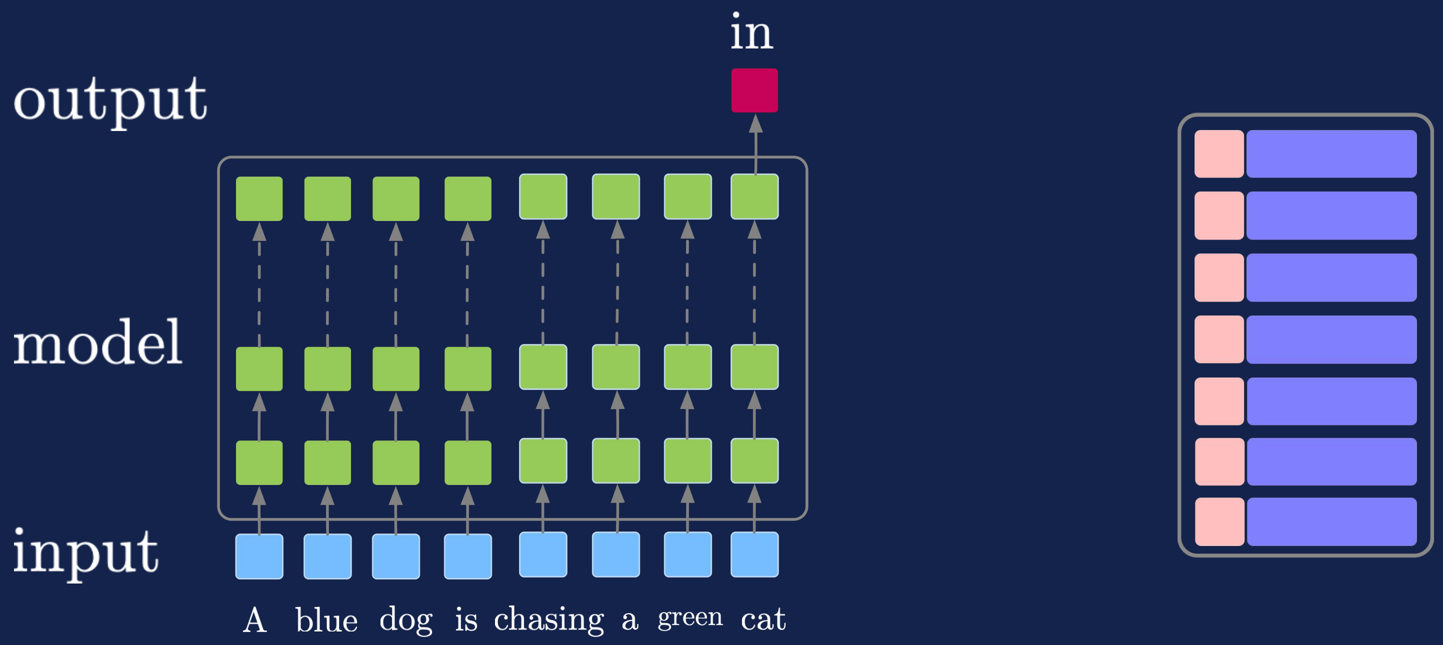
Computation: a neural network



Memory (storage): a key-value database

Nematzadeh, Ruder, Yogatama, ICLR Bridging AI and Cognitive Science Workshop 2020

Semiparametric Language Models



Problem Setup

University of Waterloo Wikipedia

The University of Waterloo (commonly referred to as Waterloo, UW, or UWaterloo) is a public research university with a main campus in Waterloo, Ontario, Canada. The main campus is on 404 hectares of land adjacent to **Uptown**

Problem Setup

University of Waterloo Wikipedia

The University of Waterloo (commonly referred to as Waterloo, UW, or UWaterloo) is a public research university with a main campus in Waterloo, Ontario, Canada. The main campus is on 404 hectares of land adjacent to Uptown **Waterloo**

Problem Setup

University of Waterloo Wikipedia

The University of Waterloo (commonly referred to as Waterloo, UW, or UWaterloo) is a public research university with a main campus in Waterloo, Ontario, Canada. The main campus is on 404 hectares of land adjacent to Uptown Waterloo **and**

Problem Setup

University of Waterloo Wikipedia

The University of Waterloo (commonly referred to as Waterloo, UW, or UWaterloo) is a public research university with a main campus in Waterloo, Ontario, Canada. The main campus is on 404 hectares of land adjacent to Uptown Waterloo and **Waterloo**

Problem Setup

University of Waterloo Wikipedia

The University of Waterloo (commonly referred to as Waterloo, UW, or UWaterloo) is a public research university with a main campus in Waterloo, Ontario, Canada. The main campus is on 404 hectares of land adjacent to Uptown Waterloo and Waterloo **???**

Language Model

University of Waterloo Wikipedia

The University of Waterloo (commonly referred to as Waterloo, UW, or UWaterloo) is a public research university with a main campus in Waterloo, Ontario, Canada. The main campus is on 404 hectares of land adjacent to Uptown Waterloo and Waterloo **???**

Current context

(computation)

Language Model

University of Waterloo Wikipedia

The University of Waterloo (commonly referred to as Waterloo, UW, or UWaterloo) is a public research university with a main campus in Waterloo, Ontario, Canada. The main campus is on 404 hectares of land adjacent to Uptown Waterloo and Waterloo **???**

Current context

(computation)

Extended context

(short-term memory)

Language Model

University of Waterloo Wikipedia

The University of Waterloo (commonly referred to as Waterloo, UW, or UWaterloo) is a public research university with a main campus in Waterloo, Ontario, Canada. The main campus is on 404 hectares of land adjacent to Uptown Waterloo and Waterloo **???**

Current context

(computation)

Extended context

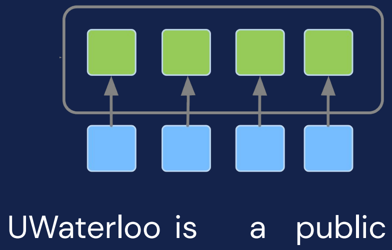
(short-term memory)

Long-term memory

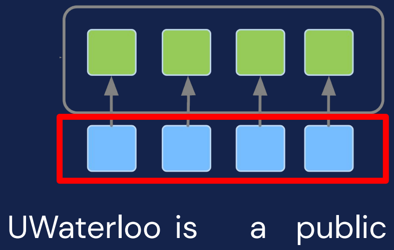
Waterloo Park Wikipedia

Waterloo Park is an urban park situated in Waterloo, Ontario, Canada.

Language Model



Language Model

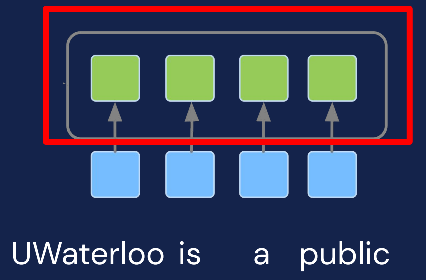


Input: a sequence of tokens.

Language Model

Encoder: transformer

(Vaswani et al., 2017)

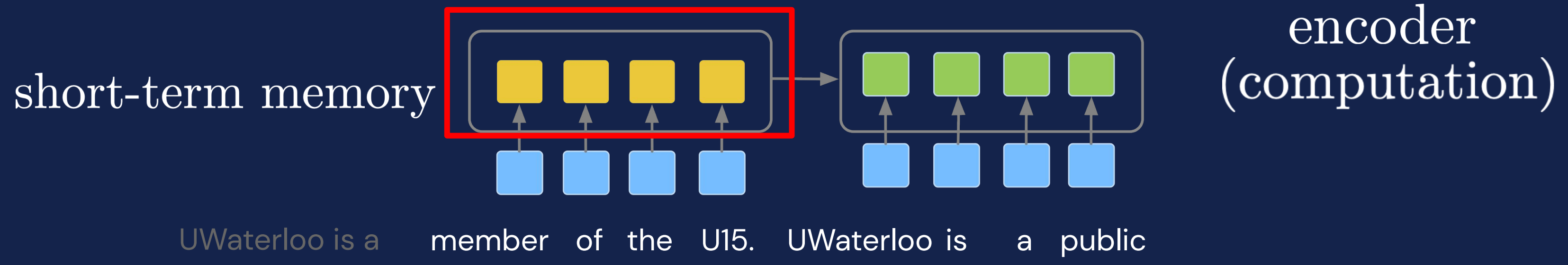


encoder
(computation)

Language Model

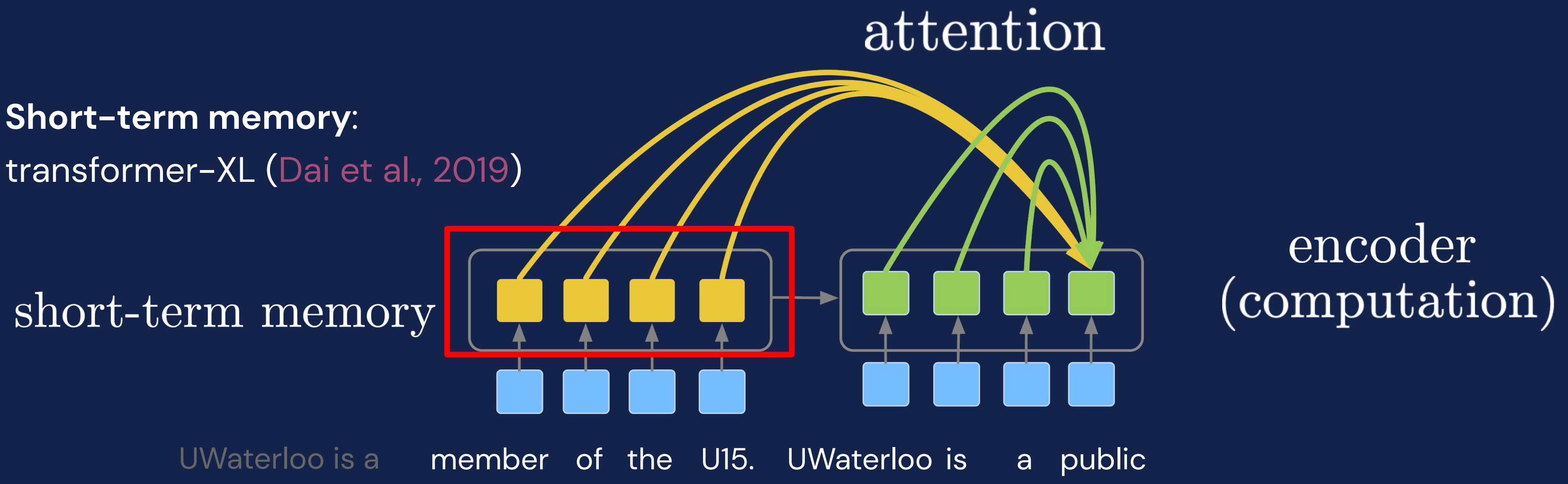
Short-term memory:

transformer-XL (Dai et al., 2019)



Language Model

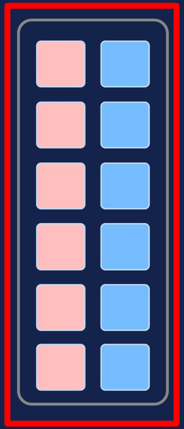
Short-term memory:
transformer-XL (Dai et al., 2019)



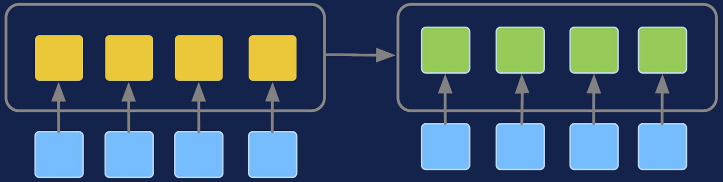
Language Model

Long-term memory:
key-value database

long-term memory



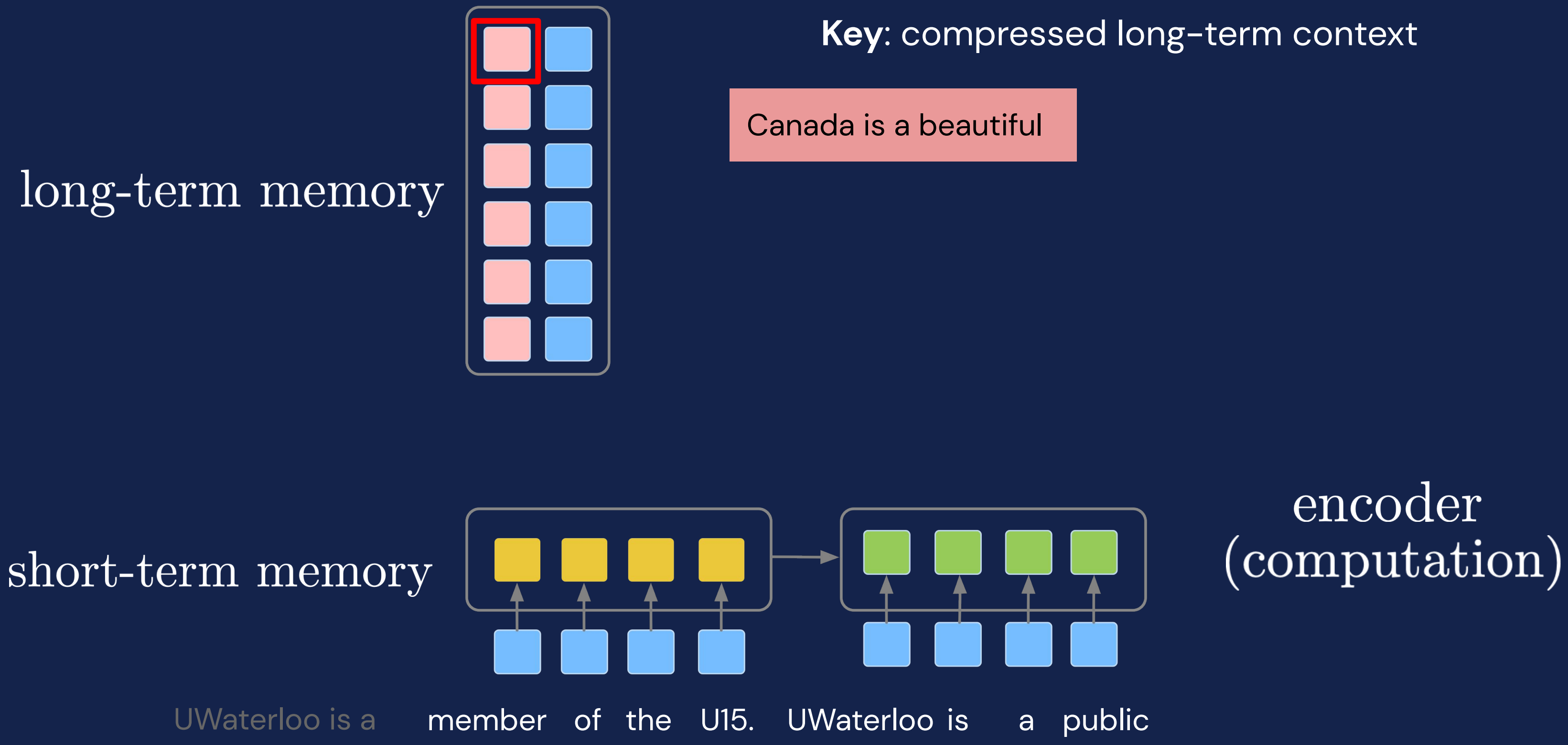
short-term memory



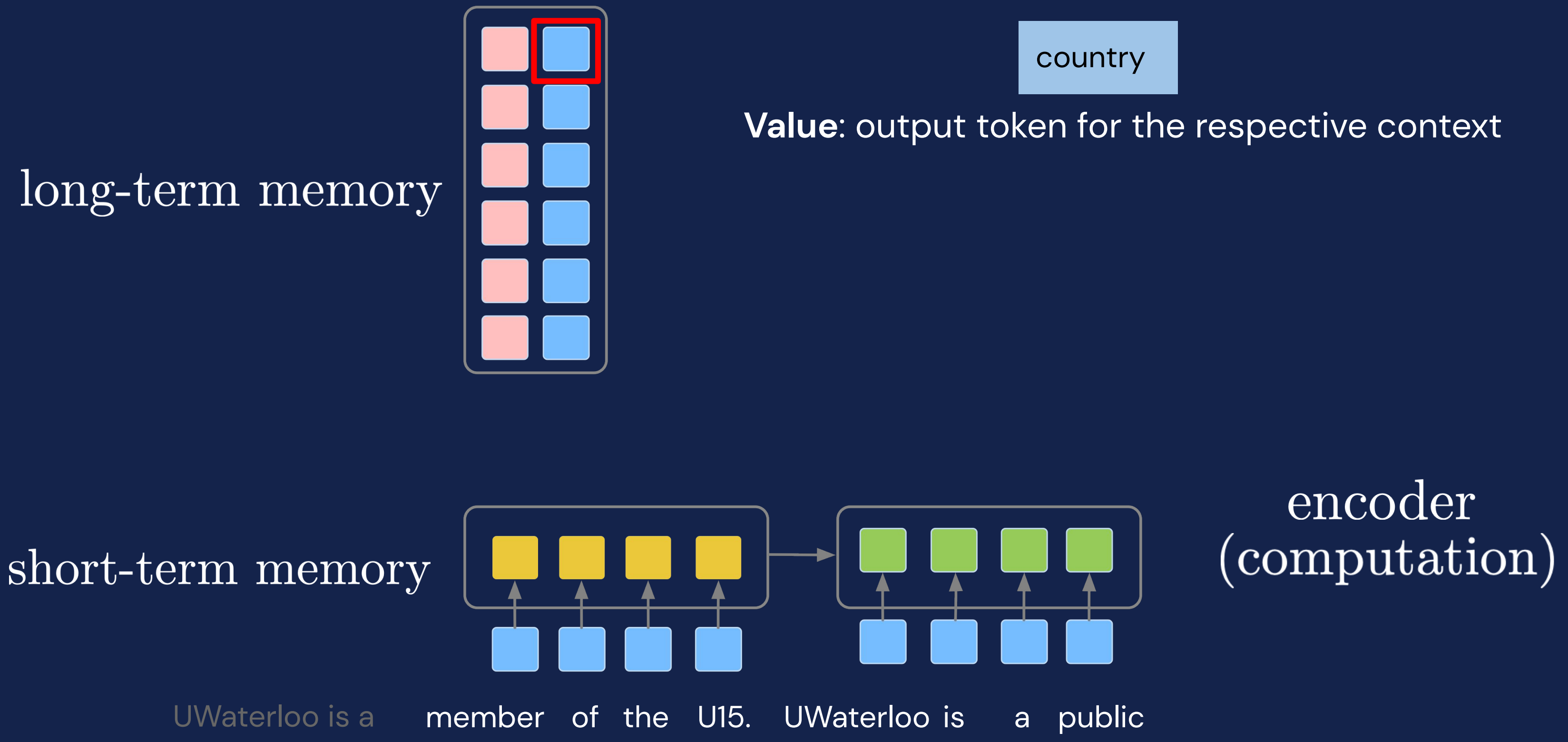
encoder
(computation)

UWaterloo is a member of the U15. UWaterloo is a public

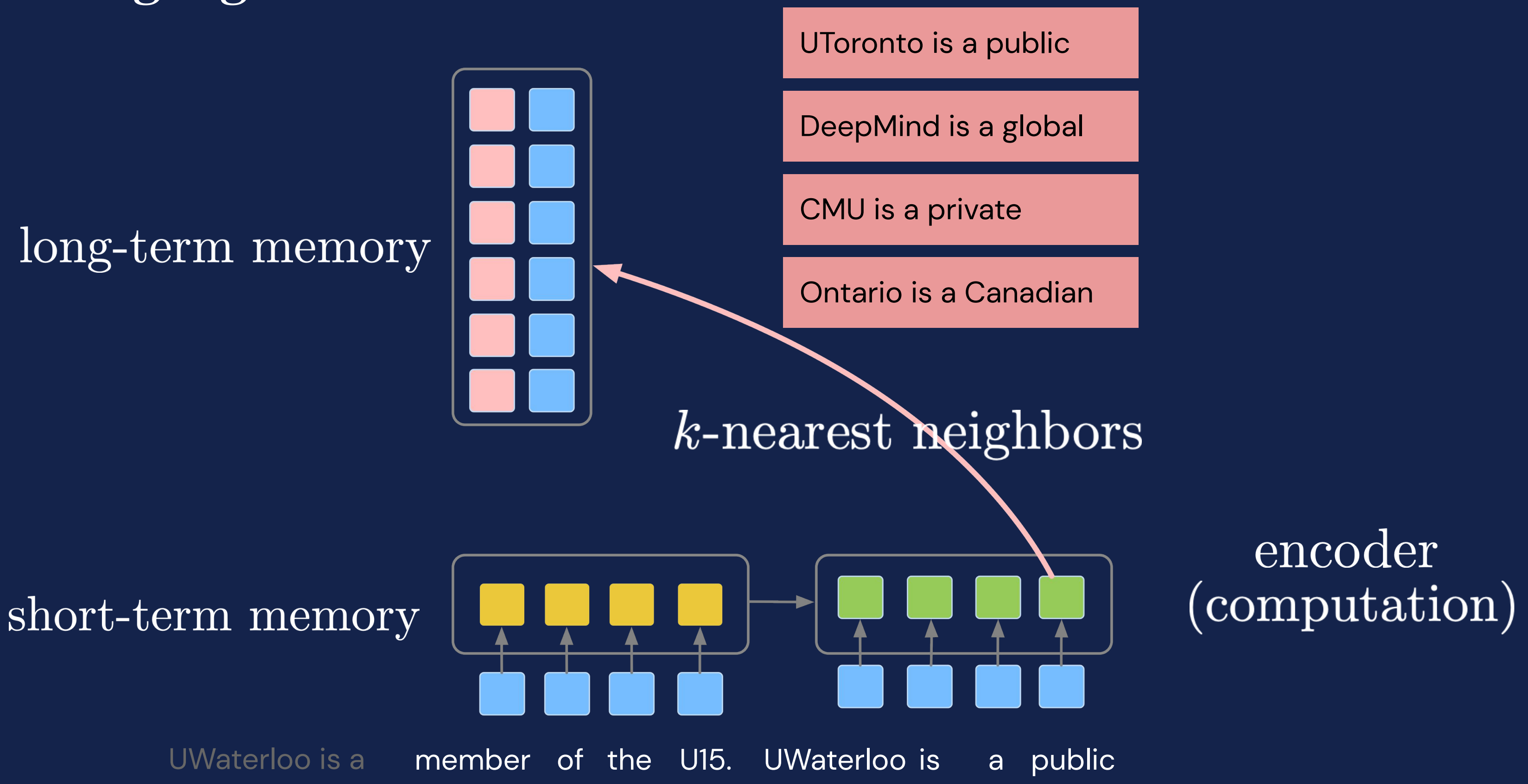
Language Model



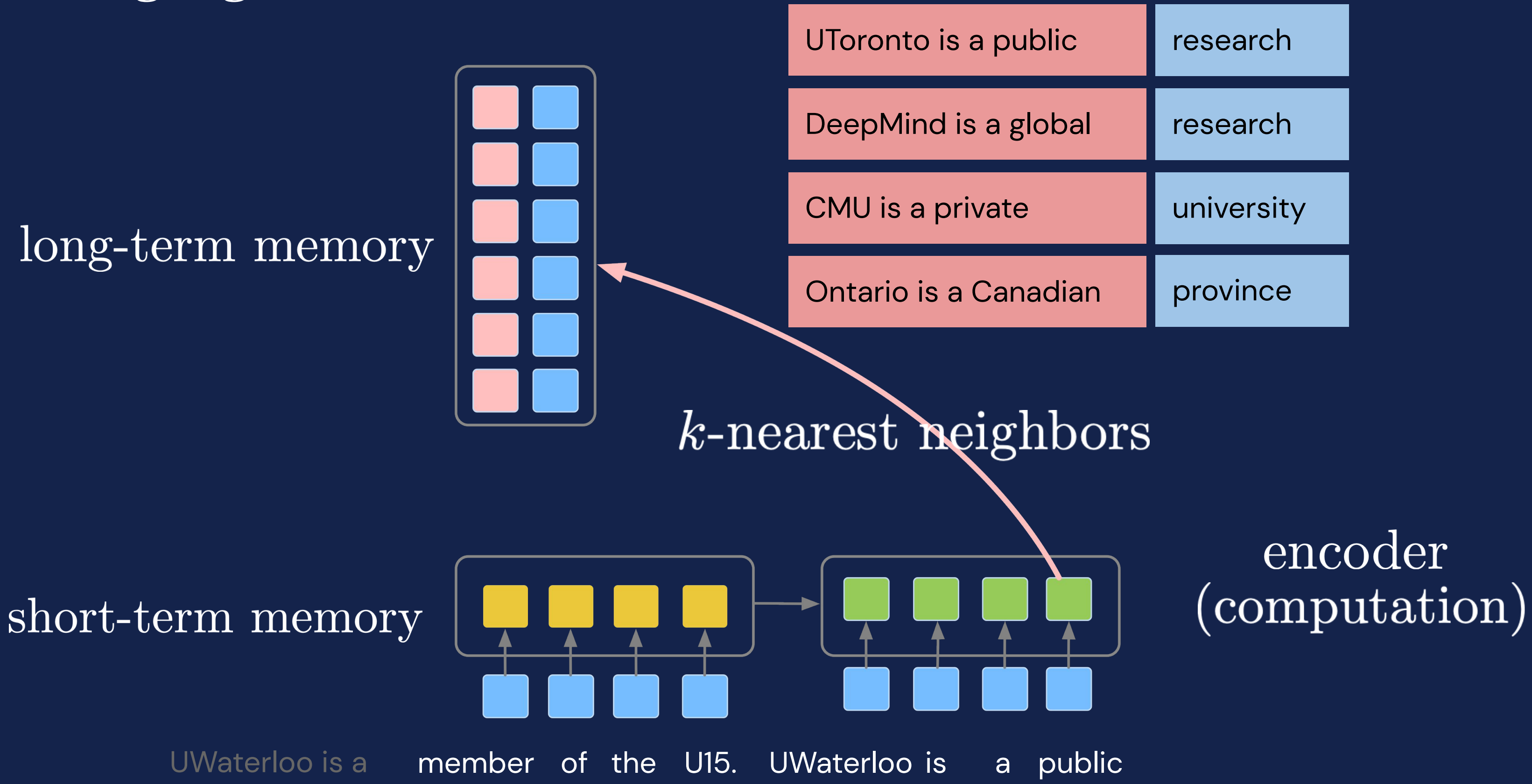
Language Model



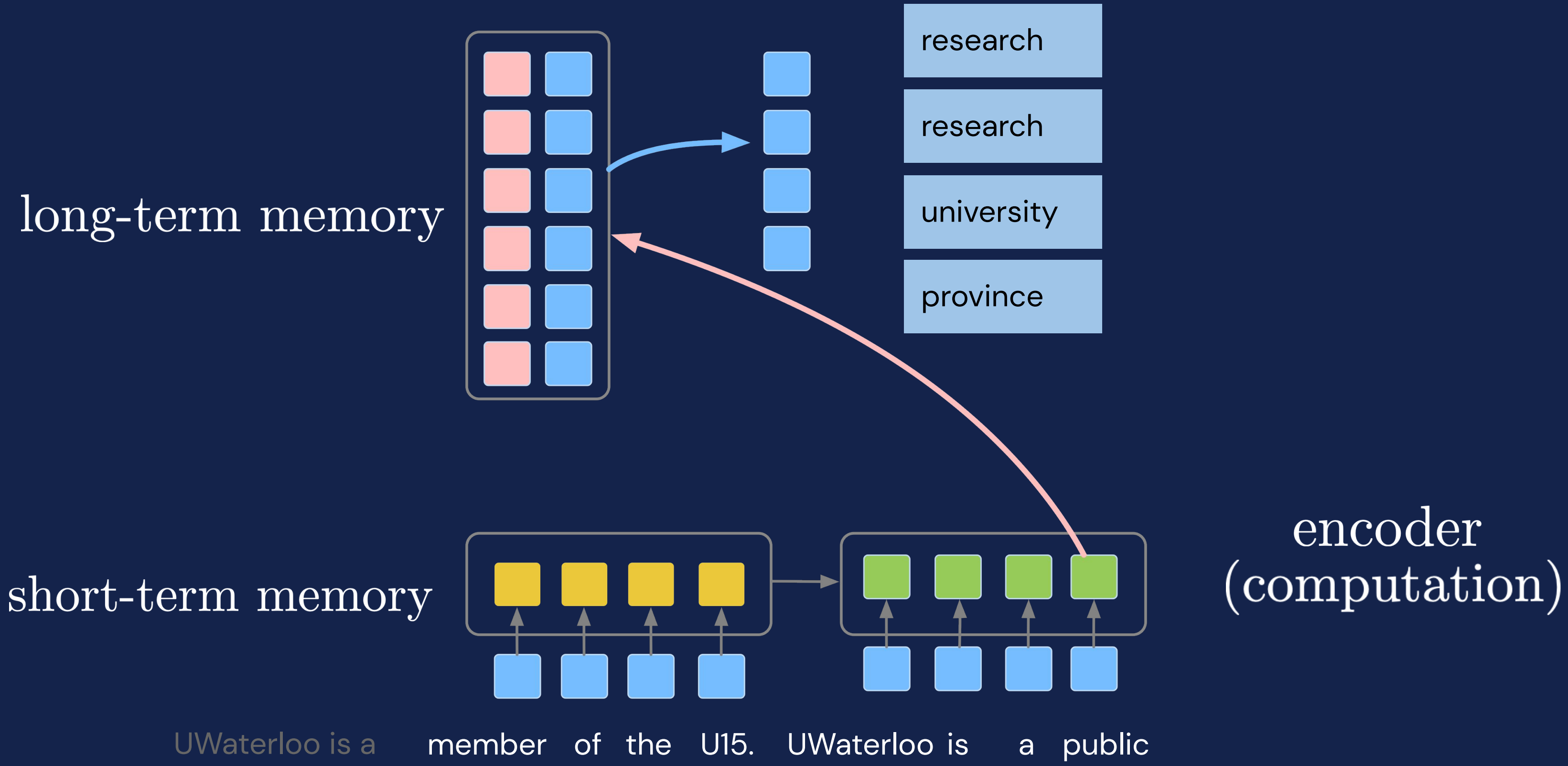
Language Model



Language Model



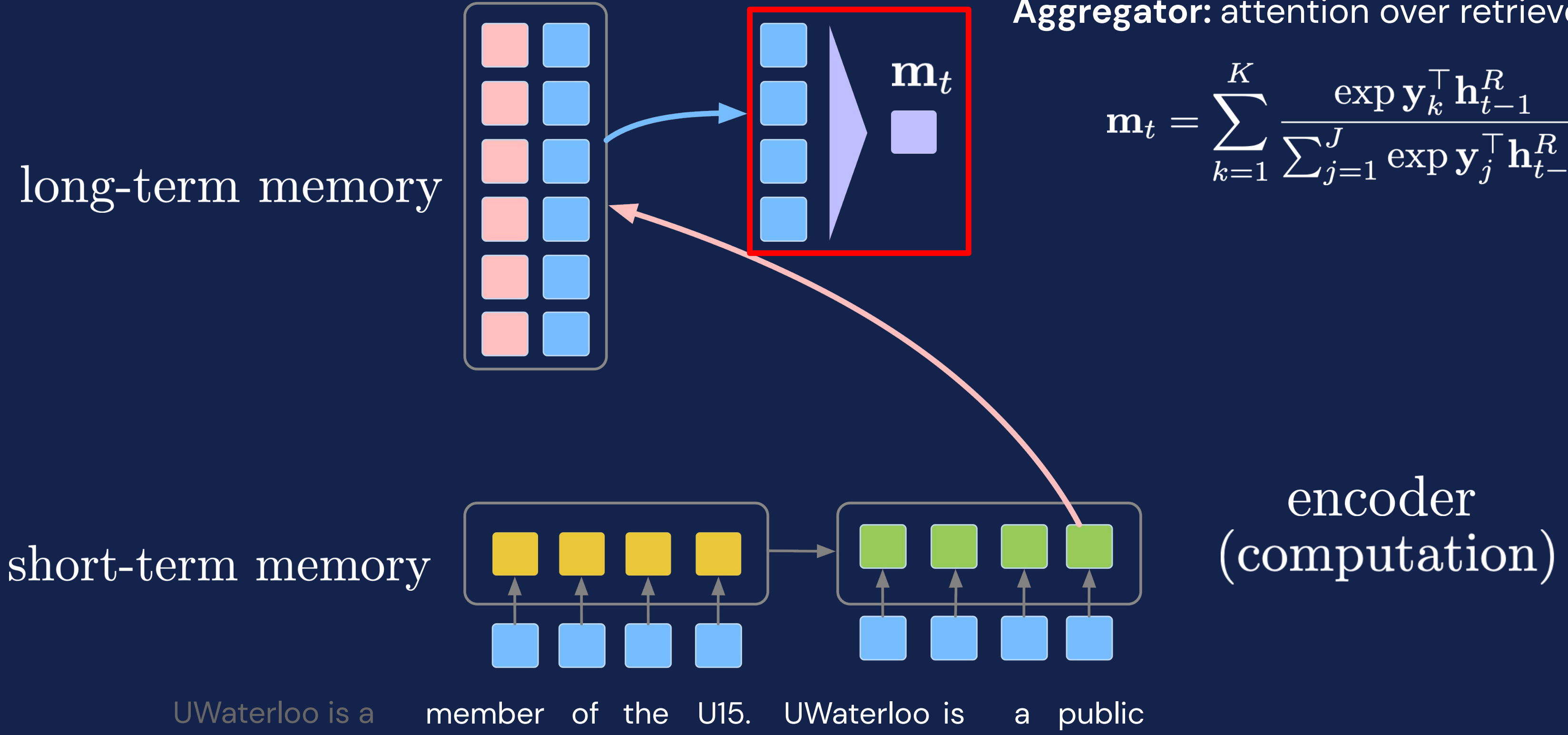
Language Model



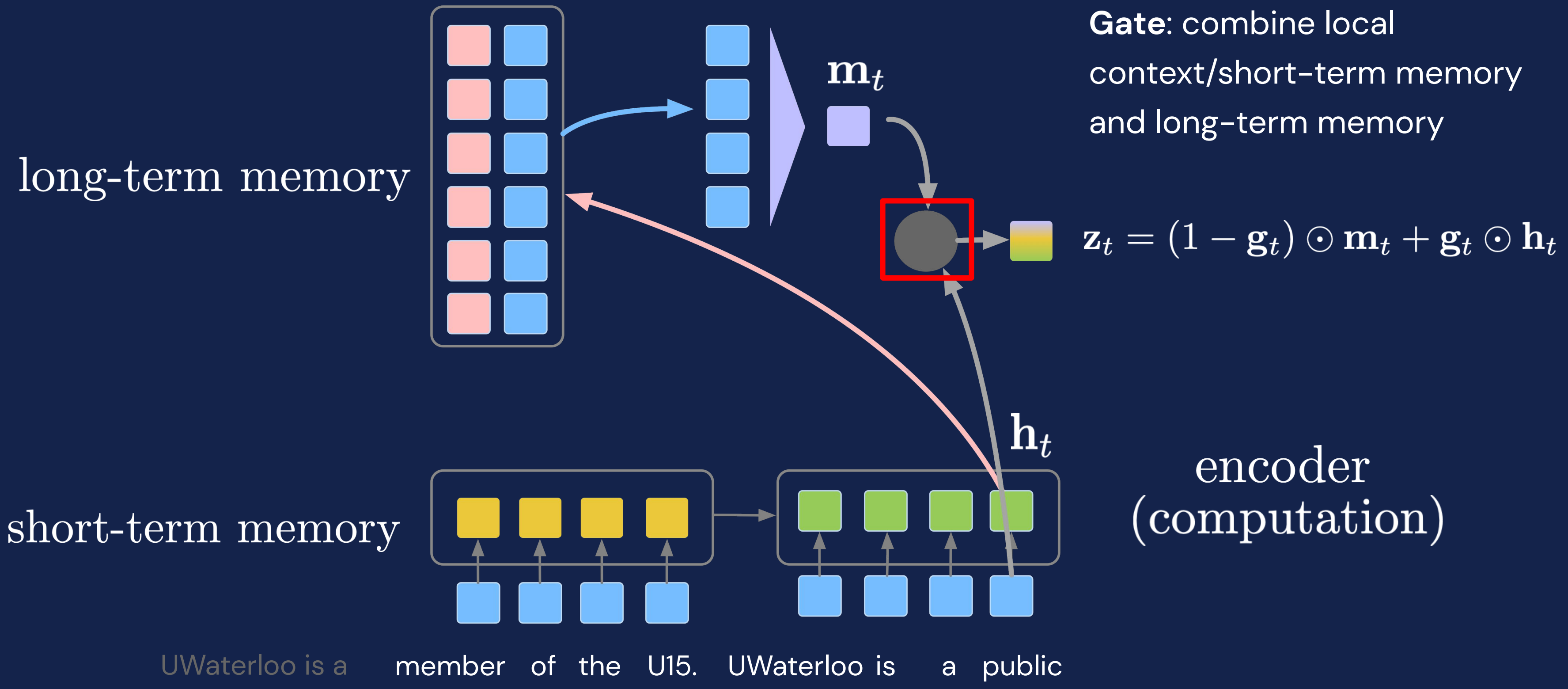
Language Model

Aggregator: attention over retrieved values

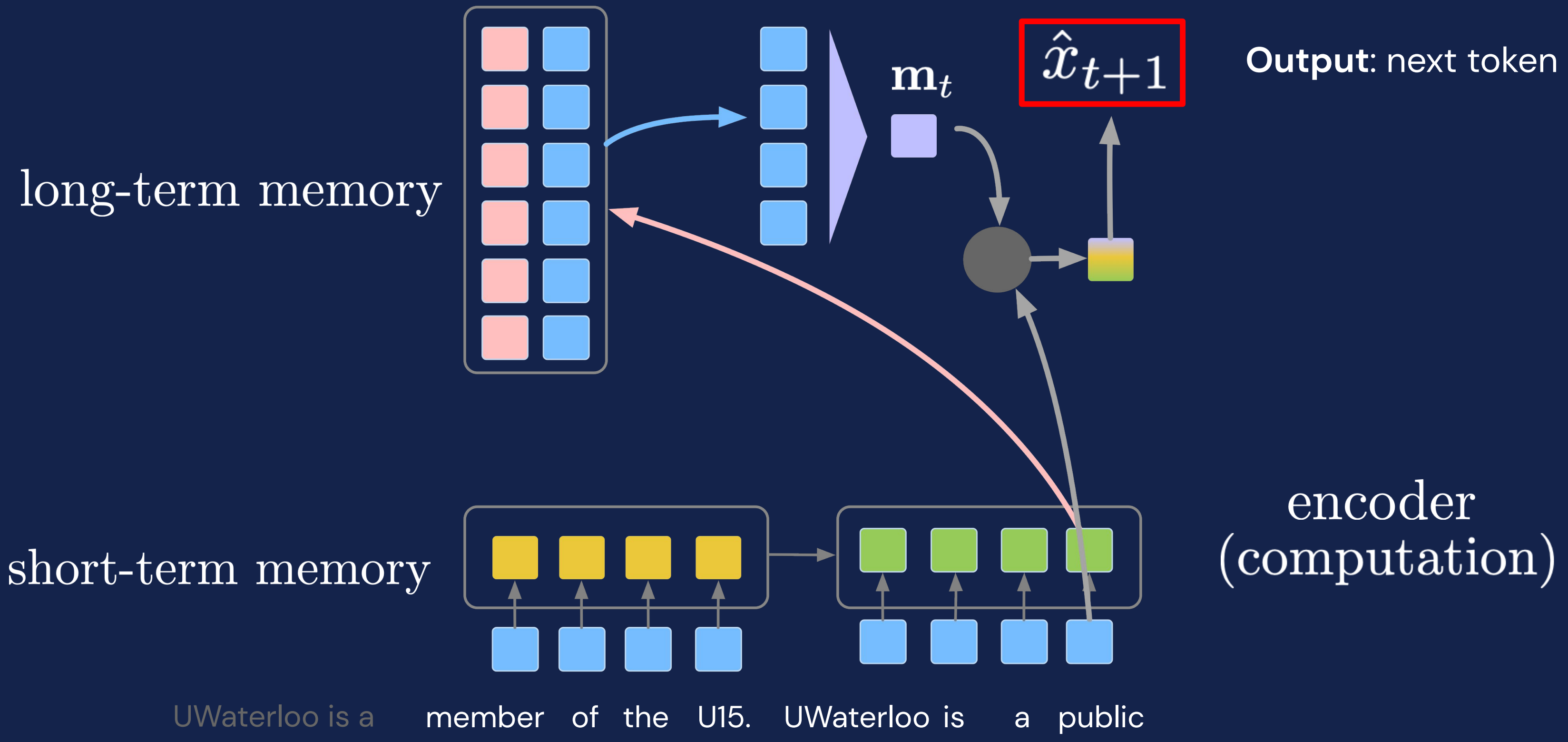
$$\mathbf{m}_t = \sum_{k=1}^{K} \frac{\exp \mathbf{y}_k^\top \mathbf{h}_{t-1}^R}{\sum_{j=1}^{J} \exp \mathbf{y}_j^\top \mathbf{h}_{t-1}^R} \mathbf{y}_k$$



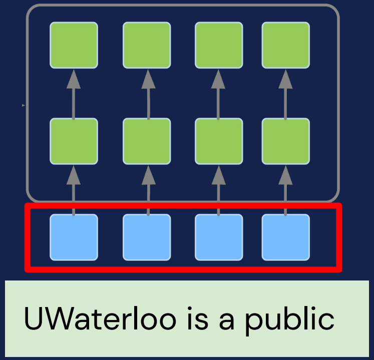
Language Model



Language Model



Language Model

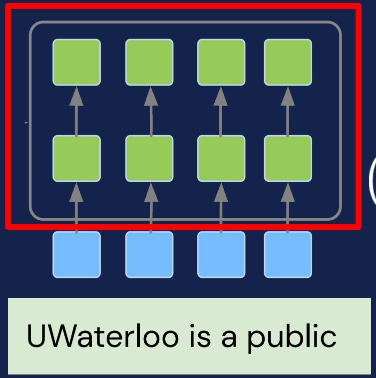


Input: a sequence of tokens.

Language Model

Encoder: transformer

(Vaswani et al., 2017)



encoder
(computation)

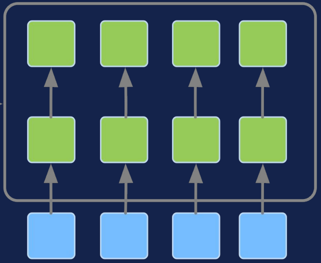
Language Model

Short-term memory:

transformer-XL (Dai et al., 2019)

Encoder: transformer

(Vaswani et al., 2017)



encoder
(computation)

UWaterloo is a

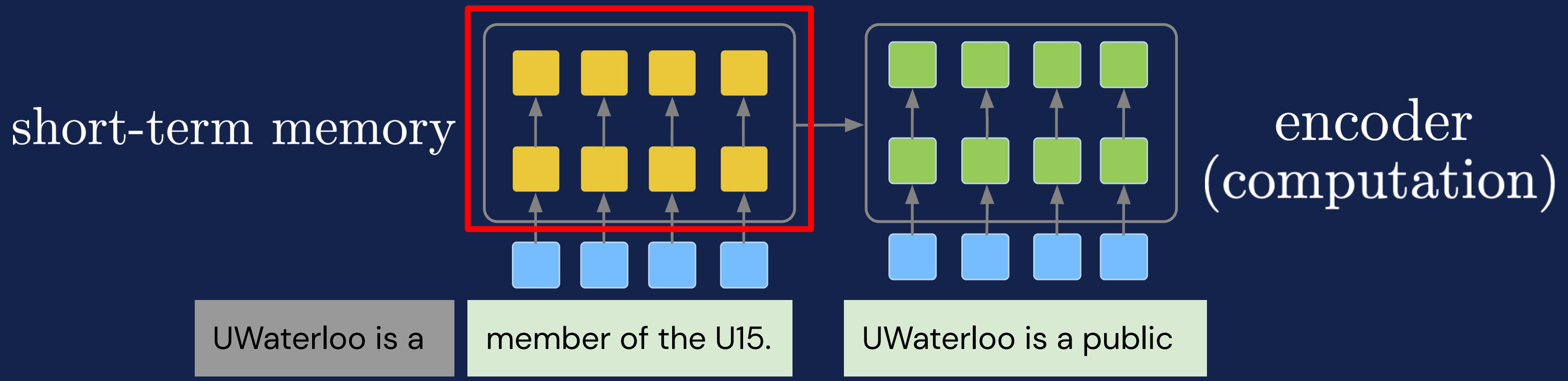
member of the U15.

UWaterloo is a public

Language Model

Short-term memory:

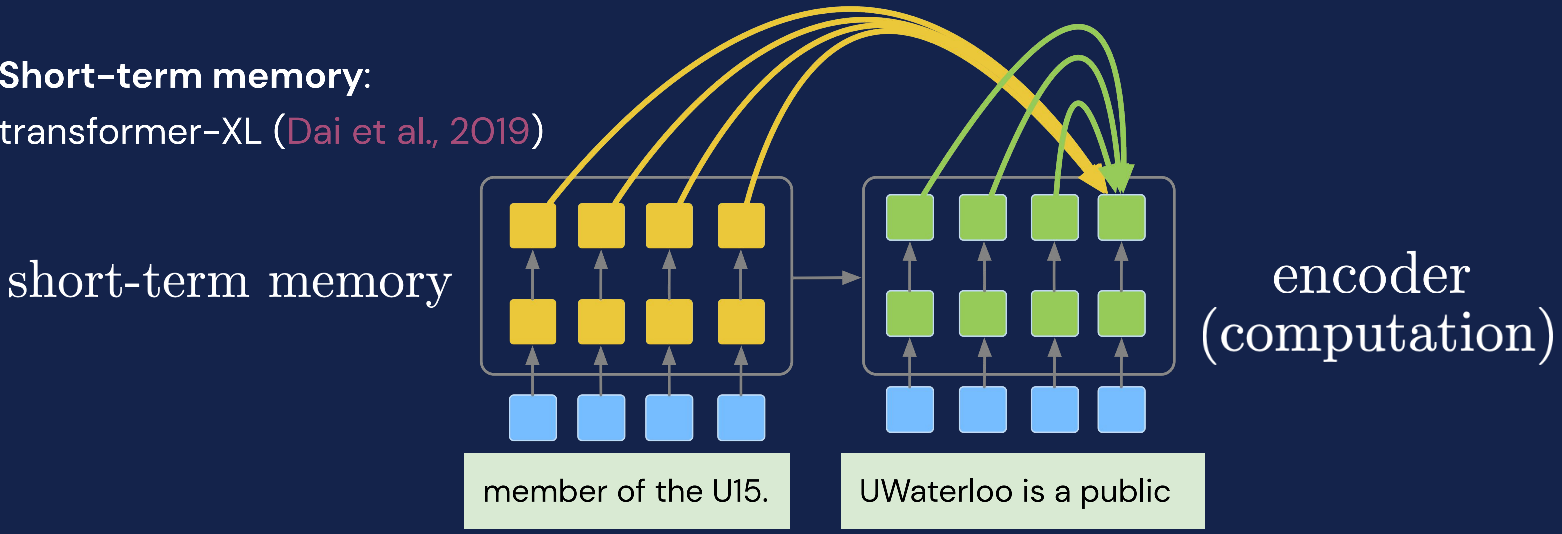
transformer-XL (Dai et al., 2019)



Language Model

Short-term memory:

transformer-XL (Dai et al., 2019)

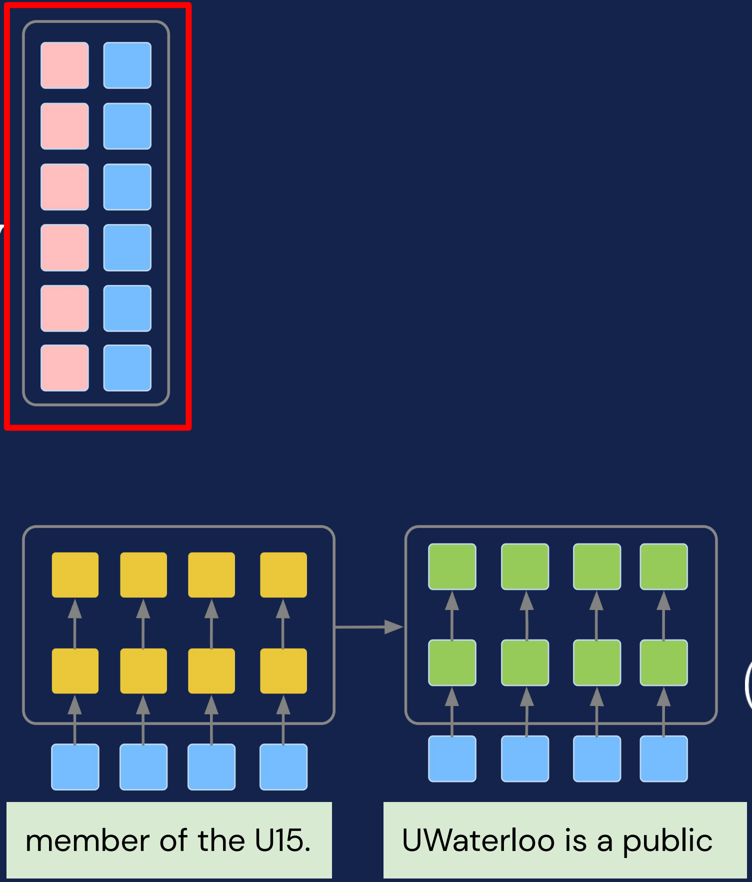


Language Model

Long-term memory:
key-value database

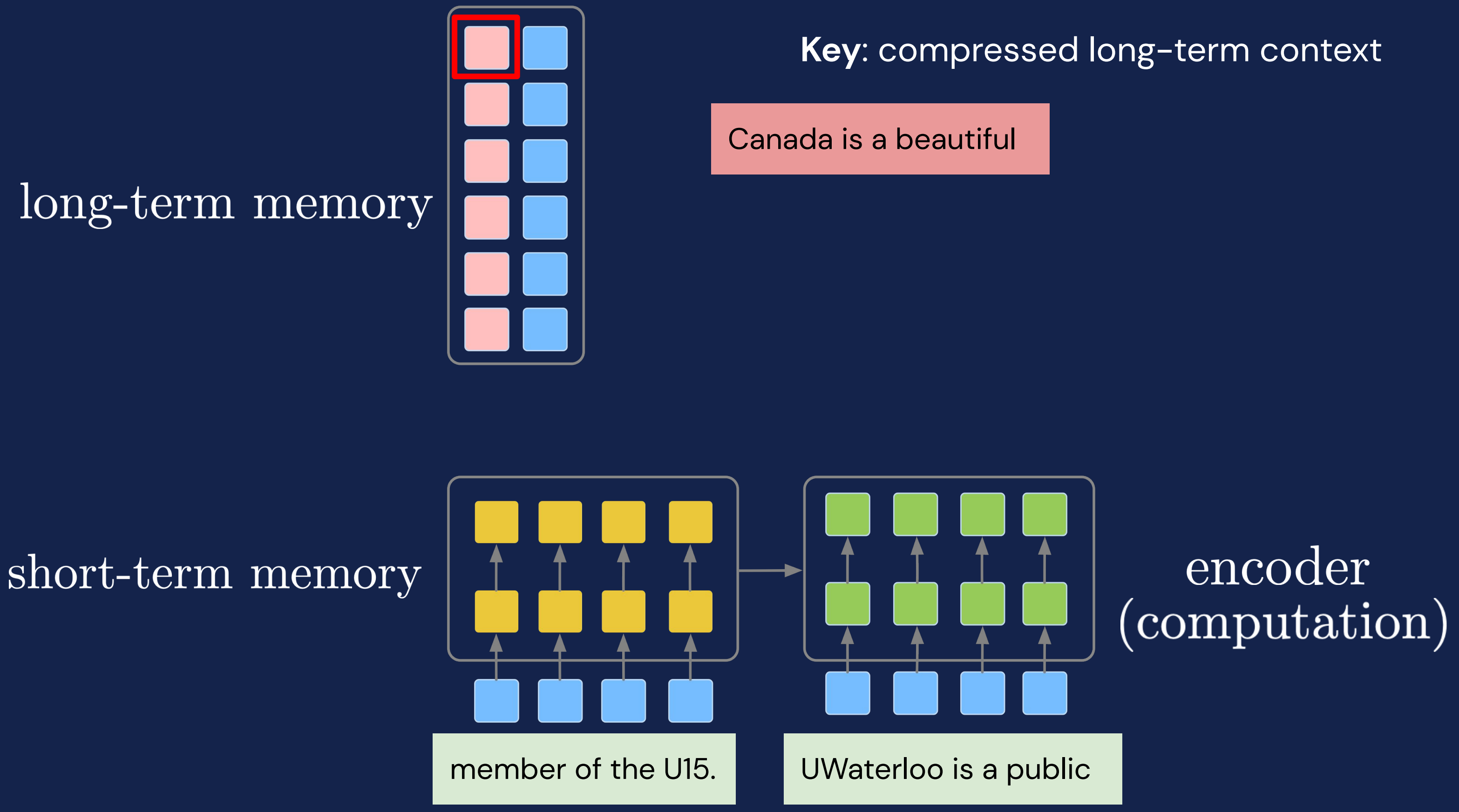
long-term memory

short-term memory

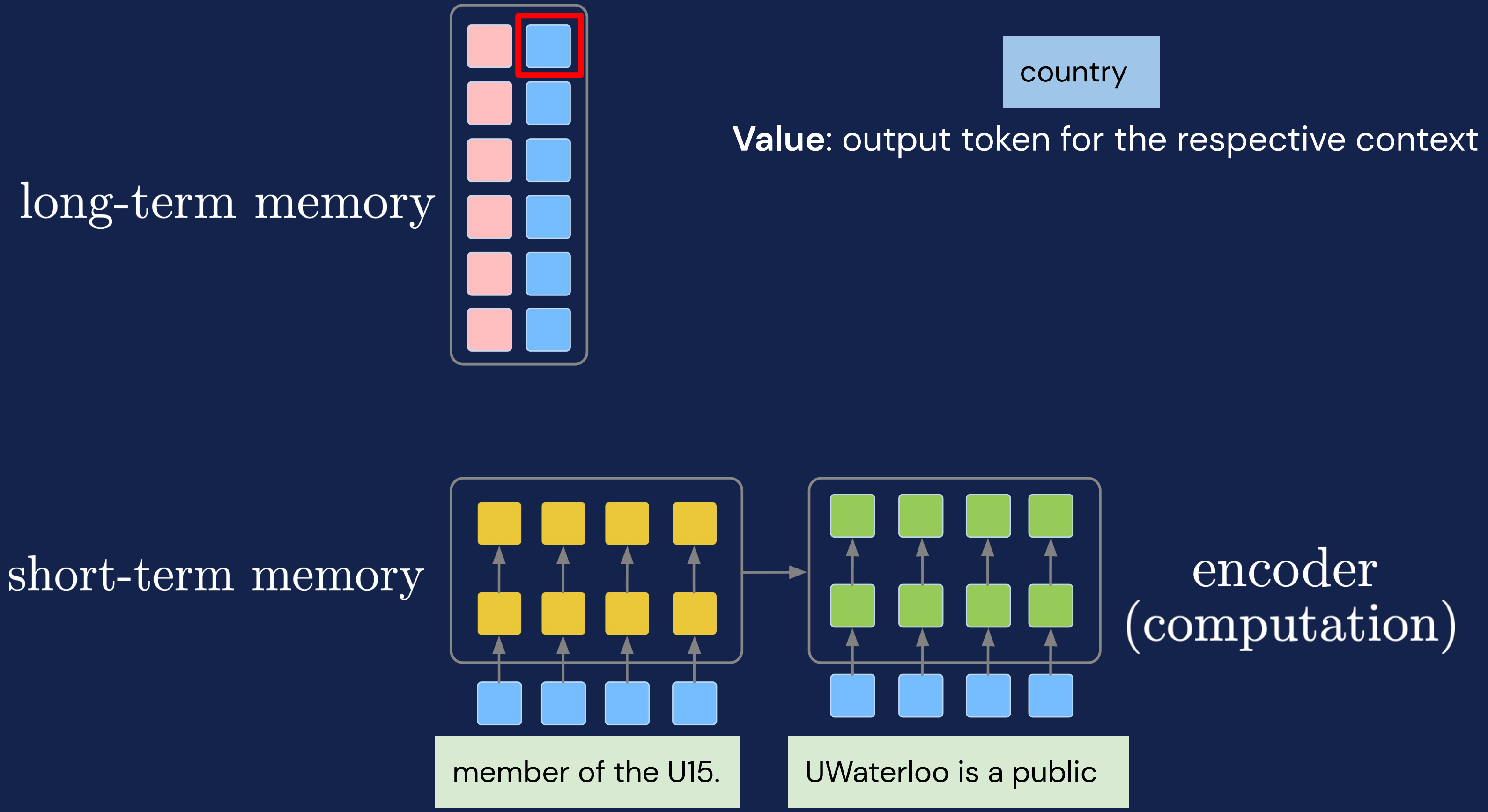


encoder
(computation)

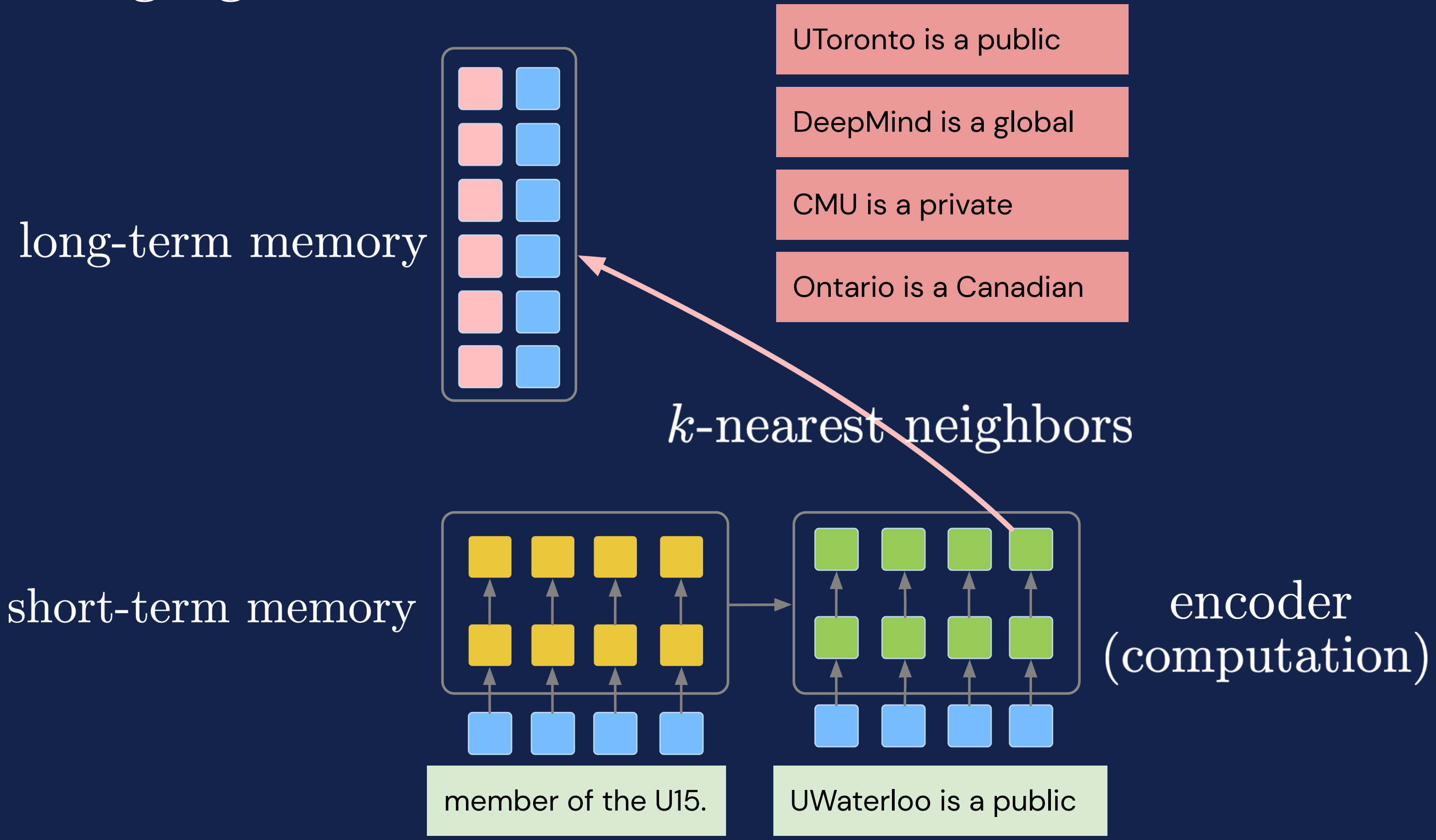
Language Model



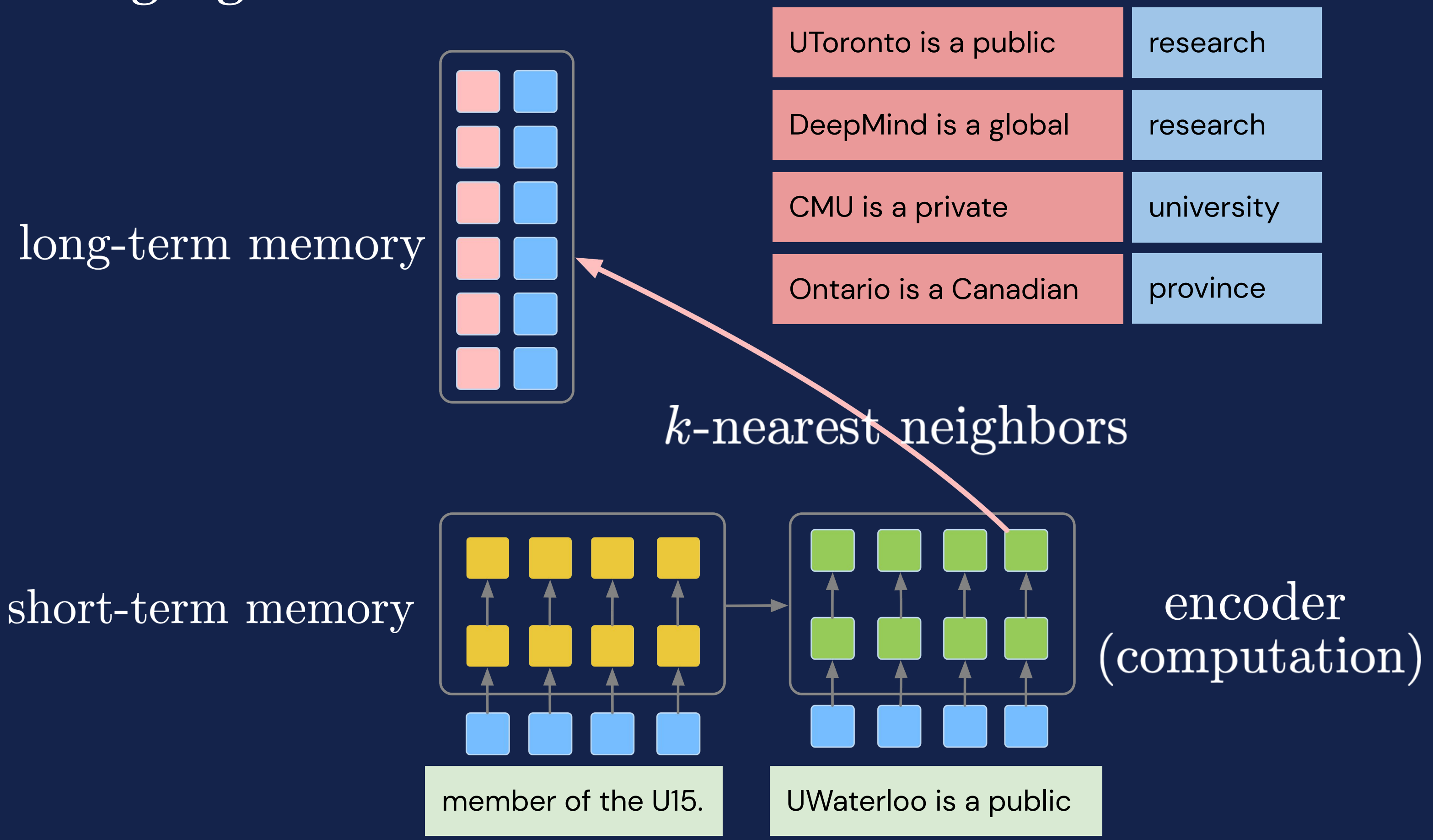
Language Model



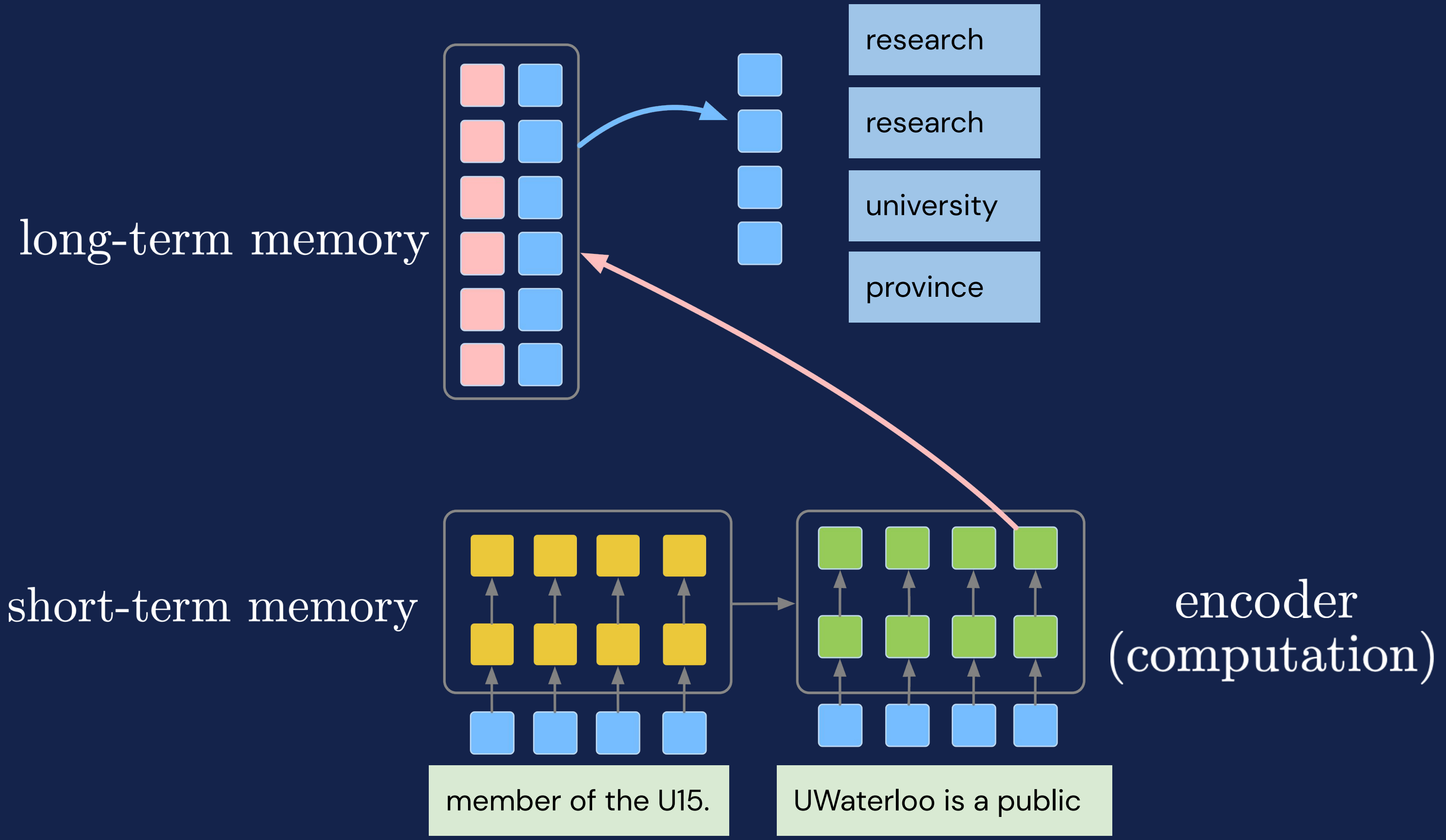
Language Model



Language Model



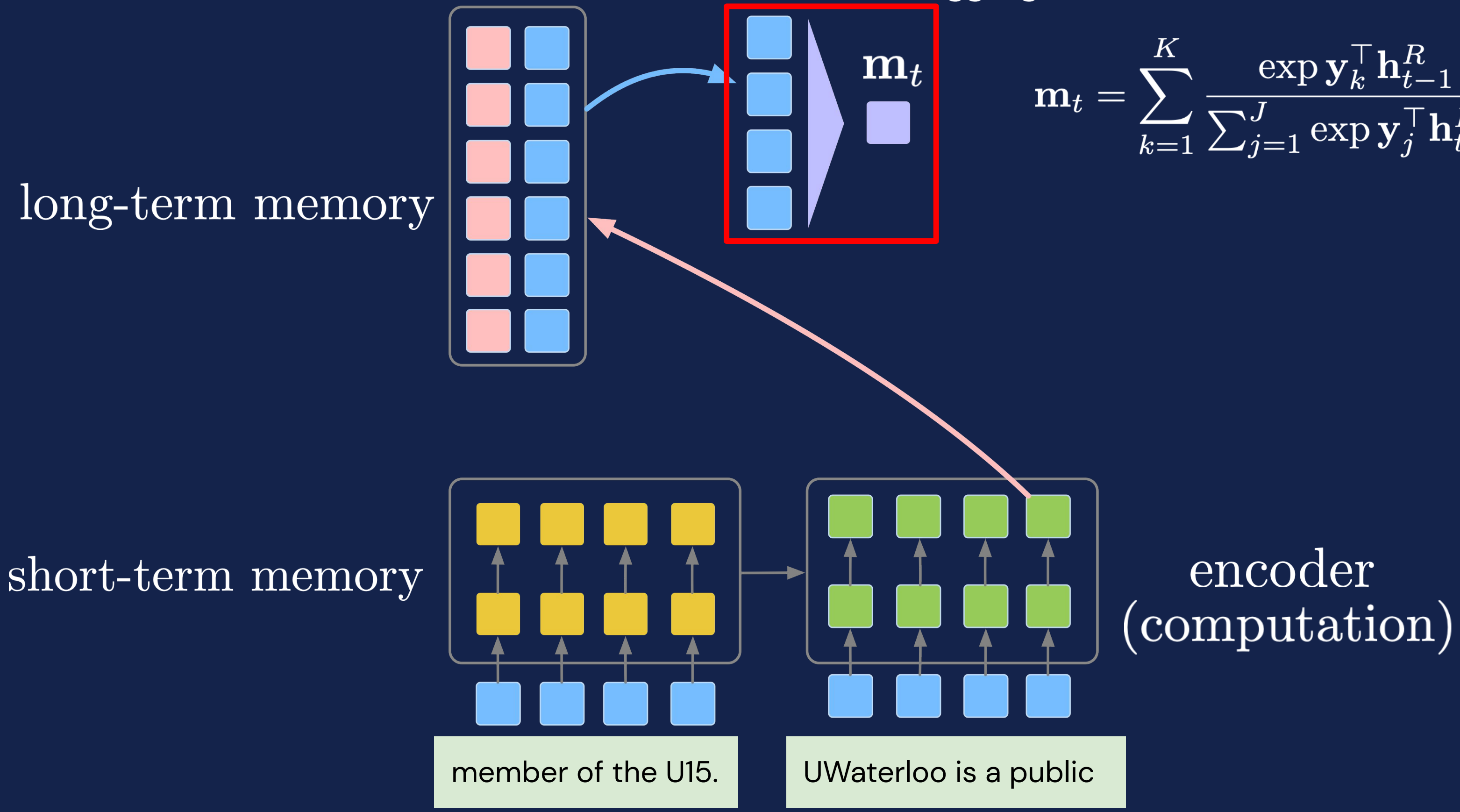
Language Model



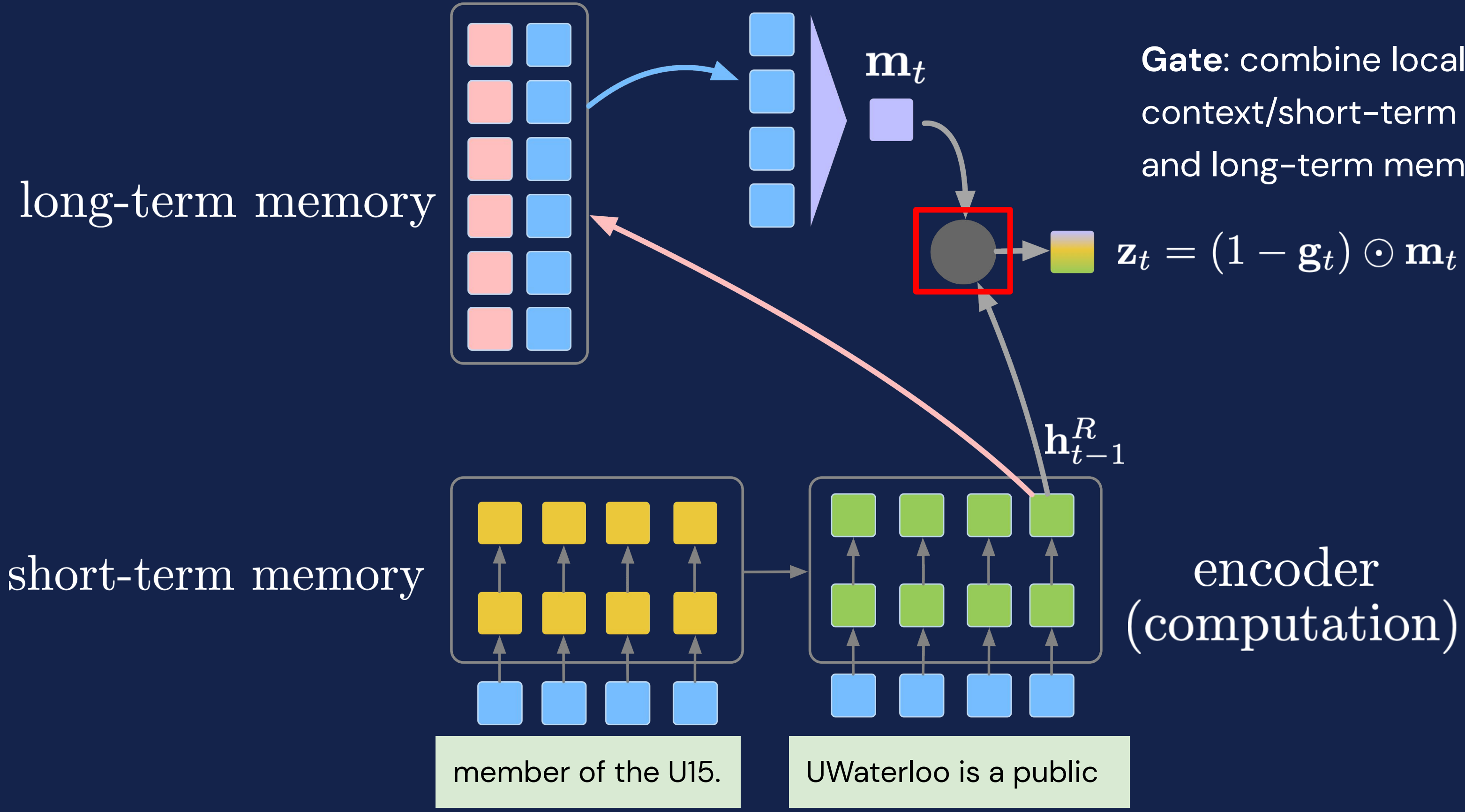
Language Model

Aggregator: attention over retrieved values

$$\mathbf{m}_t = \sum_{k=1}^{K} \frac{\exp \mathbf{y}_k^\top \mathbf{h}_{t-1}^R}{\sum_{j=1}^{J} \exp \mathbf{y}_j^\top \mathbf{h}_{t-1}^R} \mathbf{y}_k$$



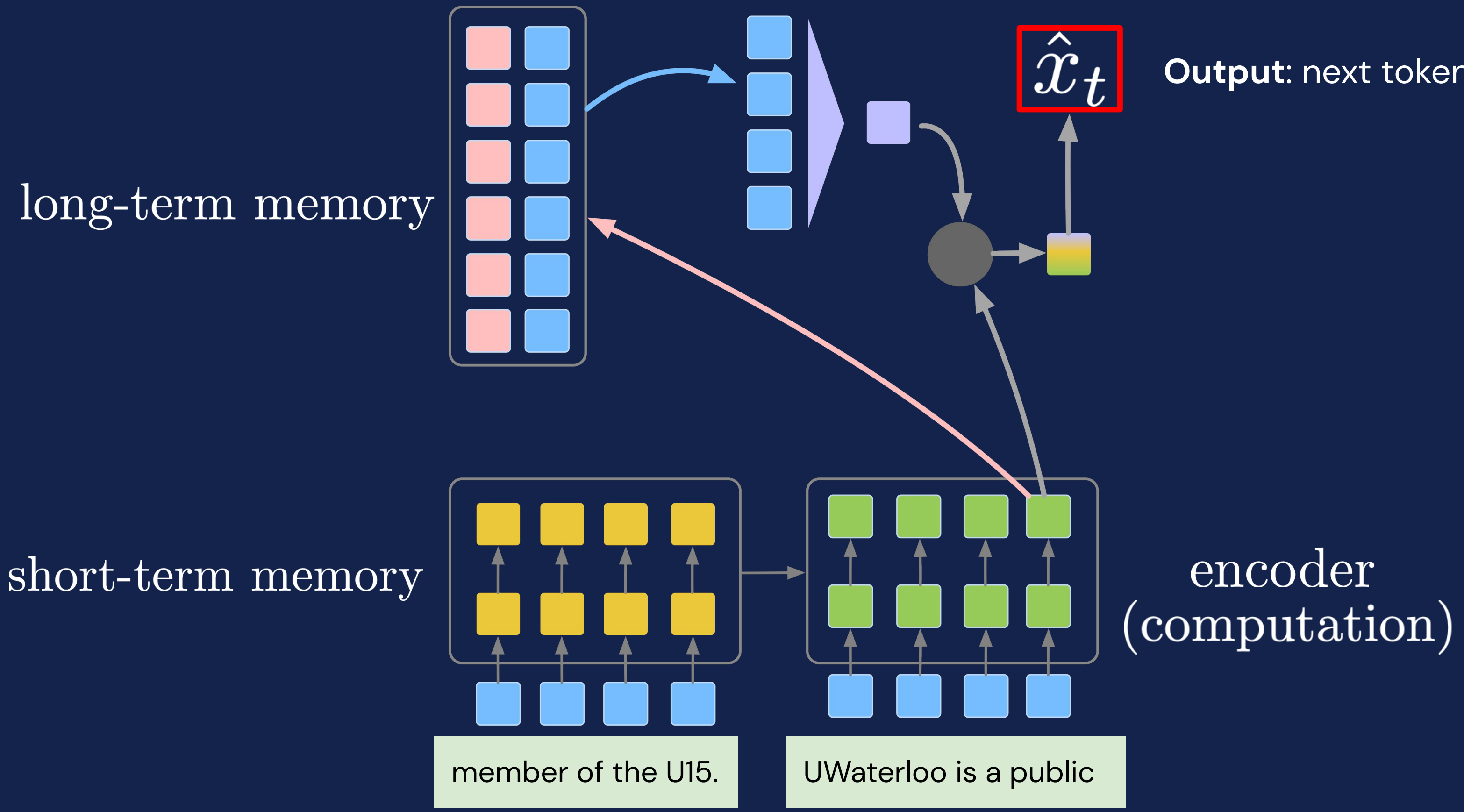
Language Model



Gate: combine local context/short-term memory and long-term memory

$$\mathbf{z}_t = (1 - \mathbf{g}_t) \odot \mathbf{m}_t + \mathbf{g}_t \odot \mathbf{h}^R_{t-1}$$

Language Model



Experiments

- Word-level language modeling.
    - WikiText-103 English (Merity et al., 2016).
    - WMT 2019 English: http://www.statmt.org/wmt19/.
- Character-level language modeling.
    - enwik8: http://prize.hutter1.net.

Experiments

Perplexity (1-inf), lower is better

	Base	TXL	kNN-LM	Ours
WikiText-103	21.8	19.1	18.0	**17.6***
WMT	16.5	15.5	15.2	**14.1**

Transformer: Vaswani et al., 2017

Transformer-XL: Dai et al., 2019

kNN-LM: Khandelwal et al., 2020

Experiments

BPC (0-inf), lower is better

	Base	TXL	kNN-LM	Ours
enwik8	1.05	1.01	1.02	**1.00**

Transformer: Vaswani et al., 2017

Transformer-XL: Dai et al., 2019

kNN-LM: Khandelwal et al., 2020

Analysis

Liberal Democrat leader Jo Swinson has said
she would work with Donald Trump in government as

Analysis

What's in the long-term memory?

Elizabeth	Warren	on	Friday	proposed	$	20	trillion	in
spending	over	the	next	decade	to	provide	health	care
every	American	without	raising	taxes	on	the	middle	class

Analysis

What's in the long-term memory?

For

Perhaps
Like

Elizabeth	Warren	on	Friday	proposed	$	20	trillion	in
spending	over	the	next	decade	to	provide	health	care
every	American	without	raising	taxes	on	the	middle	class

Analysis

What's in the long-term memory?

For	Warren							
	Warren							
Perhaps	Warren							
Like	Warren							
Elizabeth	Warren	on	Friday	proposed	$	20	trillion	in
spending	over	the	next	decade	to	provide	health	care
every	American	without	raising	taxes	on	the	middle	class

Analysis

What's in the long-term memory?

For	Warren	&						
	Warren	may						
Perhaps	Warren	has						
Like	Warren	,						
Elizabeth	Warren	on	Friday	proposed	$	20	trillion	in
spending	over	the	next	decade	to	provide	health	care
every	American	without	raising	taxes	on	the	middle	class

Analysis

What's in the long-term memory?

For	Warren	&	Wednesday					
	Warren	may	Tuesday					
Perhaps	Warren	has	Sunday					
Like	Warren	,	Monday					
Elizabeth	Warren	on	Friday	proposed	$	20	trillion	in
spending	over	the	next	decade	to	provide	health	care
every	American	without	raising	taxes	on	the	middle	class

Analysis

What's in the long-term memory?

For	Warren	&	Wednesday	briefly	a	5	billion	to
	Warren	may	Tuesday	praised	wiping	16	trillion	in
Perhaps	Warren	has	Sunday	stood	breaking	10	billion	for
Like	Warren	,	Monday	defended	using	166	trillion	in
Elizabeth	Warren	on	Friday	proposed	$	20	trillion	in
spending	over	the	next	decade	to	provide	health	care
every	American	without	raising	taxes	on	the	middle	class

Analysis

What's in the long-term memory?

For	Warren	&	Wednesday	briefly	a	5	billion	to
	Warren	may	Tuesday	praised	wiping	16	trillion	in
Perhaps	Warren	has	Sunday	stood	breaking	10	billion	for
Like	Warren	,	Monday	defended	using	166	trillion	in
Elizabeth	Warren	on	Friday	proposed	$	20	trillion	in

grants	in	10	course	eight	.	fight	even	care
funding	over	the	next	three	.	upgrade	them	cover
funds	over	10	next	five	in	improve	American	-
,	over	a	next	10	,	invest	a	insur
spending	over	the	next	decade	to	provide	health	care

every	American	without	raising	taxes	on	the	middle	class

Analysis

What's in the long-term memory?

For	Warren	&	Wednesday	briefly	a	5	billion	to
	Warren	may	Tuesday	praised	wiping	16	trillion	in
Perhaps	Warren	has	Sunday	stood	breaking	10	billion	for
Like	Warren	,	Monday	defended	using	166	trillion	in
Elizabeth	Warren	on	Friday	proposed	$	20	trillion	in

grants	in	10	course	eight	.	fight	even	care
funding	over	the	next	three	.	upgrade	them	cover
funds	over	10	next	five	in	improve	American	-
,	over	a	next	10	,	invest	a	insur
spending	over	the	next	decade	to	provide	health	care

more	community	as	the	rates	.	the	middle	class
everyone	child	,	a	taxes	on	the	wealthy	class
some	baby	,	co	taxes	.	the	middle	class
every	American	by	triggering	taxes	on	all	middle	class
every	American	without	raising	taxes	on	the	middle	class

Takeaways

- A language model that adaptively combines local context, short-term memory, and long-term memory.

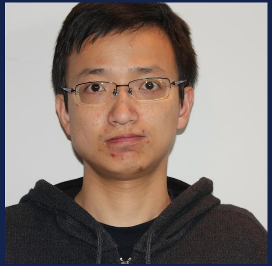
Takeaways

- A language model that adaptively combines local context, short-term memory, and long-term memory.

- A variant of the models for question answering (de Masson d'Autume et al., NeurIPS 2019)


Cyprien


Sebastian


Lingpeng

Dani

Future Directions

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

Future Directions

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

Training Paradigms

Model Architectures

Future Directions

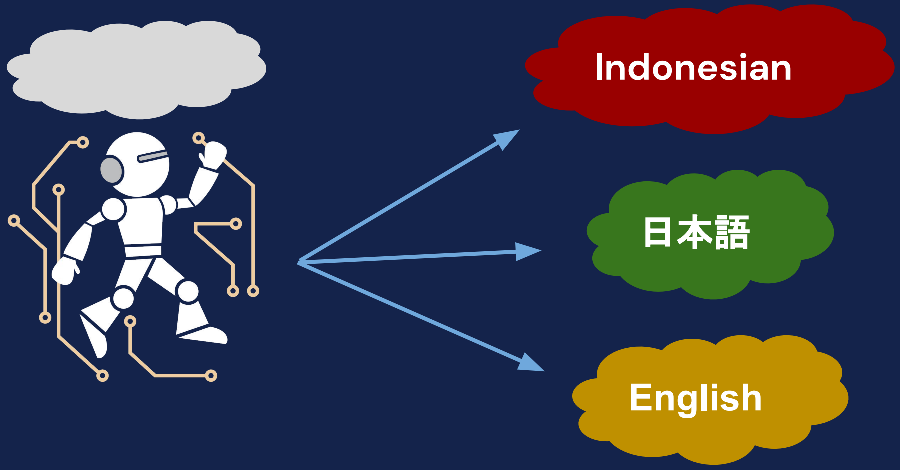
A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

Training Paradigms

Model Architectures

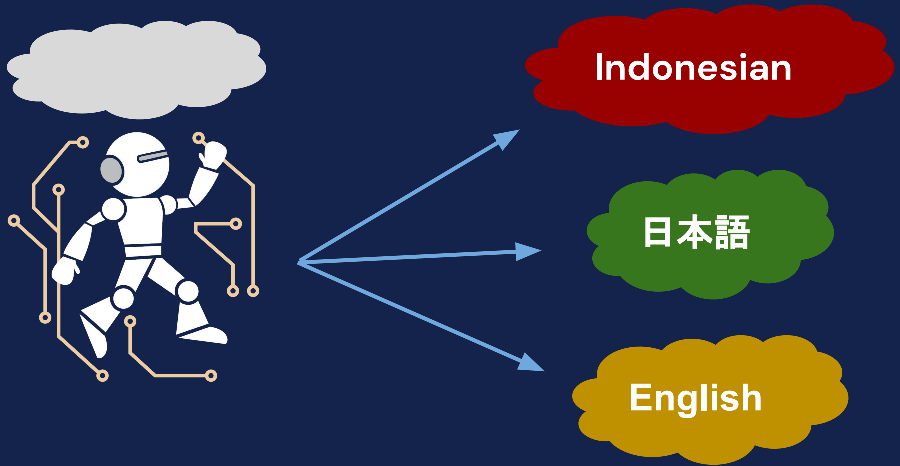
Future Directions

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.



Future Directions

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.



Learning cross-lingual transferable representations

Artetxe et al., ACL 2020

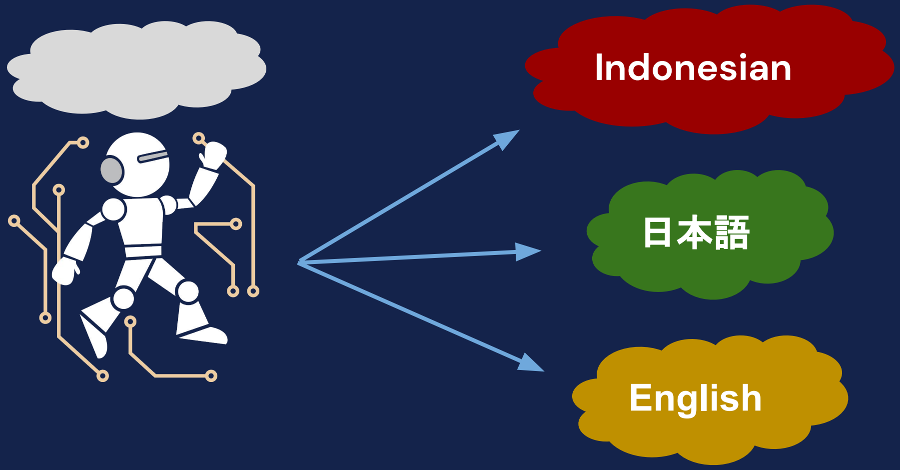
Mikel

Sebastian

Dani

Future Directions

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

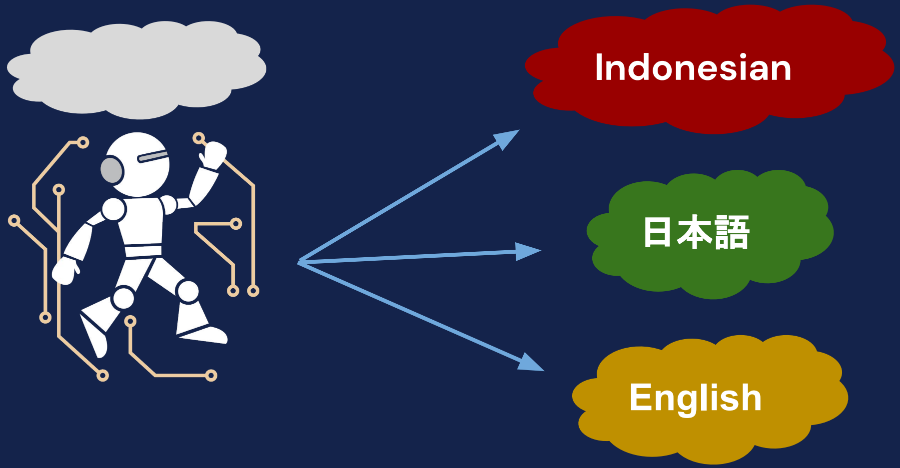


Distributionally Robust Optimization

$$\min_{\theta} \sup_{q} \mathbb{E}_{(x,y)\sim q} \mathcal{L}_{\theta}(x, y)$$

Future Directions

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.



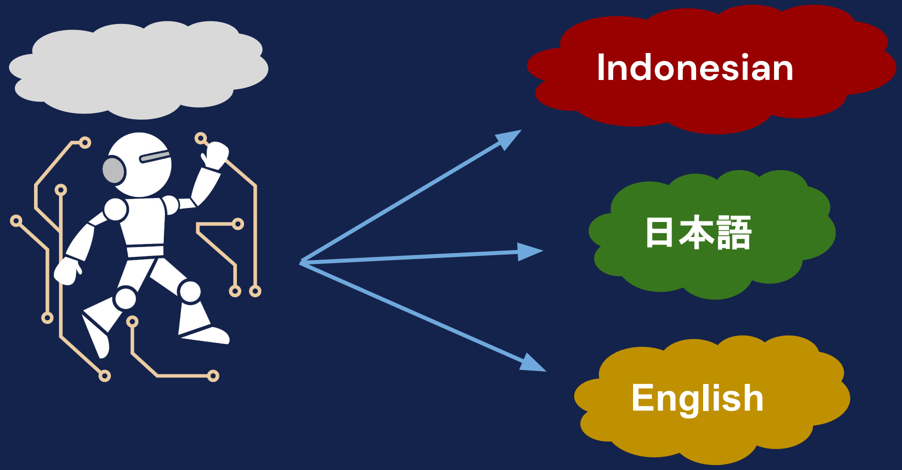
Distributionally Robust Optimization

$$\min_{\theta} \sup_{q} \mathbb{E}_{(x,y)\sim q} \mathcal{L}_{\theta}(x, y)$$

Language

Future Directions

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.



Distributionally Robust Optimization

$$\min_{\theta} \sup_{q} \mathbb{E}_{(x,y)\sim q}\mathcal{L}_{\theta}(x, y)$$

Ensuring that a language model works equally well across languages (important for fairness)

Future Directions

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

Training Paradigms

Model Architectures

Future Directions

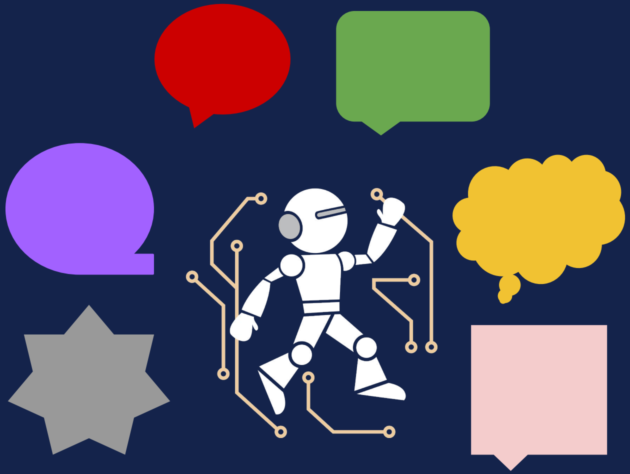
A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

Training Paradigms

Model Architectures

Future Directions

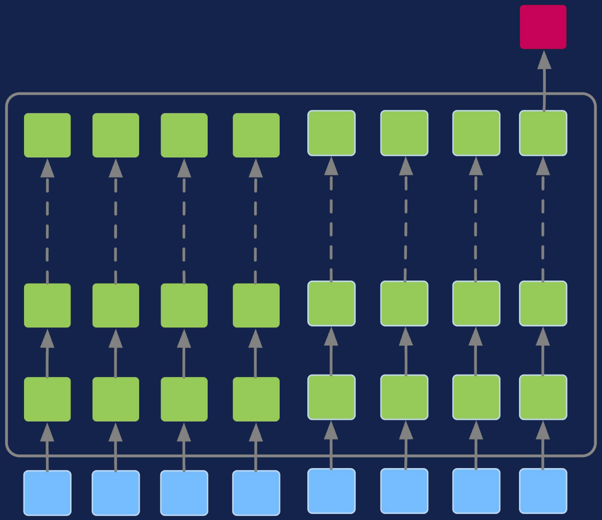
A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.



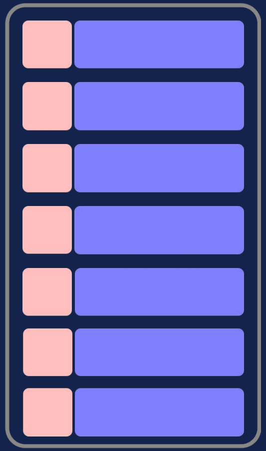
Integration of data from various sources and modalities.

Future Directions

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.



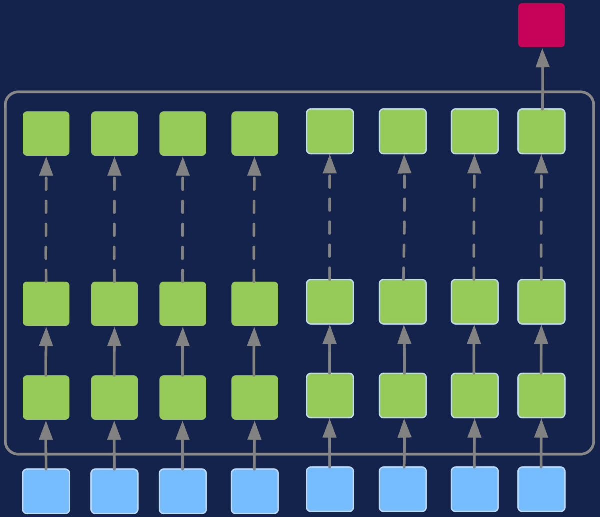
Computation



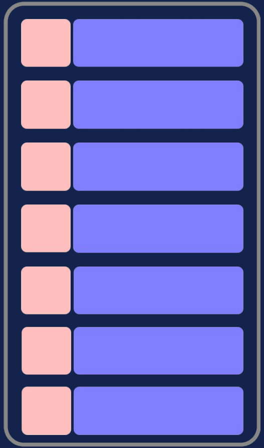
Storage

Future Directions

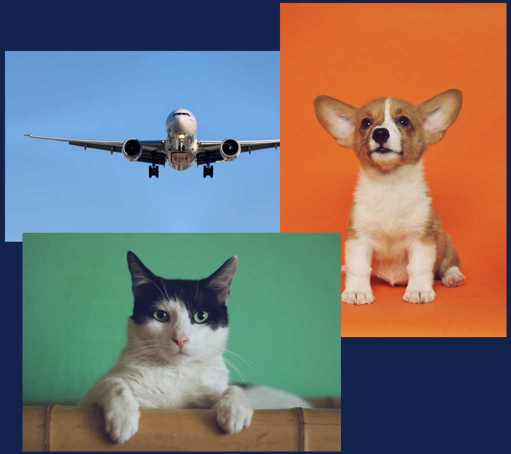
A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.



Computation

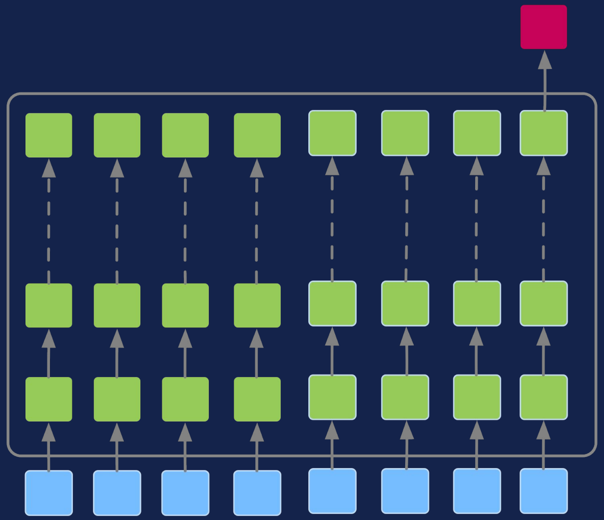


Storage

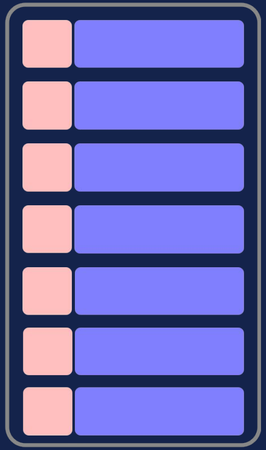


Future Directions

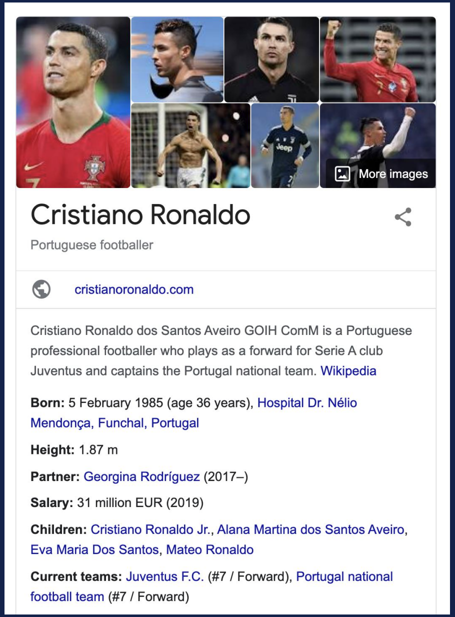
A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.



Computation

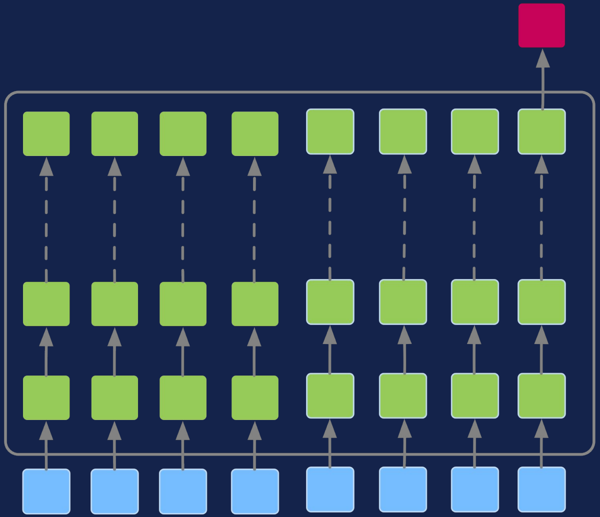


Storage

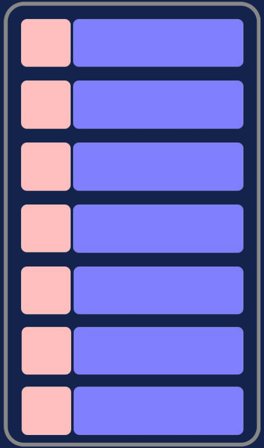


Future Directions

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.



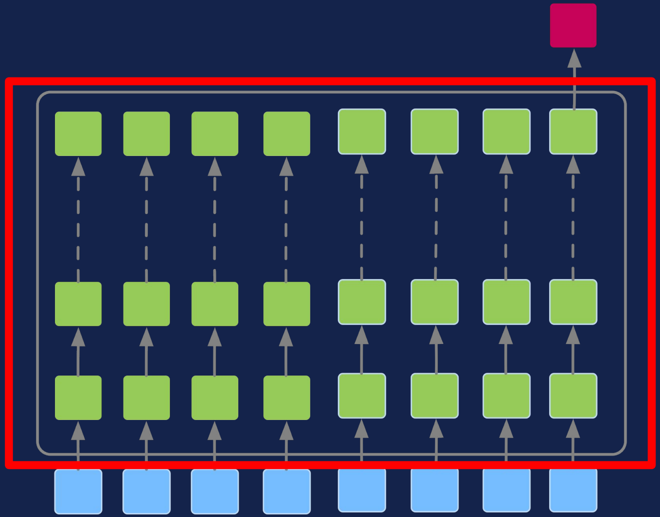
Computation



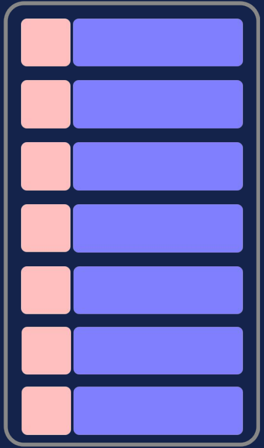
Storage

Future Directions

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.



Computation

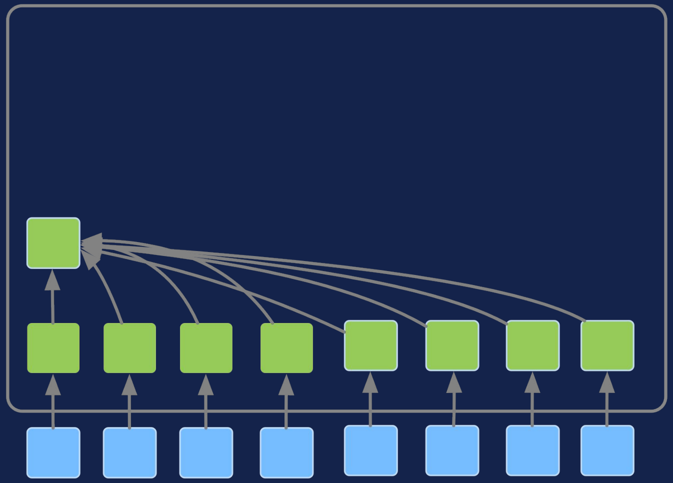


Storage

⬆ computational efficiency

Future Directions

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

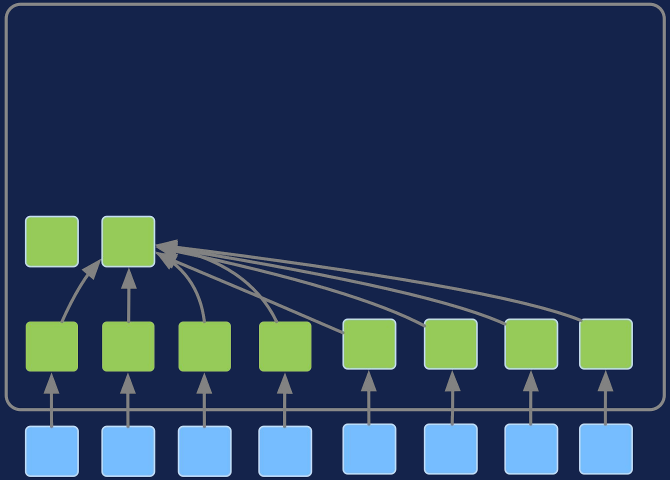


Random Feature Attention

Peng et al., ICLR 2021

Future Directions

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

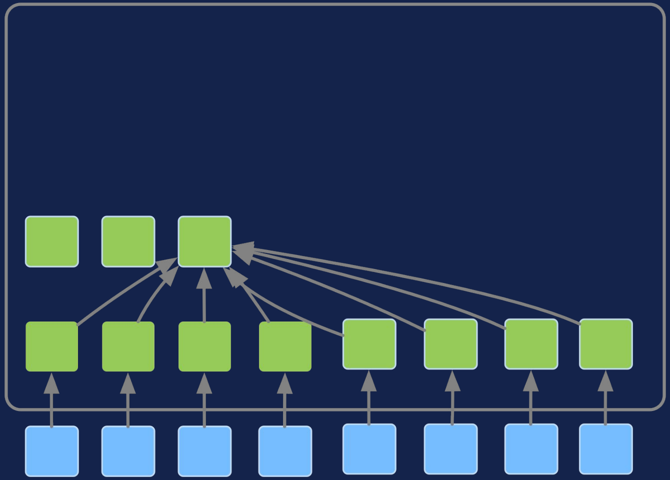


Random Feature Attention

Peng et al., ICLR 2021

Future Directions

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

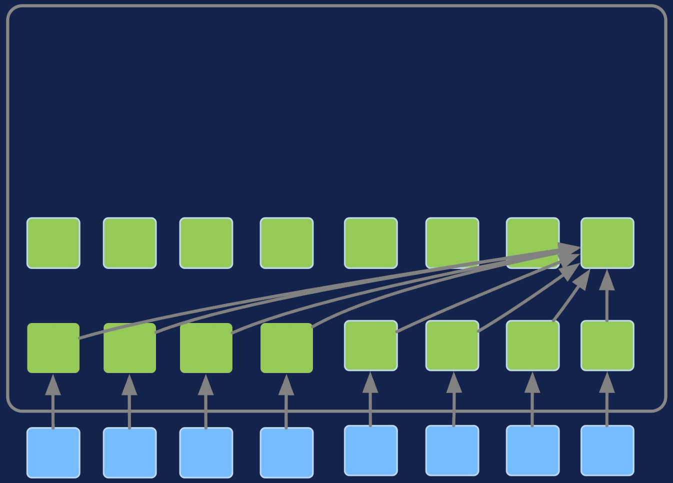


Random Feature Attention

Peng et al., ICLR 2021

Future Directions

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

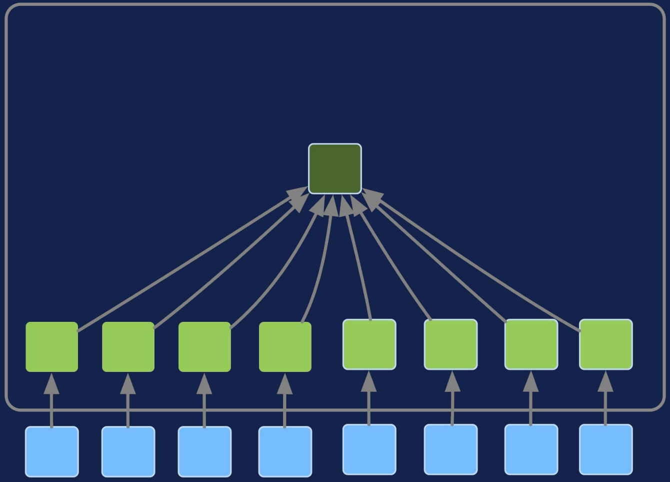


Random Feature Attention

Peng et al., ICLR 2021

Future Directions

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

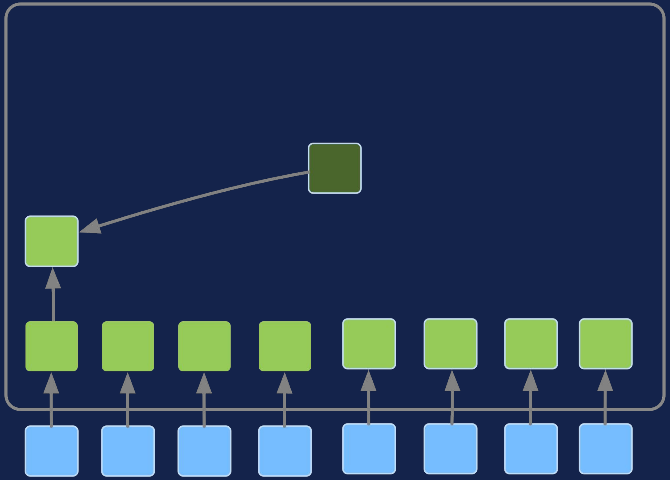


Random Feature Attention

Peng et al., ICLR 2021

Future Directions

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

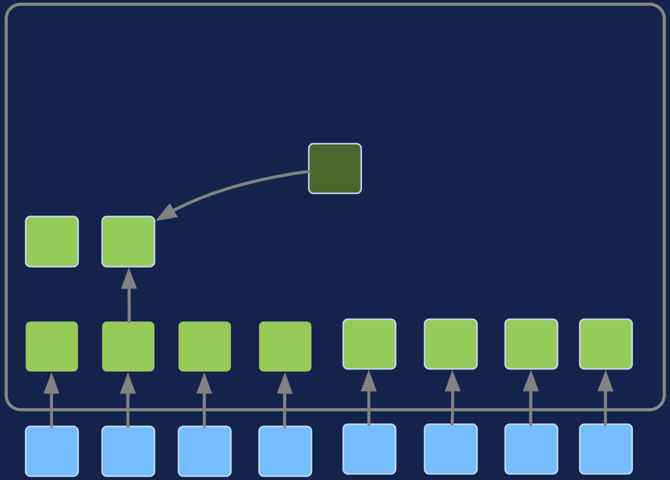


Random Feature Attention

Peng et al., ICLR 2021

Future Directions

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

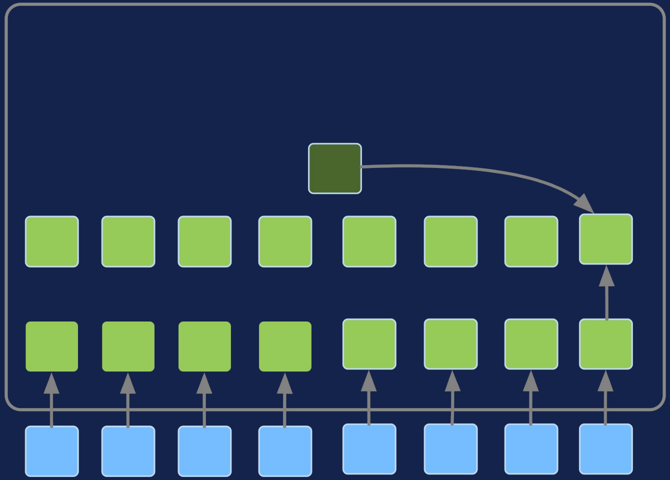


Random Feature Attention

Peng et al., ICLR 2021

Future Directions

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

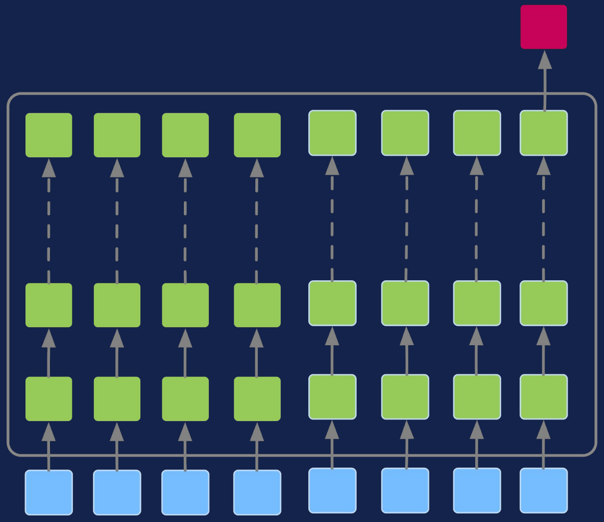


Random Feature Attention

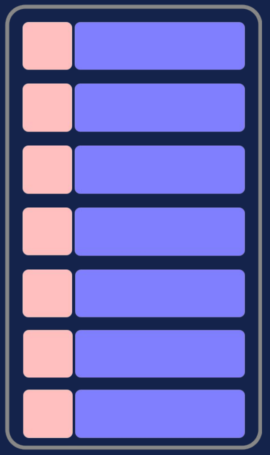
Peng et al., ICLR 2021

Future Directions

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.



Computation

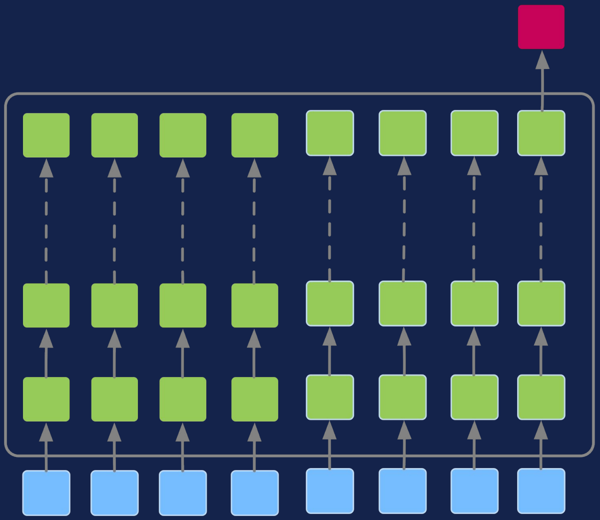


Storage

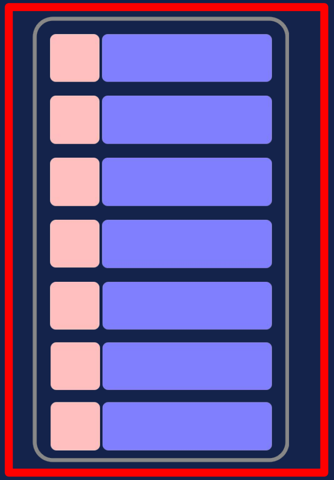
⬆ computational efficiency

Future Directions

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.



Computation

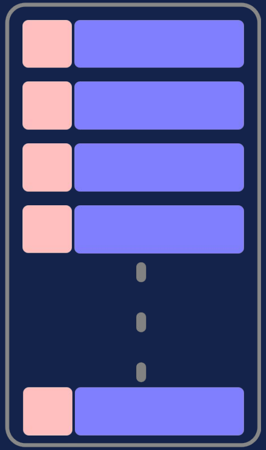


Storage

computational efficiency

Future Directions

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.



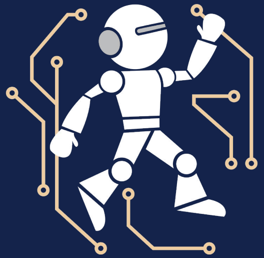
Storage

Learning what to remember and forget

Constant-size memory

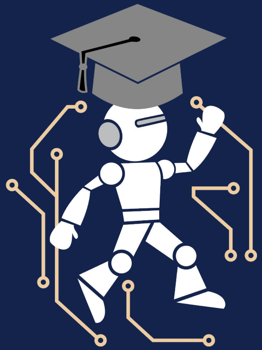
Future Directions

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.



Future Directions

A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.



tack շնորհակալություն Danke

ありがとうございました Salamat

grazie **Thank you** mulțumesc

धन्यवाद நன்றி

Terima kasih Dankie 감사합니다 Merci

Спасибо شكرا جزيلا σας ευχαριστώ

teşekkür ederim 谢谢 cảm ơn bạn

https://dyogatama.github.io

dyogatama@google.com

DeepMind

Memory in Humans

Human language processing is facilitated by specialized memory systems.

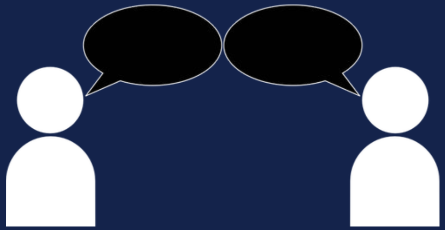
(Tulving, 1985; Rolls, 2000; Eichenbaum, 2012)

Memory in Humans

Human language processing is facilitated by specialized memory systems.

(Tulving, 1985; Rolls, 2000; Eichenbaum, 2012)

Short term



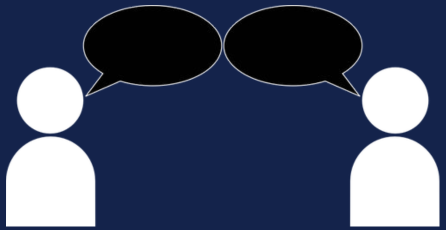
Working

Memory in Humans

Human language processing is facilitated by specialized memory systems.

(Tulving, 1985; Rolls, 2000; Eichenbaum, 2012)

Short term



Working

Long term

Implicit

ML is fun

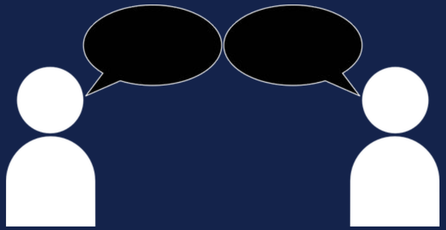
Procedural

Memory in Humans

Human language processing is facilitated by specialized memory systems.

(Tulving, 1985; Rolls, 2000; Eichenbaum, 2012)

Short term



Working

Long term

Implicit

Procedural

Explicit

Semantic

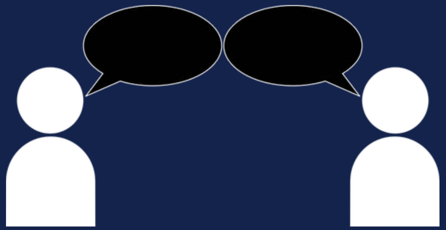
Episodic

Memory in Humans

Human language processing is facilitated by specialized memory systems.

(Tulving, 1985; Rolls, 2000; Eichenbaum, 2012)

Short term



Working

Long term

Explicit

Procedural

Semantic

Episodic

Memory in AI

Short term	Long term
LSTM (Hochreiter and Schmidhuber, 1997)	Memory Networks (Weston et al, 2015)
Differentiable Neural Computers (Graves et al, 2016)	Never-Ending Language Learning (Mitchell et al, 2015)
Reformer (Kitaev et al., 2020)	Matching Networks (Vinyals et al, 2016)
Transformer XL (Dai et al., 2019)	REALM (Guu et al, 2020)

Memory in AI

Short term	Long term
LSTM (Hochreiter and Schmidhuber, 1997)	Memory Networks (Weston et al, 2015)
Differentiable Neural Computers (Graves et al, 2016)	Never-Ending Language Learning (Mitchell et al, 2015)
Reformer (Kitaev et al., 2020)	Matching Networks (Vinyals et al, 2016)
Transformer XL (Dai et al., 2019)	REALM (Guu et al, 2020)

Stack LSTM

Yogatama et al., ICLR 2018

Memory-based Parameter Adaptation ++

de Masson d'Autume, Ruder, Kong, Yogatama, NeurIPS 2019

Memory in AI

Short term	Long term
LSTM (Hochreiter and Schmidhuber, 1997)	Memory Networks (Weston et al, 2015)
Differentiable Neural Computers (Graves et al, 2016)	Never-Ending Language Learning (Mitchell et al, 2015)
Reformer (Kitaev et al., 2020)	Matching Networks (Vinyals et al, 2016)
Transformer XL (Dai et al., 2019)	REALM (Guu et al, 2020)

Stack LSTM

Yogatama et al., ICLR 2018

Memory-based Parameter Adaptation ++

de Masson d'Autume, Ruder, Kong, Yogatama, NeurIPS 2019

A language model with short-term and long-term memory.

Background

Knowledge is encoded in the weights of a parametric neural network.

Interpretations via cloze-style questions (Petroni et al., 2020) or prompts (Brown et al., 2020).

Dante was born in [MASK] .

Q: Where was Dante born in?

A:

Experiments

Perplexity (1-inf), lower is better

	Base	TXL	kNN-LM	Ours
WikiText-103	21.8	19.1	18.0	**17.6***
WMT	16.5	15.5	15.2	**14.1**

$$\lambda p_{k\mathrm{NN}}(x_t \mid \boldsymbol{x}_{<t}) + (1-\lambda) p_{\mathrm{LM}}(x_t \mid \boldsymbol{x}_{<t})$$

kNN-LM: Khandelwal et al., 2020

Experiments

BPC (0-inf), lower is better

	Base	TXL	kNN-LM	Ours
enwik8	1.05	1.01	1.02	**1.00**

Transformer: Vaswani et al., 2017

Transformer-XL: Dai et al., 2019

kNN-LM: Khandelwal et al., 2020

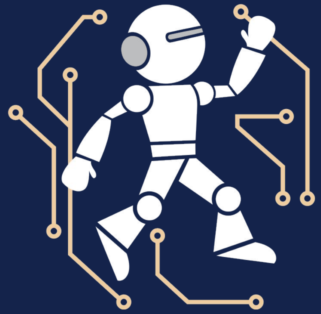
Takeaway and Limitation

- A language model that adaptively combines local context, short-term memory, and long-term memory.

- Retrieving from long-term memory is expensive.

	CPUs	Hours
WikiText-103	1,000	6
WMT	9,000	18
enwik8	1,000	8

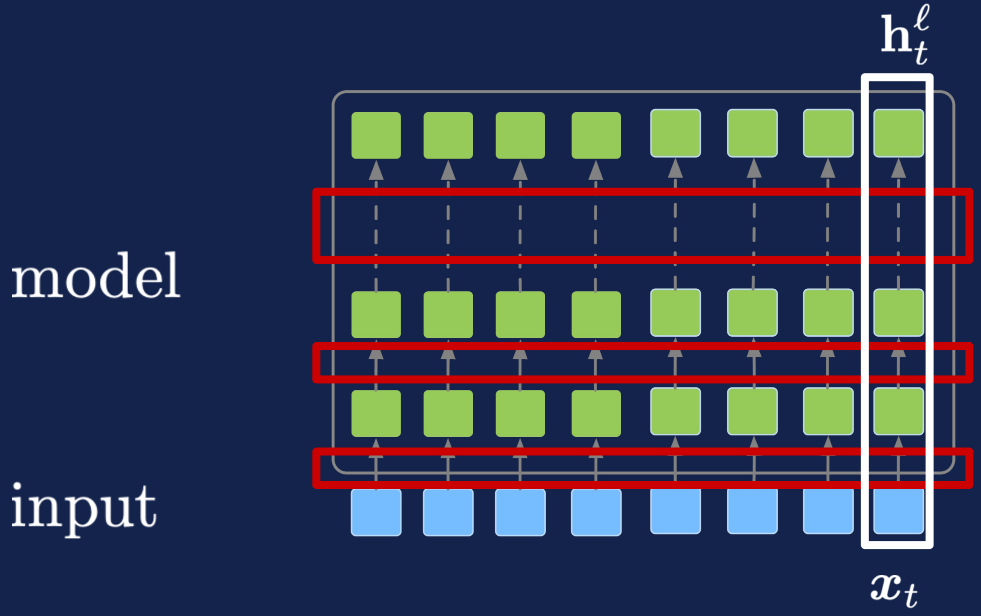
Challenges: Human Learning vs. Machine Learning



	Machine
Acquisition	Large datasets (representation learning)
Task Training	Large datasets (supervised fine tuning)
Linguistic knowledge	Dataset specific
Generalization	Forget previous tasks given a new task

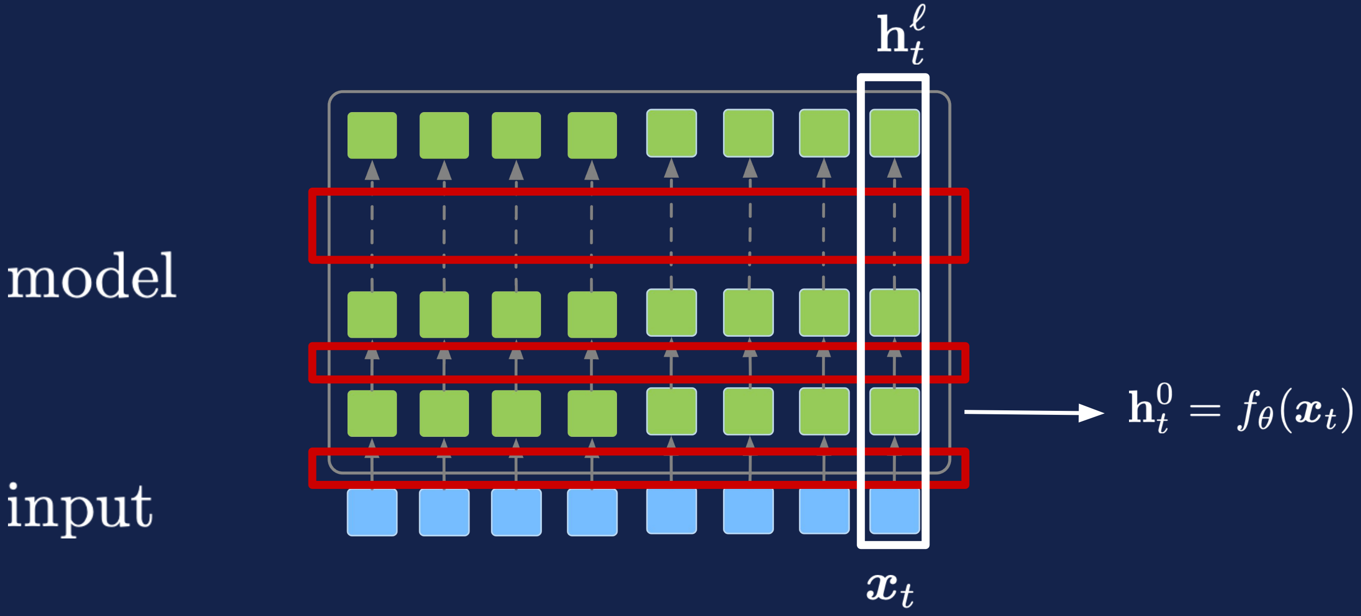
Background

Knowledge is encoded in the weights of a parametric neural network.



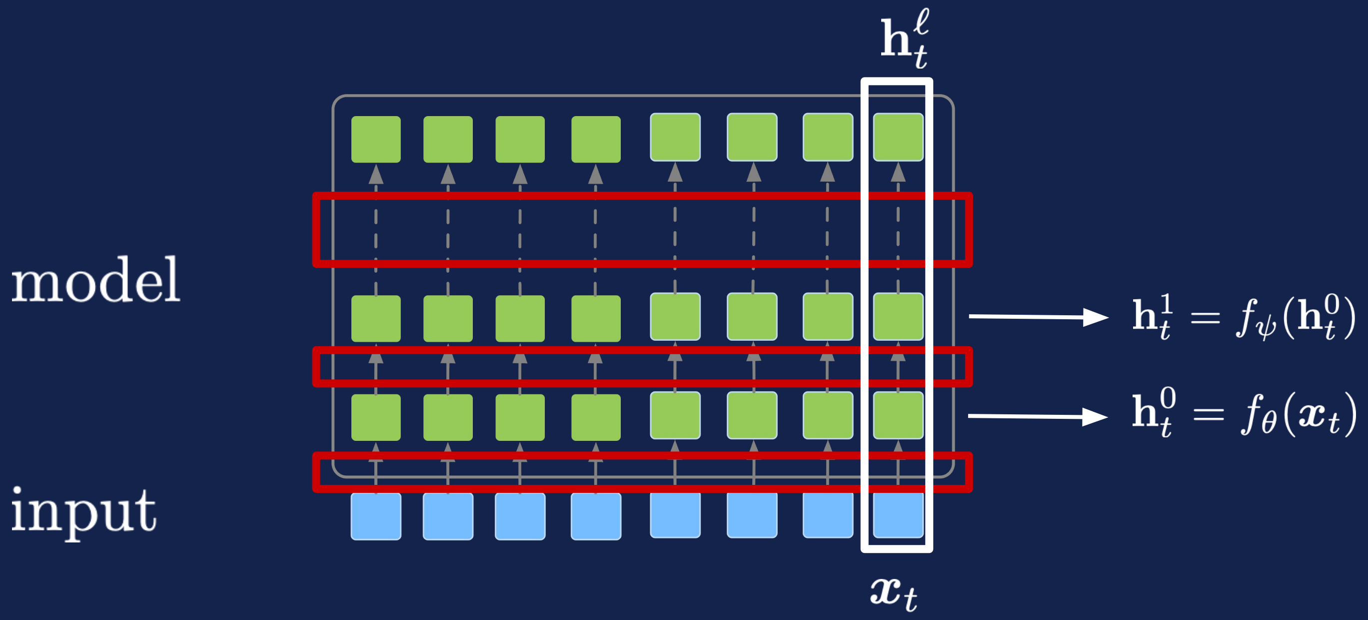
Background

Knowledge is encoded in the weights of a parametric neural network.



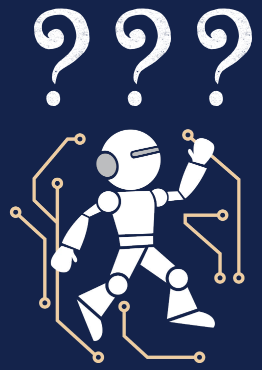
Background

Knowledge is encoded in the weights of a parametric neural network.



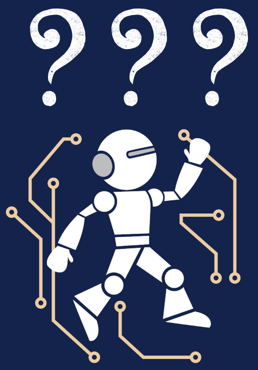
Background

Current models are prone to forgetting



Background

Current models are prone to forgetting



- Incoherent text generations.
- Hallucinating answers in open-domain QA.
- Performance degradation over time.

Background

Our semiparametric language model architecture is designed to mitigate these problems