

# Toward General Linguistic Intelligence

Dani Yogatama

# 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).

హల్లు Përhëndetje Halo  
Aloha こんにちは Sveiki שָׁלוּם  
Ciao Ahoj Hello Сайн уу  
নমস্কার Ahoj Hello വണ്ണക്കമ்  
**KAMUSTA** Γειά σου 여보세요 Salve  
Здравствуйте ابْحَرْم Merhaba  
Hej 你好 Hola xin chào



# 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**.



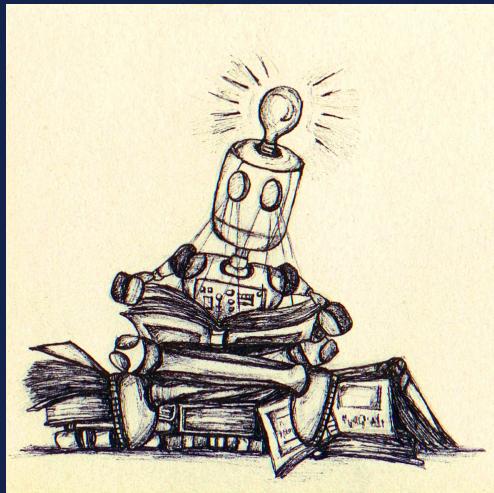
# General Linguistic Intelligence

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



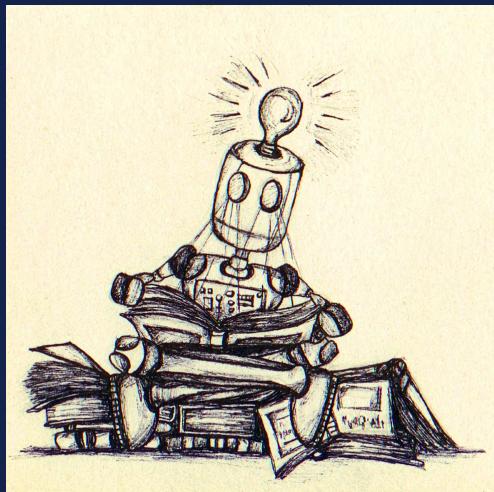
# General Linguistic Intelligence

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



హలో Përhëndetje Halo  
Aloha こんにちは Sveiki שָׁלוּם  
Ciao Ahoj Hello Сайн уу  
नमस्कार Ahoj Hello വണ്ണക്കമ்  
**KAMUSTA** Γειά σου 여보세요 Salve  
Здравствуйте اب حرم Merhaba  
Hej 你好 Hola xin chào



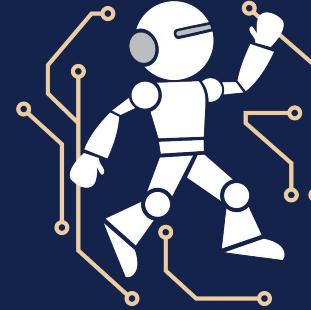
# Challenges: Human Learning vs. Machine Learning



Human	
“Large” datasets	<b>Acquisition</b>
Few examples	<b>Task Training</b>
Dataset agnostic	<b>Linguistic knowledge</b>
Generalizable to new tasks	<b>Generalization</b>



# Challenges: Human Learning vs. Machine Learning



Human		Machine
“Large” datasets	<b>Acquisition</b>	Large datasets (representation learning)
Few examples	<b>Task Training</b>	Large datasets (supervised fine tuning)
Dataset agnostic	<b>Linguistic knowledge</b>	Dataset specific
Generalizable to new tasks	<b>Generalization</b>	Forget previous tasks given a new task



# The State of Natural Language Processing

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

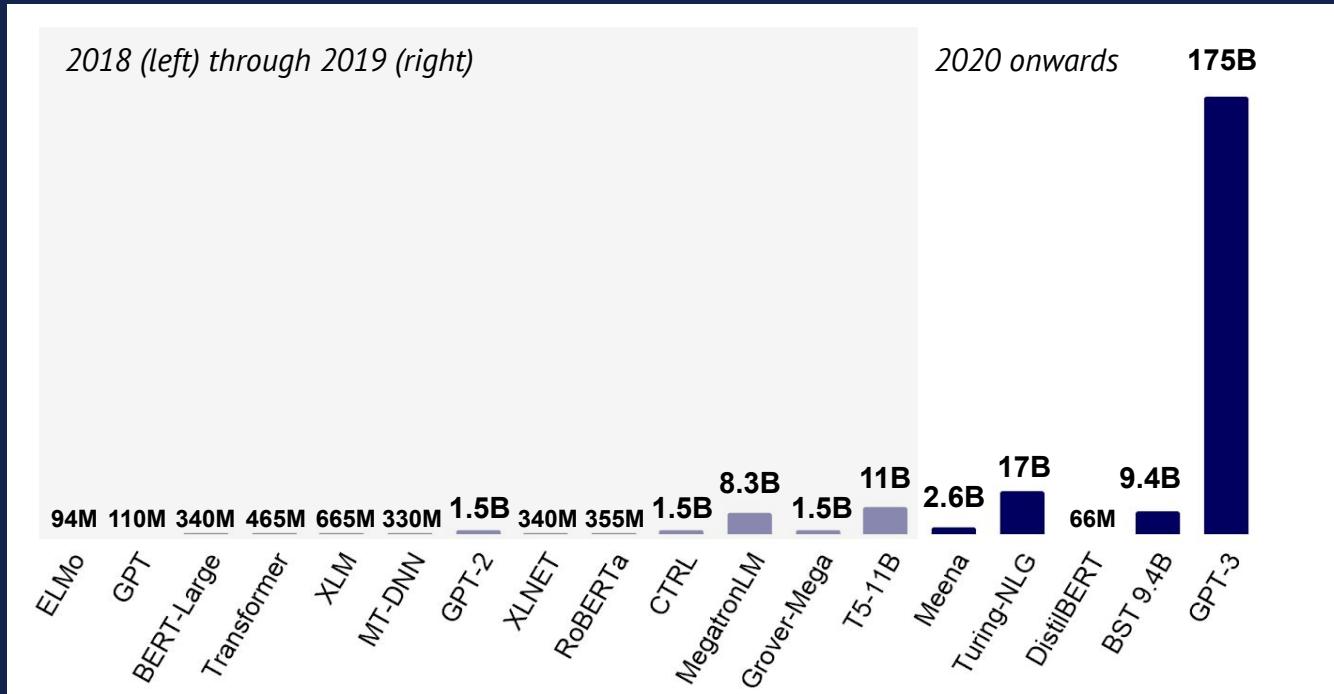
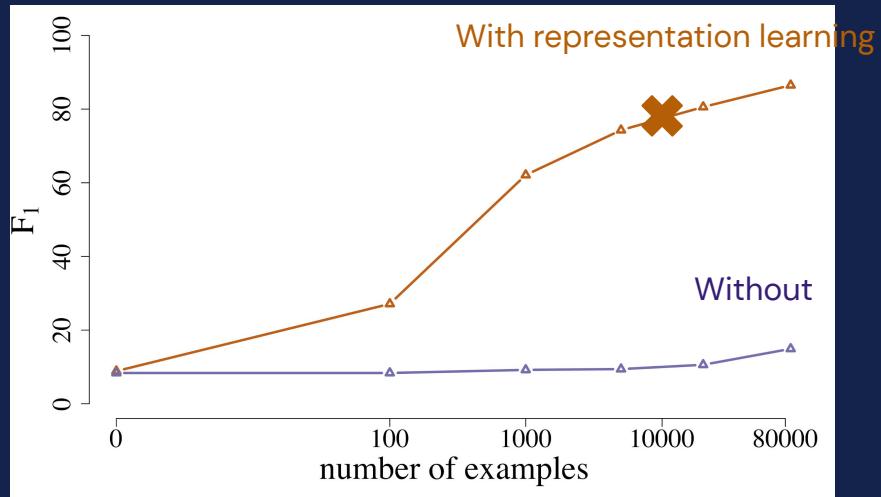


Figure taken from [State of AI Report 2020](#).



# The State of Natural Language Processing

- Great progress, but current models still require many in-domain training examples.



**Yogatama et al., arXiv 2019**

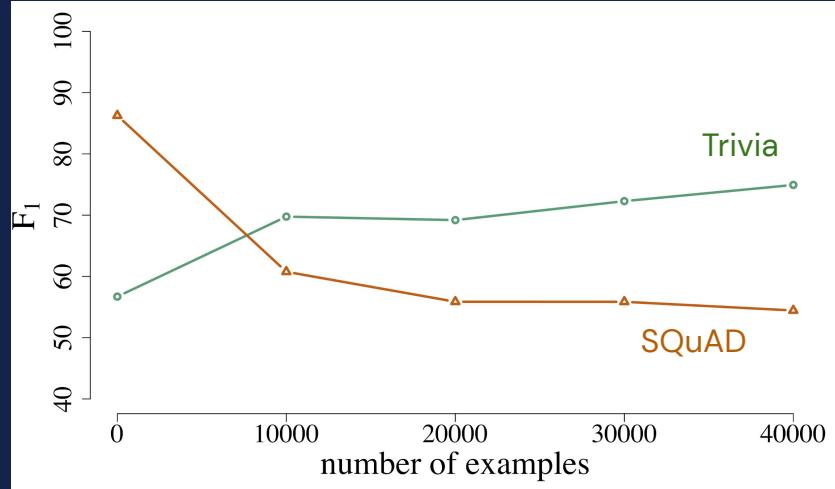
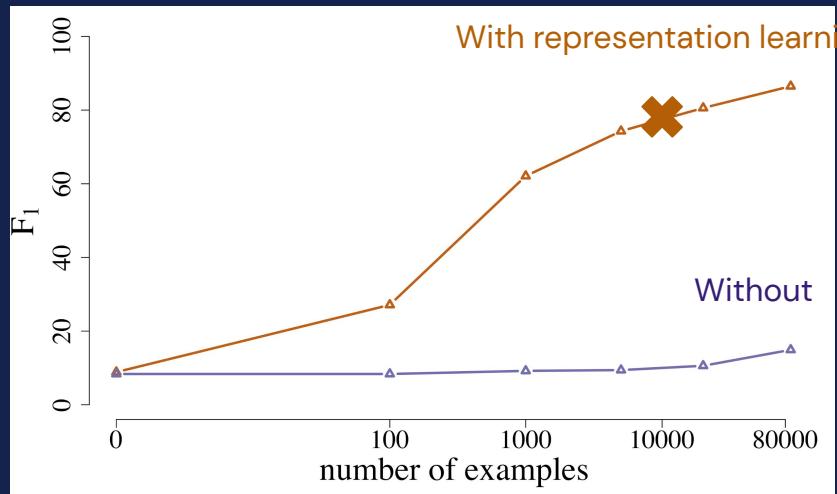
Model: BERT, Devlin et al. 2019

QA dataset: SQuAD, Rajpurkar et al., 2016



# The State of Natural Language Processing

- Great progress, but current models still require many in-domain training examples.
- They overfit to a specific dataset (task) and often forget.



**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.

## 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., in review**



# 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., in review**

## 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., in review

## 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.; in review

## 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.



# 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

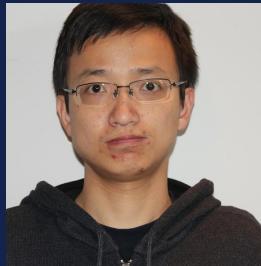
de Masson d'Autume et al., NeurIPS 2019



Cyprien



Sebastian



Lingpeng



Dani



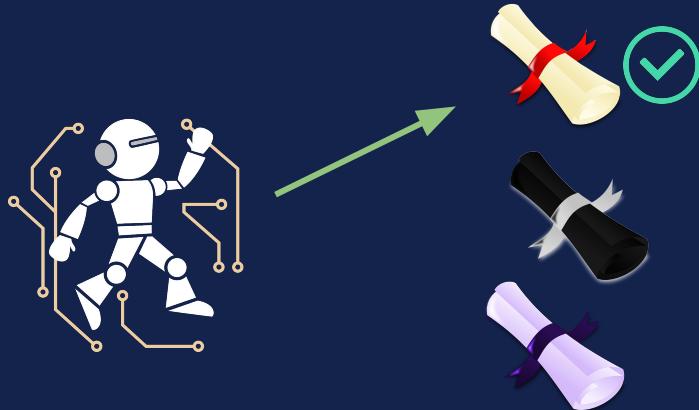
# Background

- 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 (McClosky and Cohen, 1989; Ratcliff, 1990).



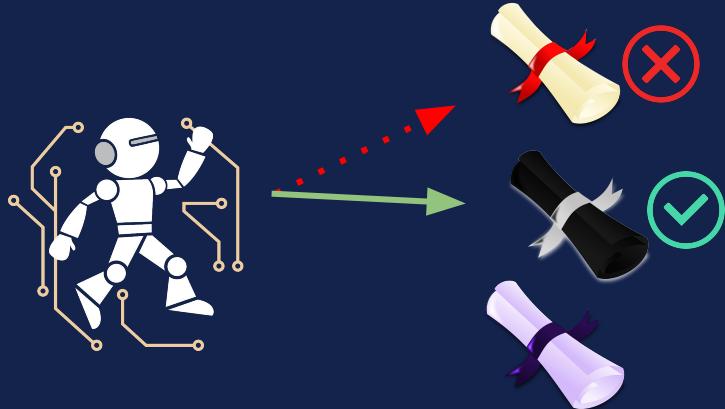
# Background

- 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 (McCloskey and Cohen, 1989; Ratcliff, 1990).



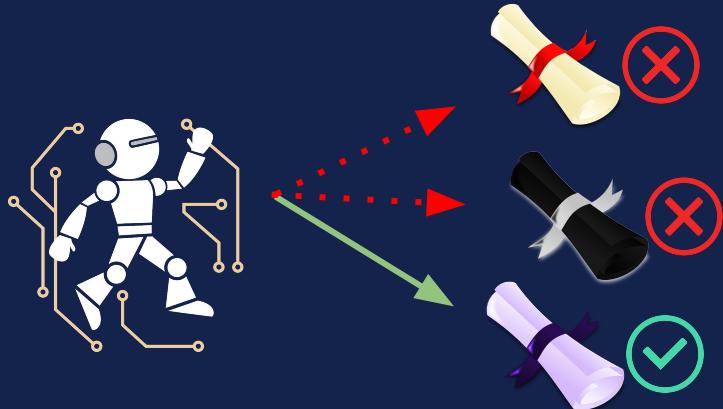
# Background

- 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 (McCloskey and Cohen, 1989; Ratcliff, 1990).



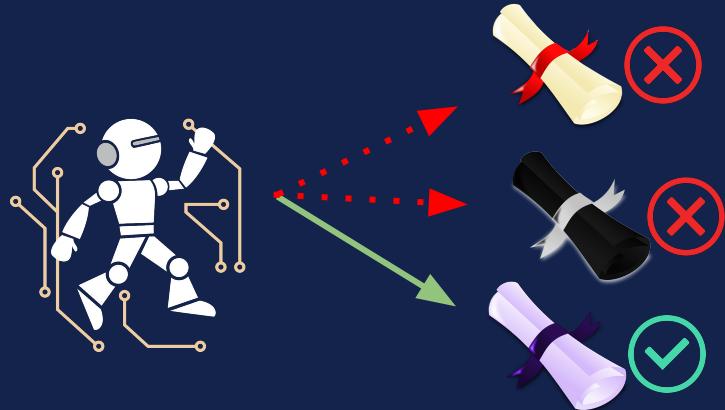
# Background

- 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 (McCloskey and Cohen, 1989; Ratcliff, 1990).



# Background

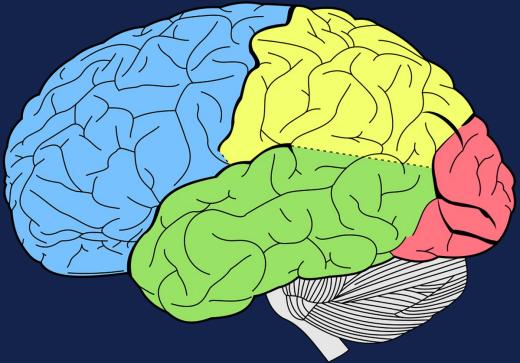
- A model should be able to reuse knowledge from related tasks to learn a new task faster.
- Current models not only fail to do this, they **catastrophically forget** previously learned tasks (McCloskey and Cohen, 1989; Ratcliff, 1990).



**Hypothesis:** episodic memory mitigates catastrophic forgetting in language learning.



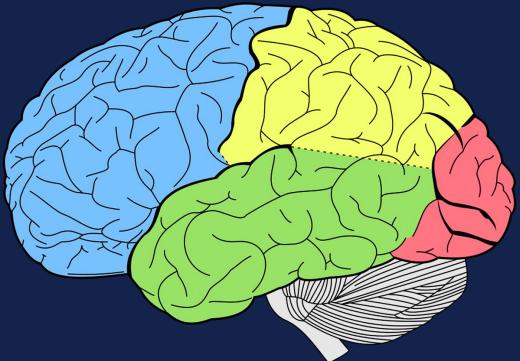
# On Memory-Augmented Neural Networks



**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



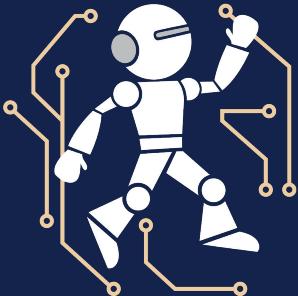
**Episodic memory** is a type of long-term memory of **events** and **experiences**. It is often associated with a module that stores training examples in neural networks..

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**

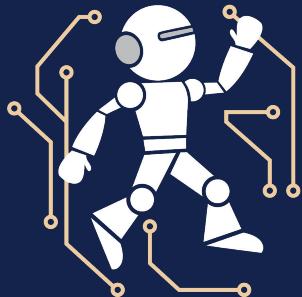
*The Independent, Friday 16 February 1996*  
A massive anti-pollution operation was underway last night after a 147,000-ton super tanker ran aground off Milford Haven, West Wales. [...]

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



# Problem Setup

SQuAD: Rajpurkar et al., 2016



Training

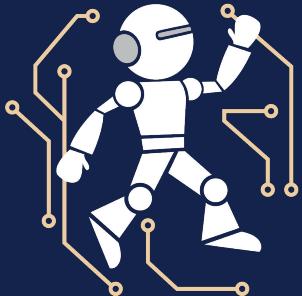


**Computational Complexity Theory.**  
Computational complexity theory is a branch of the theory of computation in theoretical computer science that focuses on classifying computational problems according to their inherent difficulty [...]

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 ...**

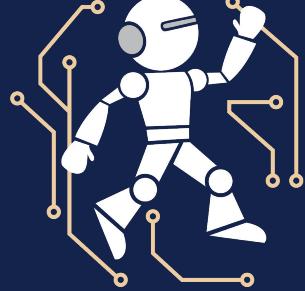
Pinochet was publicly known as a man with a lack of culture. This image was reinforced by the fact [...]

**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



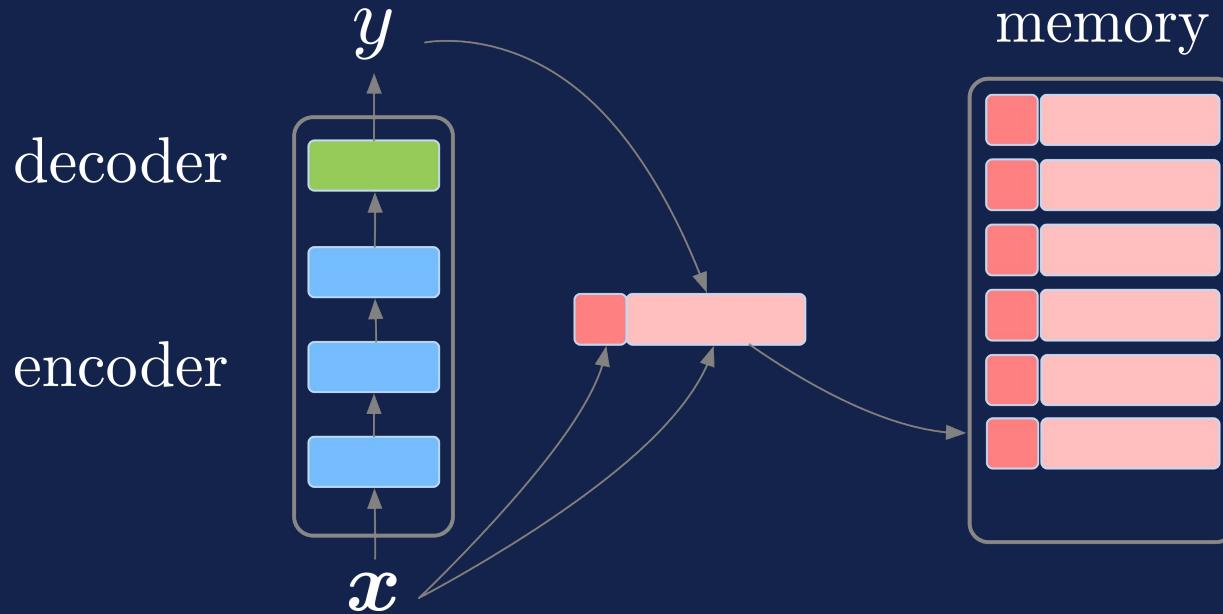
Training



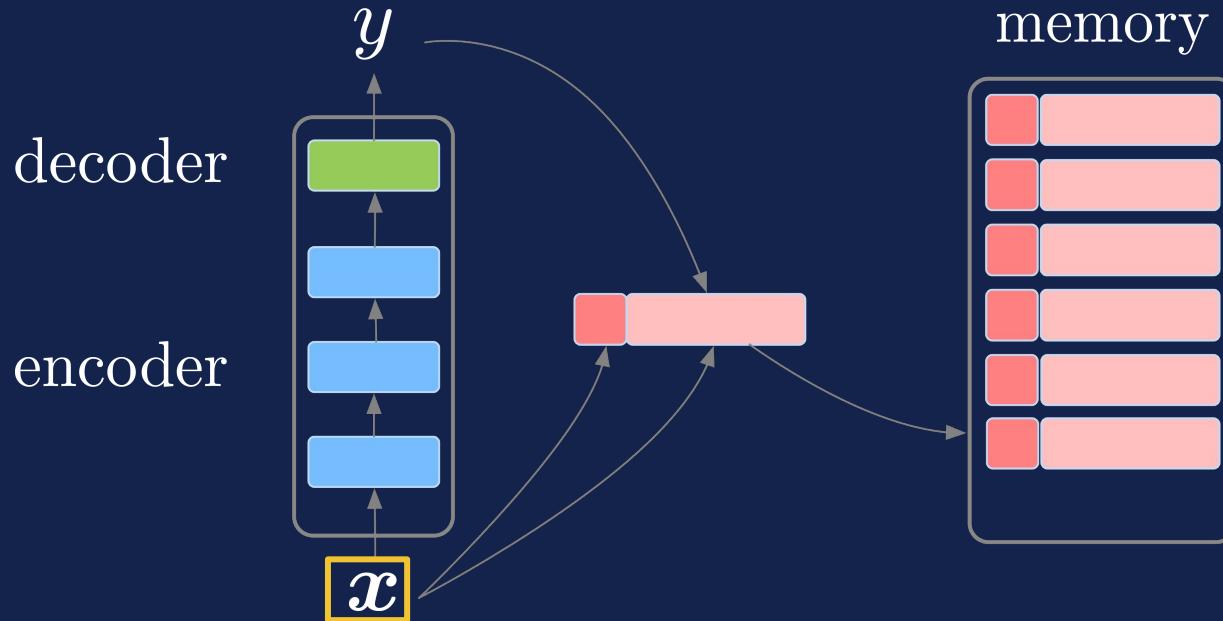
Test



# Question Answering Model



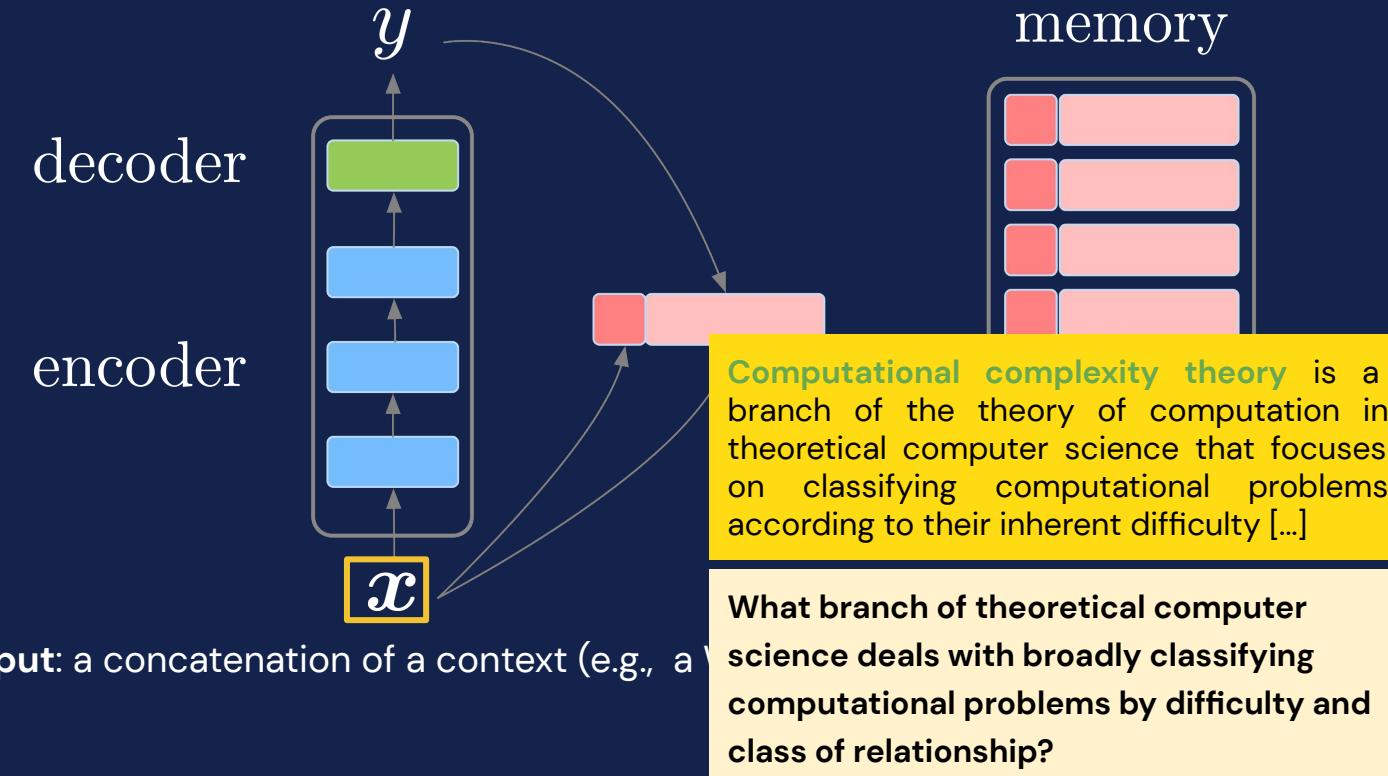
# Question Answering Model



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

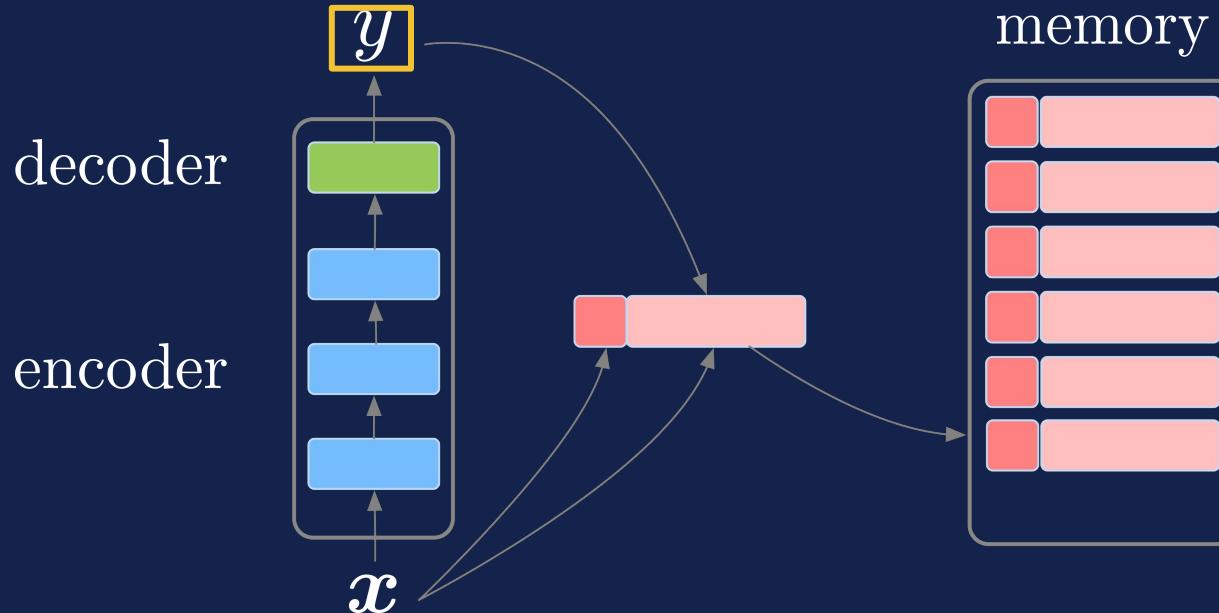


# Question Answering Model



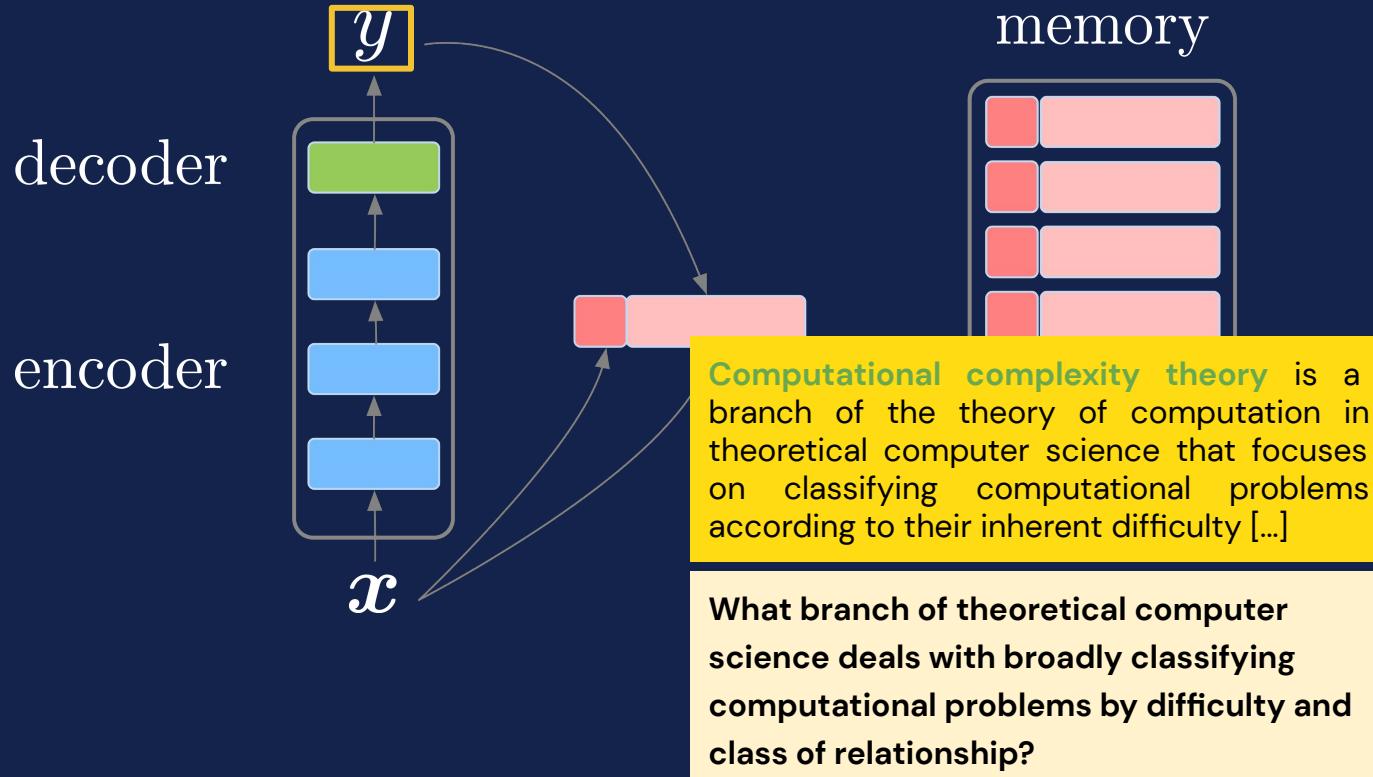
# Question Answering Model

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

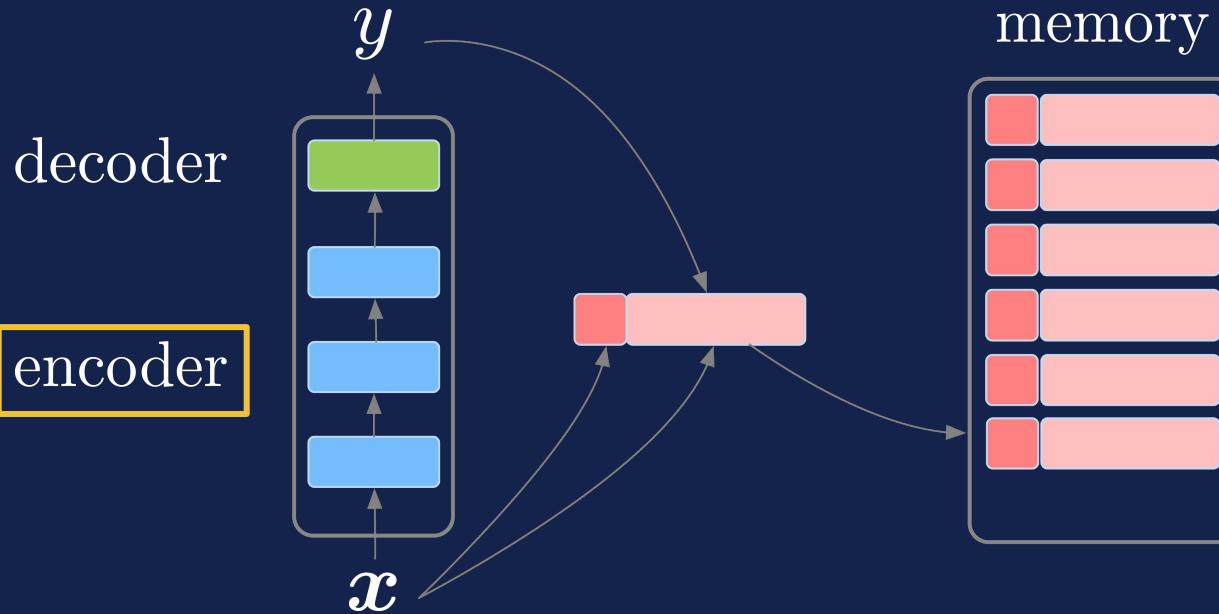


# Question Answering Model

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



# 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).



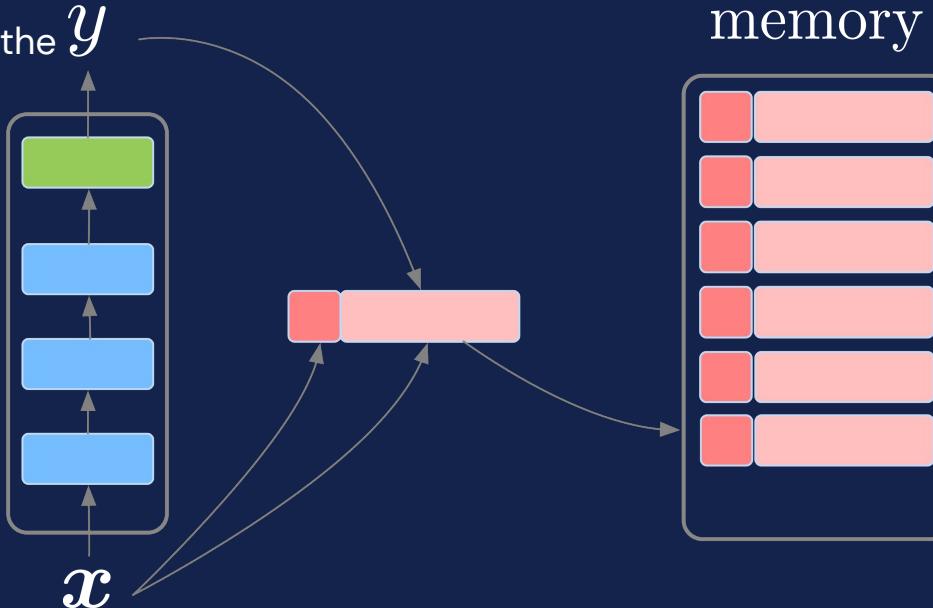
# Question Answering Model

**Decoder:** a linear function

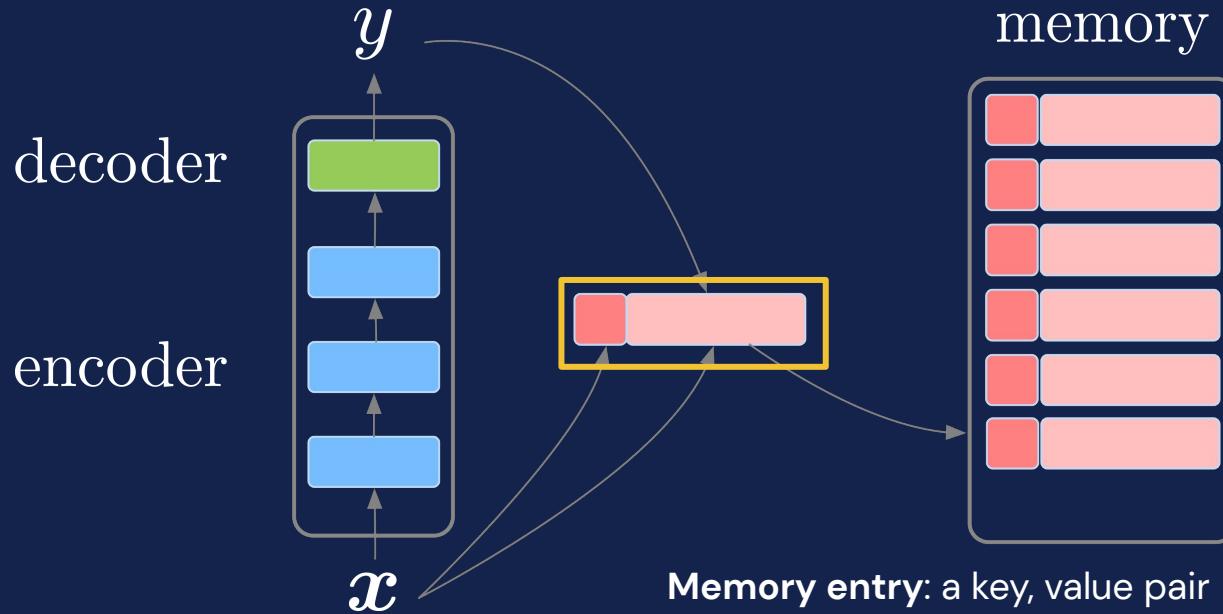
that predicts start and end  
indices of the answer in the  $y$   
context.

decoder

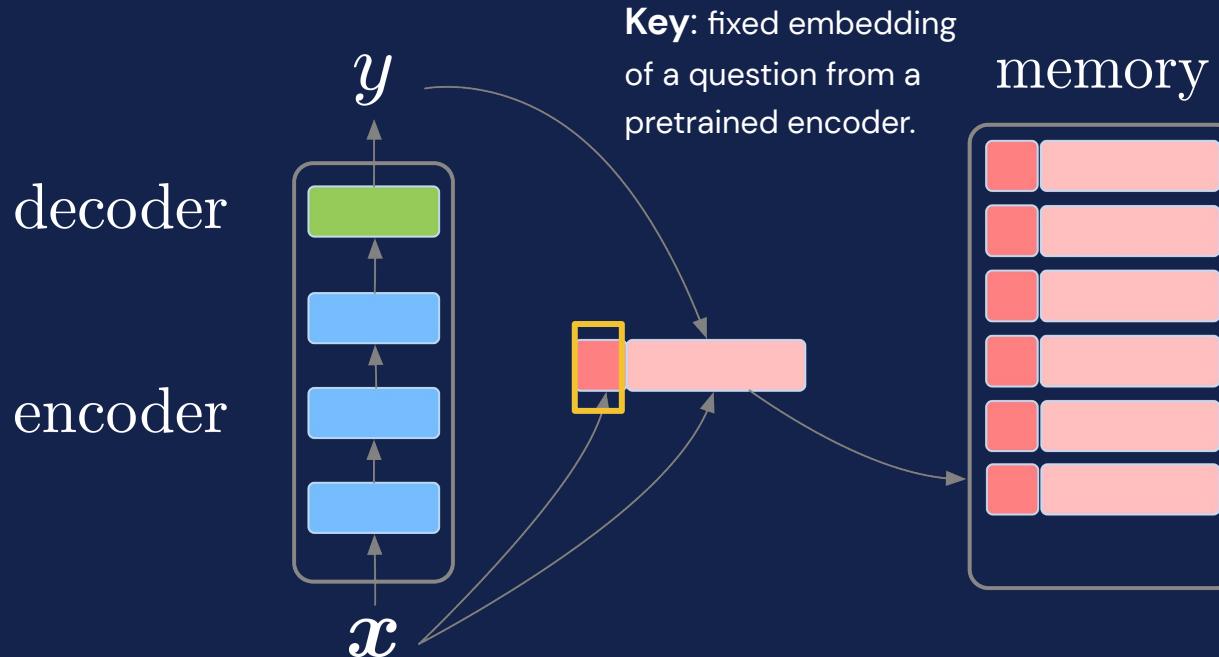
encoder



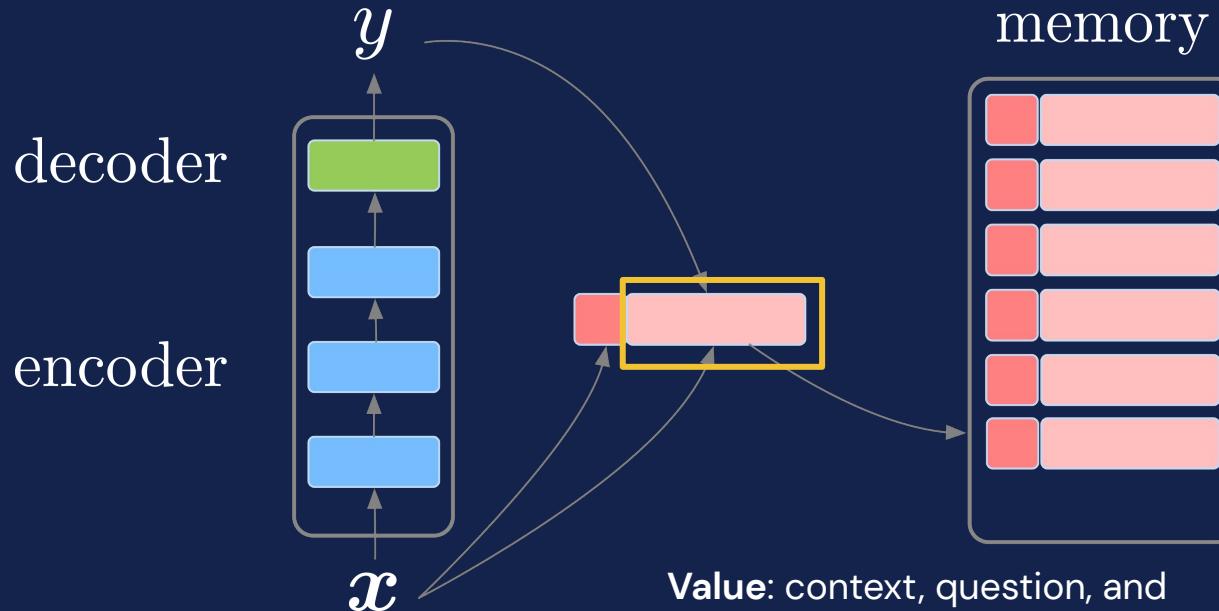
# Question Answering Model



# Question Answering Model



# Question Answering Model

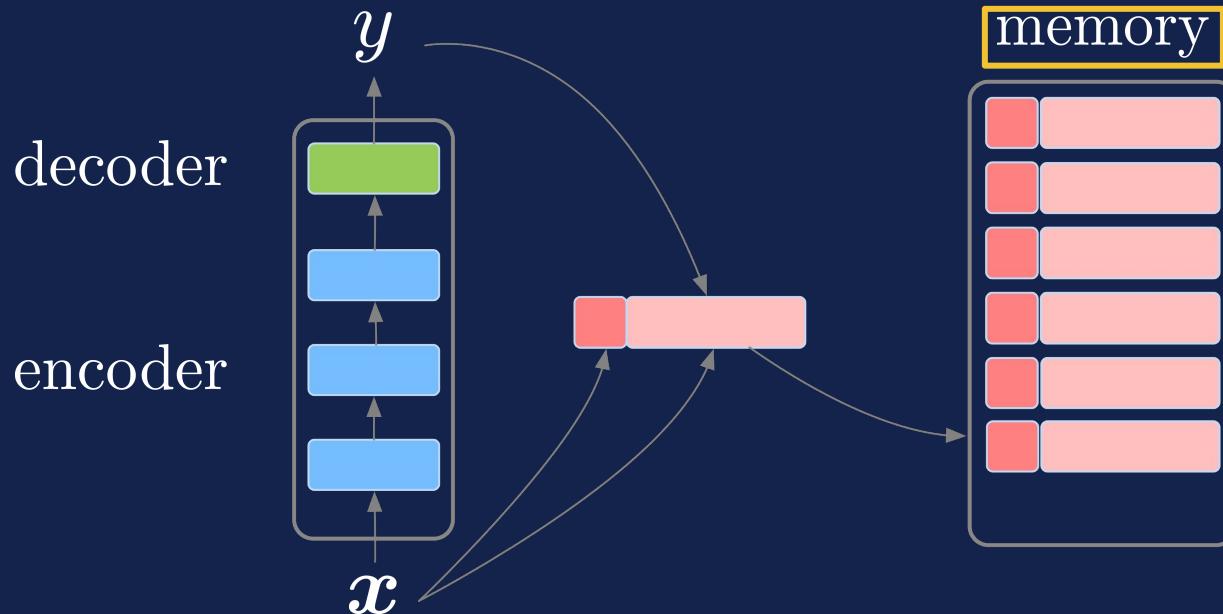


**Value:** context, question, and answer in textual forms (strings).



# Question Answering Model

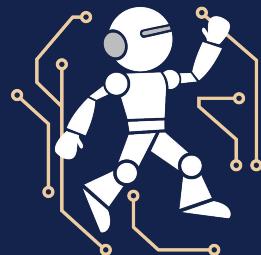
**Memory:** stores memory entries



# Training



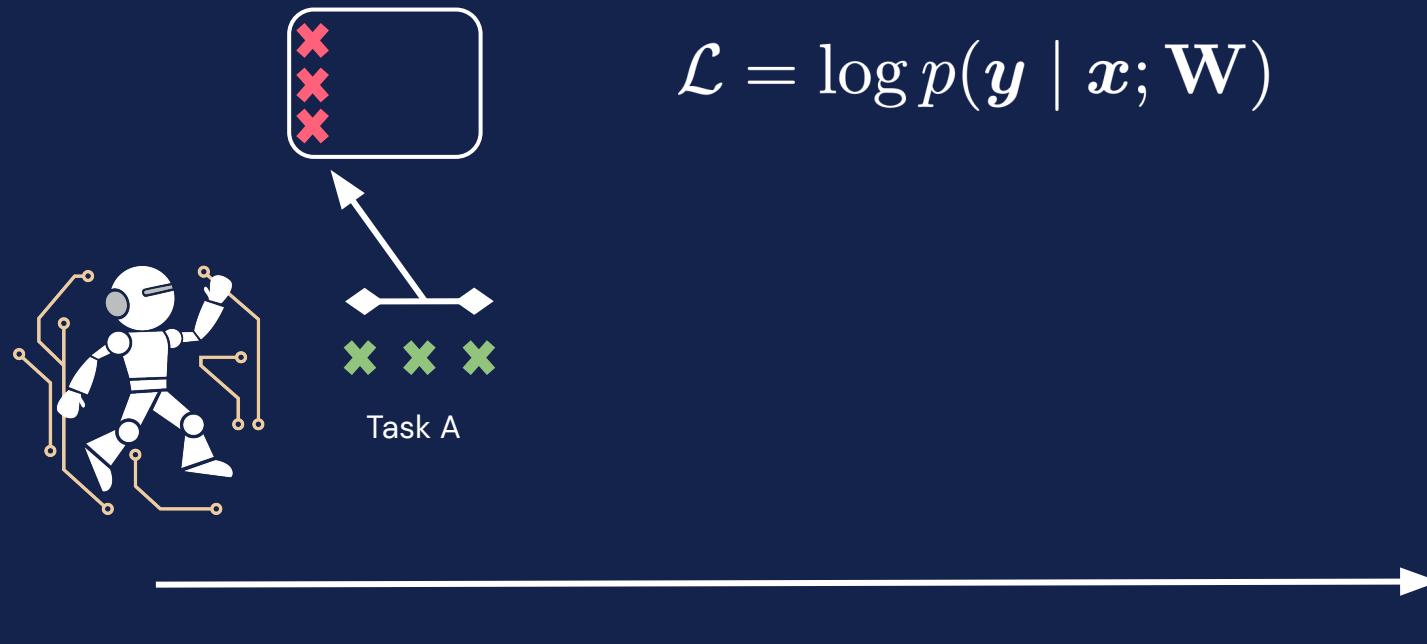
$$\mathcal{L} = \log p(\mathbf{y} \mid \mathbf{x}; \mathbf{W})$$



Task A

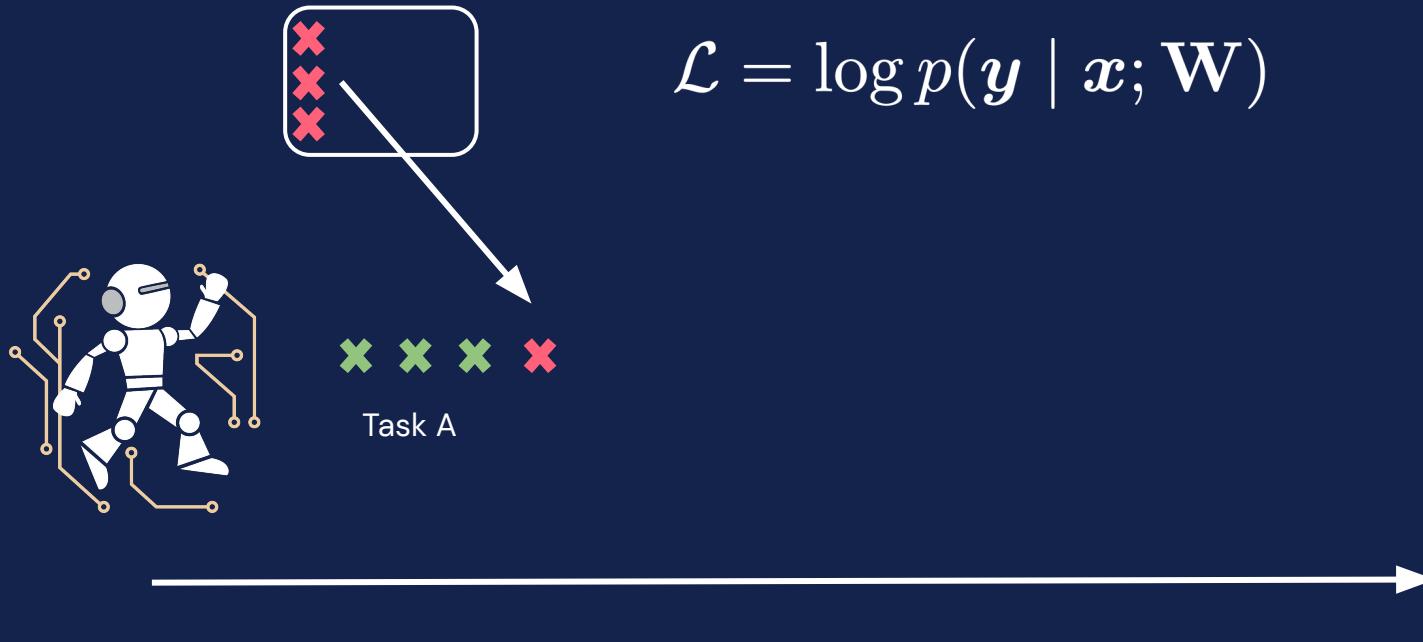


# Training



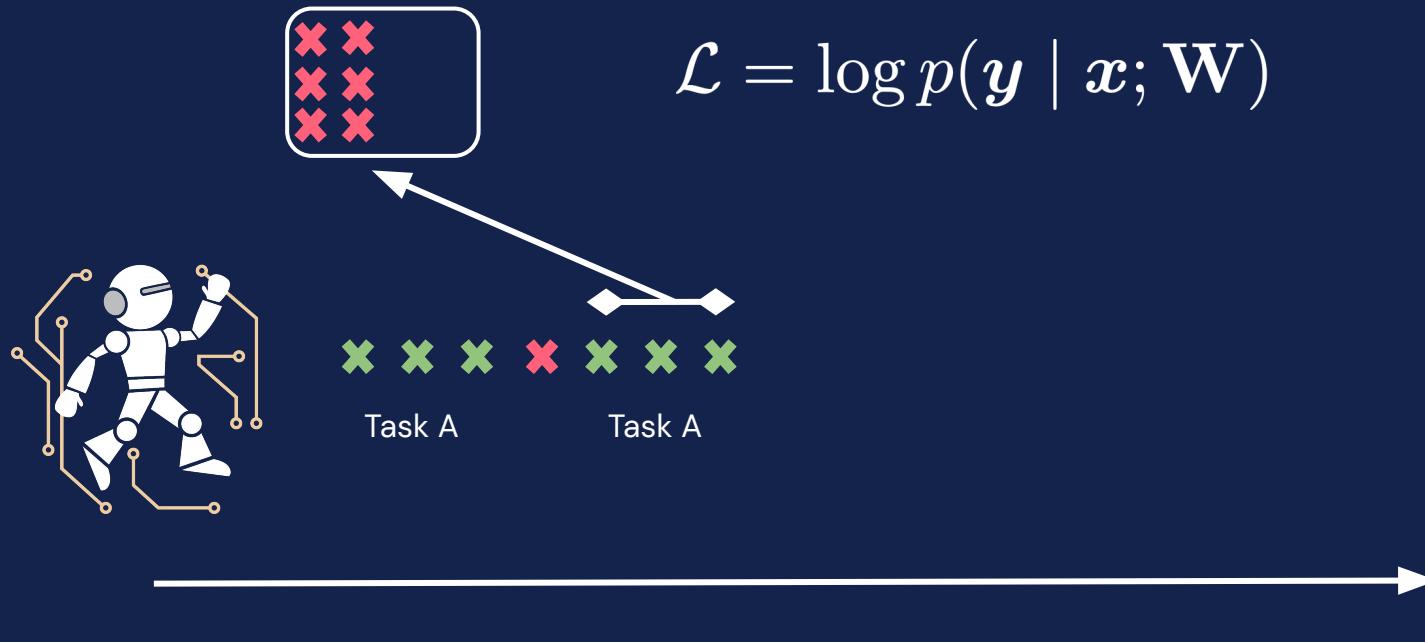
# Training

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



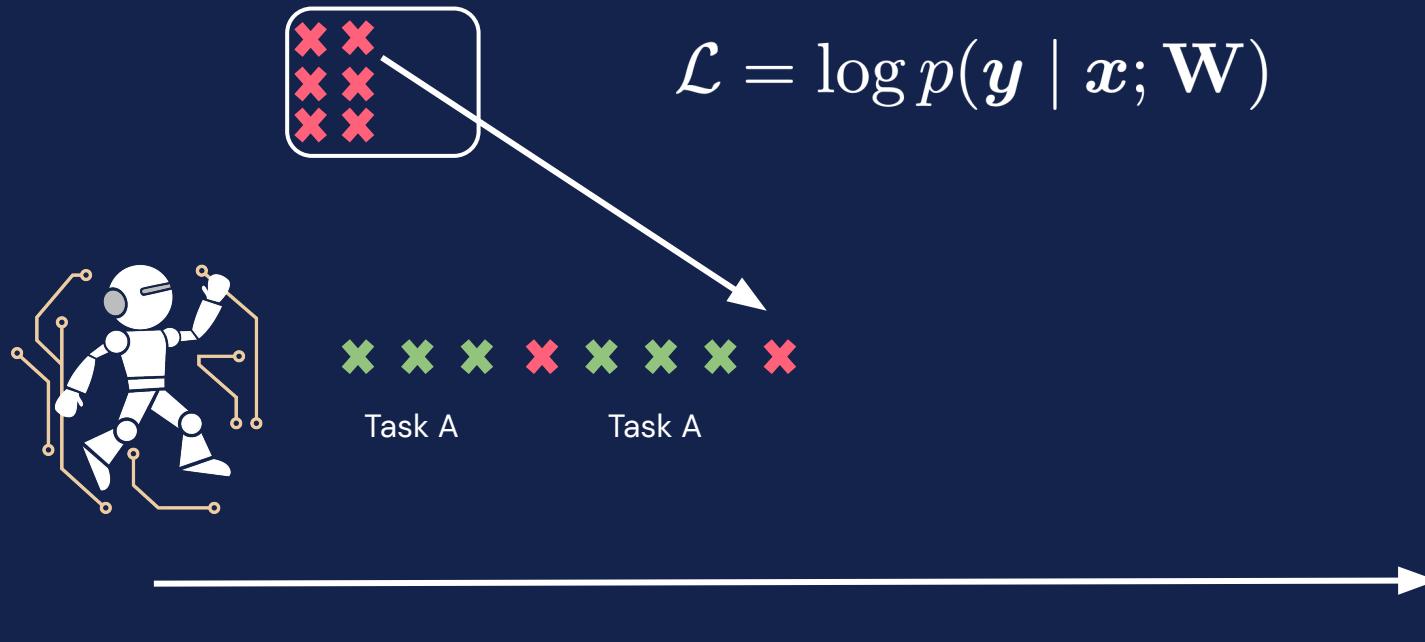
# Training

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



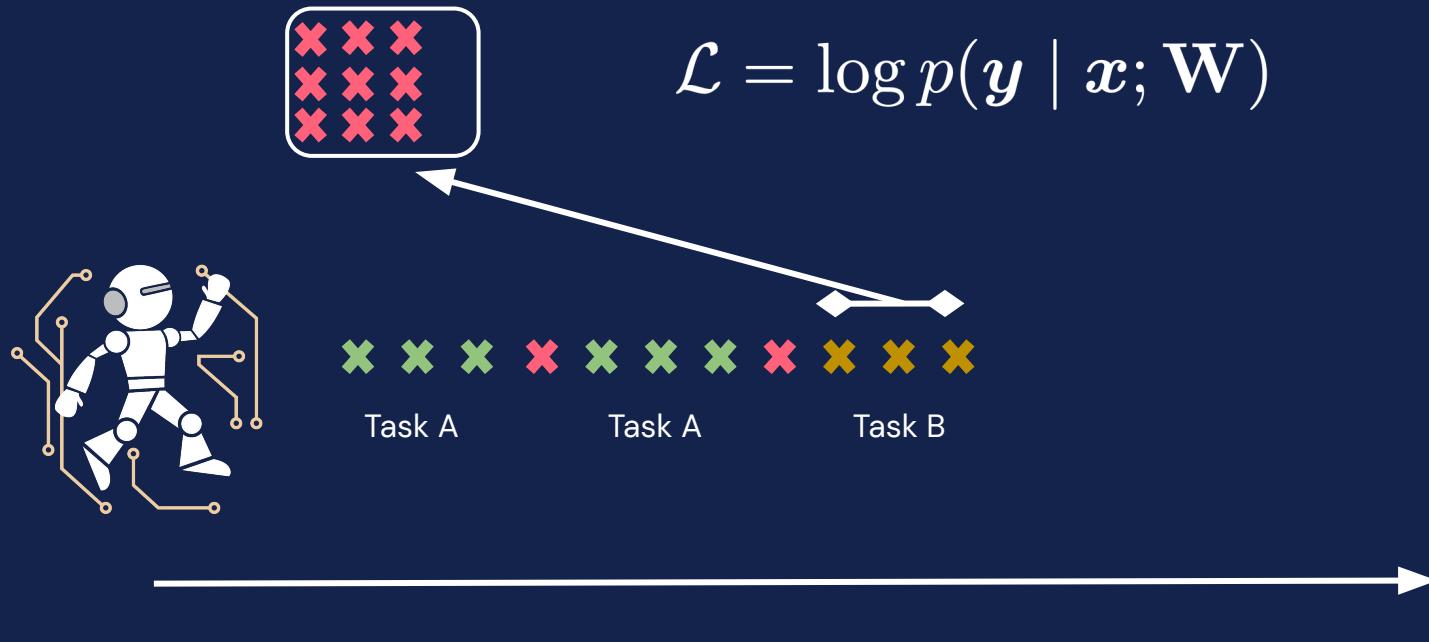
# Training

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



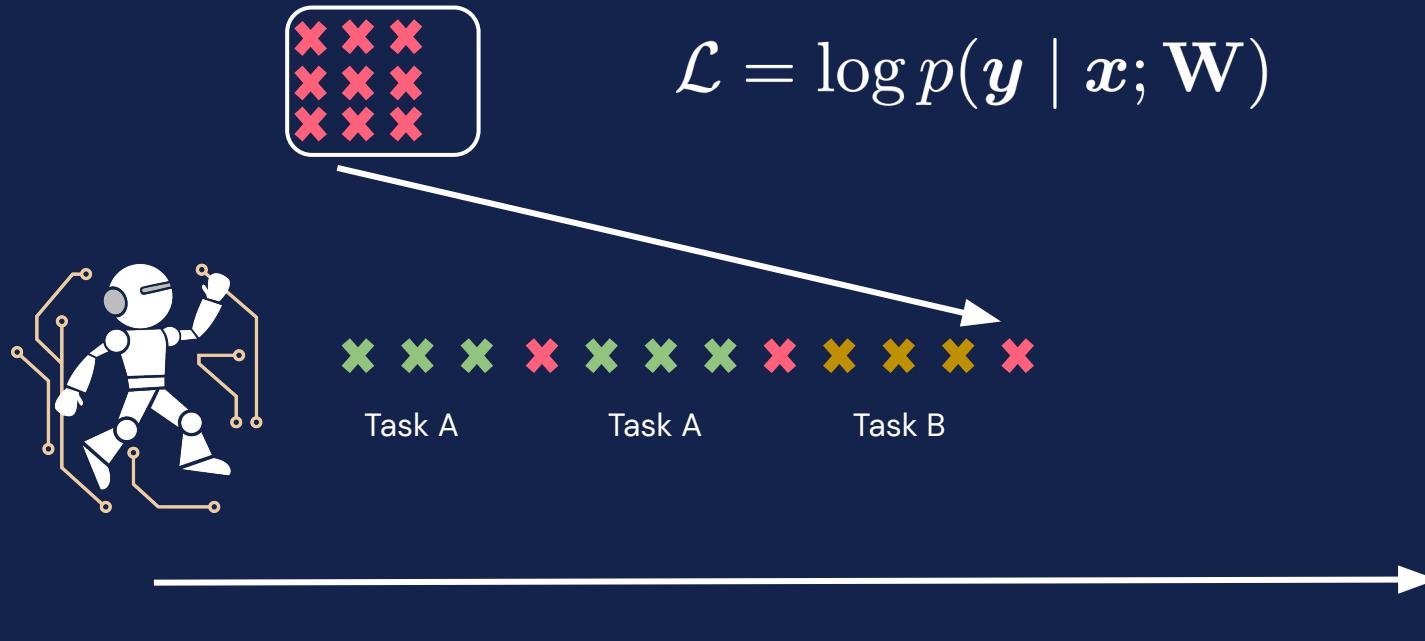
# Training

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



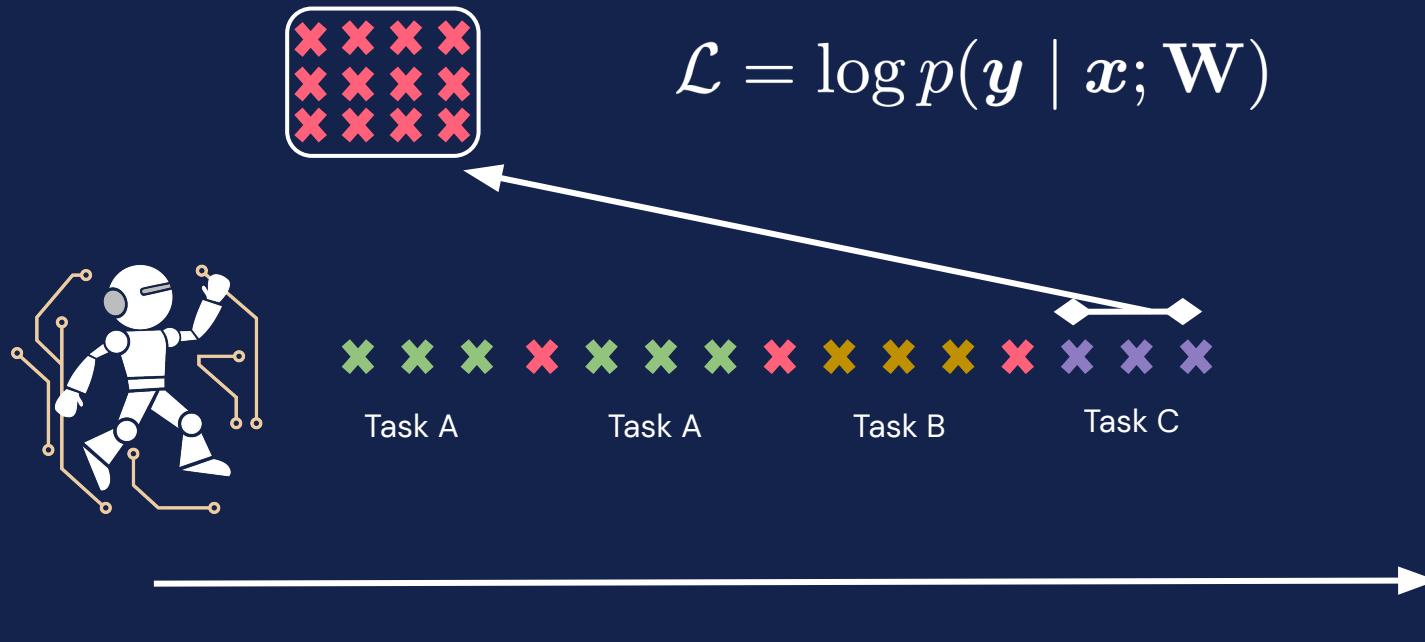
# Training

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



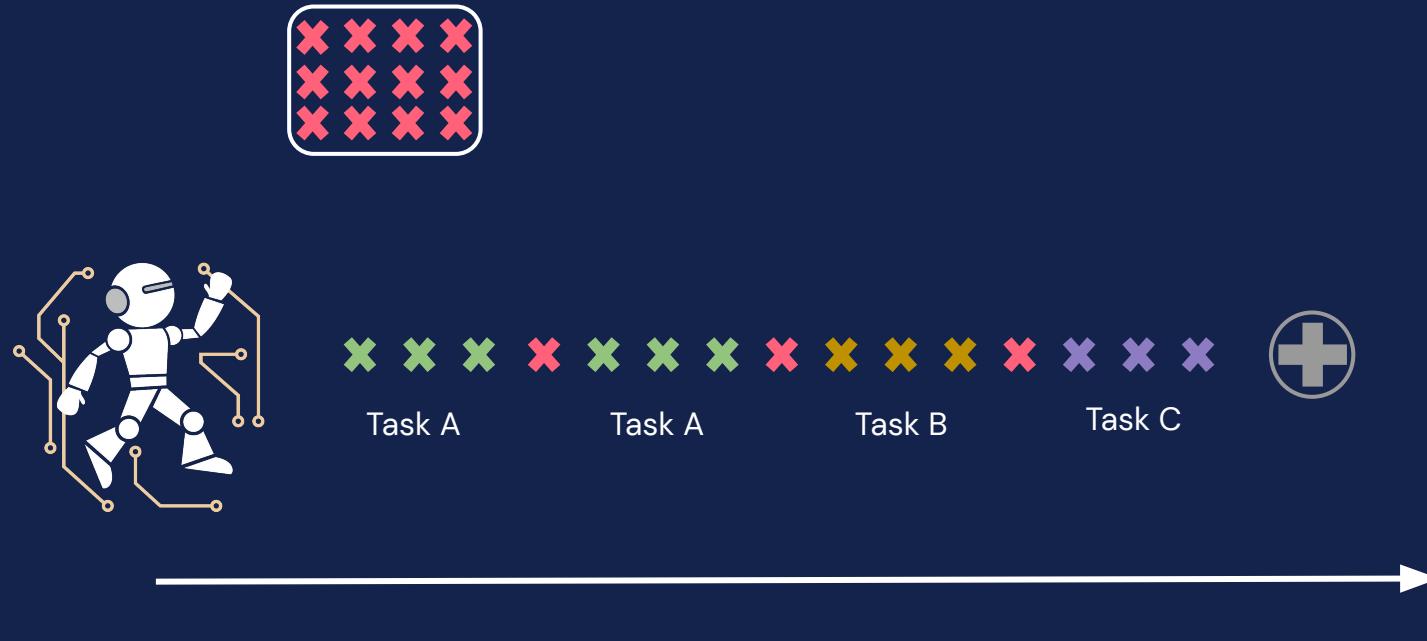
# Training

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



# Inference (Prediction)

Local adaptation (Sprechmann et al., 2018).



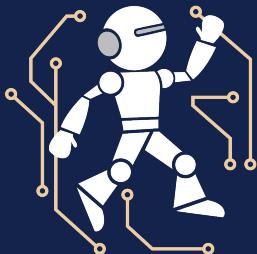
# Inference (Prediction)

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?



Task A

Task A

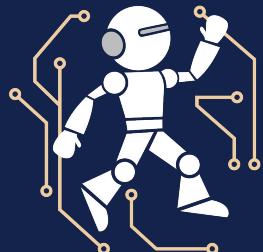
Task B

Task C



# 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?

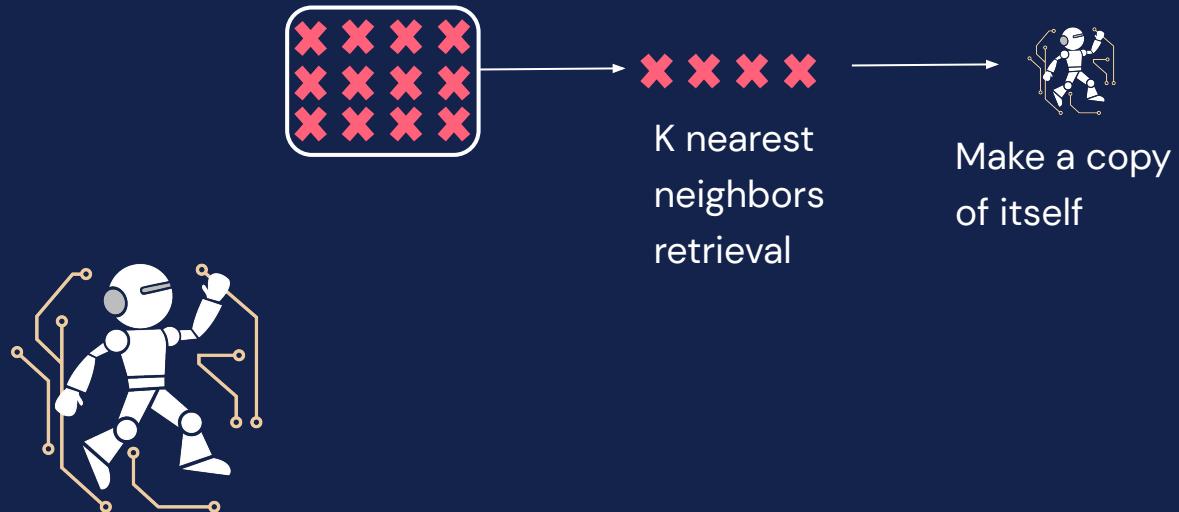
In what country is Spoleto located?

In what part of Africa is Palermo located?



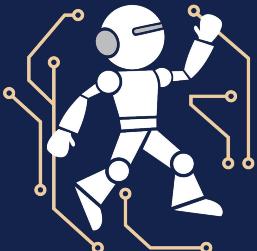
# Inference (Prediction)

Local adaptation (Sprechmann et al., 2018).



# Inference (Prediction)

Local adaptation (Sprechmann et al., 2018).

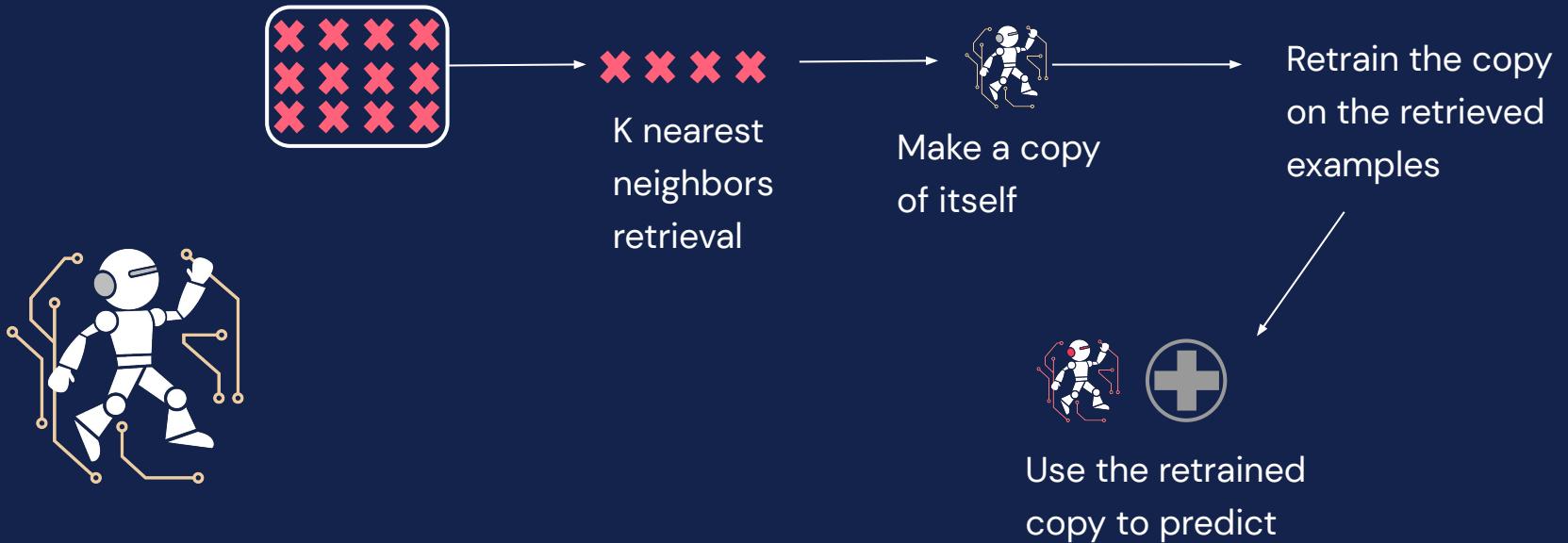


$$\mathbf{W}_i = \arg \min_{\tilde{\mathbf{W}}} \lambda \|\tilde{\mathbf{W}} - \mathbf{W}\|_2^2 - \sum_{k=1}^K \alpha_k \log p(y_i^k \mid \mathbf{x}_i^k; \tilde{\mathbf{W}})$$



# 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	<b>62.4</b>

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

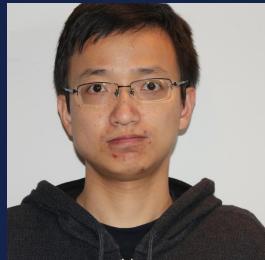
- Episodic memory allows a language model to deal with changes in data distribution.
- Linear **space complexity** in the number of examples, **constant** is more realistic.

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



# A Mutual Information Maximization Perspective of Language Representation Learning

Kong et al., ICLR 2020



Lingpeng



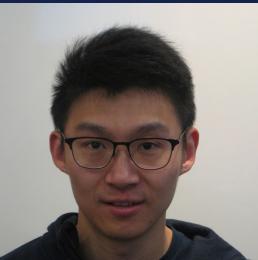
Cyprien



Wang



Lei



Zihang



Dani

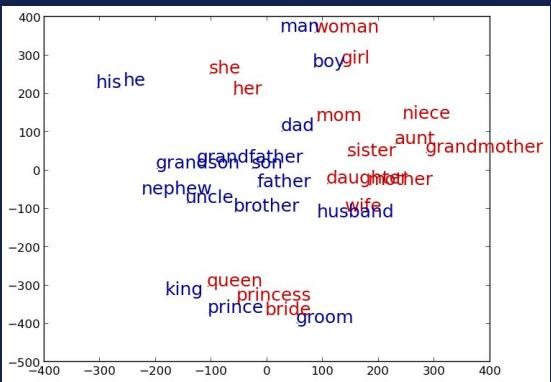


# Text Representations



<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.

XLNet, Yang et al., 2019.

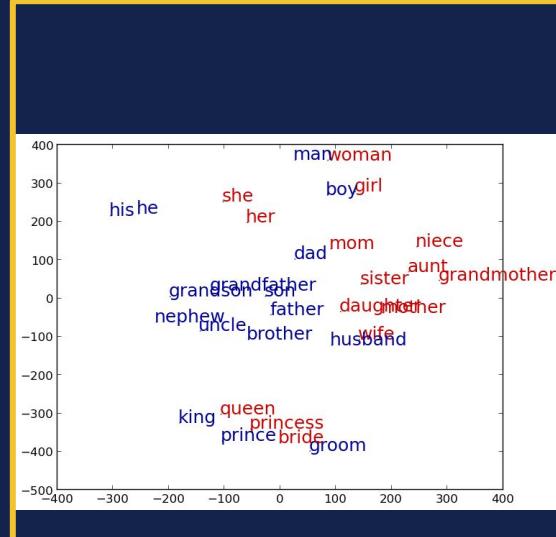


# Text Representations



[https://twitter.com/SmithaMilli/status/837153616116985856/](https://twitter.com/SmithaMilli/status/837153616116985856)

**Bag of words**



**Hypothesis:** these methods are different instances of one framework.

**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.

XLNet, Yang 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).

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

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).

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

InfoNCE objective  
Logeswaran and Lee, 2018  
van den Oord, et al., 2019



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

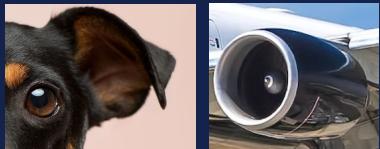


# Contrastive Learning with InfoNCE

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

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

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

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

The National University of Singapore is located in Singapore



# Contrastive Learning with InfoNCE

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

*a*

*b*

The National University of Singapore is located in Singapore



# Contrastive Learning with InfoNCE

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

*a*                    *b*                    *a*  
The National University of Singapore is located in Singapore



# Contrastive Learning with InfoNCE

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

*a* The National University of Singapore *b* is located in Singapore



# Contrastive Learning with InfoNCE

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

*a*                    *b*  
The National University of Singapore is located in Singapore

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

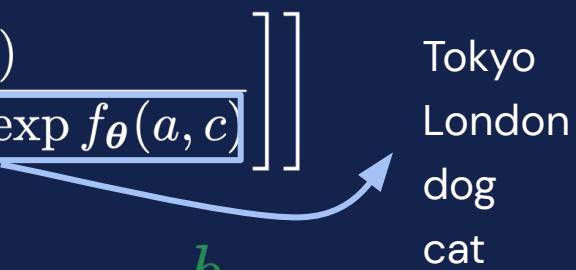


# Contrastive Learning with InfoNCE

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

*a*   *b*

The National University of Singapore is located in Singapore



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



# Contrastive Learning with InfoNCE

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

*a*                    *b*  
The National University of Singapore is located in Singapore

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



# Contrastive Learning with InfoNCE

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

*a*                    *b*  
The National University of Singapore is located in Singapore

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



# Contrastive Learning with InfoNCE

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

*a*                    *b*  
The National University of Singapore is located in Singapore

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



# Skip-gram

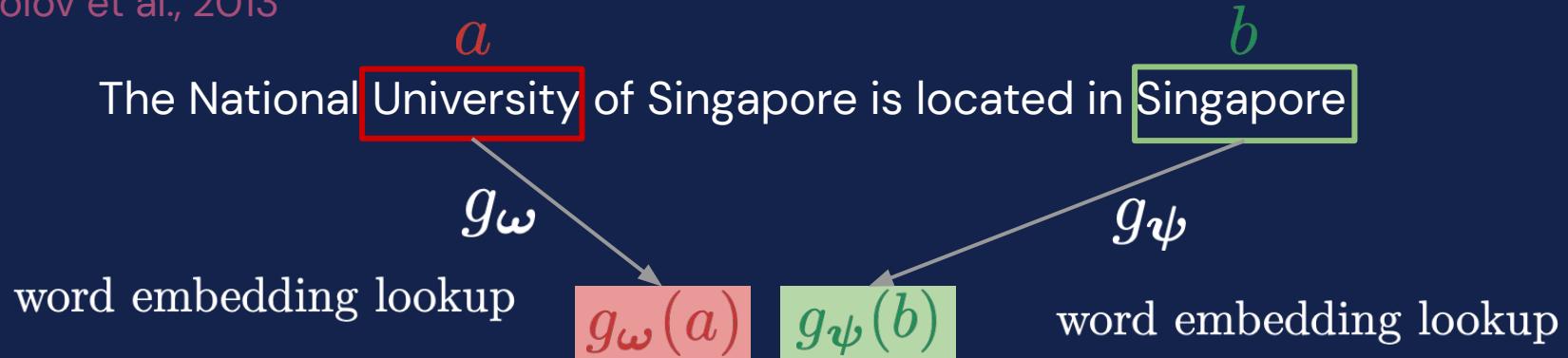
Mikolov et al., 2013

The National University of Singapore is located in Singapore



# Skip-gram

Mikolov et al., 2013



# BERT

Devlin et al., 2019

The National University of Singapore is located in Singapore



# BERT

Devlin et al., 2019

The National University of Singapore is located in  $b$  Singapore

$$g_{\psi}(b)$$

$$g_{\psi}$$

word embedding lookup



# BERT

Devlin et al., 2019

The National University of Singapore is located in Singapore

$a$

$g_{\omega}$

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.

		$a$	$b$	$g_\omega$	$g\psi$
Mikolov et al., 2013	<b>Skip-gram</b>	word	word	lookup	lookup
Devlin et al., 2019	<b>BERT</b>	context	word	transformer	lookup
Yang et al., 2019	<b>XLNet</b>	context	word	TXL++	lookup



# Why is this interesting?

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

		$a$	$b$	$g_\omega$	$g_\psi$
Mikolov et al., 2013	<b>Skip-gram</b>	word	word	lookup	lookup
Devlin et al., 2019	<b>BERT</b>	context	word	transformer	lookup
Yang et al., 2019	<b>XLNet</b>	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.

		$a$	$b$	$g_\omega$	$g_\psi$
Mikolov et al., 2013	<b>Skip-gram</b>	word	word	lookup	lookup
Devlin et al., 2019	<b>BERT</b>	context	word	transformer	lookup
Yang et al., 2019	<b>XLNet</b>	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)

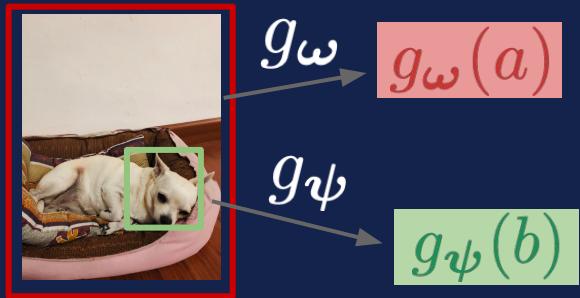


$$g_{\omega} \rightarrow g_{\omega}(a)$$



# Model

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



# Model

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



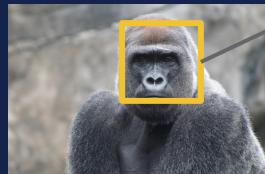
$$g_{\omega} \rightarrow g_{\omega}(a)$$

$$g_{\psi} \rightarrow g_{\psi}(b)$$



$$g_{\psi} \rightarrow g_{\psi}(c_1)$$

$$g_{\psi} \rightarrow g_{\psi}(c_2)$$



# Model

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



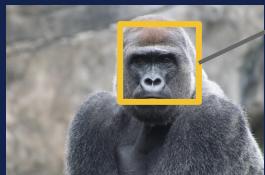
$$g_{\omega} \rightarrow g_{\omega}(a)$$

$$g_{\psi} \rightarrow g_{\psi}(b)$$



$$g_{\psi} \rightarrow g_{\psi}(c_1)$$

$$g_{\psi} \rightarrow g_{\psi}(c_2)$$

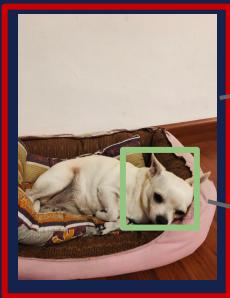


$$\mathcal{I}_{\text{DIM}} = \mathbb{E}_{p(A,B)} \left[ \mathbb{E}_{p(C)} \left[ \log \frac{\exp[g_{\omega}(a)^\top g_{\psi}(b)]}{\exp[g_{\omega}(a)^\top g_{\psi}(b)] + \sum_{c \neq b} \exp[g_{\omega}(a)^\top g_{\psi}(c)]} \right] \right]$$



# Model

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



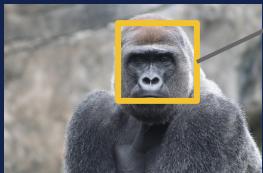
$$g_{\omega} \rightarrow g_{\omega}(a)$$

$$g_{\psi} \rightarrow g_{\psi}(b)$$



$$g_{\psi} \rightarrow g_{\psi}(c_1)$$

$$g_{\psi} \rightarrow g_{\psi}(c_2)$$

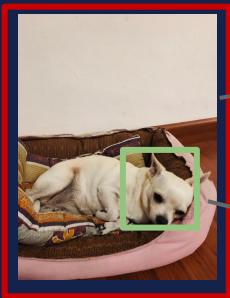


NUS is located in Singapore



# Model

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



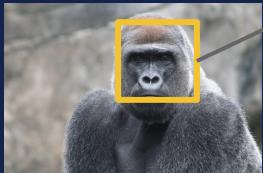
$$g_{\omega} \rightarrow g_{\omega}(a)$$

$$g_{\psi} \rightarrow g_{\psi}(b)$$



$$g_{\psi} \rightarrow g_{\psi}(c_1)$$

$$g_{\psi} \rightarrow g_{\psi}(c_2)$$



NUS is located in Singapore

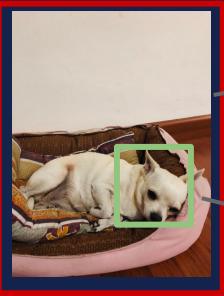
$g_{\psi}$  transformer

$$g_{\psi}(b)$$



# Model

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



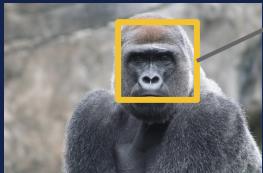
$$g_{\omega} \rightarrow g_{\omega}(a)$$

$$g_{\psi} \rightarrow g_{\psi}(b)$$



$$g_{\psi} \rightarrow g_{\psi}(c_1)$$

$$g_{\psi} \rightarrow g_{\psi}(c_2)$$



NUS is located in Singapore

$g_{\omega}$   
transformer

$$g_{\omega}(a) \quad g_{\psi}(b)$$



# Model

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



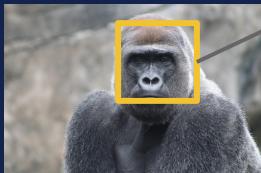
$$g_{\omega} \rightarrow g_{\omega}(a)$$

$$g_{\psi} \rightarrow g_{\psi}(b)$$



$$g_{\psi} \rightarrow g_{\psi}(c_1)$$

$$g_{\psi} \rightarrow g_{\psi}(c_2)$$



NUS is located in Singapore

$$g_{\omega}(a) \quad g_{\psi}(b)$$

Starcraft II is a fun game

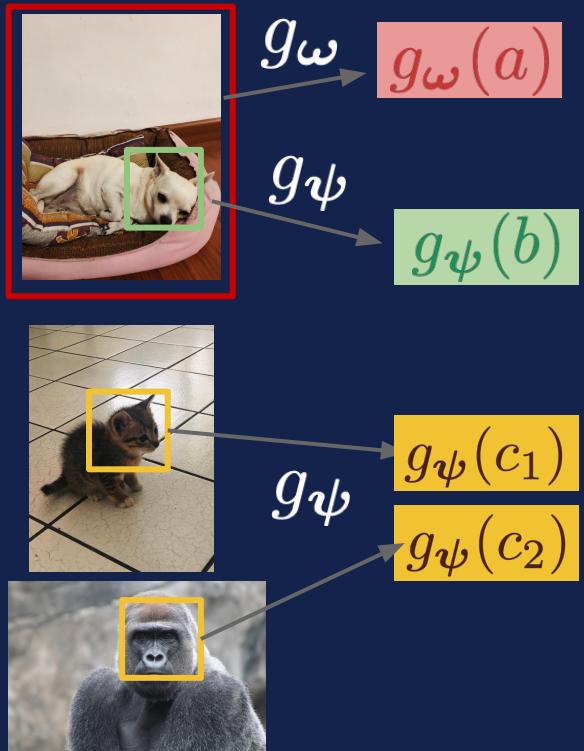
Cristiano Ronaldo scores an own goal

Machine learning is transforming drug discovery

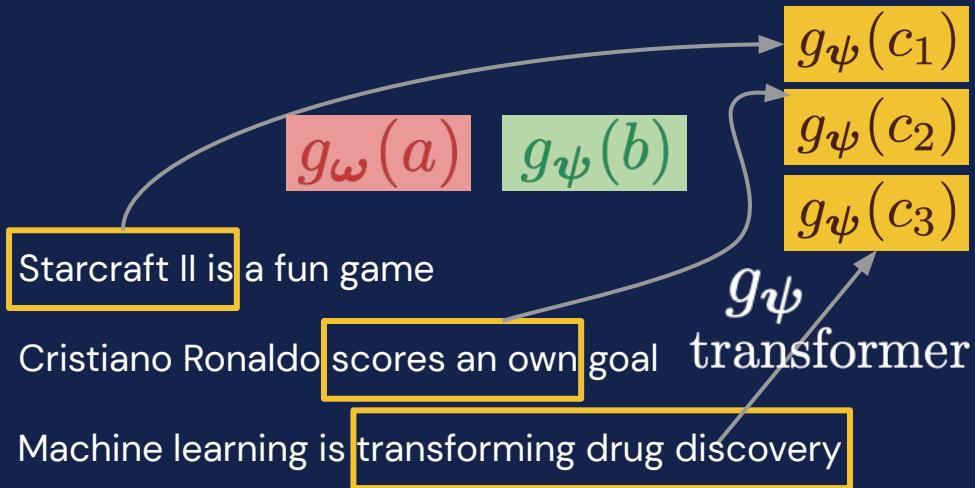


# Model

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

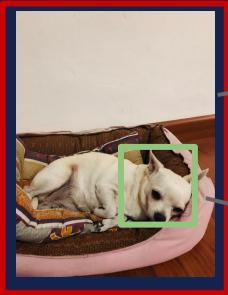


NUS is located in Singapore



# Model

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



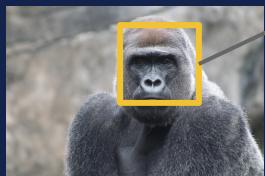
$$g_{\omega} \rightarrow g_{\omega}(a)$$

$$g_{\psi} \rightarrow g_{\psi}(b)$$



$$g_{\psi} \rightarrow g_{\psi}(c_1)$$

$$g_{\psi} \rightarrow g_{\psi}(c_2)$$



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

NUS is located in Singapore

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

$$\begin{bmatrix} g_{\psi}(c_1) \\ g_{\psi}(c_2) \\ g_{\psi}(c_3) \end{bmatrix}$$

Starcraft II is a fun game

Cristiano Ronaldo scores an own goal

Machine learning is transforming drug discovery



# Experiments

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

		F1
Small Model	BERT	90.9
	Ours	<b>91.4</b>
Large Model	BERT	92.7
	Ours	<b>93.1</b>

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



# 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.

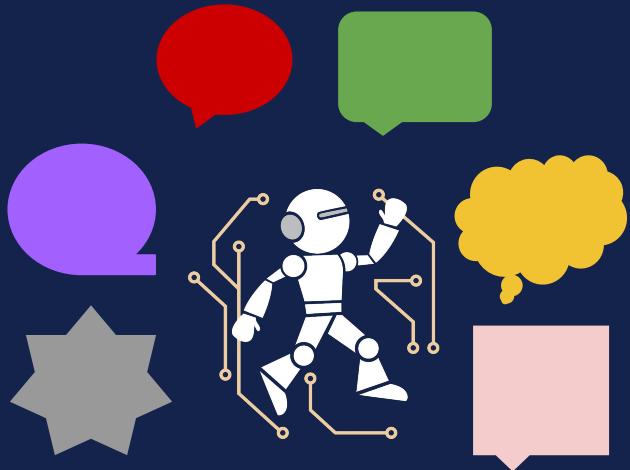
**Generative Models**



# 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.

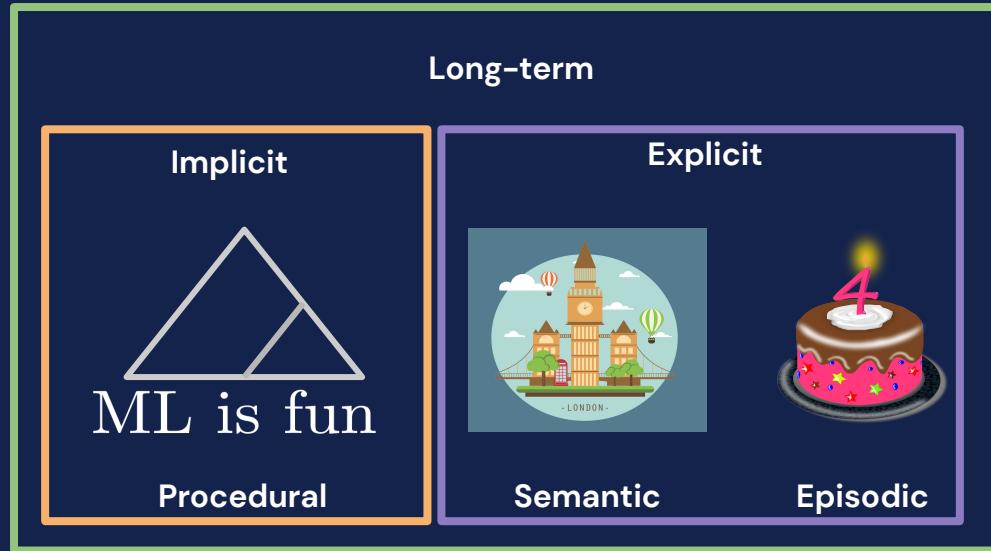
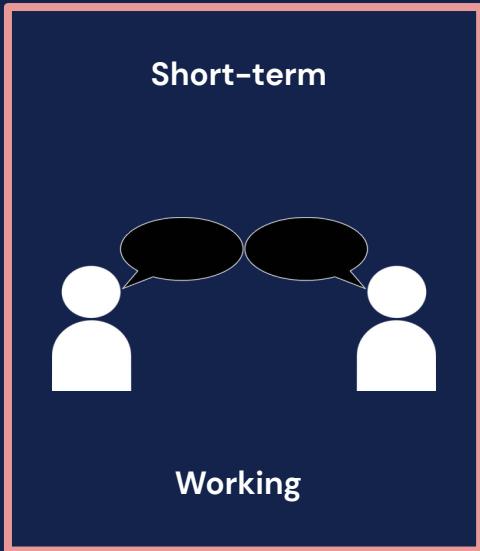
Memory



# Future Directions



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



# Future Directions



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

## Adaptive Semiparametric Language Models

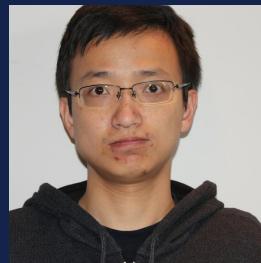
**Yogatama et al., in review**



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

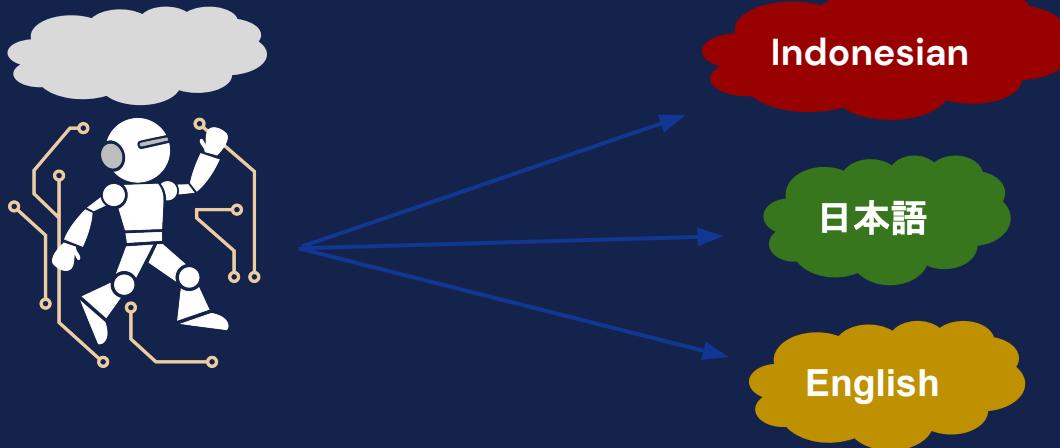


# Future Directions



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

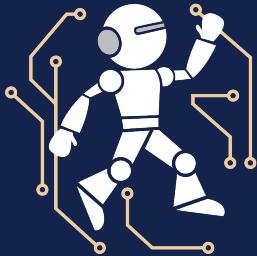


A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

Memory

Representation Learning

Generative Models



# Future Directions

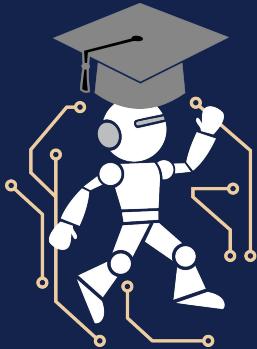


A language model that continually learns in an efficient way to perform multiple complex tasks in many languages.

Memory

Representation Learning

Generative Models



tack ՀԱՌԻԱԿԱԼՈՒԹՅՈՒՆ Danke  
ありがとうございます Salamat  
**grazie** **Thank you** multumesc  
ধন্যবাদ **Thank you** ଧନ୍ୟର୍ଥି  
Terima kasih Dankie 감사합니다 Merci  
Спасибо مکارکش σας ευχαριστώ  
teşekkür ederim 谢谢 cảm ơn bạn

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