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

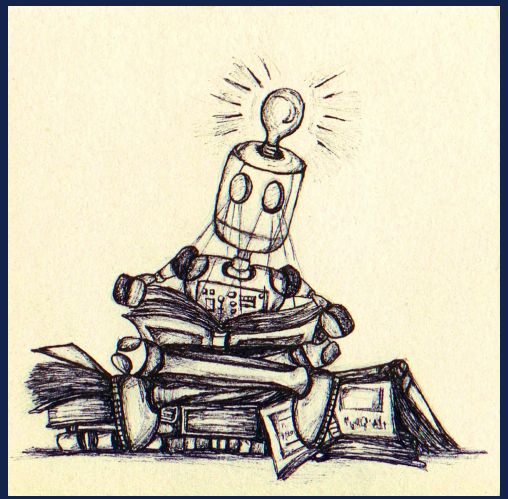
Language is key to **human intelligence** and is important for **artificial intelligence**.

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.

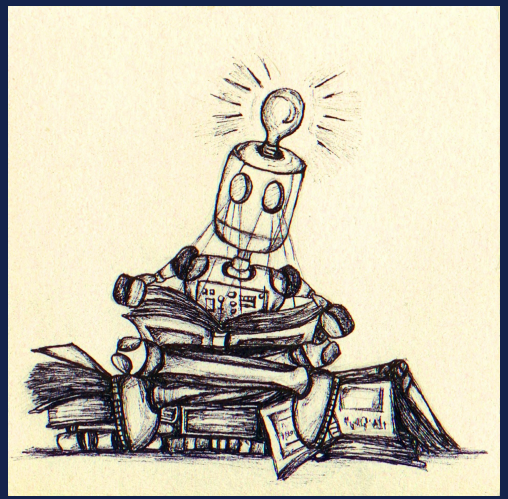
General Linguistic Intelligence

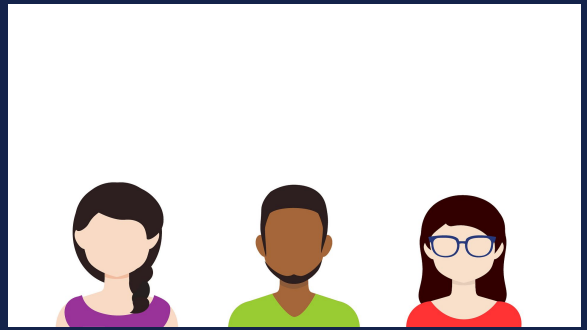
The ability to **acquire**, **store**, and **reuse** knowledge (about a language's lexicon, syntax, semantics, and pragmatic conventions) to **adapt** to new tasks **quickly without forgetting** old ones.



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.

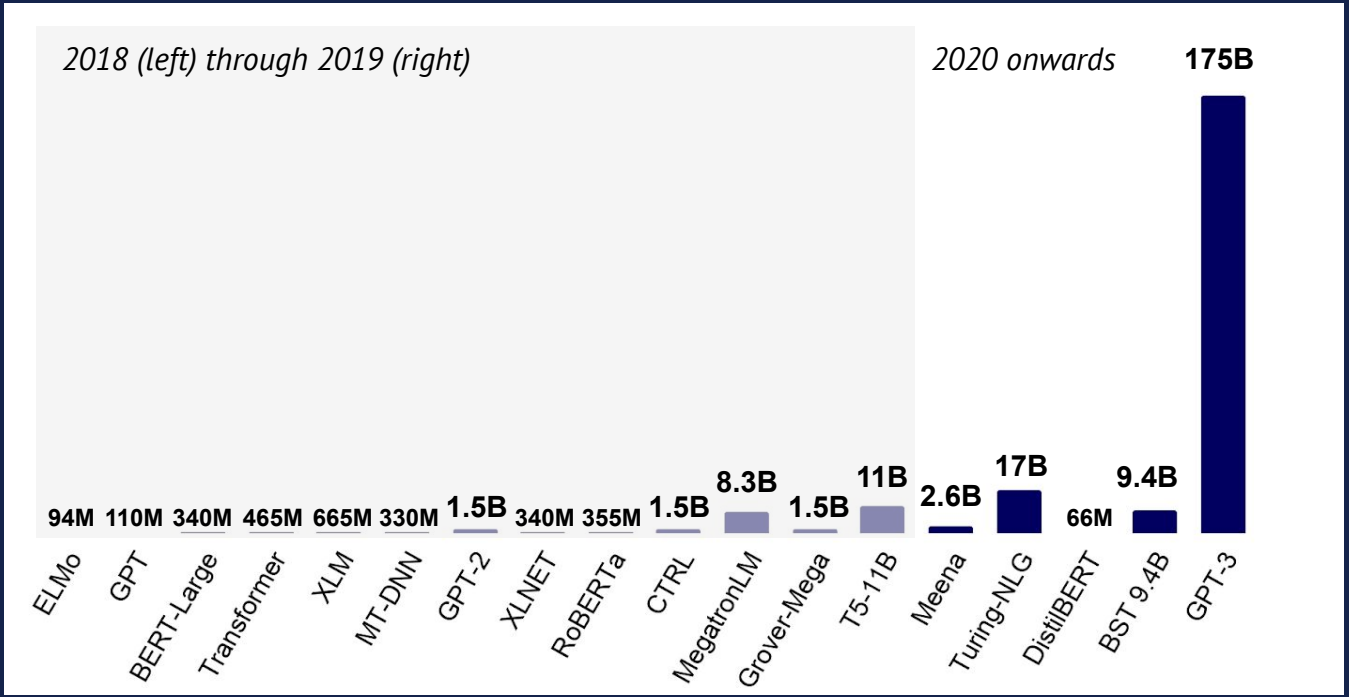


Figure taken from State of AI Report 2020.

Challenges: Human Learning vs. Machine Learning

	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

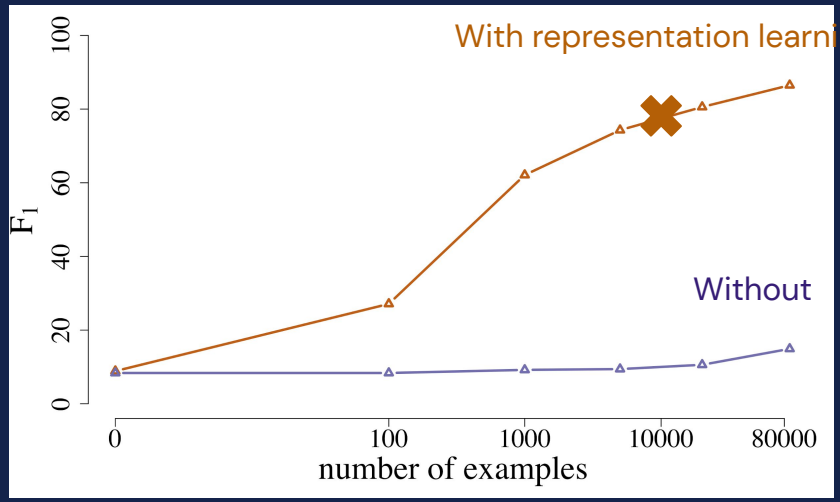
Challenges: Human Learning vs. Machine Learning

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**	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

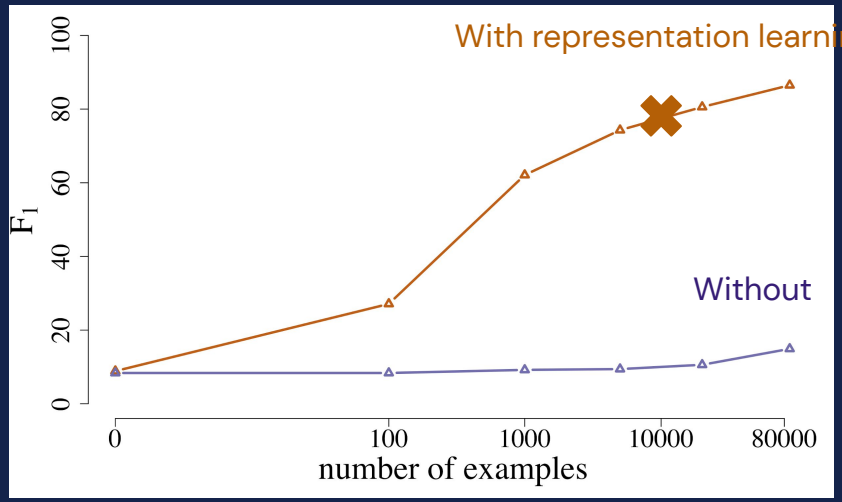


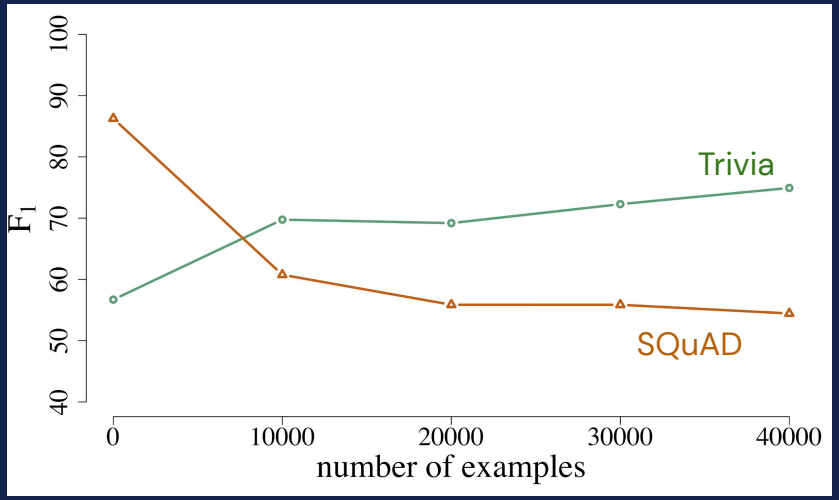
Yogatama et al., arXiv 2019

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

Better 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

Model Architectures

Research Areas

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

Training Paradigms

Generative Training

Yogatama et al., TACL 2014
Yogatama et al., arXiv 2017
Kong, Melis, Ling, Yu, and Yogatama, ICLR 2018
Cao and Yogatama, arXiv 2020

Model Architectures

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

Model Architectures

Research Areas

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

Training Paradigms

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

This Talk

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

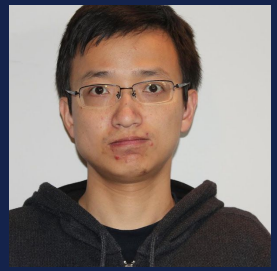
  Kong et al., ICLR 2020

- Semiparametric (memory-augmented) language models.

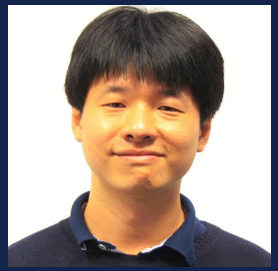
  Yogatama et al., TACL 2021

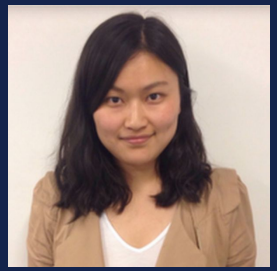
A Mutual Information Maximization Perspective of Language Representation Learning

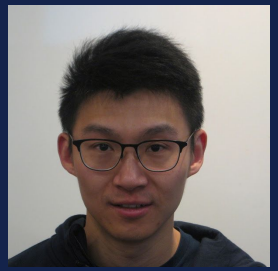
Kong et al., ICLR 2020


Lingpeng


Cyprien


Wang

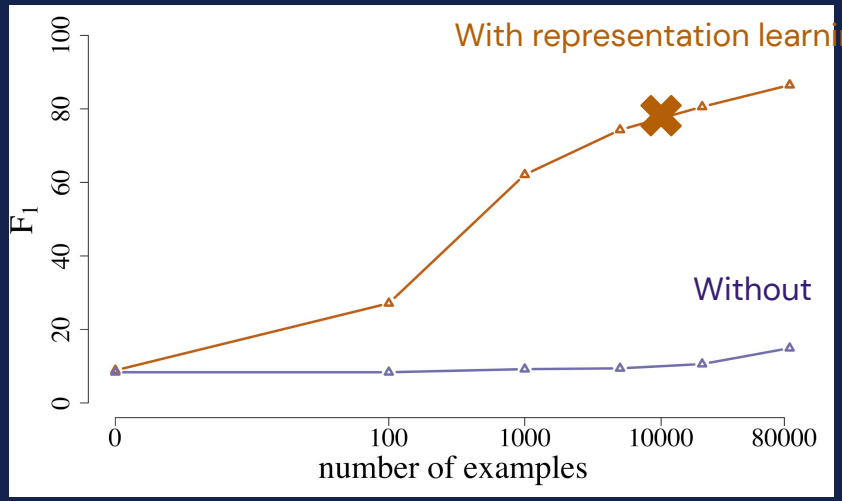

Lei


Zihang

Dani

Text Representations

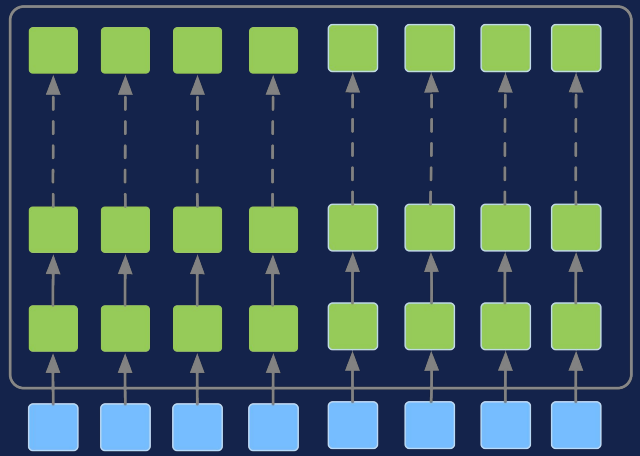
Good representations facilitate more efficient transfer.



Text Representations

model

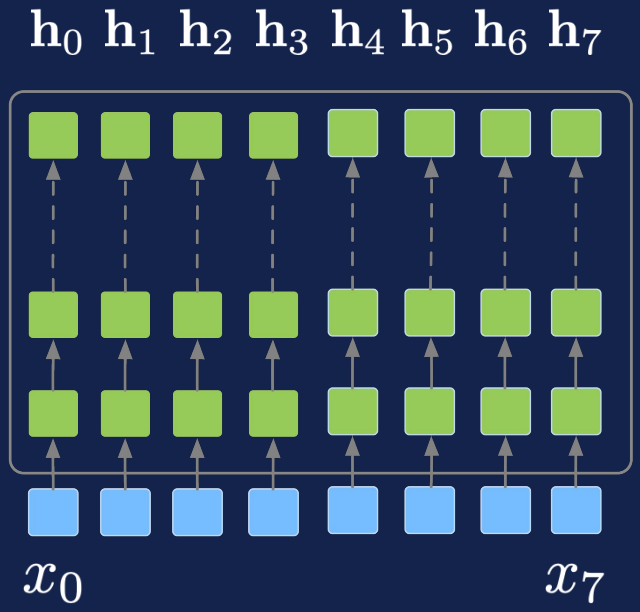
input



Text Representations

model

input

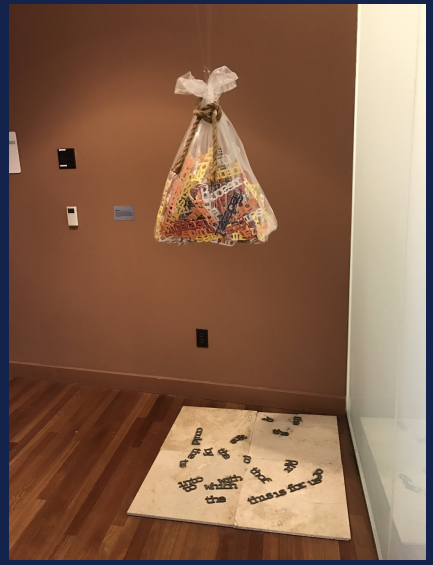


$$f(x_0, x_1, \ldots, x_t) \rightarrow \mathbf{h}_0, \mathbf{h}_1, \ldots, \mathbf{h}_t$$

Text Representations

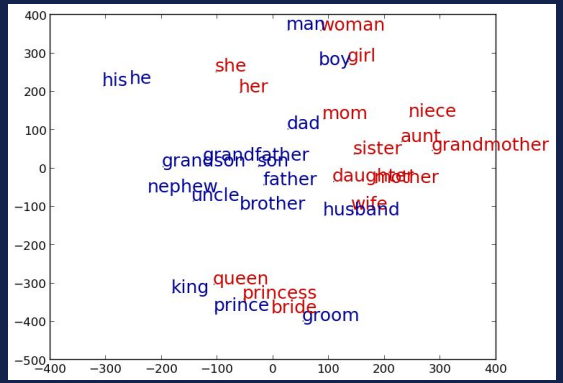
Representation learning started showing promise ~8 years ago.

Text Representations



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

Representation learning started showing promise ~8 years ago.



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

Skip gram, Mikolov et al., 2013.

GloVe, Pennington et al., 2014.

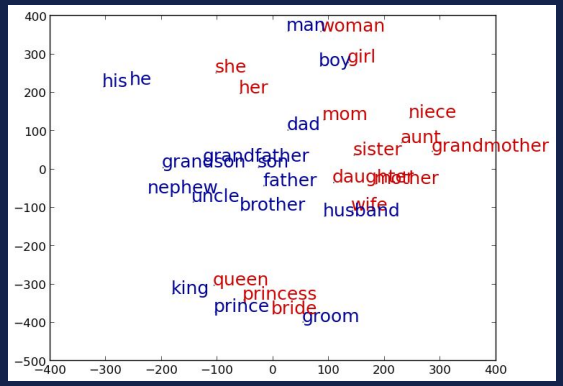
ELMo, Peters et al., 2018.

BERT, Devlin et al., 2019.

Text Representations

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

What has been the main driver of progress so far?



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

Skip gram, Mikolov et al., 2013.

GloVe, Pennington et al., 2014.

ELMo, Peters et al., 2018.

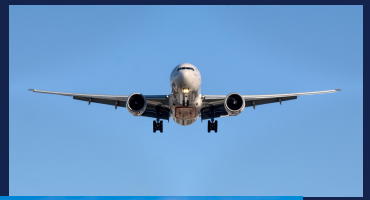
BERT, Devlin et al., 2019.

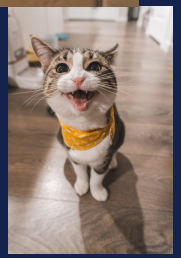
Contrastive Learning

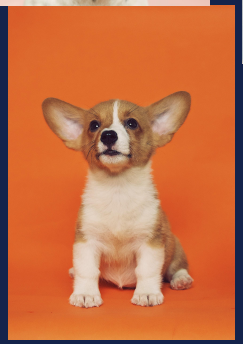
Main assumption: representations should capture similarity (Arora et al., 2019).

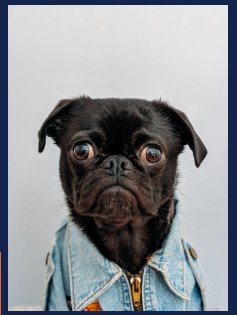
Contrastive Learning

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).

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

InfoNCE objective

Logeswaran and Lee, 2018

van den Oord, et al., 2019

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

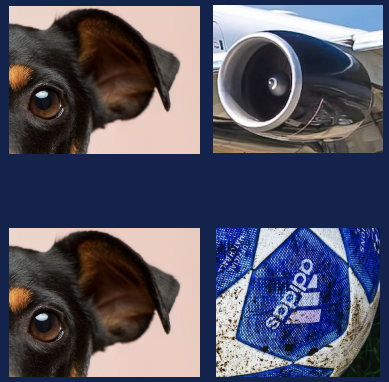
Contrastive Learning with InfoNCE

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

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

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

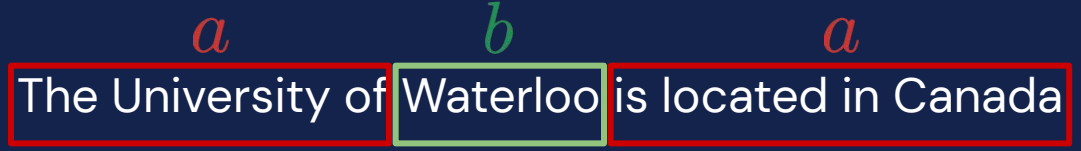
$$\mathbb{E}_{p(A,B)}\left[\mathbb{E}_{p(C)}\left[\log \frac{\exp f_{\boldsymbol{\theta}}(a,b)}{\exp f_{\boldsymbol{\theta}}(a,b) + \sum_{c \neq b} \exp f_{\boldsymbol{\theta}}(a,c)}\right]\right]$$

a b

The University of Waterloo is located in Canada

Contrastive Learning with InfoNCE

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



Contrastive Learning with InfoNCE

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

a b

The University of Waterloo is located in Canada

Contrastive Learning with InfoNCE

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

a b

The University of Waterloo is located in Canada

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

Contrastive Learning with InfoNCE

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

Tokyo
London
dog
cat

a b

The University of Waterloo is located in Canada

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

Contrastive Learning with InfoNCE

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

a b

The University of Waterloo is located in Canada

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

Contrastive Learning with InfoNCE

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

a b

The University of Waterloo is located in Canada

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

Contrastive Learning with InfoNCE

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

a b

The University of Waterloo is located in Canada

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

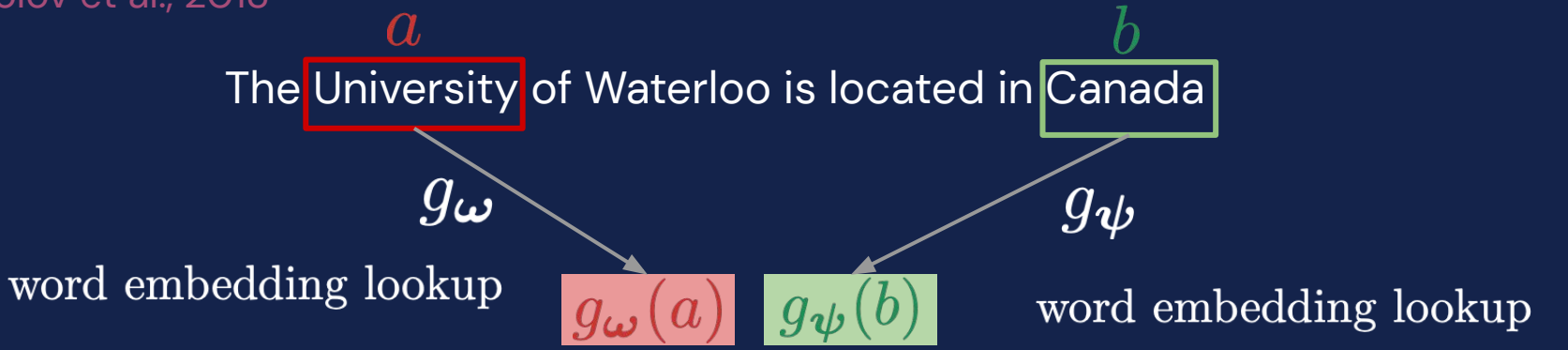
Skip-gram

Mikolov et al., 2013

The University of Waterloo is located in Canada

Skip-gram

Mikolov et al., 2013



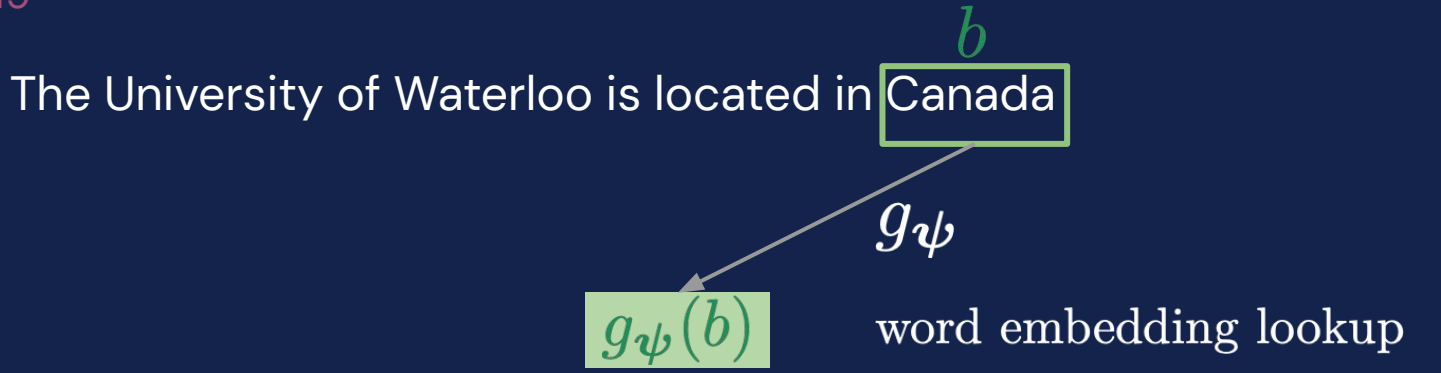
BERT

Devlin et al., 2019

The University of Waterloo is located in Canada

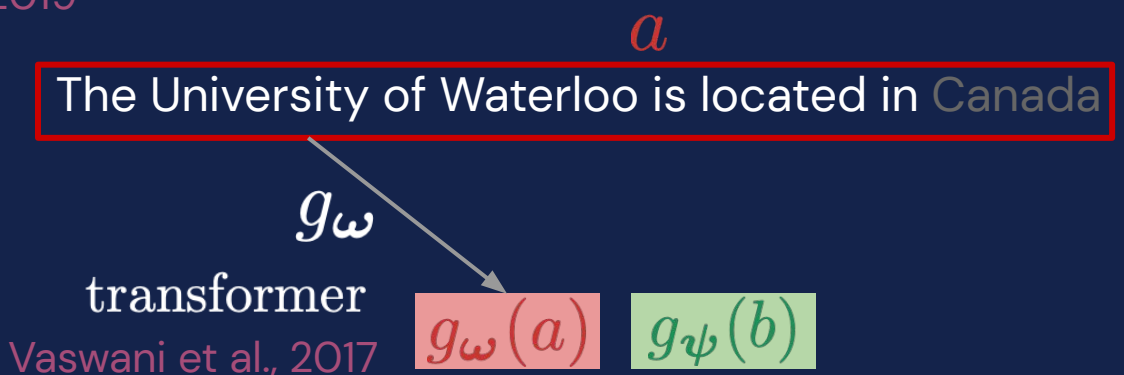
BERT

Devlin et al., 2019



BERT

Devlin et al., 2019



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).

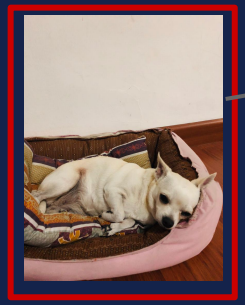
- Facilitates exchanges of ideas on how to improve representation learning models.

Model

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

Model

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

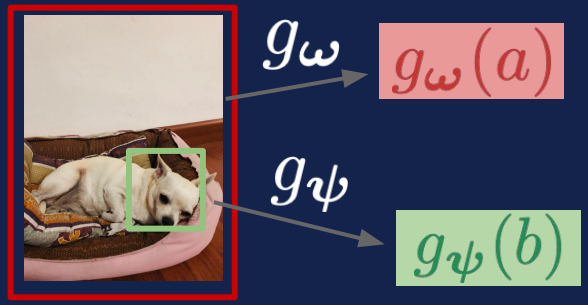


g_ω

$g_\omega(a)$

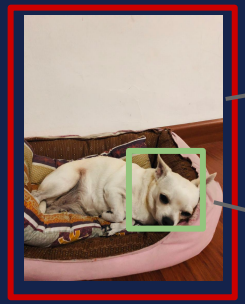
Model

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



Model

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

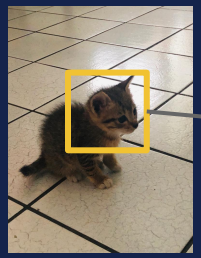


g_ω

$g_\omega(a)$

g_ψ

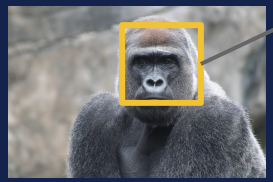
$g_\psi(b)$



g_ψ

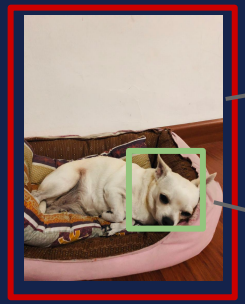
$g_\psi(c_1)$

$g_\psi(c_2)$



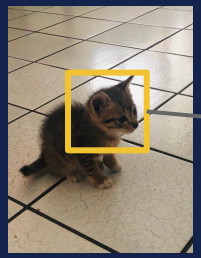
Model

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

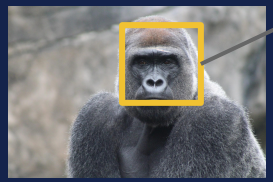


g_ω $g_\omega(a)$

g_ψ $g_\psi(b)$



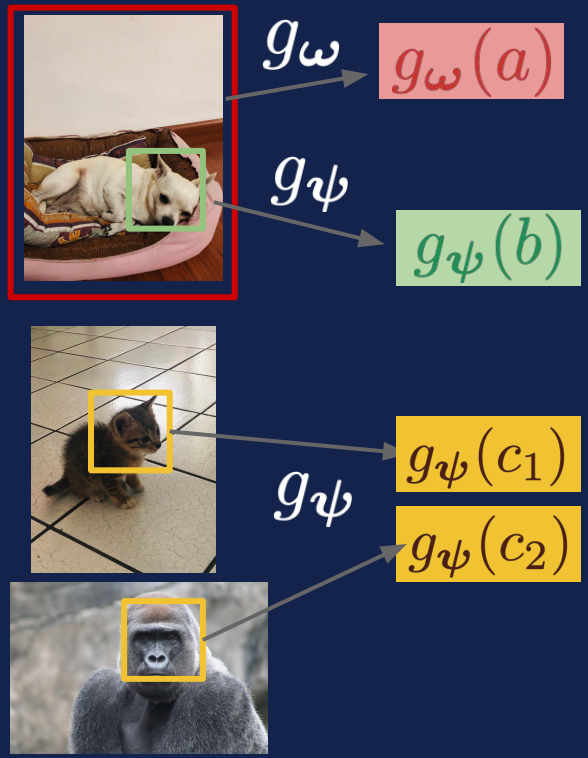
g_ψ $g_\psi(c_1)$ $g_\psi(c_2)$



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

Model

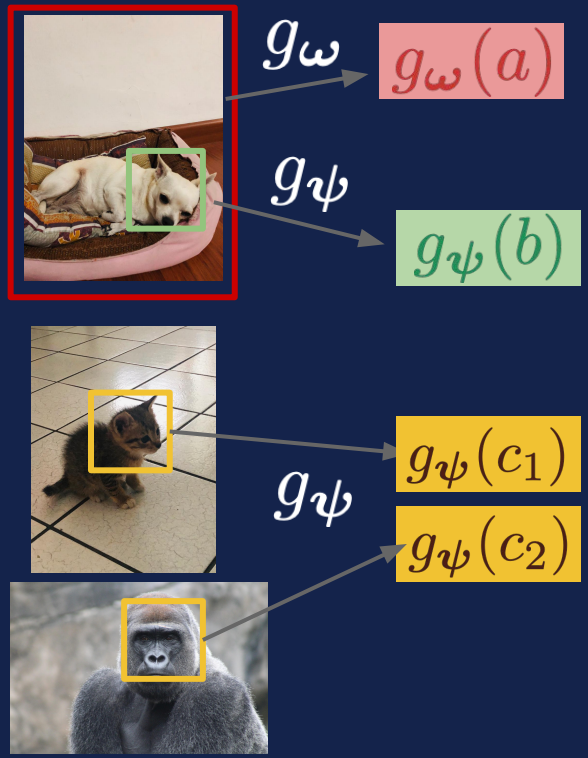
Deep InfoMax (DIM; Hjelm et al., 2019)

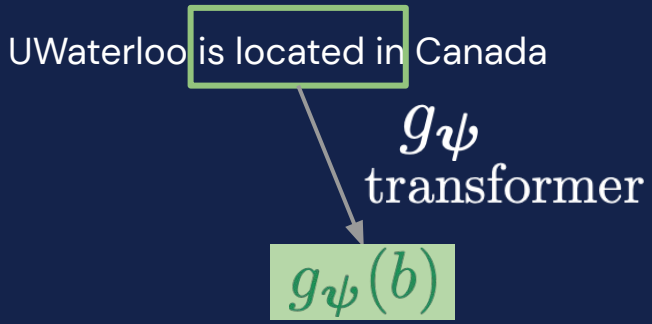


UWaterloo is located in Canada

Model

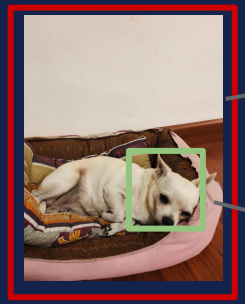
Deep InfoMax (DIM; Hjelm et al., 2019)





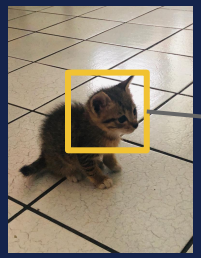
Model

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

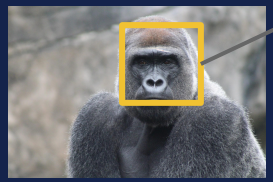


g_ω $g_\omega(a)$

g_ψ $g_\psi(b)$



g_ψ $g_\psi(c_1)$ $g_\psi(c_2)$



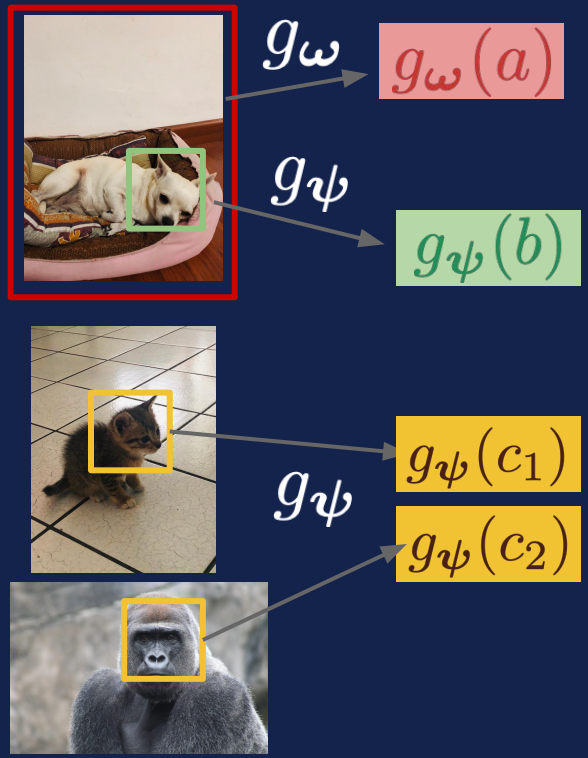
UWaterloo is located in Canada

g_ω
transformer

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

Model

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



UWaterloo is located in Canada

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

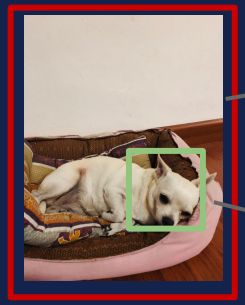
Starcraft II is a fun game

Cristiano Ronaldo scores an own goal

Machine learning is transforming drug discovery

Model

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

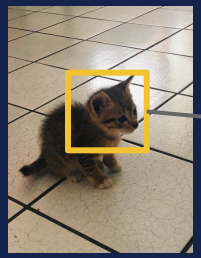


g_ω

$g_\omega(a)$

g_ψ

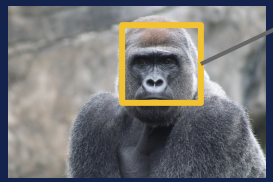
$g_\psi(b)$



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

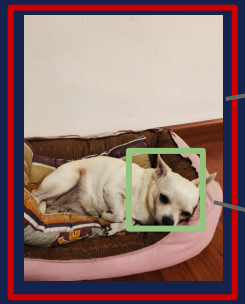
Cristiano Ronaldo scores an own goal

Machine learning is transforming drug discovery

g_ψ transformer

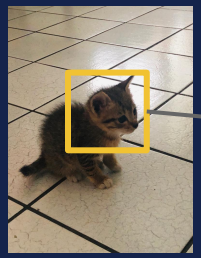
Model

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

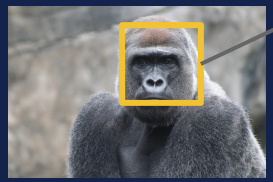


g_ω $g_\omega(a)$

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

- Progress in language representation learning has largely been driven by advances in model architectures (and training objectives).

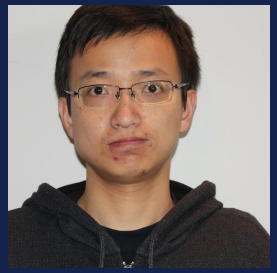
- It is possible to transfer ideas across domains when designing self-supervised tasks.

Adaptive Semiparametric Language Models

Yogatama et al., TACL 2021

Dani

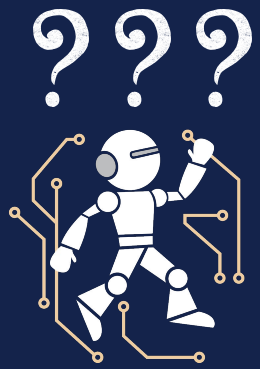
Cyprien



Lingpeng

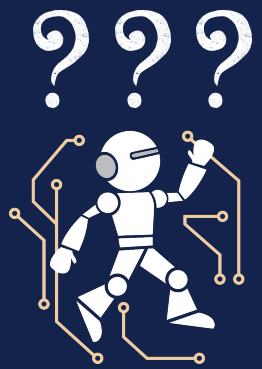
Background

Why does a model forget?



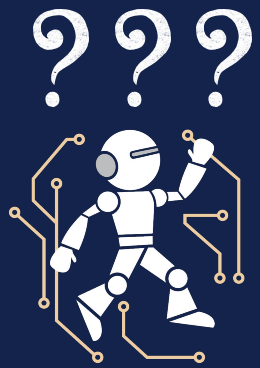
Background

Why does a model forget?



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

Background



Why does a model forget?

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

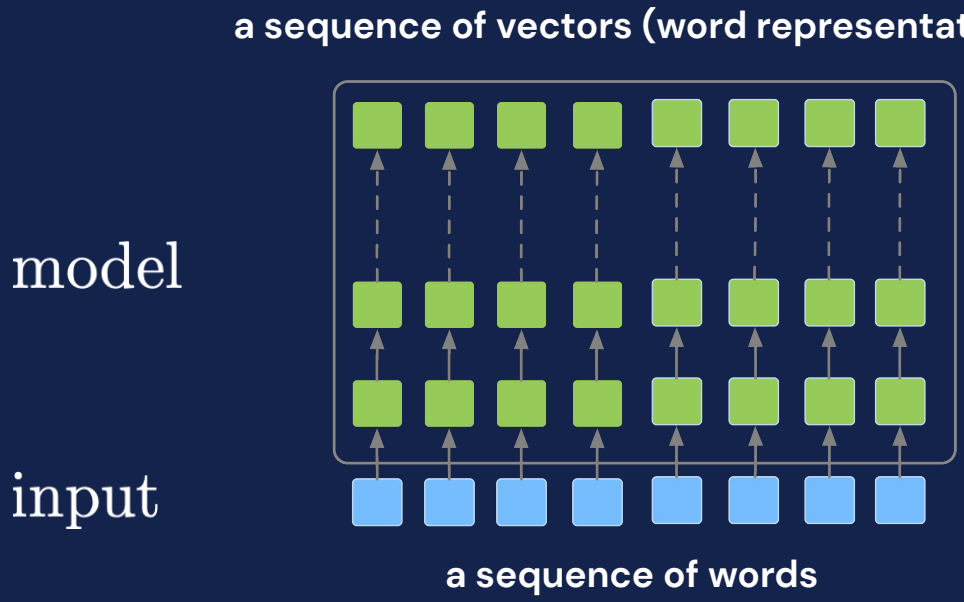
Inability to deal with long-term context

Background

Knowledge is implicitly represented in the weights of a parametric neural network.

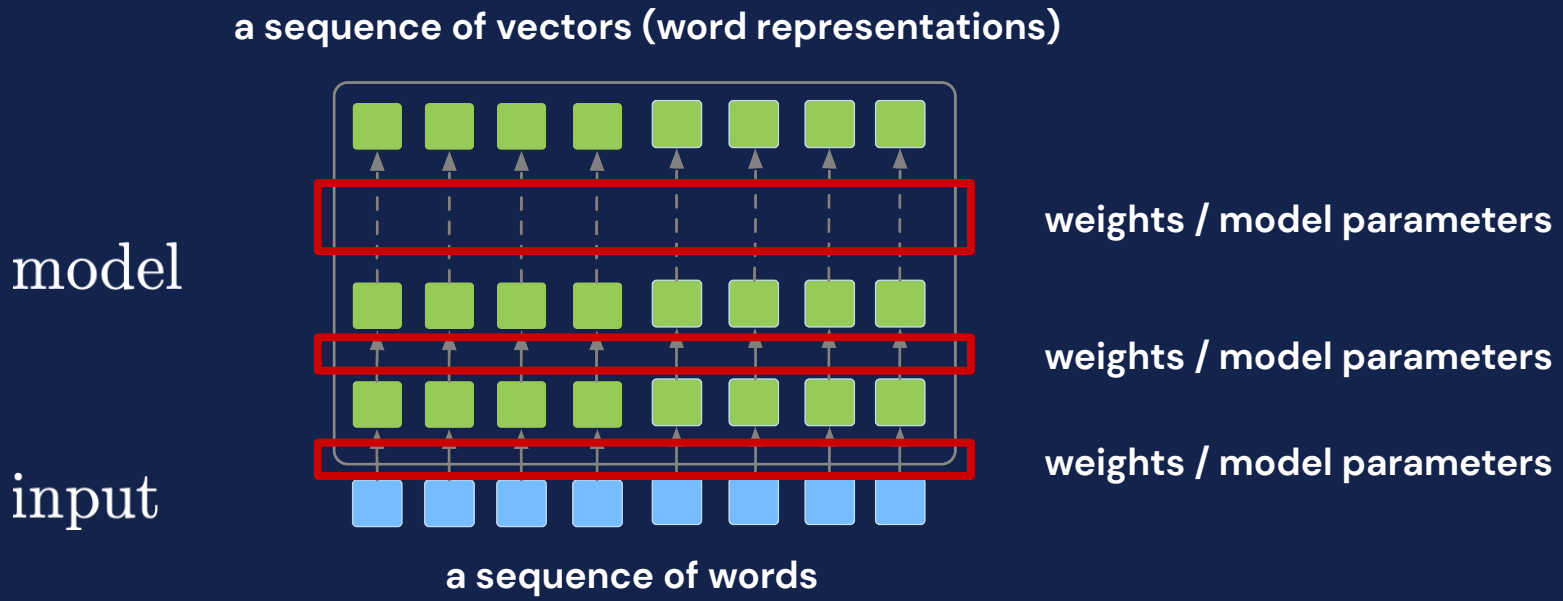
Background

Knowledge is implicitly represented in the weights of a parametric neural network.



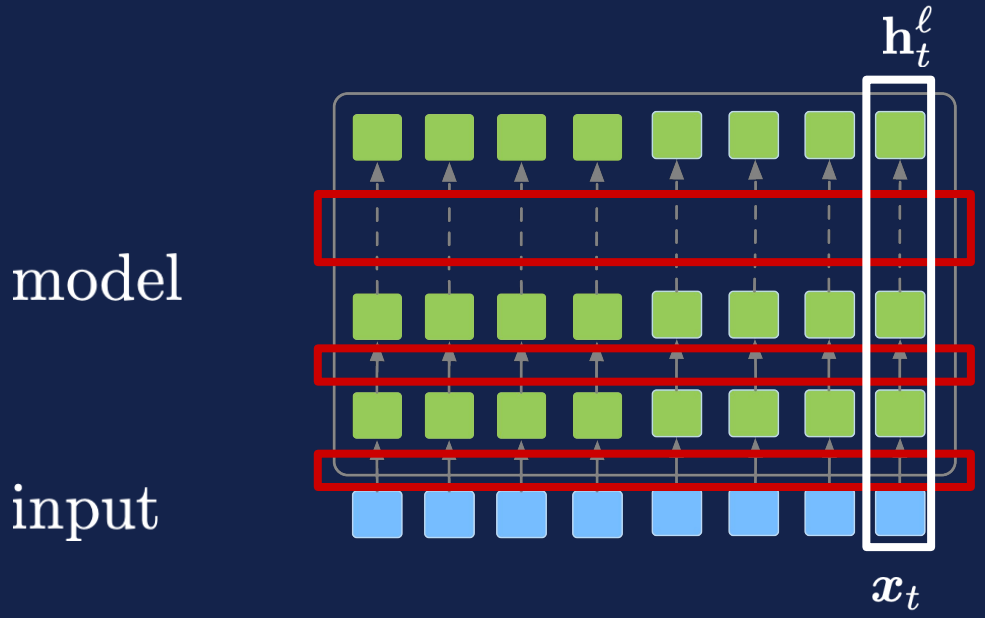
Background

Knowledge is implicitly represented in the weights of a parametric neural network.



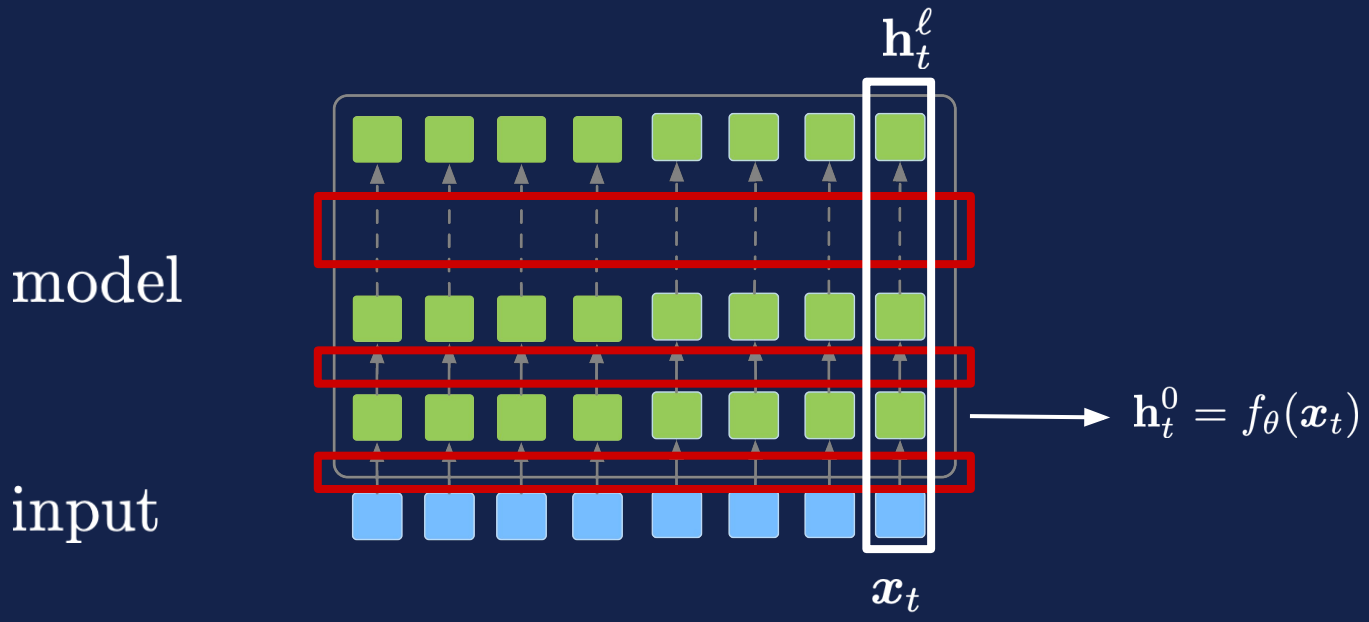
Background

Knowledge is implicitly represented in the weights of a parametric neural network.



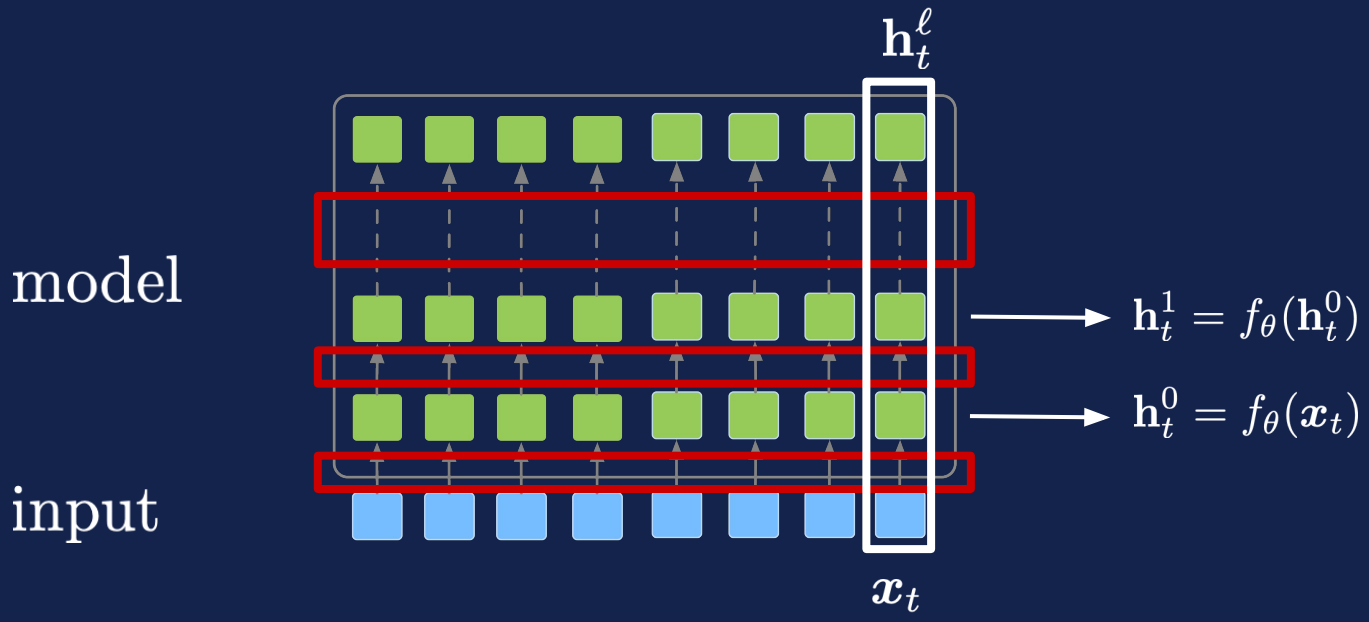
Background

Knowledge is implicitly represented in the weights of a parametric neural network.



Background

Knowledge is implicitly represented in the weights of a parametric neural network.



Background

Knowledge is implicitly represented in the weights of a parametric neural network.

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

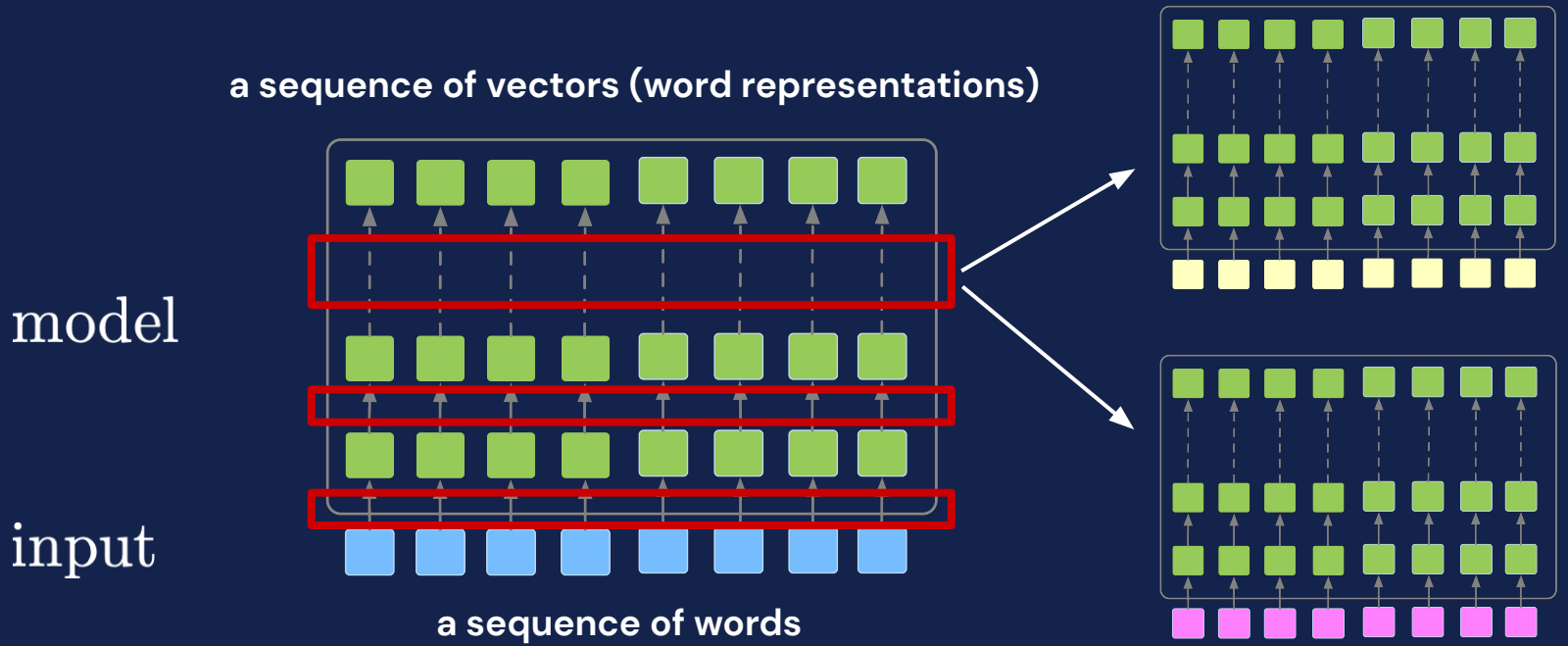
Dante was born in [MASK] .

Q: Where was Dante born in?

A:

Background

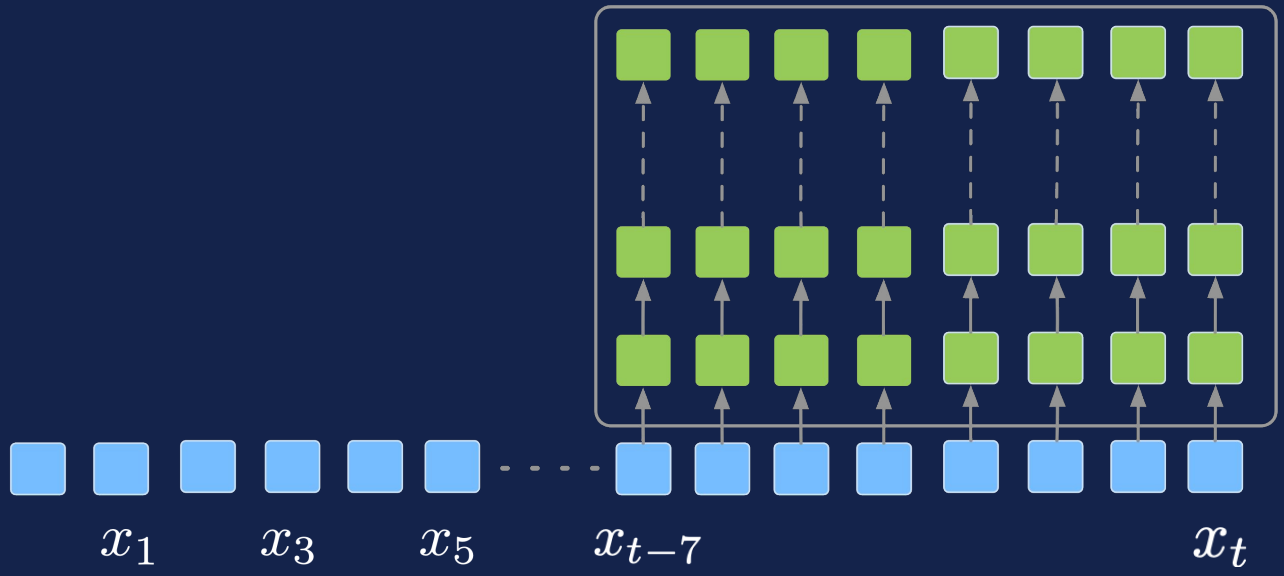
Knowledge is implicitly represented in the weights of a parametric neural network.



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

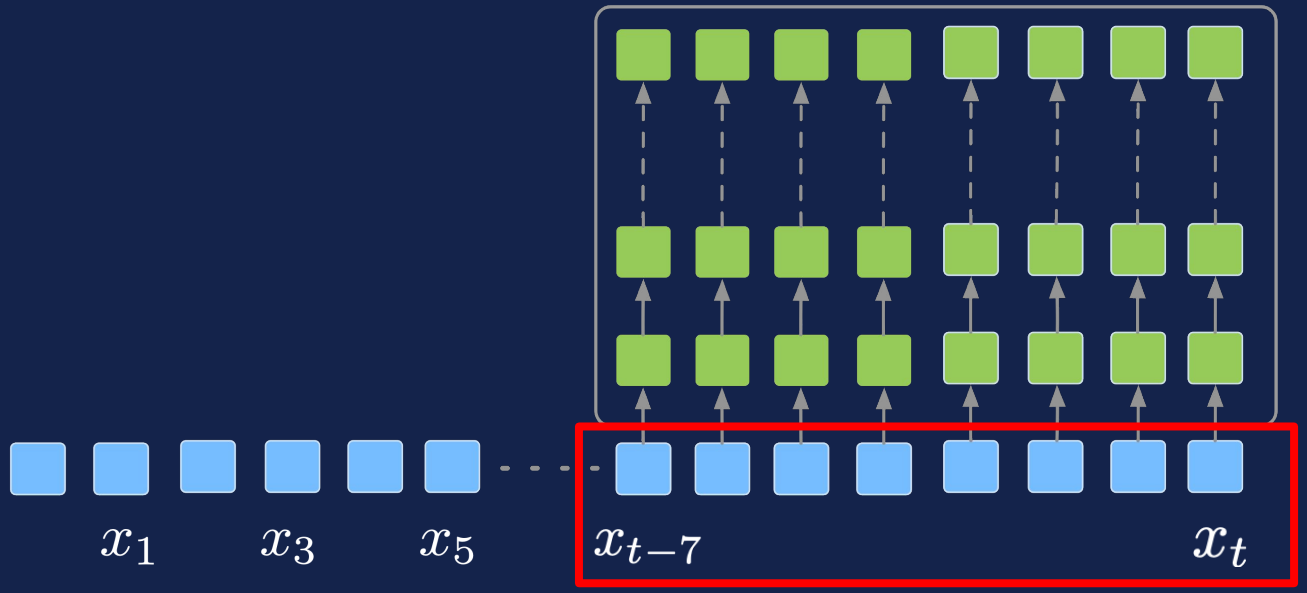
Background

Transformers, no matter how large, are limited by the input sequence length.



Background

Transformers, no matter how large, are limited by the input sequence length.



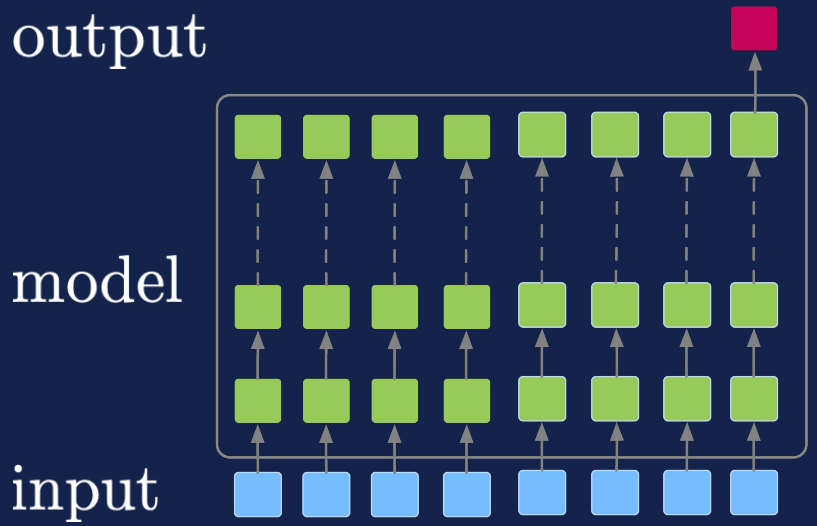
Semiparametric Language Models

Separation of computation and storage as an architectural bias.

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

Semiparametric Language Models

Separation of computation and storage as an architectural bias.

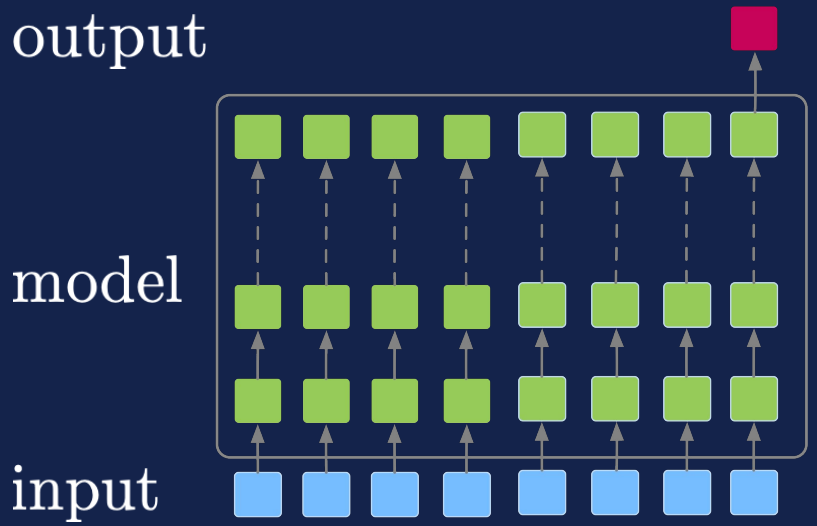


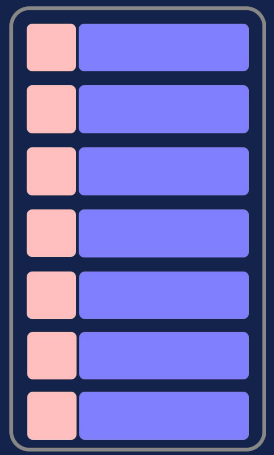
Computation: a neural network

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

Semiparametric Language 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

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 **???**

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 **???**

Current context

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 **???**

Current context

Extended short-term context

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 **???**

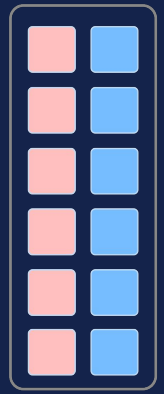
Current context | Extended short-term context | Long-term context

Waterloo Park Wikipedia

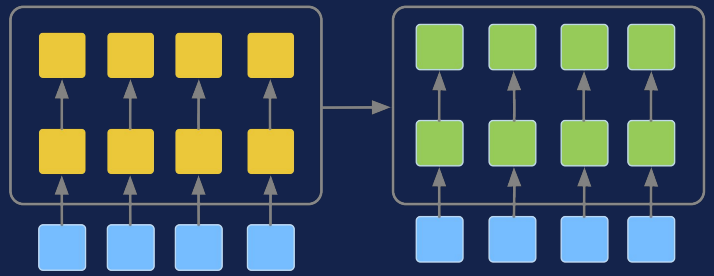
Waterloo Park is an urban park situated in Waterloo, Ontario, Canada.

Language Model

long-term memory

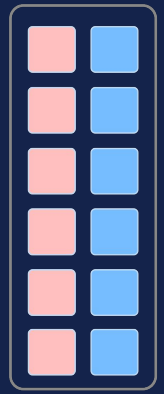


short-term memory

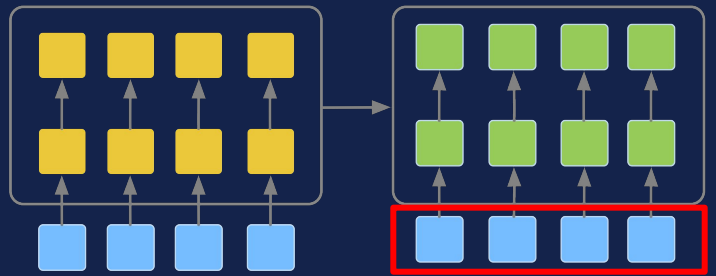


Language Model

long-term memory



short-term memory



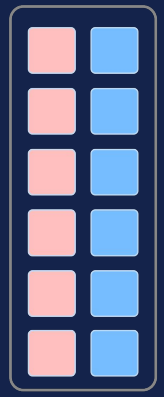
encoder

UWaterloo is a public

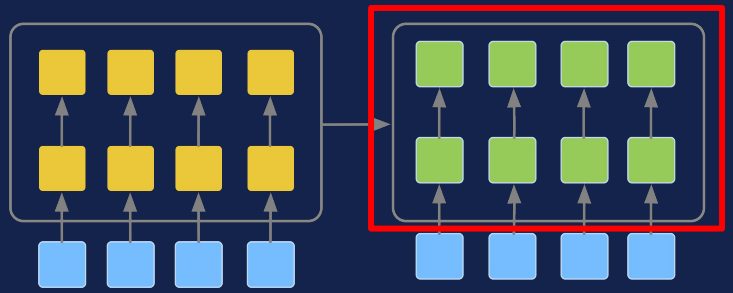
Input: a sequence of context tokens.

Language Model

long-term memory



short-term memory



encoder

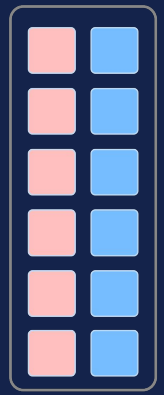
Encoder: transformer

(Vaswani et al., 2017)

UWaterloo is a public

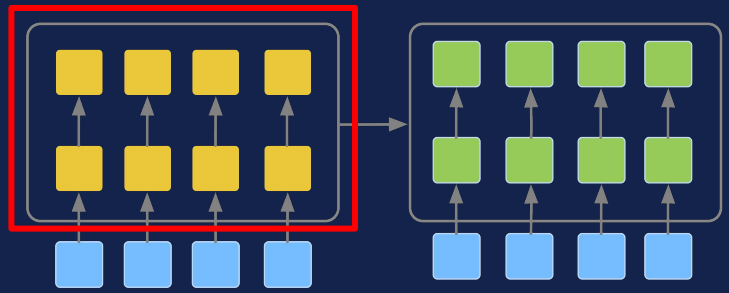
Language Model

long-term memory



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

short-term memory

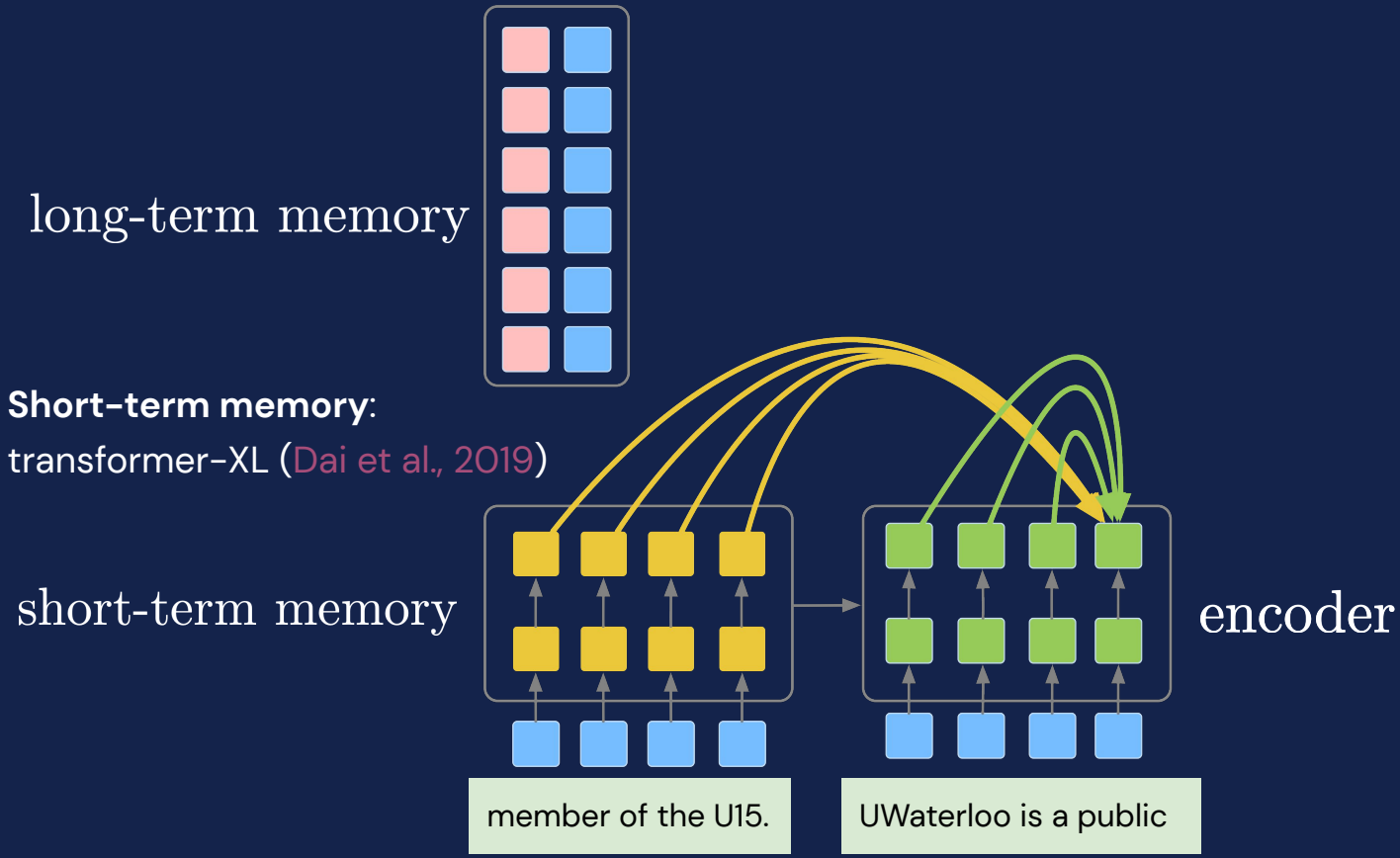


encoder

member of the U15.

UWaterloo is a public

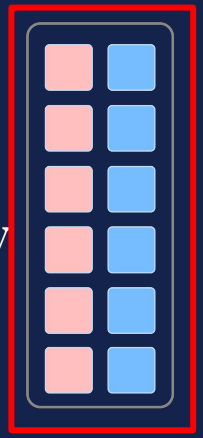
Language Model



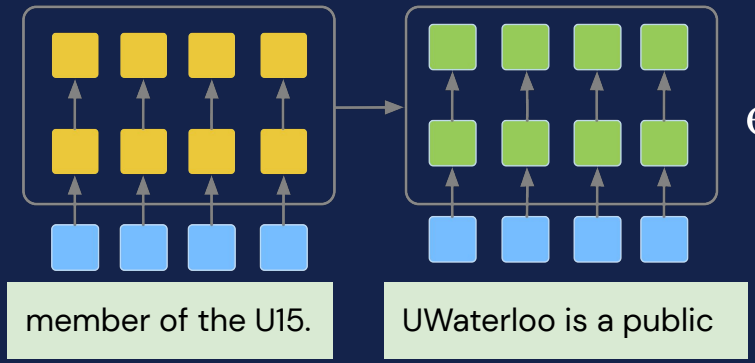
Language Model

Long-term memory:
key-value database

long-term memory



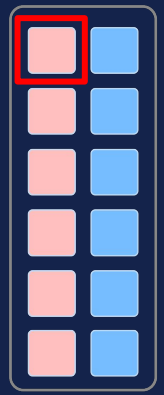
short-term memory



encoder

Language Model

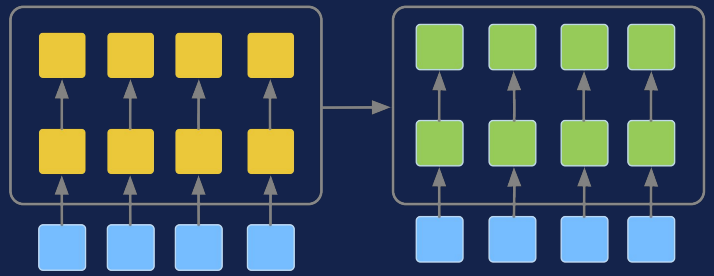
long-term memory



Key: compressed long-term context

Canada is a country in the northern part of North

short-term memory



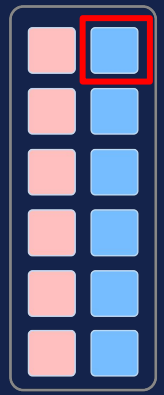
encoder

member of the U15.

UWaterloo is a public

Language Model

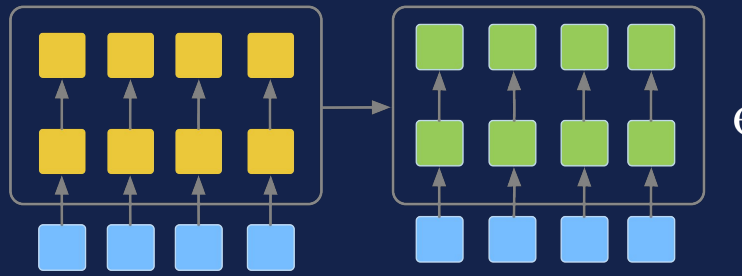
long-term memory



America

Value: output token for the respective context

short-term memory

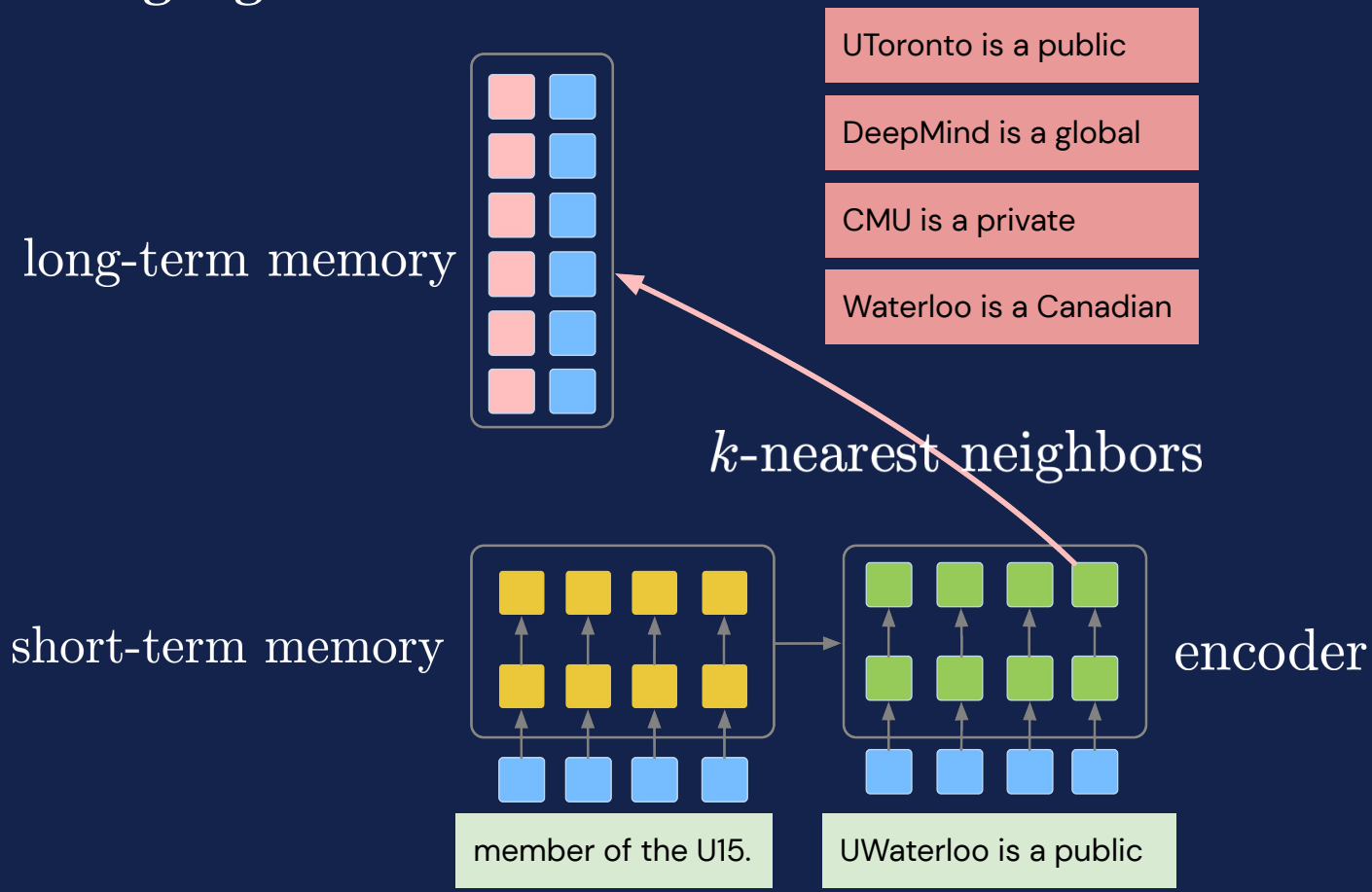


encoder

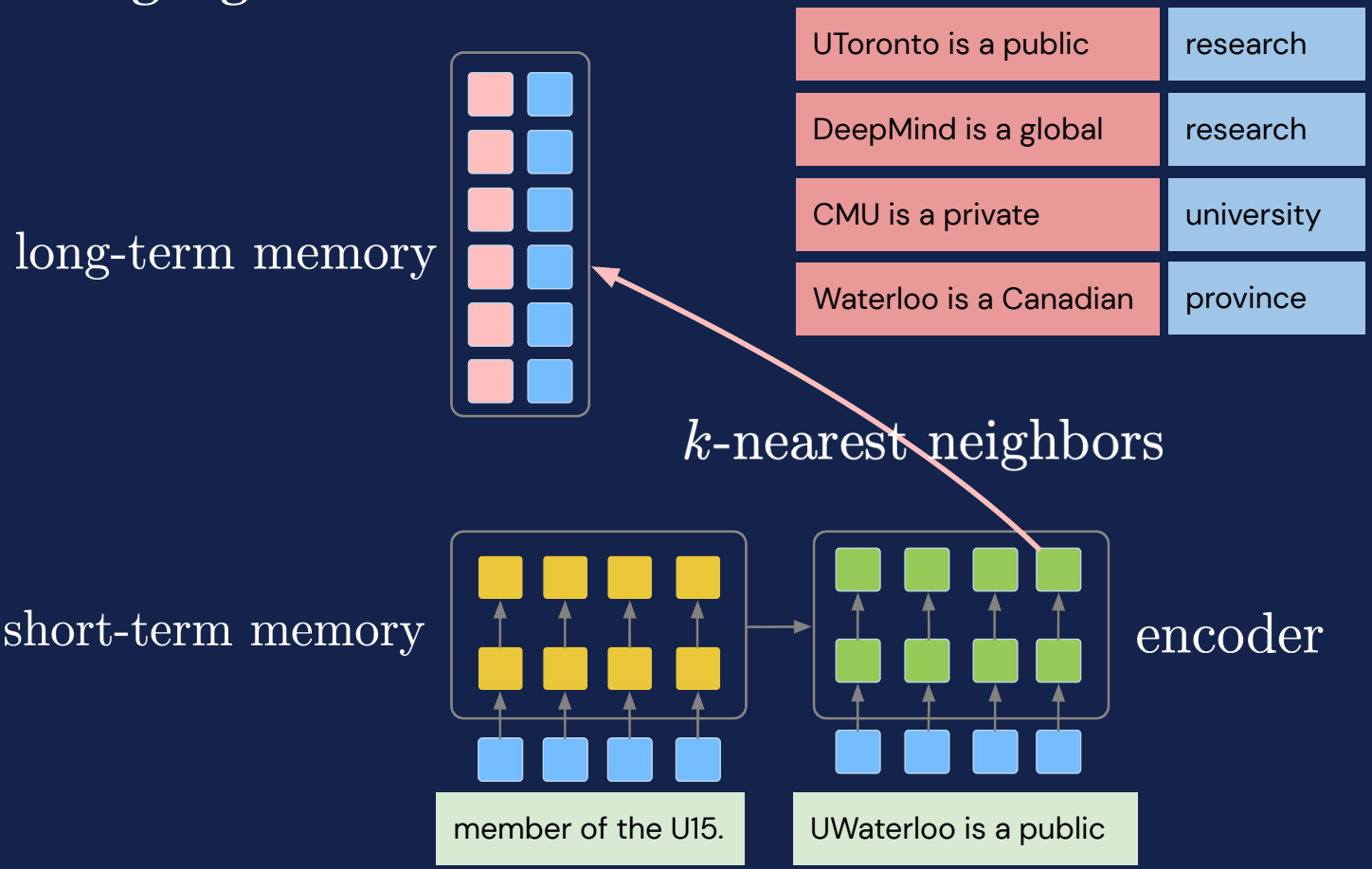
member of the U15.

UWaterloo is a public

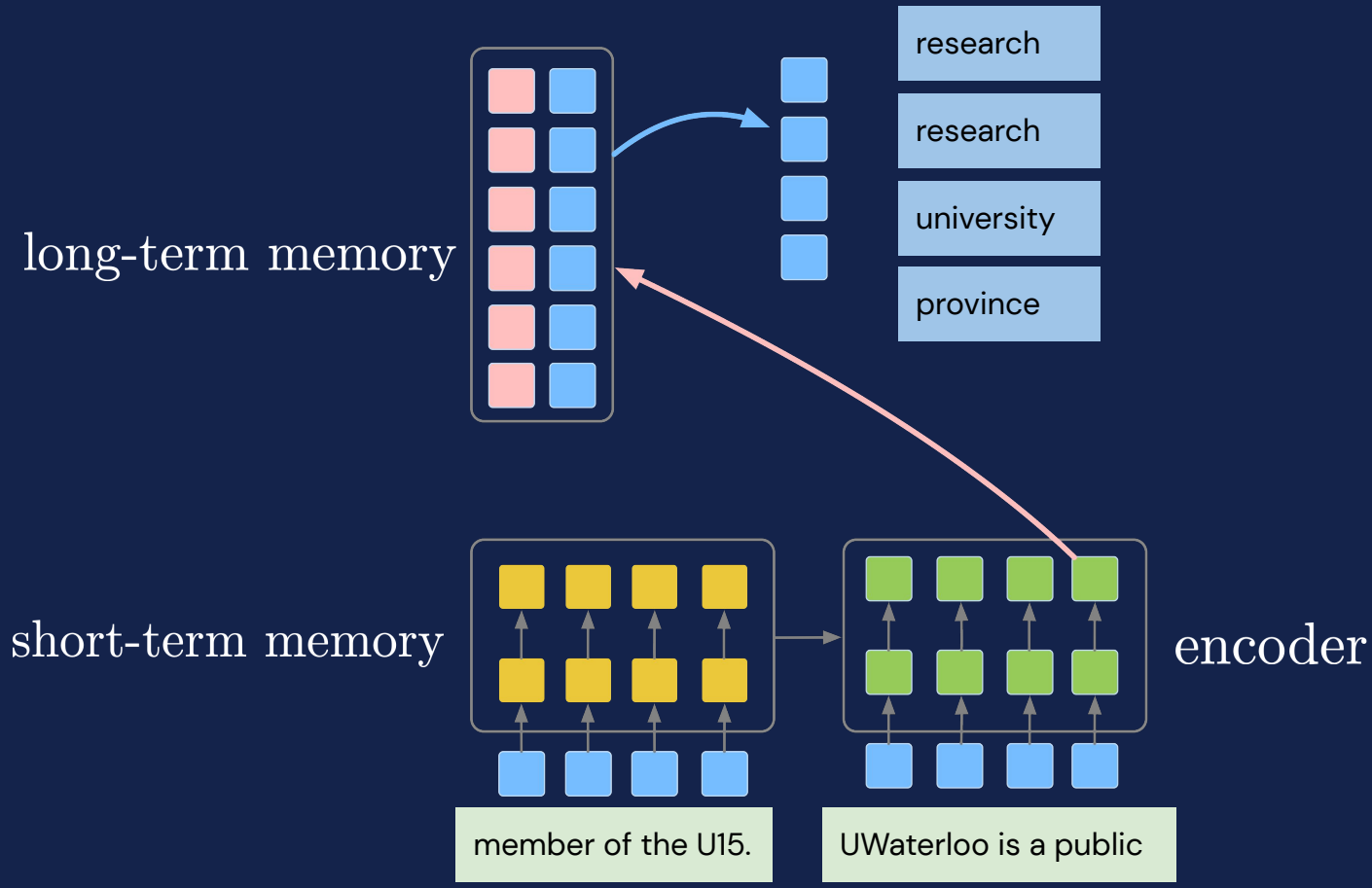
Language Model



Language Model



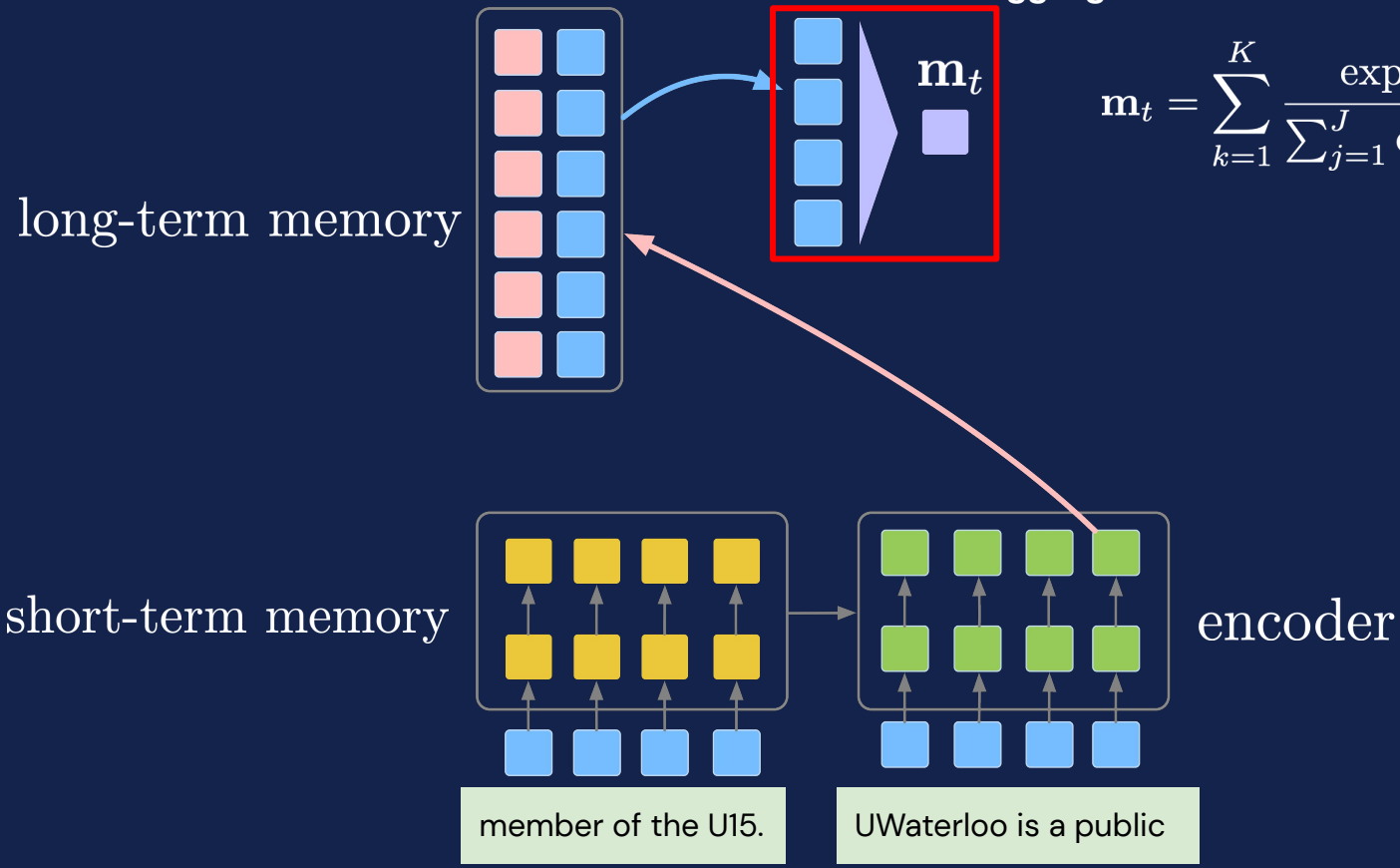
Language Model



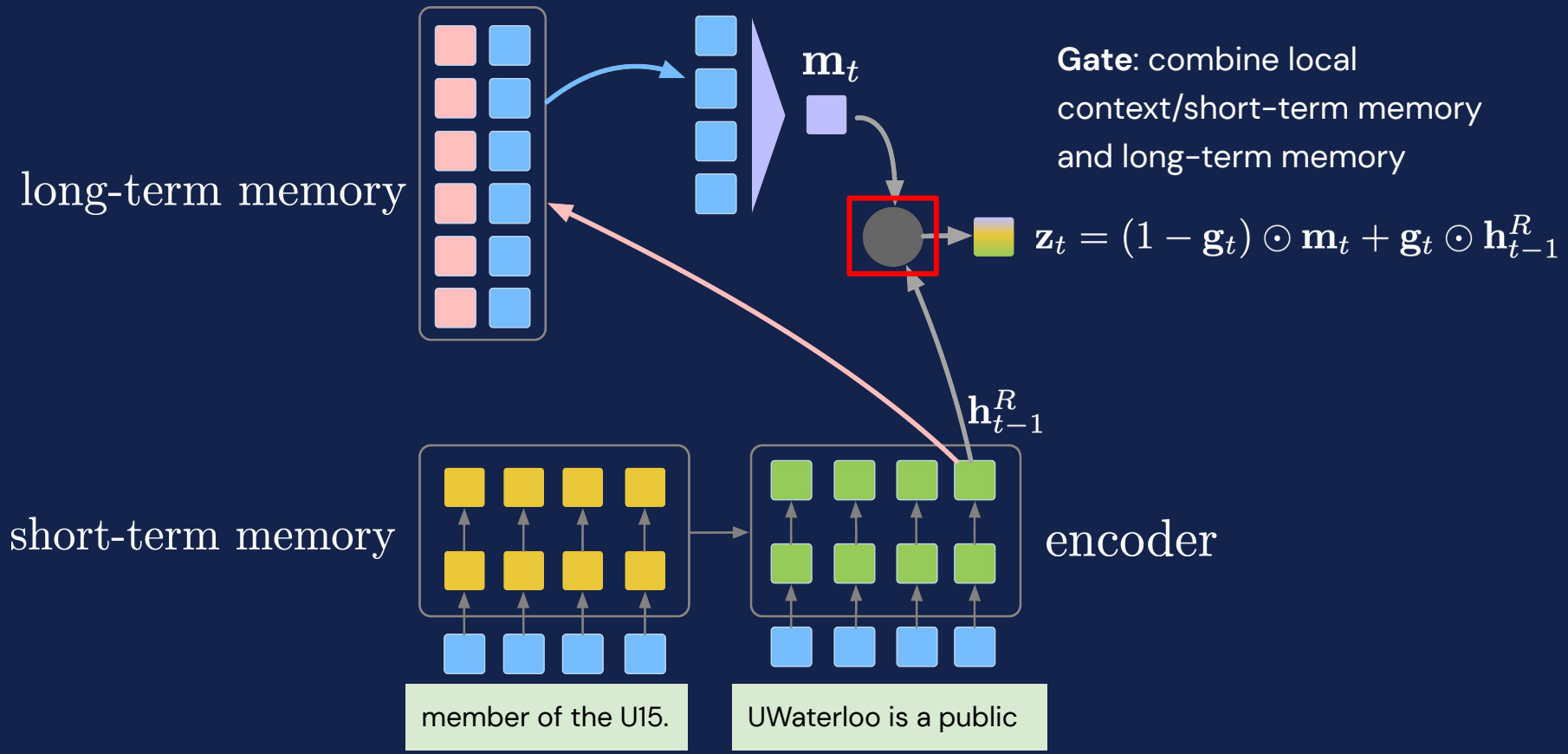
Language Model

Aggregator: attention over retrieved values

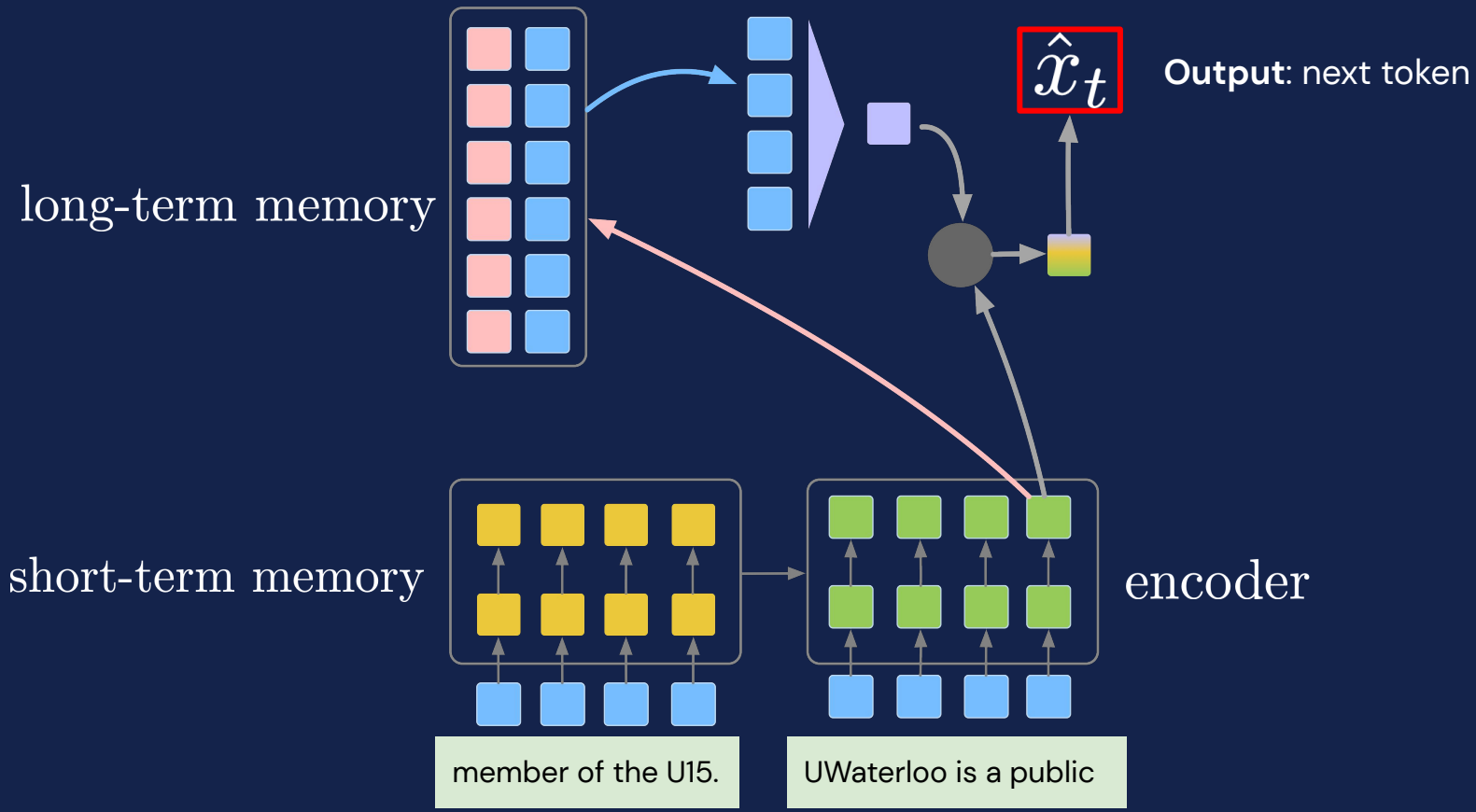
$$\mathbf{m}_t = \sum_{k=1}^{K} \frac{\exp \mathbf{y}_k^\top \mathbf{h}_{t-1}^R}{\sum_{j=1}^{J} \exp \mathbf{y}_j^\top \mathbf{h}_{t-1}^R} \mathbf{y}_k$$



Language Model



Language Model



Experiments

- Word-level language modeling.
  - 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

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	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

Takeaway and Limitation

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

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

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

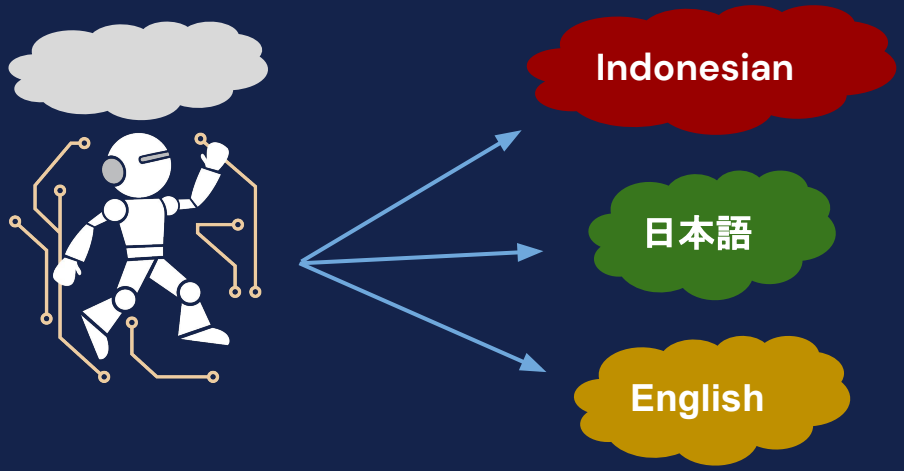
A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

Training Paradigms

Model Architectures

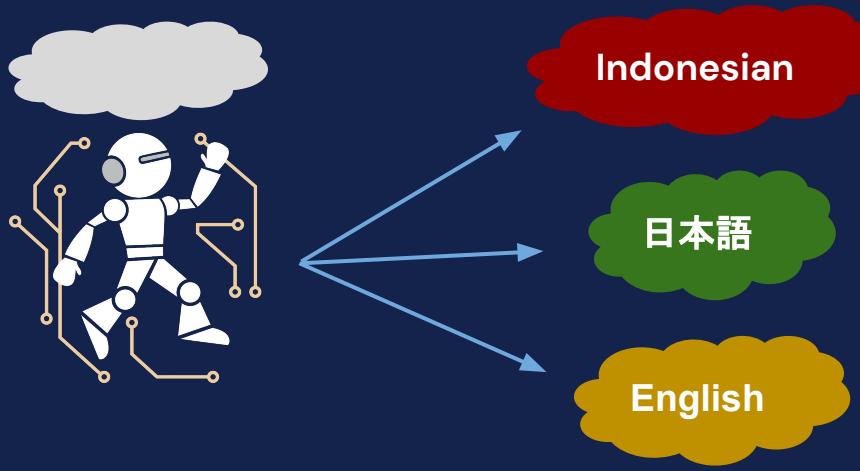
Future Directions

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.

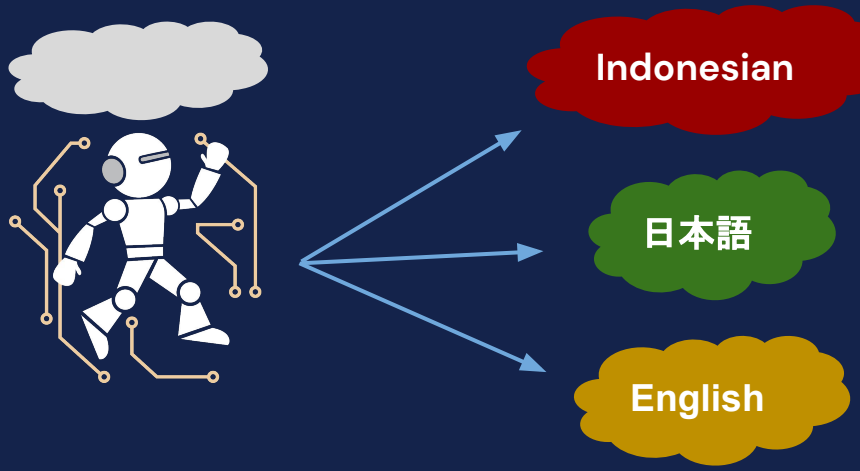


Learning cross-lingual transferable representations

Artetxe et al., ACL 2020

Future Directions

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

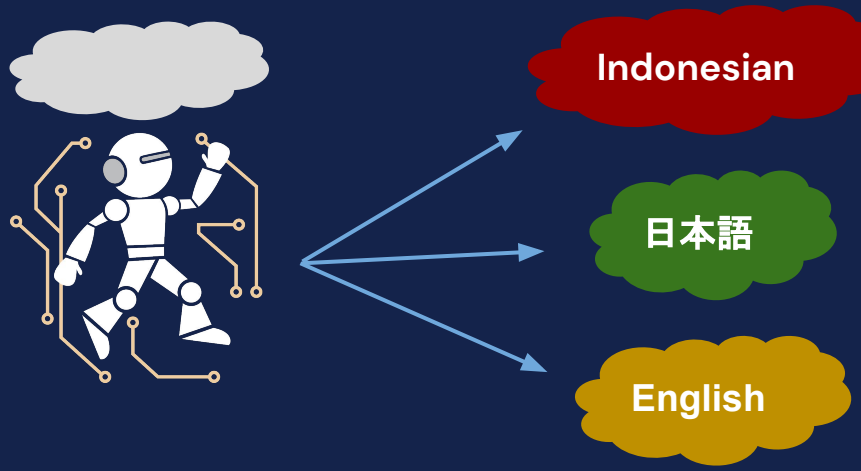


Distributionally Robust Optimization

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

Future Directions

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



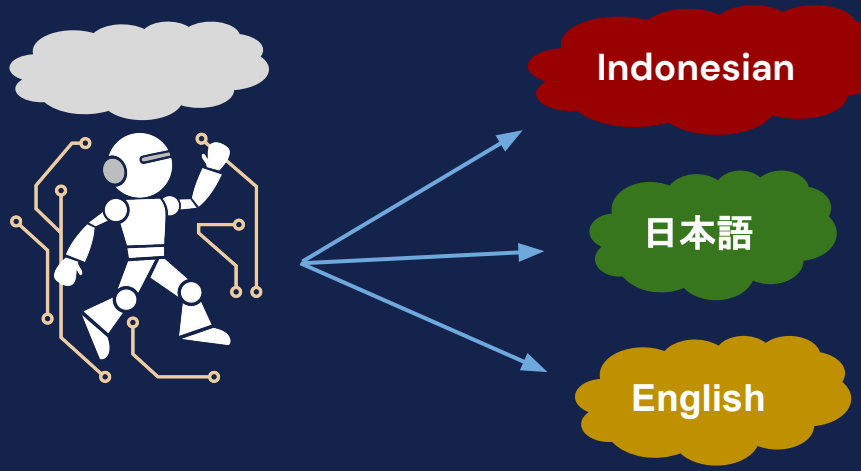
Distributionally Robust Optimization

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

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

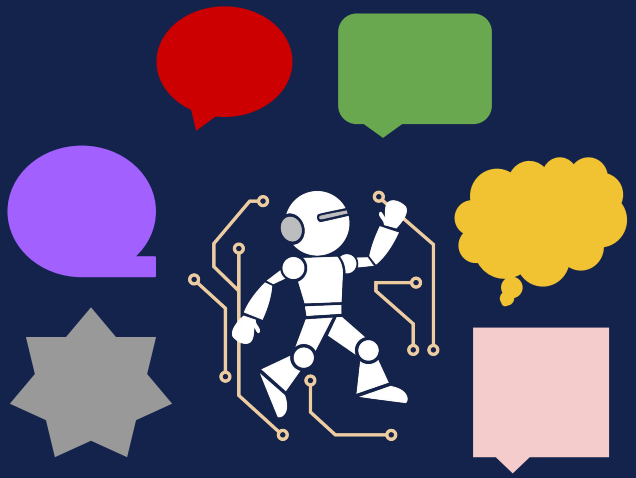
A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

Training Paradigms

Model Architectures

Future Directions

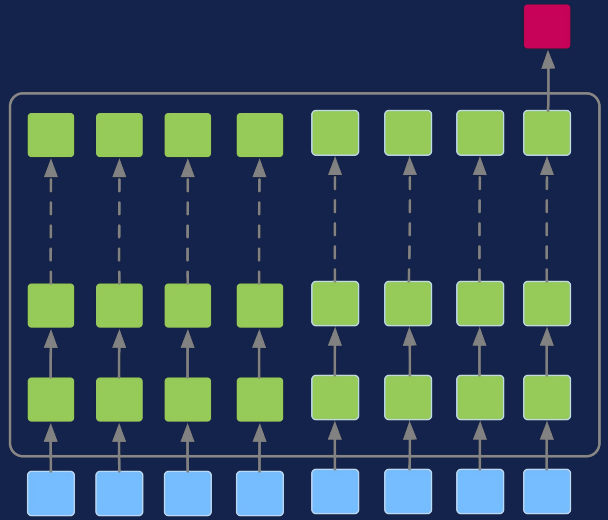
A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.



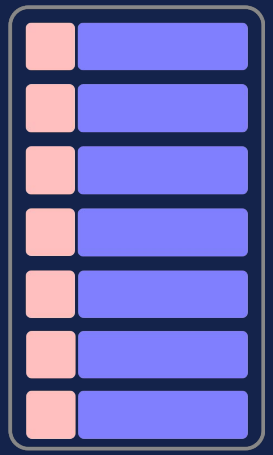
Integration of data from various sources and modalities.

Future Directions

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



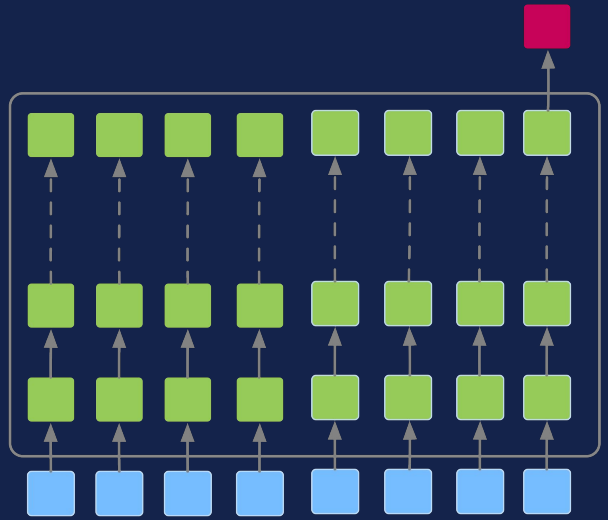
Computation



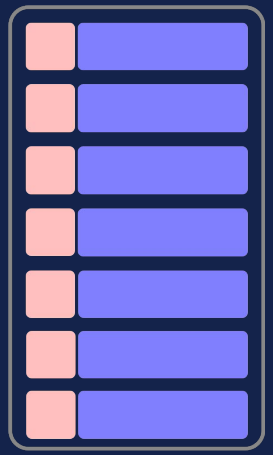
Storage

Future Directions

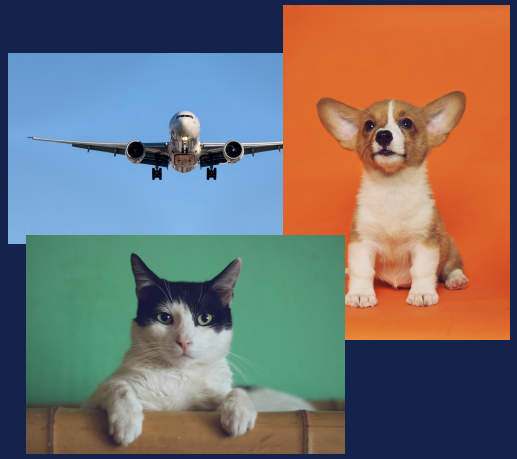
A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.



Computation

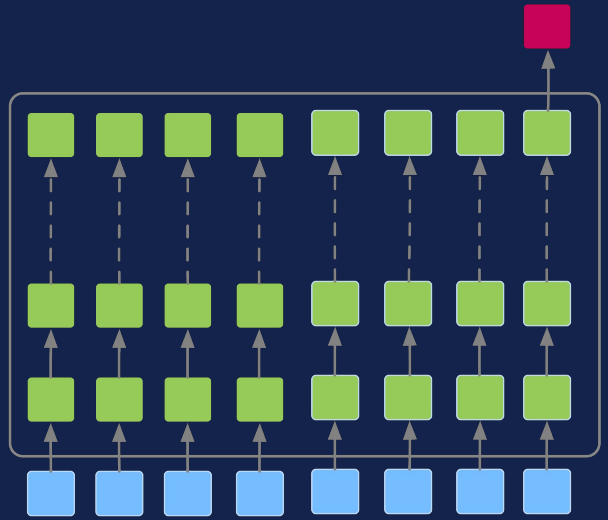


Storage

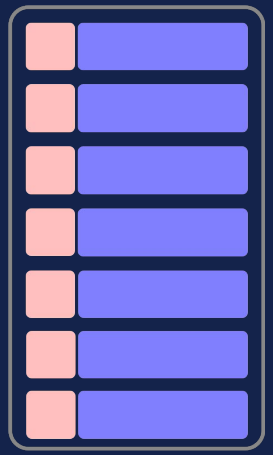


Future Directions

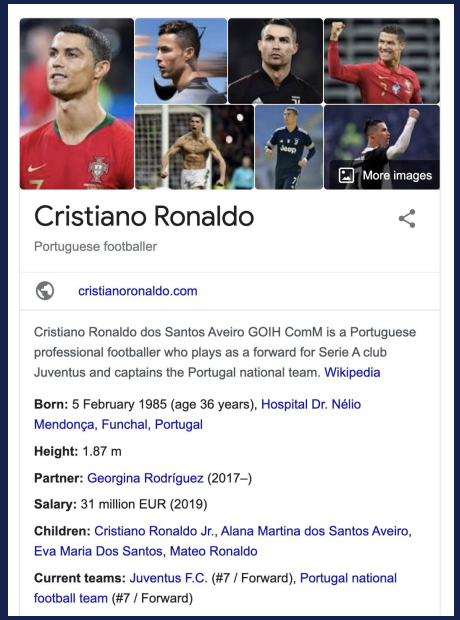
A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.



Computation

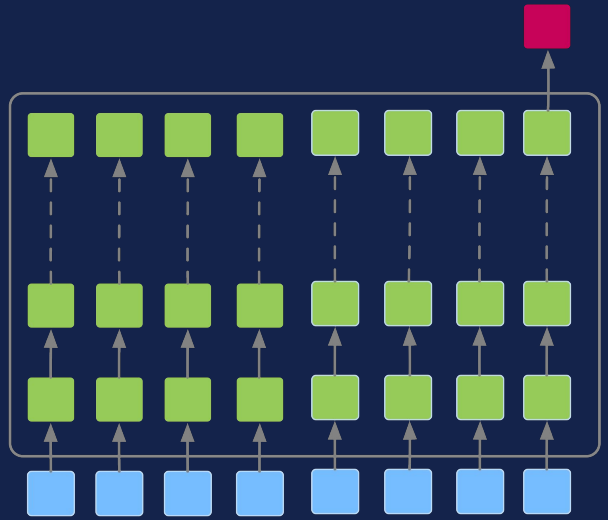


Storage

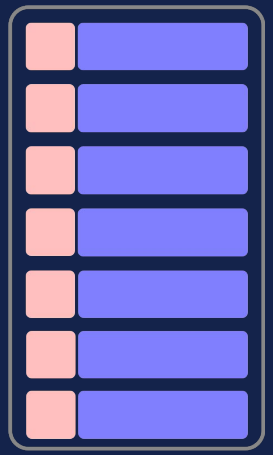


Future Directions

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



Computation

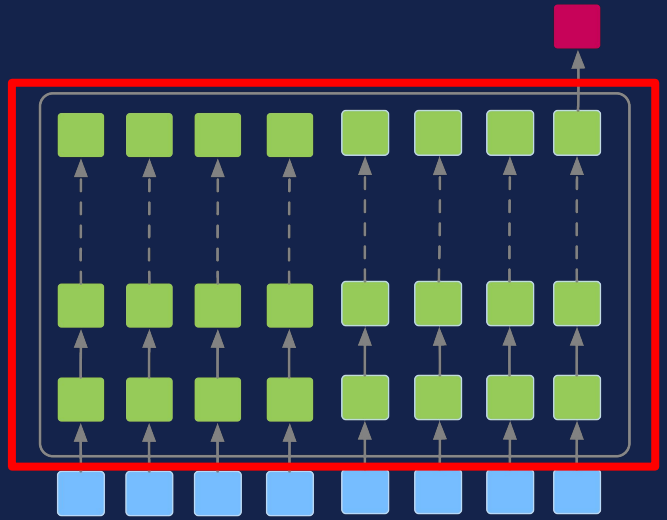


Storage

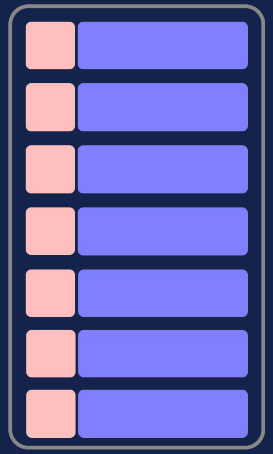
computational efficiency

Future Directions

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



Computation

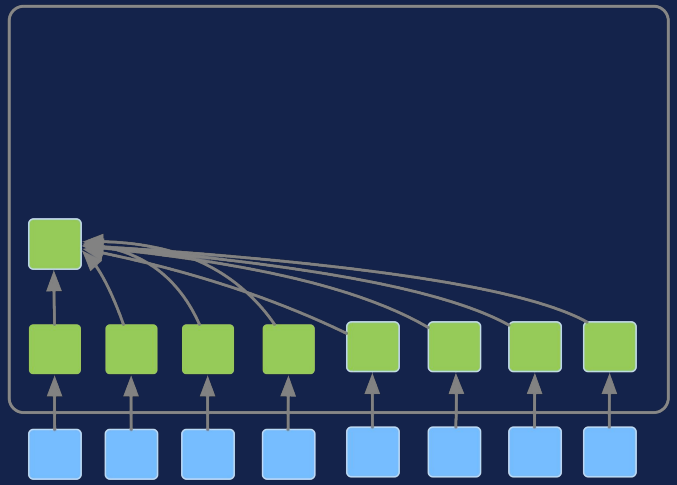


Storage

computational efficiency

Future Directions

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

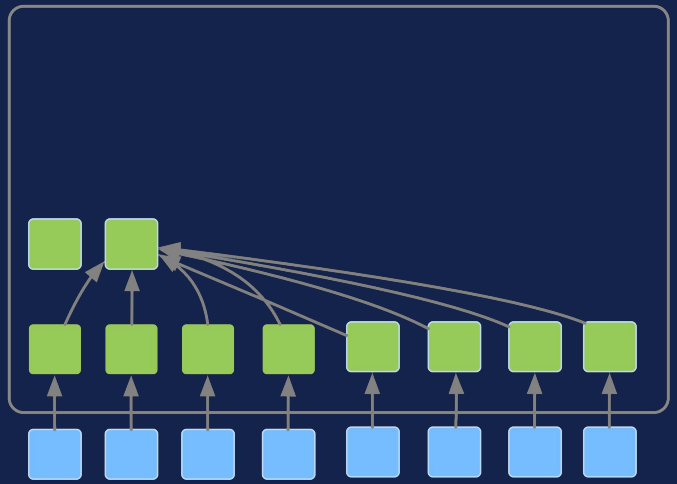


Random Feature Attention

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.

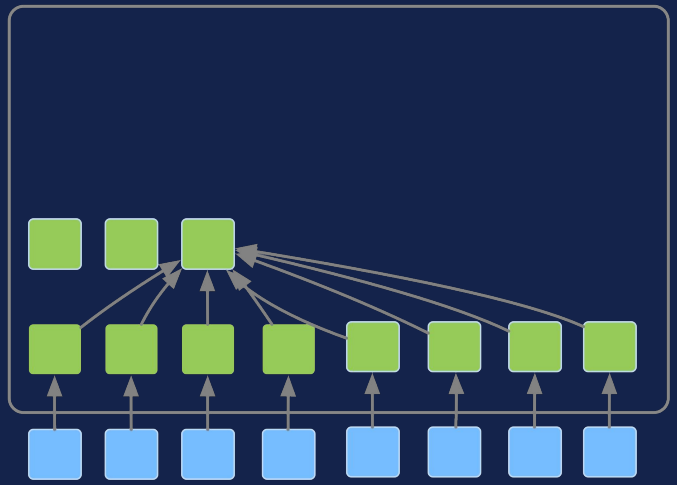


Random Feature Attention

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.

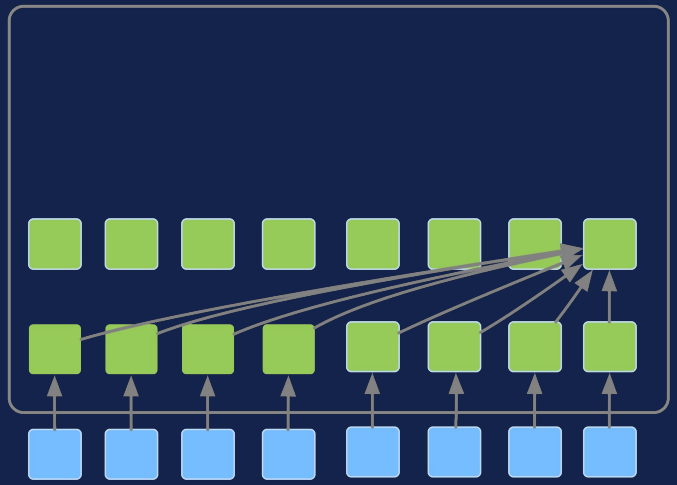


Random Feature Attention

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.

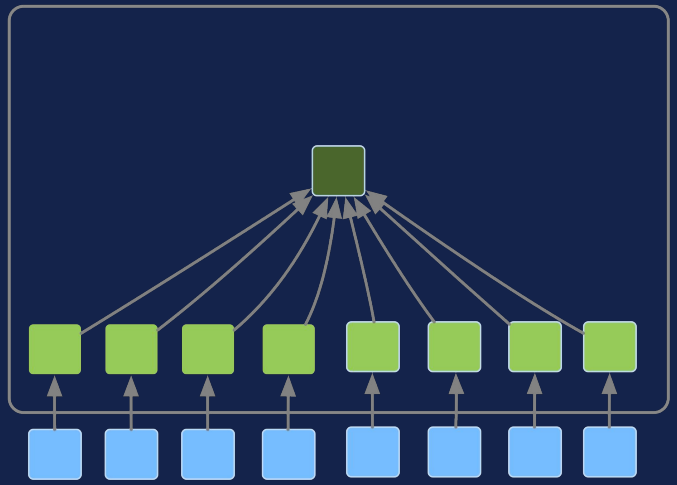


Random Feature Attention

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.

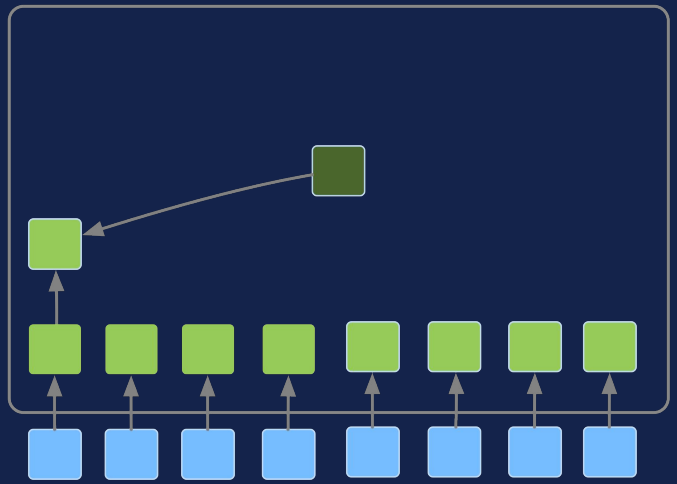


Random Feature Attention

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.

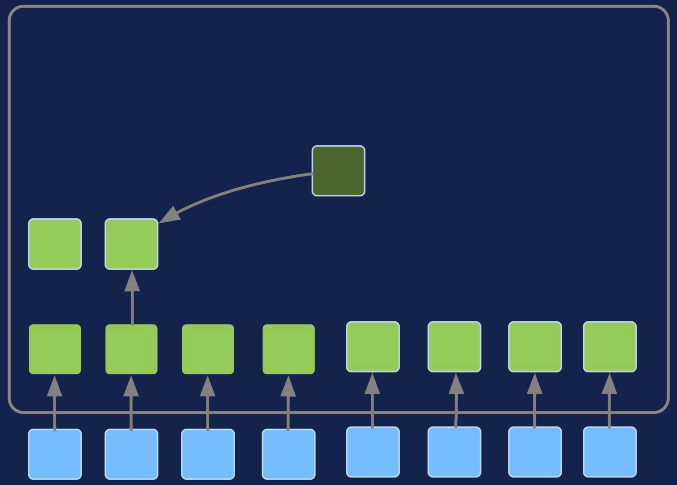


Random Feature Attention

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.

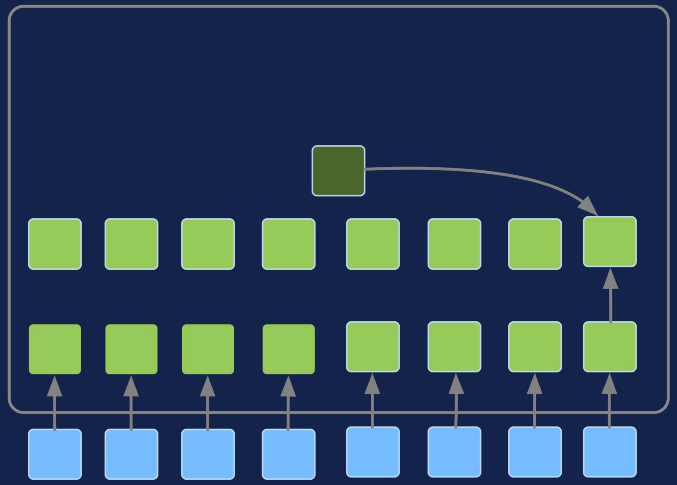


Random Feature Attention

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.

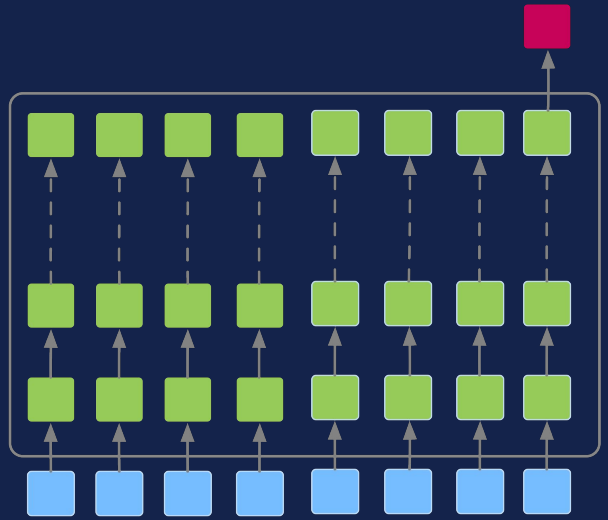


Random Feature Attention

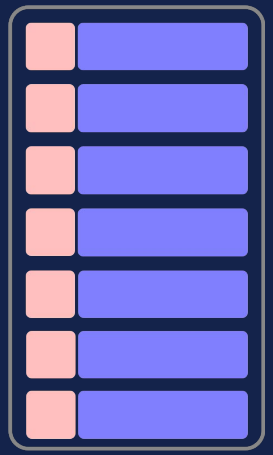
Peng et al., ICLR 2021

Future Directions

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



Computation

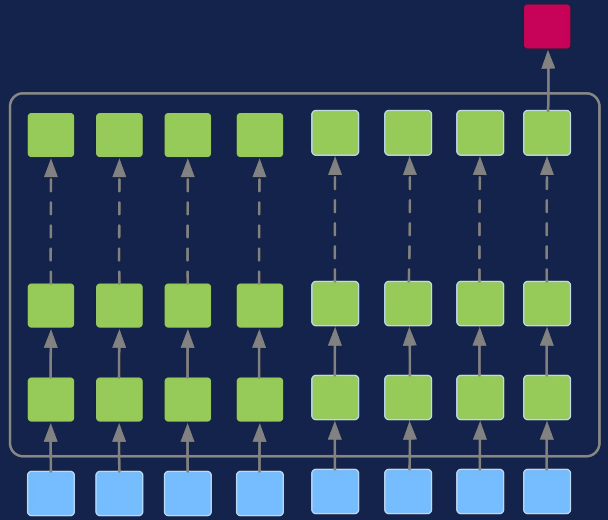


Storage

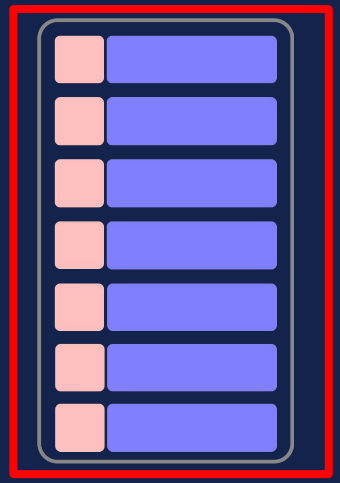
computational efficiency

Future Directions

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



Computation

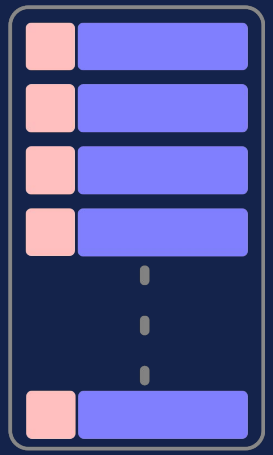


Storage

computational efficiency

Future Directions

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



Storage

Learning to what to remember and forget

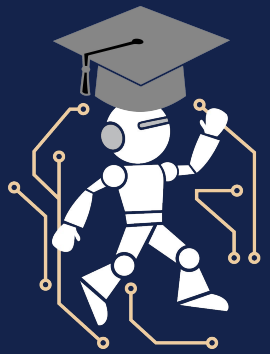
Constant-size memory

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.



https://dyogatama.github.io
dyogatama@google.com

DeepMind

Challenges: Human Learning vs. Machine Learning

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

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

This Talk

- Episodic memory in lifelong language learning.

  de Masson d'Autume et al., NeurIPS 2019

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

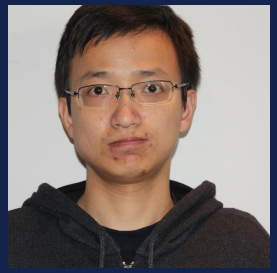
  Kong et al., ICLR 2020

Episodic Memory in Lifelong Language Learning

de Masson d'Autume et al., NeurIPS 2019

Cyprien

Sebastian



Lingpeng

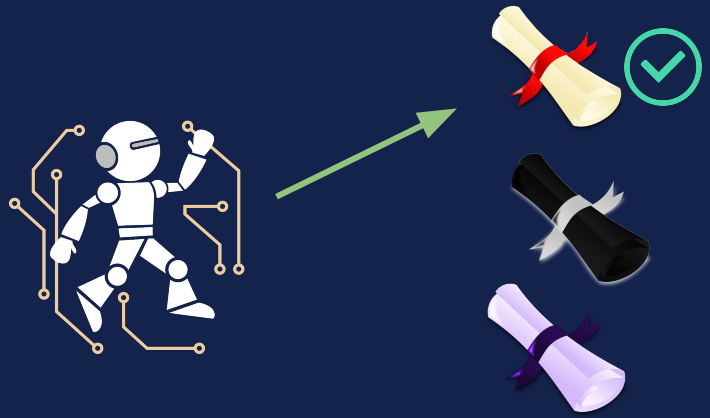
Dani

Background

- A model should be able to reuse knowledge from related tasks to learn a new task faster.
- Current models not only fail to do this, they **catastrophically forget** previously learned tasks (McCloseky and Cohen, 1989; Ratcliff, 1990).

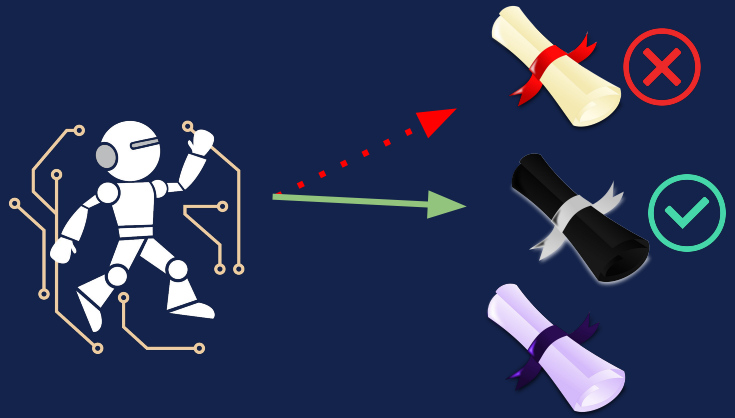
Background

- A model should be able to reuse knowledge from related tasks to learn a new task faster.
- Current models not only fail to do this, they **catastrophically forget** previously learned tasks (McCloseky and Cohen, 1989; Ratcliff, 1990).



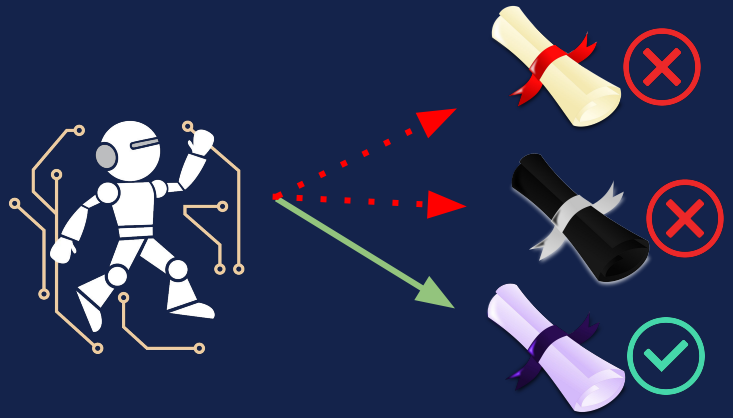
Background

- A model should be able to reuse knowledge from related tasks to learn a new task faster.
- Current models not only fail to do this, they **catastrophically forget** previously learned tasks (McCloseky and Cohen, 1989; Ratcliff, 1990).



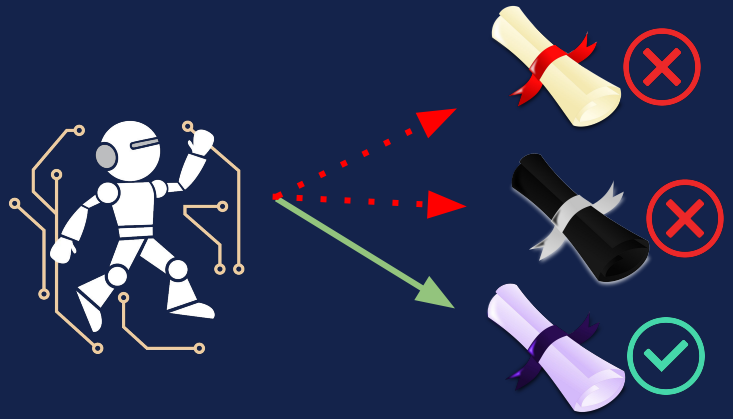
Background

- A model should be able to reuse knowledge from related tasks to learn a new task faster.
- Current models not only fail to do this, they **catastrophically forget** previously learned tasks (McCloseky and Cohen, 1989; Ratcliff, 1990).



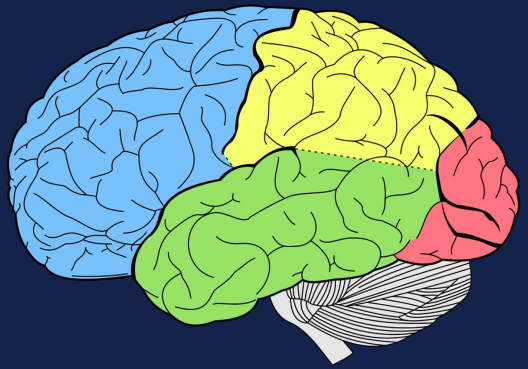
Background

- A model should be able to reuse knowledge from related tasks to learn a new task faster.
- Current models not only fail to do this, they **catastrophically forget** previously learned tasks (McCloseky and Cohen, 1989; Ratcliff, 1990).



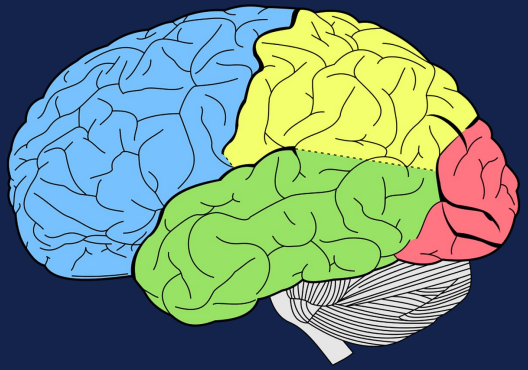
Hypothesis: episodic memory mitigates catastrophic forgetting in language learning.

On Memory-Augmented Neural Networks



Episodic memory is a type of long-term memory of **events** and **experiences**. It is often associated with a module that stores training examples in neural networks..

On Memory-Augmented Neural Networks



Episodic memory is a type of long-term memory of **events** and **experiences**. It is often associated with a module that stores training examples in neural networks..

Contrast this with short-term **working memory** in LSTMs (Hochreiter and Schmidhuber, 1997) and DNCs (Graves et al., 2016) that e.g., remembers context.

See **Yogatama et al., ICLR 2018** for comparisons of working memory models for language models.

Problem Setup

Training

TriviaQA: Joshi et al., 2017

Tanker leaks 6,000 tons of oil after running aground

The Independent, Friday 16 February 1996

A massive anti-pollution operation was underway last night after a 147,000-ton super tanker ran aground off Milford Haven, West Wales. [...]

Which super-tanker ran aground near Milford Haven in 1996?

Problem Setup

Training

SQuAD: Rajpurkar et al., 2016

Computational Complexity Theory. Computational complexity theory is a branch of the theory of computation in theoretical computer science that focuses on classifying computational problems according to their inherent difficulty [...]

What branch of theoretical computer science deals with broadly classifying computational problems by difficulty and class of relationship?

Problem Setup

Training

QuAC: Choi et al., 2018

Augusto Pinochet --- Intellectual life ...
Pinochet was publicly known as a man with a lack of culture. This image was reinforced by the fact [...]

Was he known for being intelligent? No, Pinochet was publicly known as a man with a lack of culture.

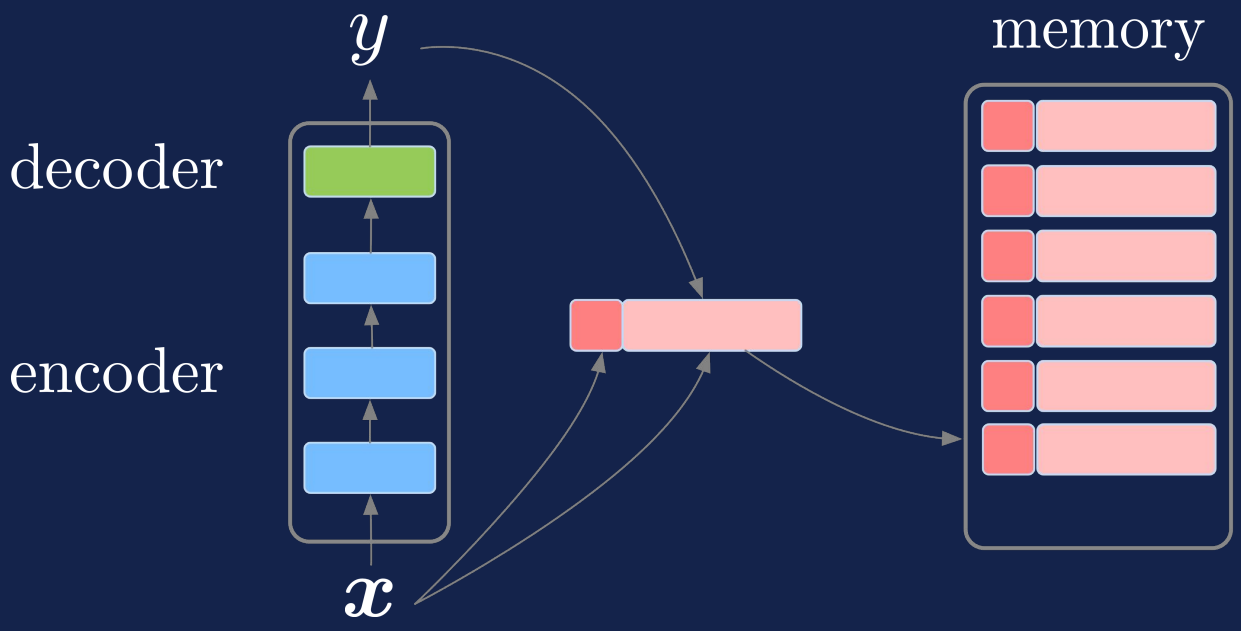
Why did people feel that way?

Problem Setup

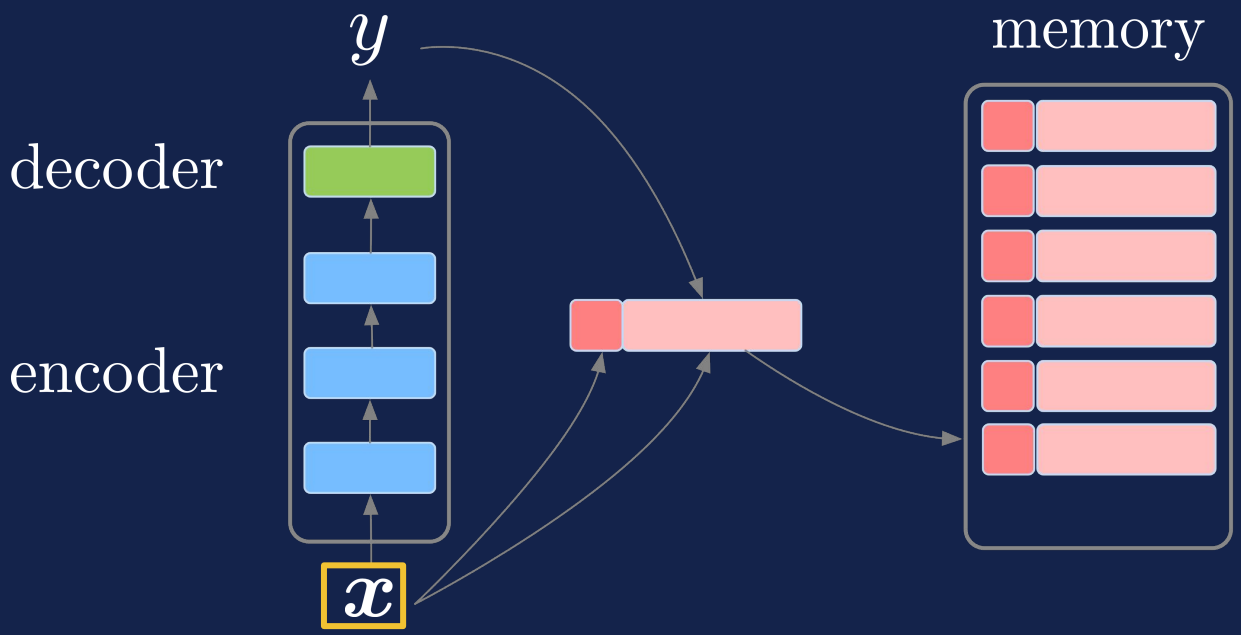
Training

Test

Question Answering Model

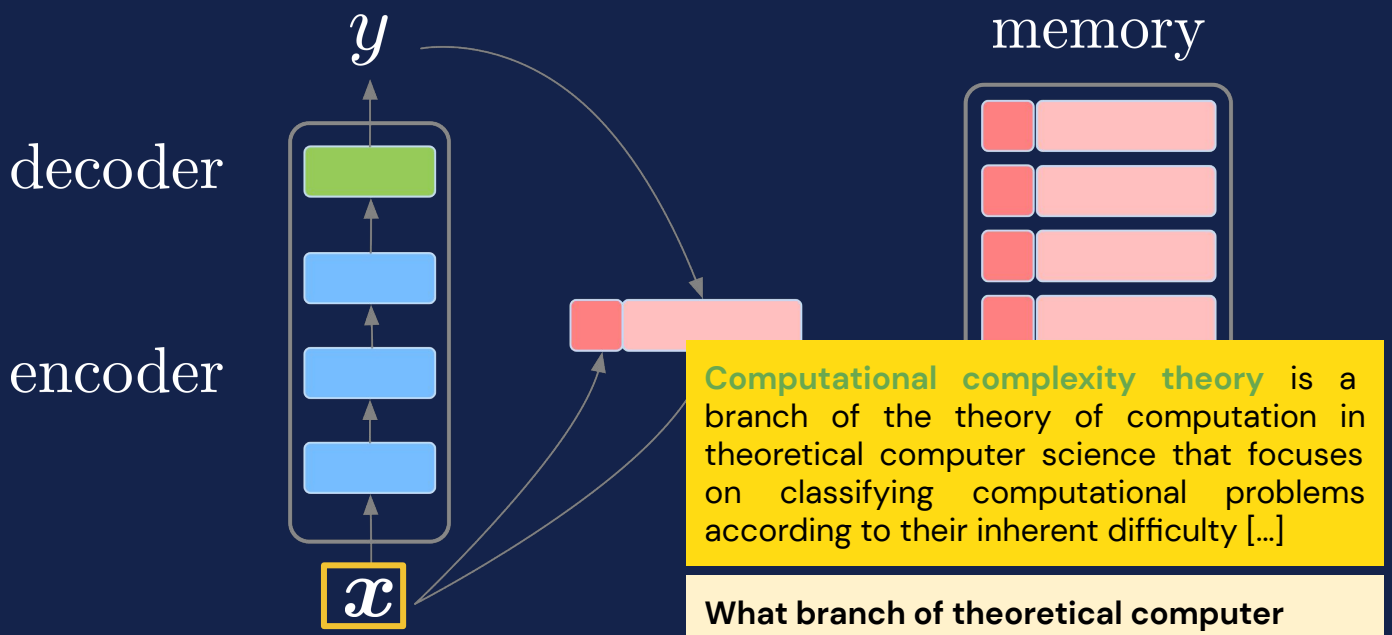


Question Answering Model



Input: a concatenation of a context (e.g., a Wikipedia article) and a question.

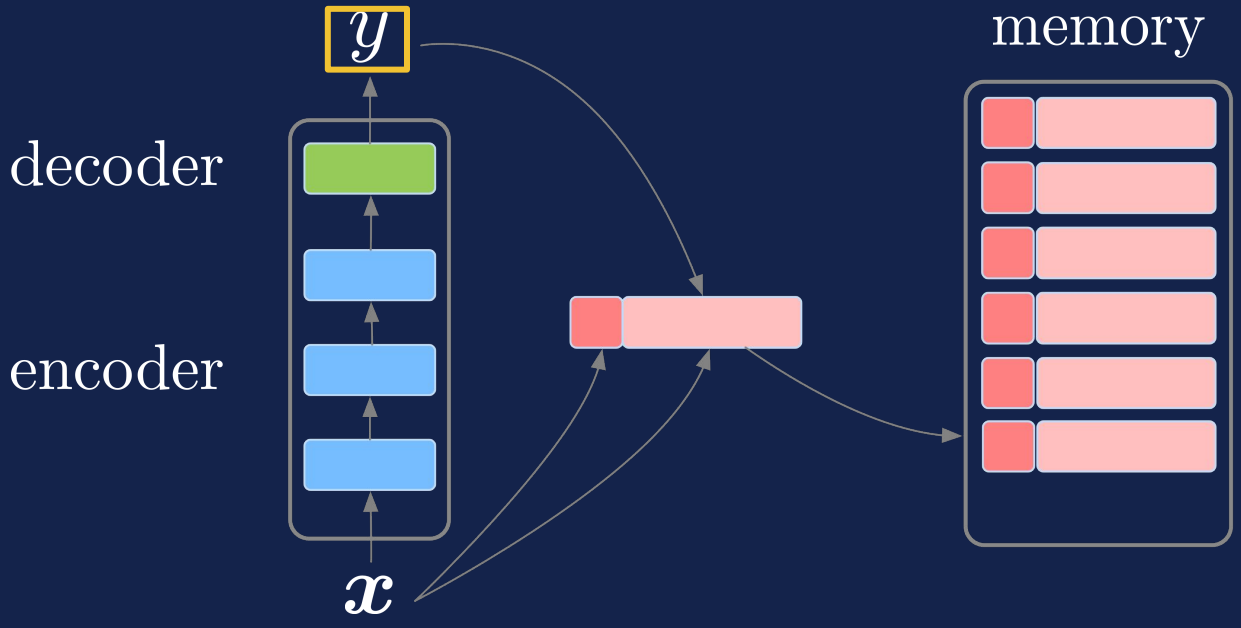
Question Answering Model



Input: a concatenation of a context (e.g., a

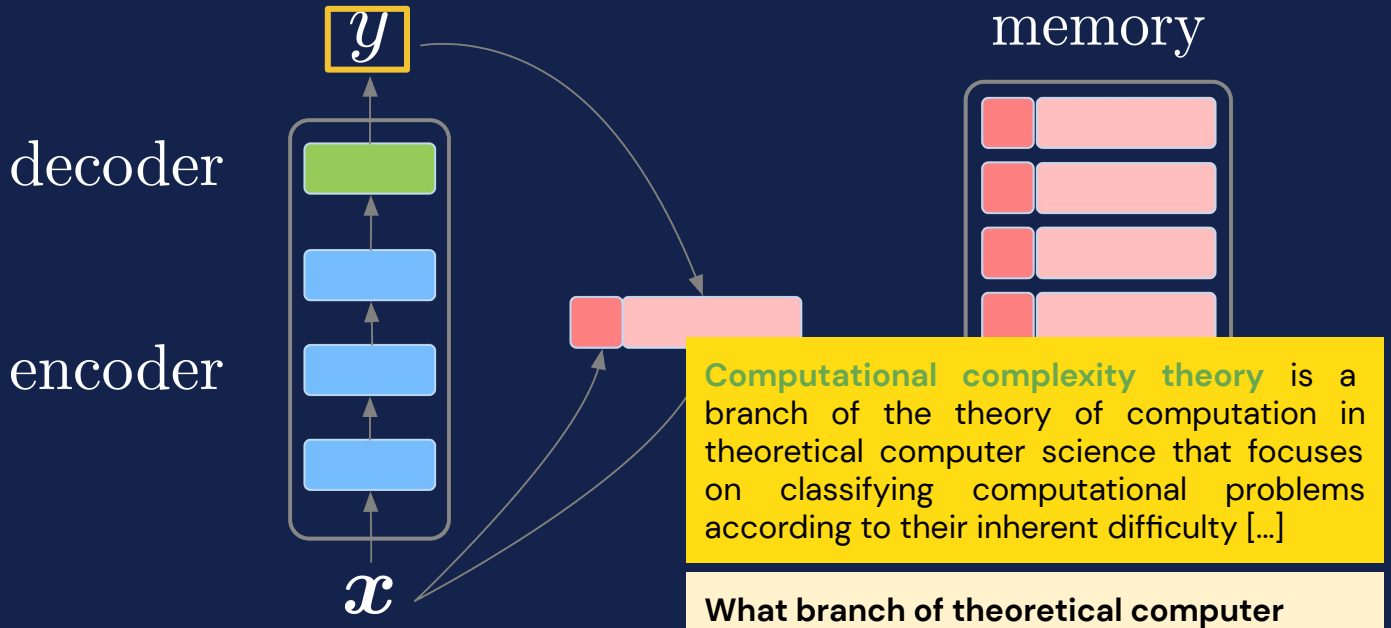
Question Answering Model

Output: an answer, predicted as start and end indices of the answer in the context.



Question Answering Model

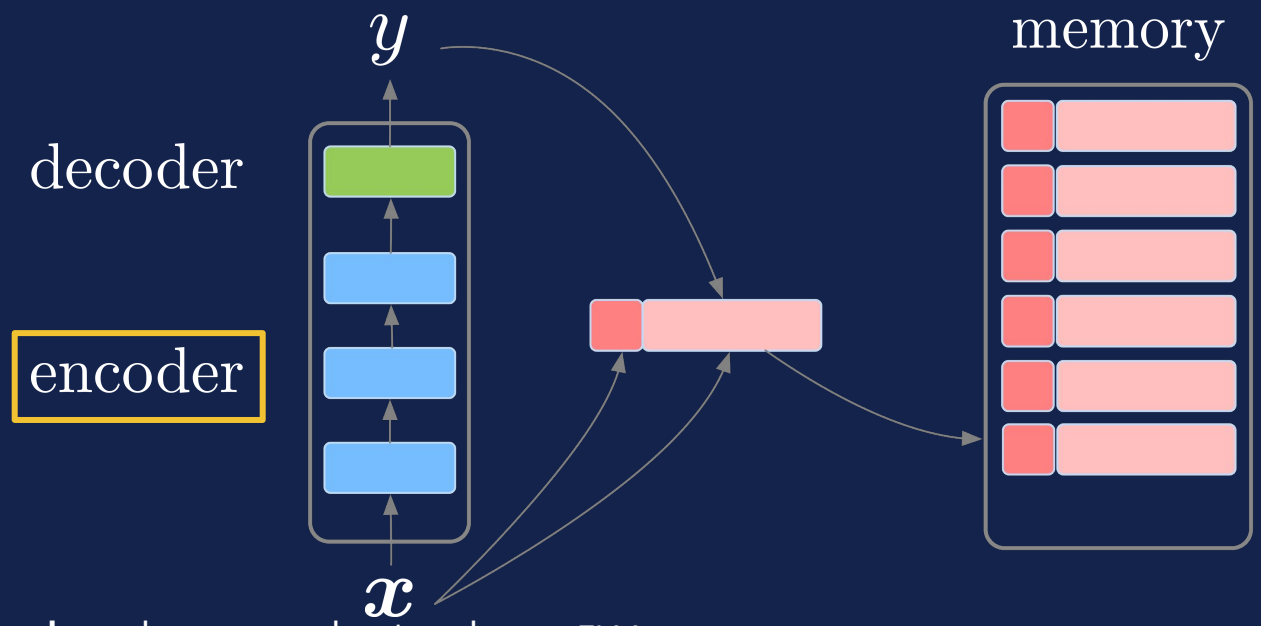
Output: an answer, predicted as start and end indices of the answer in the context.



Computational complexity theory is a branch of the theory of computation in theoretical computer science that focuses on classifying computational problems according to their inherent difficulty [...]

What branch of theoretical computer science deals with broadly classifying computational problems by difficulty and class of relationship?

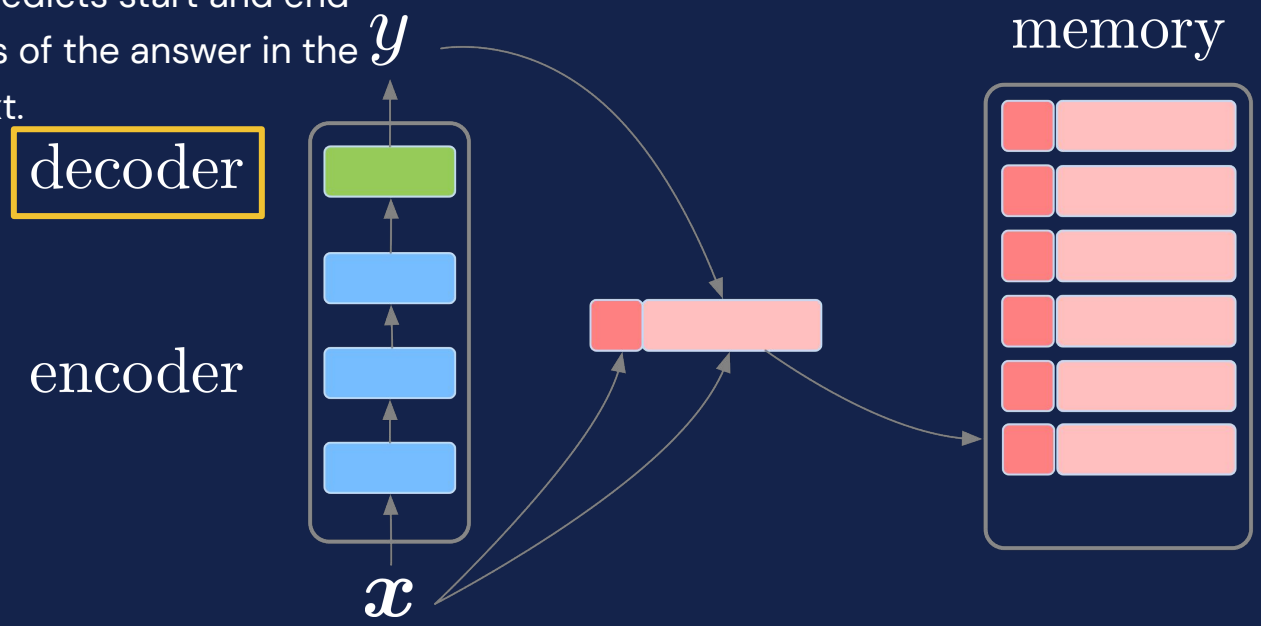
Question Answering Model



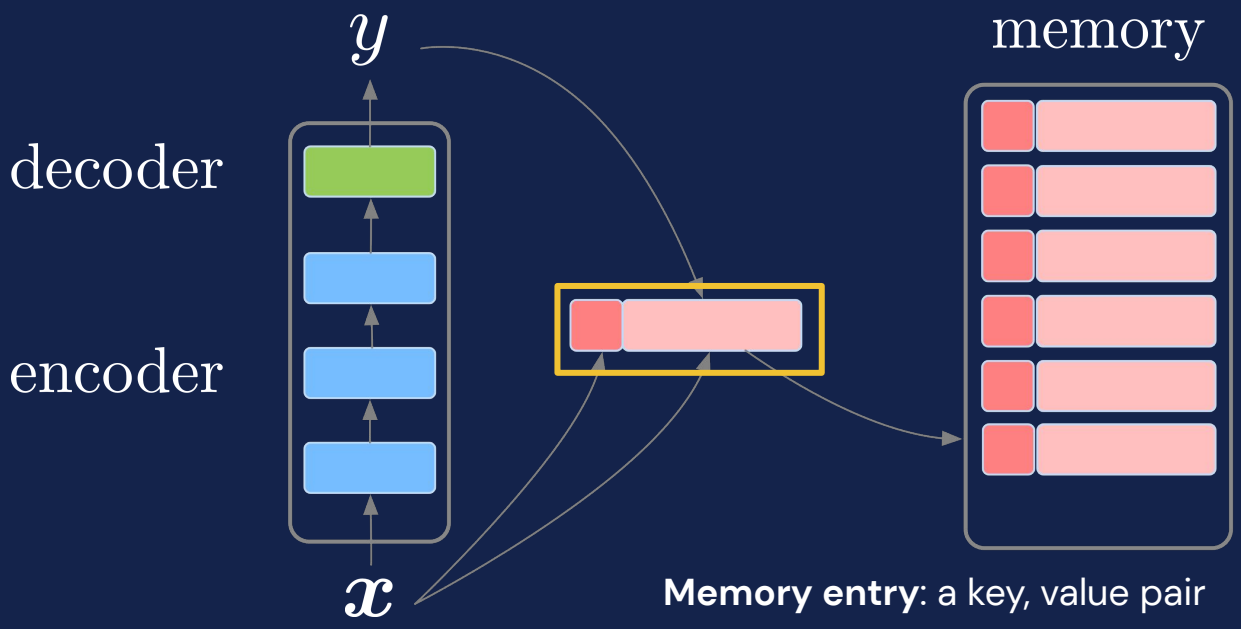
Encoder: a large neural network, e.g., ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), XLNet (Yang et al., 2019).

Question Answering Model

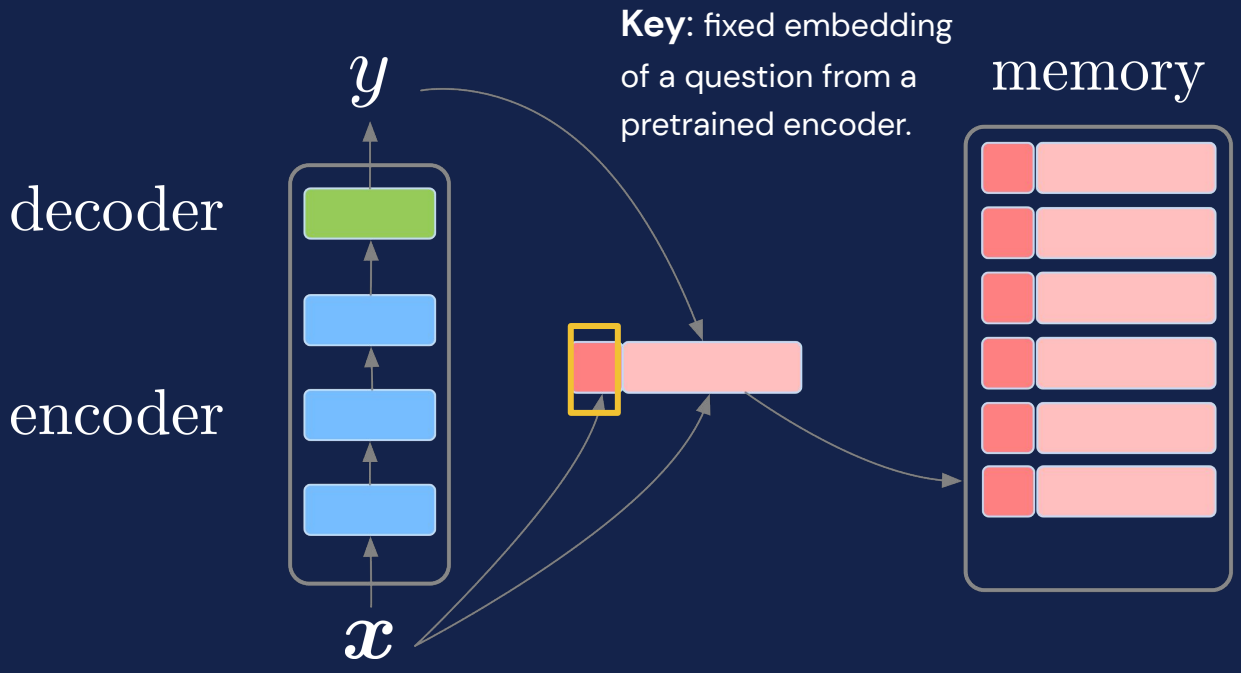
Decoder: a linear function that predicts start and end indices of the answer in the context.



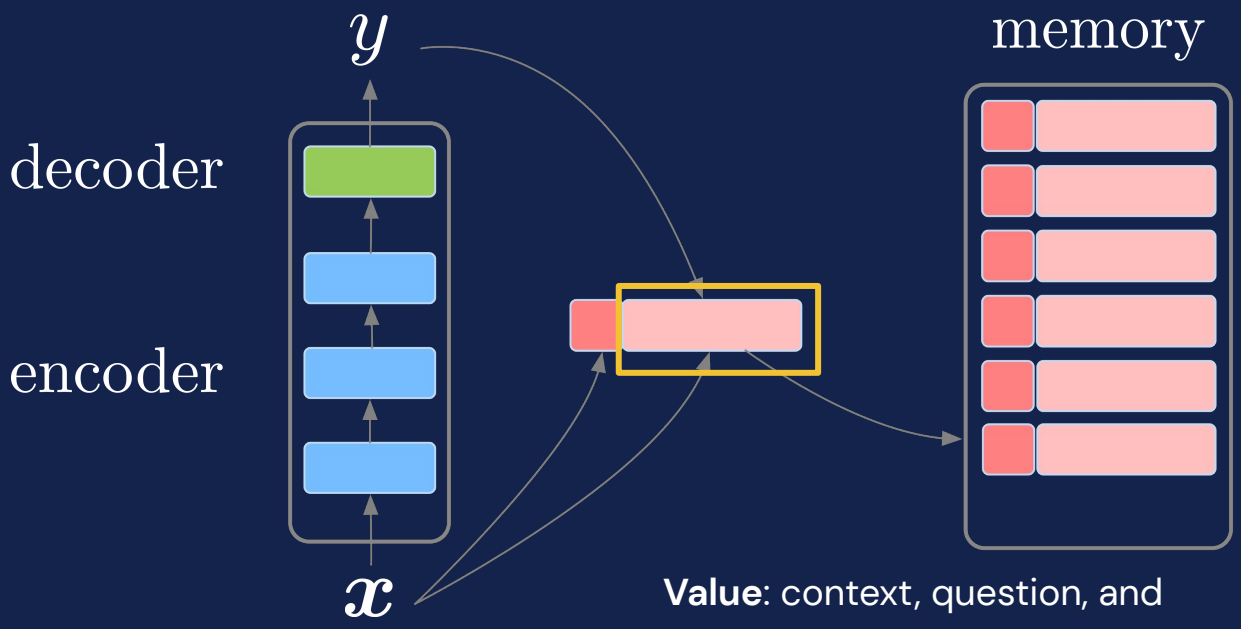
Question Answering Model



Question Answering Model



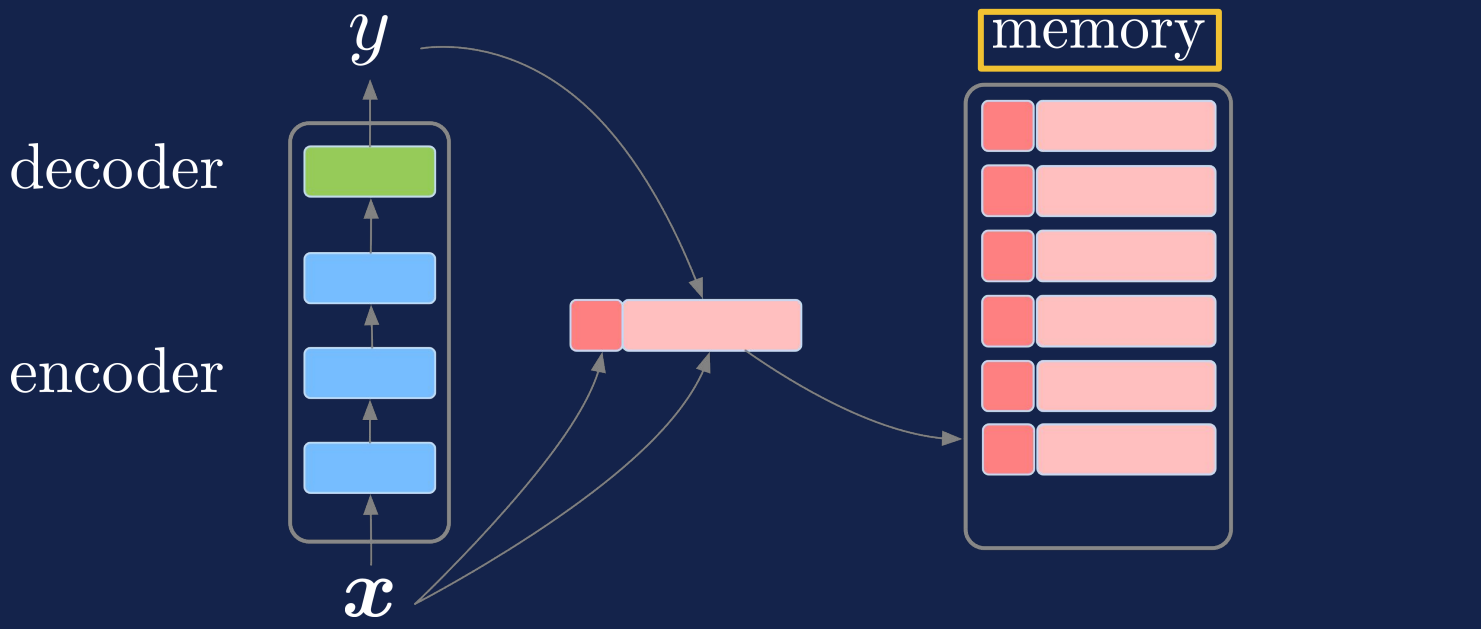
Question Answering Model



Value: context, question, and answer in textual forms (strings).

Question Answering Model

Memory: stores memory entries

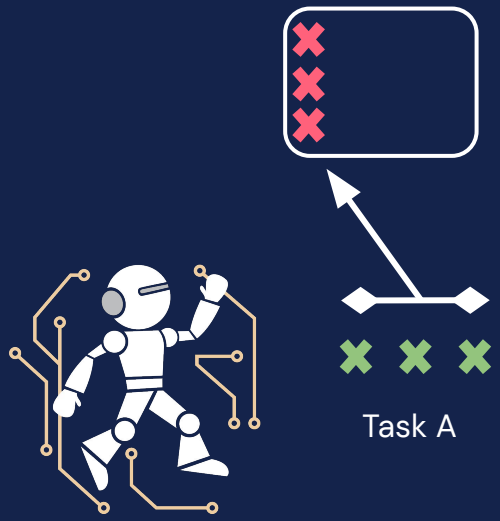


Training

Task A

$$\mathcal{L} = \log p(\boldsymbol{y} \mid \boldsymbol{x}; \mathbf{W})$$

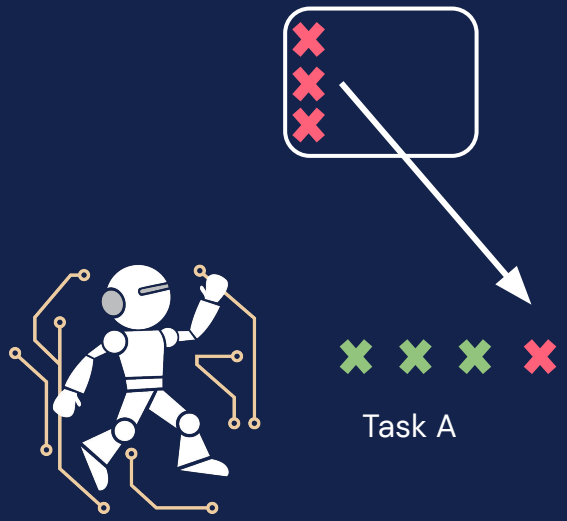
Training



$$\mathcal{L} = \log p(\boldsymbol{y} \mid \boldsymbol{x}; \mathbf{W})$$

Training

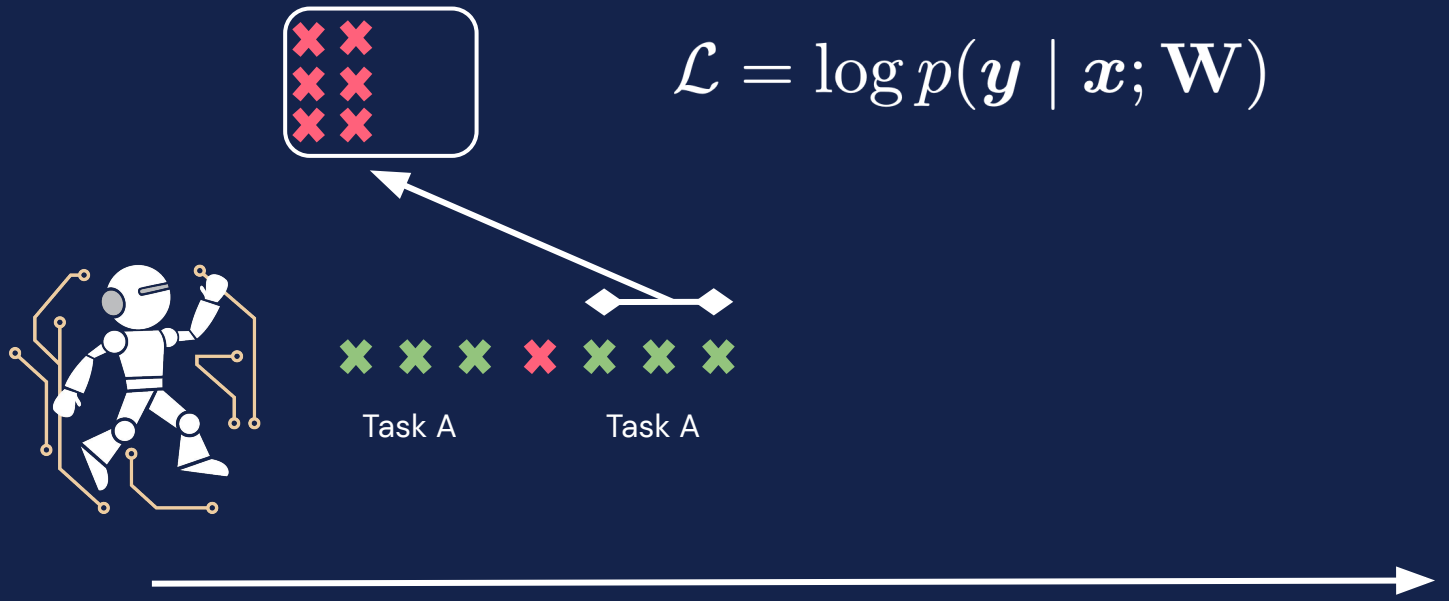
***Sparse* experience replay**: retrain on randomly sampled examples from the memory at a 1% rate.



$$\mathcal{L} = \log p(\boldsymbol{y} \mid \boldsymbol{x}; \mathbf{W})$$

Training

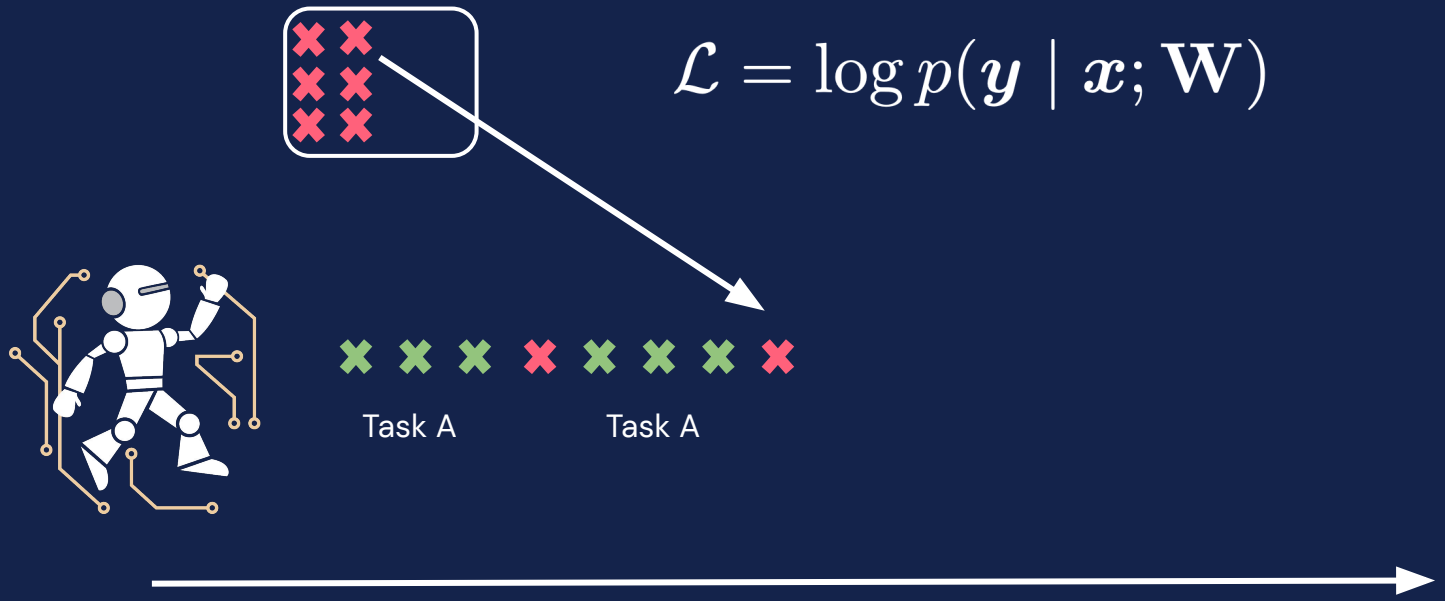
***Sparse* experience replay**: retrain on randomly sampled examples from the memory at a 1% rate.



$$\mathcal{L} = \log p(\boldsymbol{y} \mid \boldsymbol{x}; \mathbf{W})$$

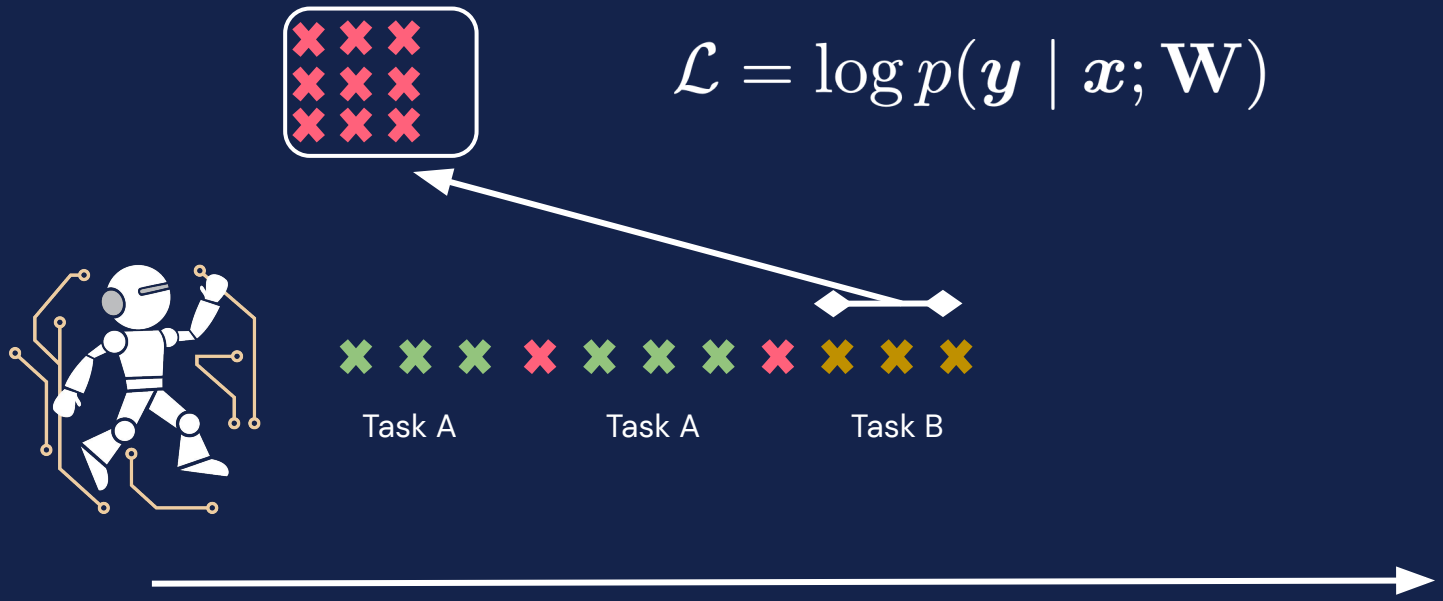
Training

***Sparse* experience replay**: retrain on randomly sampled examples from the memory at a 1% rate.



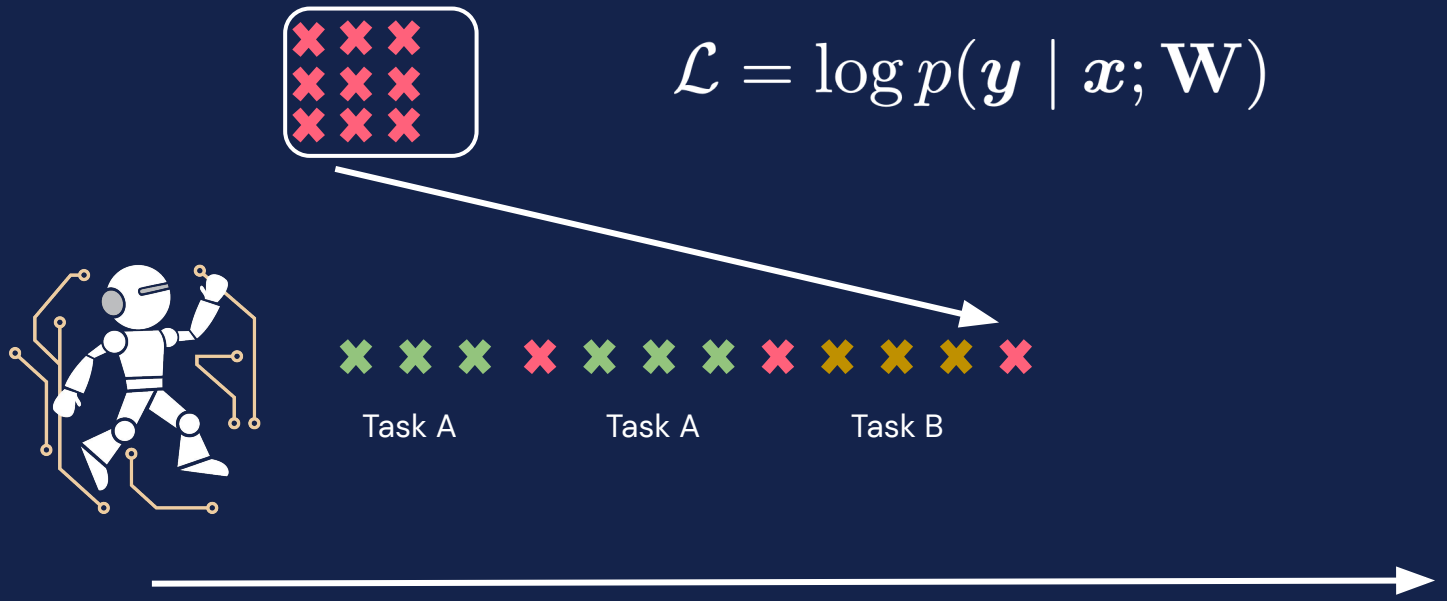
Training

***Sparse* experience replay**: retrain on randomly sampled examples from the memory at a 1% rate.



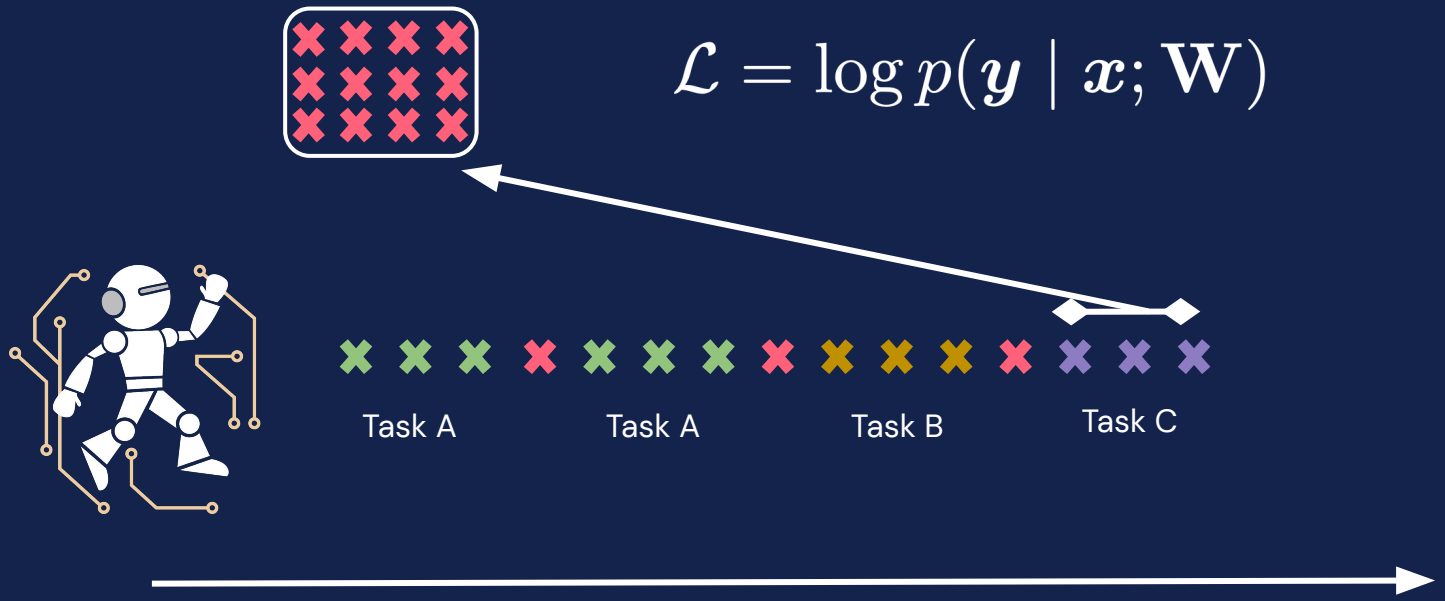
Training

***Sparse* experience replay**: retrain on randomly sampled examples from the memory at a 1% rate.



Training

***Sparse* experience replay**: retrain on randomly sampled examples from the memory at a 1% rate.



Inference (Prediction)

Local adaptation (Sprechmann et al., 2018).

Inference (Prediction)

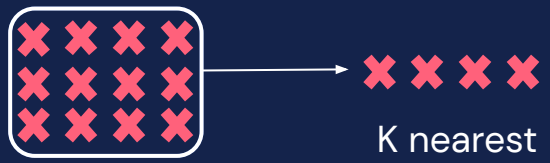
Local adaptation (Sprechmann et al., 2018).

Normans. The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. [...]

In what country is Normandy located?

Inference (Prediction)

Local adaptation (Sprechmann et al., 2018).



K nearest neighbors retrieval

Normans. The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. [...]

In what country is Normandy located?

In what area of France is Calais located?

In what country is St John's located?

In what country is Spoleto located?

In what part of Africa is Palermo located?

Inference (Prediction)

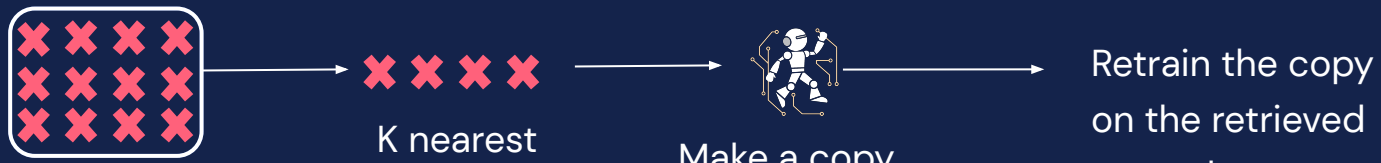
Local adaptation (Sprechmann et al., 2018).

K nearest neighbors retrieval

Make a copy of itself

Inference (Prediction)

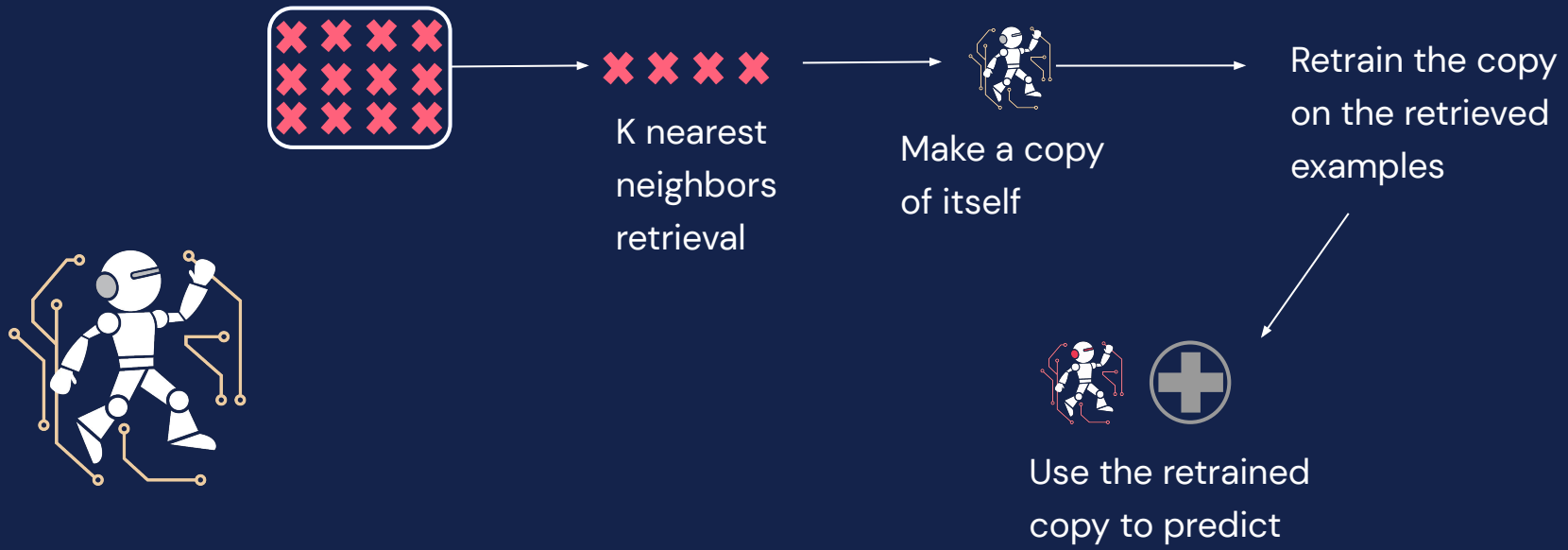
Local adaptation (Sprechmann et al., 2018).



$$\mathbf{W}_i = \arg\min_{\tilde{\mathbf{W}}} \lambda \|\tilde{\mathbf{W}} - \mathbf{W}\|_2^2 - \sum_{k=1}^{K} \alpha_k \log p(y_i^k \mid \boldsymbol{x}_i^k; \tilde{\mathbf{W}})$$

Inference (Prediction)

Local adaptation (Sprechmann et al., 2018).



Experiments

- Four question answering datasets.
  - SQuAD: Rajpurkar et al., 2016.
  - TriviaQA-Web: Joshi et al., 2017.
  - TriviaQA-Wiki: Joshi et al., 2017.
  - QuAC: Choi et al., 2018.

- The contexts come from **different domains** (e.g., Wikipedia articles, web pages).

- The questions are posed in **different styles** (e.g., information seeking, trivia questions).

Experiments

F1 scores (0-100), higher is better

	Enc-Dec	A-GEM	MbPA	Ours
QA	53.1	56.2	60.3	**62.4**

A-GEM: Chaudhry et al., 2019

MbPA: Sprechmann et al., 2018

Takeaways and Limitations

- Episodic memory allows a language model to deal with changes in data distribution.

Takeaways and Limitations

- Episodic memory allows a language model to deal with changes in data distribution.

- Linear **space complexity** in the number of examples, **constant** is more realistic.

% of stored examples in memory	10%	100%
Performance	61.5	62.0

Memory in Humans

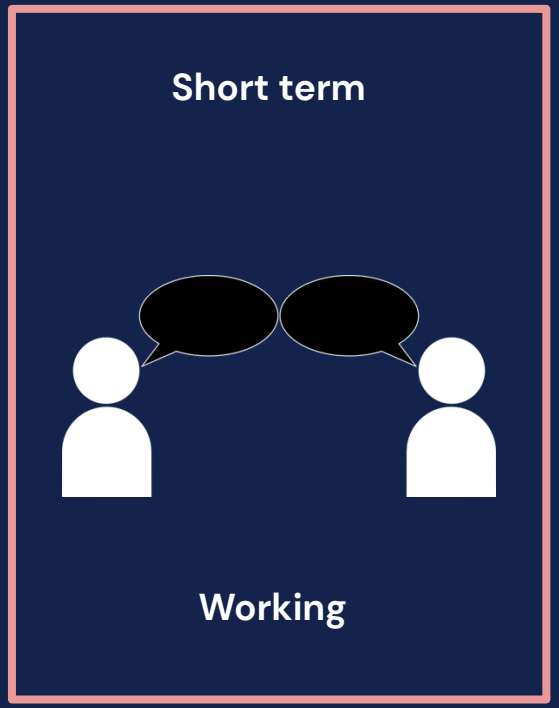
Human language processing is facilitated by specialized memory systems.

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

Memory in Humans

Human language processing is facilitated by specialized memory systems.

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

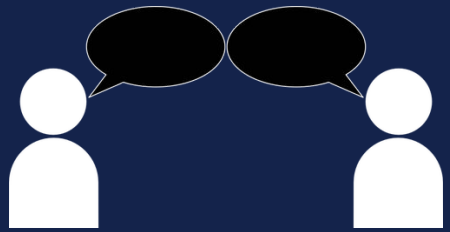


Memory in Humans

Human language processing is facilitated by specialized memory systems.

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

Short term



Working

Long term

Implicit

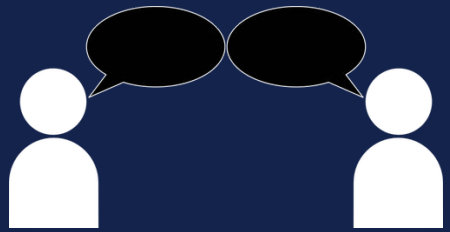
Procedural

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 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.

Future Directions

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

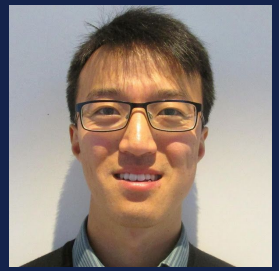
Generative Models

Future Directions

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

Modelling Latent Skills for Multitask Language Generation

Cao and Yogatama, arXiv 2020



Kris

Dani

Future Directions

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

Memory

Future Directions

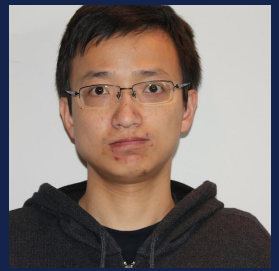
A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

Adaptive Semiparametric Language Models

Yogatama et al., TACL 2021

Dani

Cyprien



Lingpeng

Future Directions

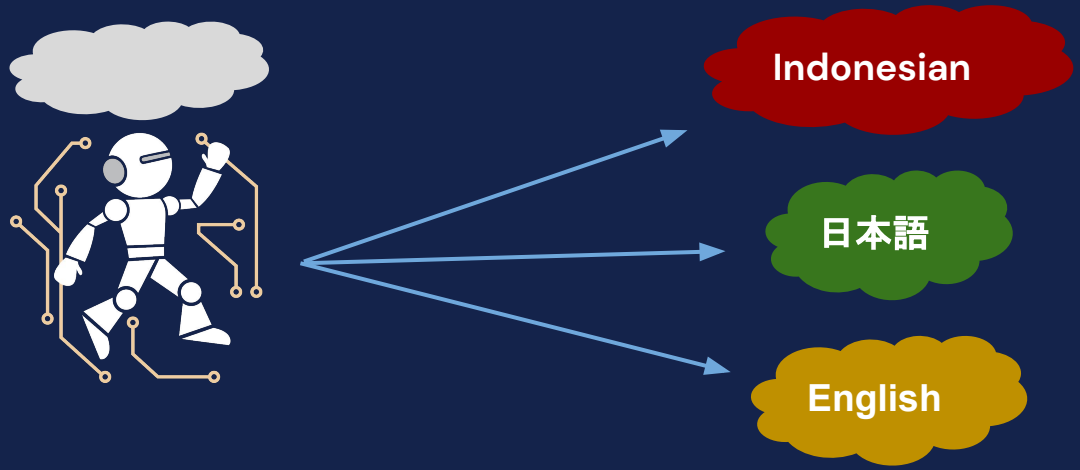
A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

Representation Learning

Future Directions

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

Representation Learning



Future Directions

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

Representation Learning

On the Crosslingual Transferability of Monolingual 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.

Memory

Representation Learning

Generative Models

Future Directions

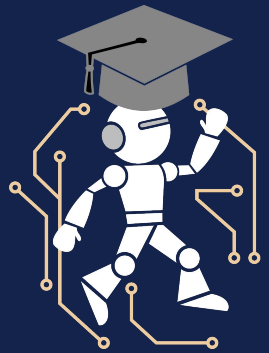
A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

Memory	Representation Learning	Generative Models

Future Directions

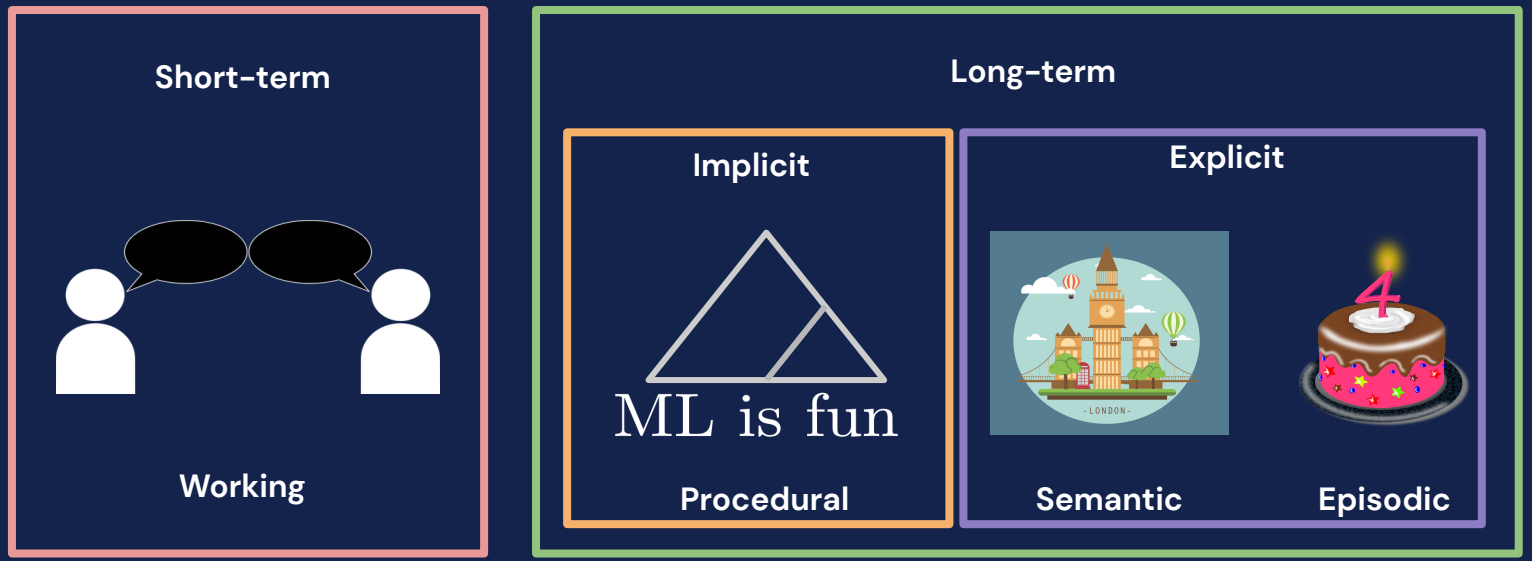
A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

Memory **Representation Learning** **Generative Models**



Future Directions

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



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

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

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.

Memory

Yogatama and Mann; AISTATS 2014

Yogatama et al., ICLR 2017

Yogatama et al., ICLR 2018

de Masson d'Autume et al.; 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.

Memory

Yogatama and Mann; AISTATS 2014
Yogatama et al., ICLR 2017
Yogatama et al., ICLR 2018
de Masson d'Autume et al.; NeurIPS 2019
Yogatama et al., TACL 2021

Representation Learning

Yogatama and Smith; ACL 2014
Yogatama and Smith; ICML 2015
Artetxe et al., ACL 2020
Kong et al., ICLR 2020

Research Areas

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

Memory

Yogatama and Mann; AISTATS 2014
Yogatama et al., ICLR 2017
Yogatama et al., ICLR 2018
de Masson d'Autume et al.; NeurIPS 2019
Yogatama et al., TACL 2021

Representation Learning

Yogatama and Smith; ACL 2014
Yogatama and Smith; ICML 2015
Artetxe et al., ACL 2020
Kong et al., ICLR 2020

Generative Models

Yogatama et al., TACL 2014
Yogatama et al., arXiv 2017
Kong et al., 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.

Memory

Yogatama and Mann; AISTATS 2014
Yogatama et al., ICLR 2017
Yogatama et al., ICLR 2018
de Masson d'Autume et al.; NeurIPS 2019
Yogatama et al., TACL 2021

Representation Learning

Yogatama and Smith; ACL 2014
Yogatama and Smith; ICML 2015
Artetxe et al., ACL 2020
Kong et al., ICLR 2020

Generative Models

Yogatama et al., TACL 2014
Yogatama et al., arXiv 2017
Kong et al., ICLR 2018
Cao and Yogatama, arXiv 2020

Reasoning, interactions with other modalities, robustness, fairness, and others.

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