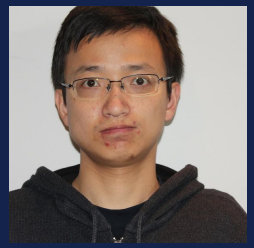
Semiparametric Language Models

Dani Yogatama


Cyprien


Lingpeng


Sebastian


Aida

DeepMind

Background

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

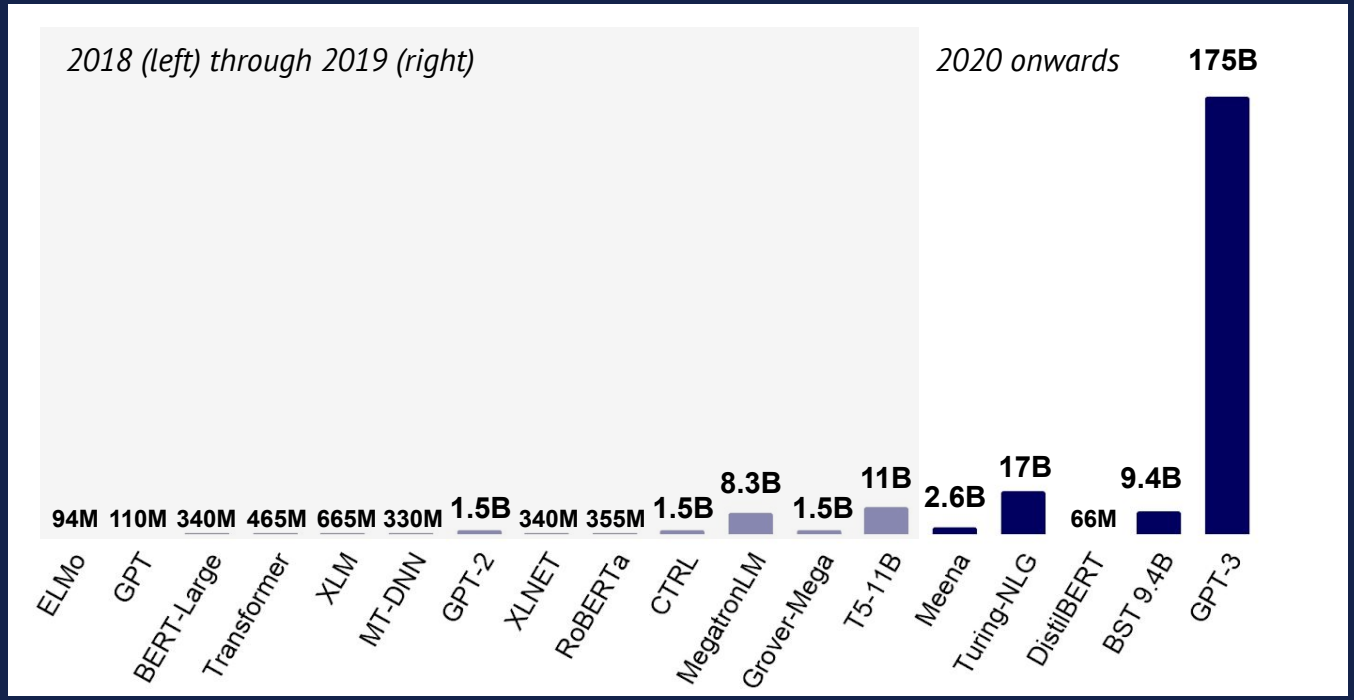
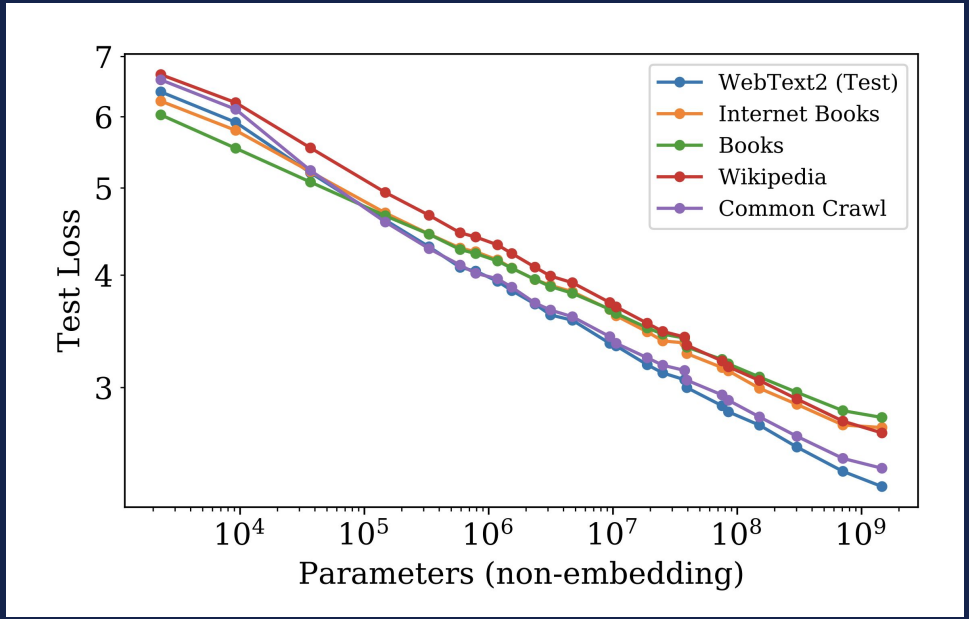


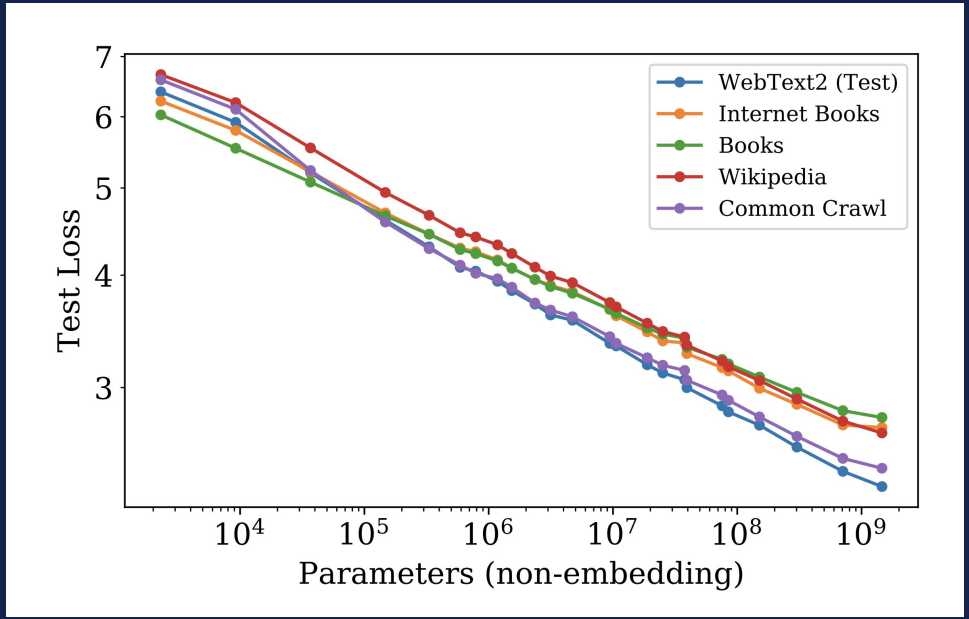
Figure taken from State of AI Report 2020.

Background



Kaplan et al., 2020

Background



Kaplan et al., 2020

Do we need structures and/or inductive biases?

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).

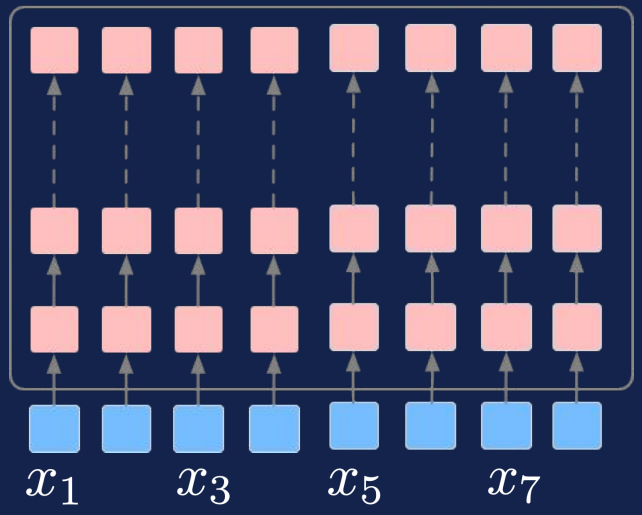
Dante was born in [MASK] .

Q: Where was Dante born in?

A:

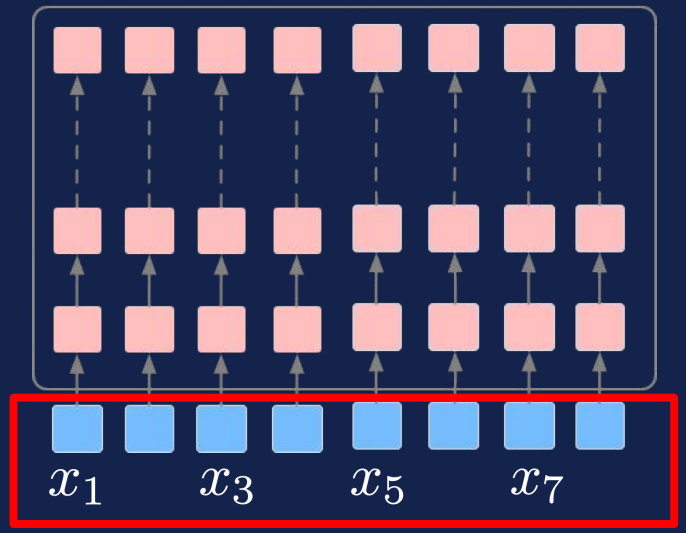
Background

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



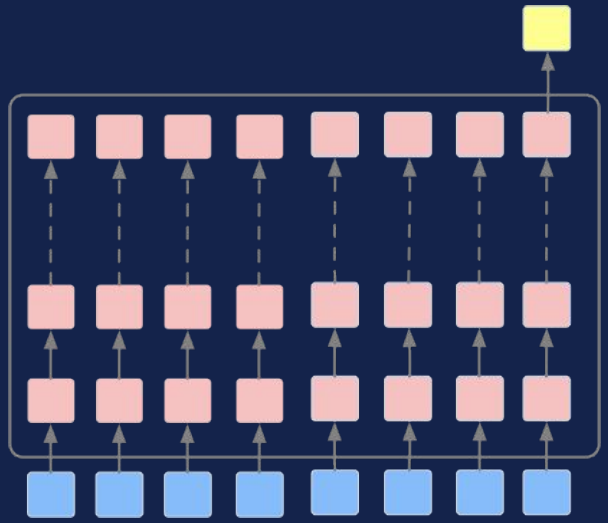
Background

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

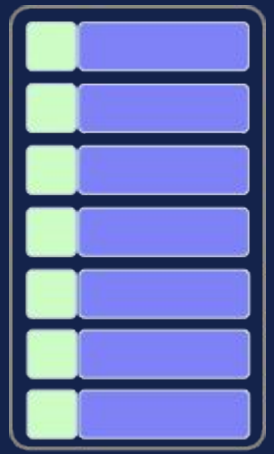


Semiparametric Language Models

Separation of computation and storage as an architectural bias.



Computation module



Storage (Memory)

Memory in AI

LSTM (Hochreiter and Schmidhuber, 1997)
Differentiable Neural Computers (Graves et al, 2016)
Reformer (Kitaev et al., 2020)
Transformer XL (Dai et al., 2019)
Never-Ending Language Learning (Mitchell et al, 2015)
Stack LSTM (Dyer et al, 2015; **Yogatama et al., 2018**)
Memory Networks (Weston et al, 2015)
*k*NN LM (Khandelwal et al, 2019)
Matching Networks (Vinyals et al, 2016)

Memory in Humans

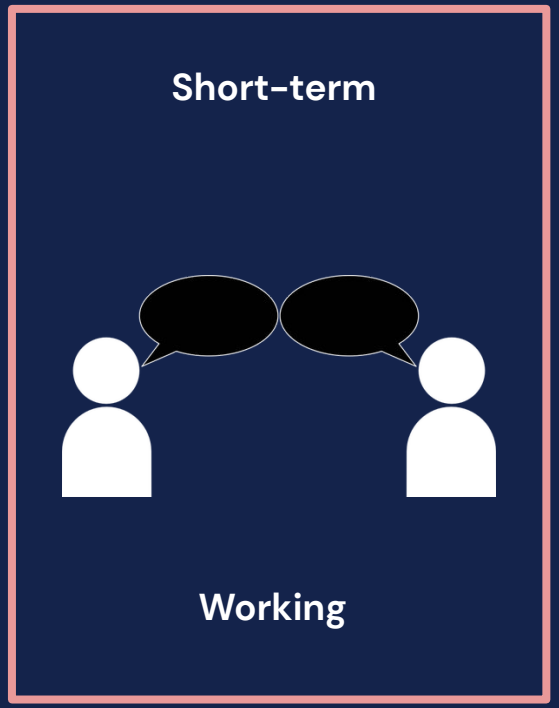
Human language processing is facilitated by specialized memory systems.

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

Memory in Humans

Human language processing is facilitated by specialized memory systems.

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

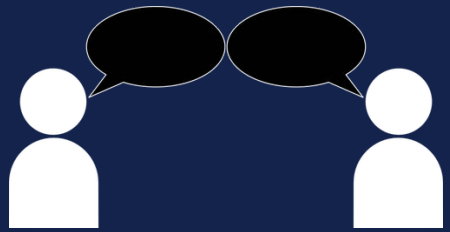


Memory in Humans

Human language processing is facilitated by specialized memory systems.

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

Short-term



Working

Long-term

Implicit

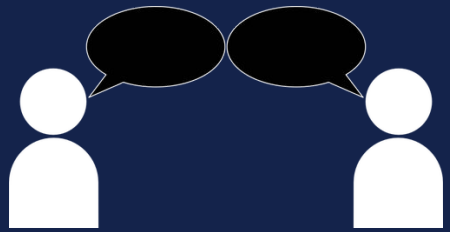
Procedural

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

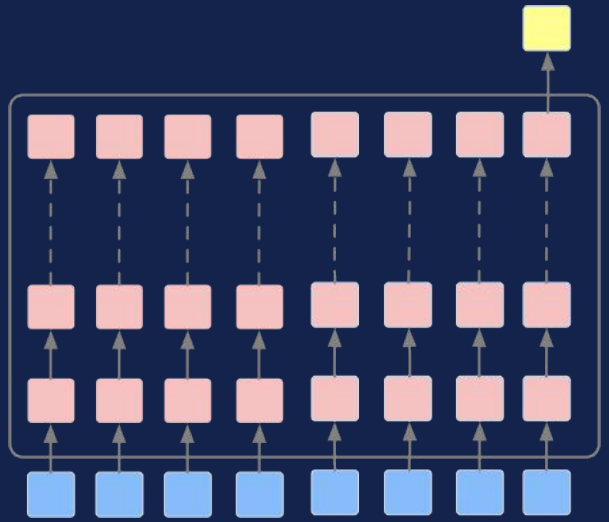
Memory in AI

LSTM (Hochreiter and Schmidhuber, 1997)
Differentiable Neural Computers (Graves et al, 2016)
Reformer (Kitaev et al., 2020)
Transformer XL (Dai et al., 2019)
Never-Ending Language Learning (Mitchell et al, 2015)
Stack LSTM (Dyer et al, 2015; **Yogatama et al., 2018**)
Memory Networks (Weston et al, 2015)
*k*NN LM (Khandelwal et al, 2019)
Matching Networks (Vinyals et al, 2016)

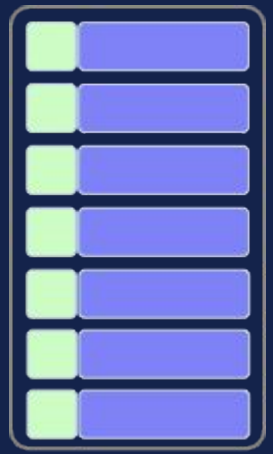
Memory in AI

LSTM (Hochreiter and Schmidhuber, 1997)	**Working**
Differentiable Neural Computers (Graves et al, 2016)	**Working**
Reformer (Kitaev et al., 2020)	**Working**
Transformer XL (Dai et al., 2019)	**Working**
Never-Ending Language Learning (Mitchell et al, 2015)	**Semantic**
Stack LSTM (Dyer et al, 2015; **Yogatama et al., 2018**)	**Procedural**
Memory Networks (Weston et al, 2015)	**Episodic**
*k*NN LM (Khandelwal et al, 2019)	**Episodic**
Matching Networks (Vinyals et al, 2016)	**Episodic**

This Talk

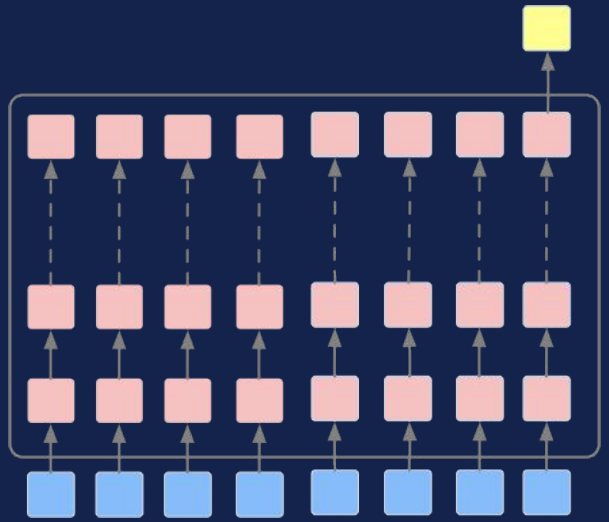


Computation module

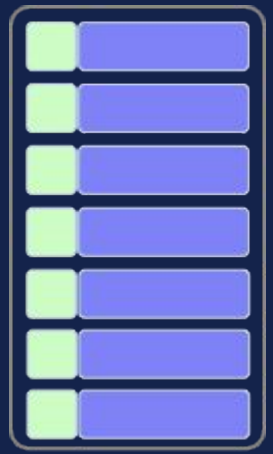


Storage (Memory)

This Talk



Computation module

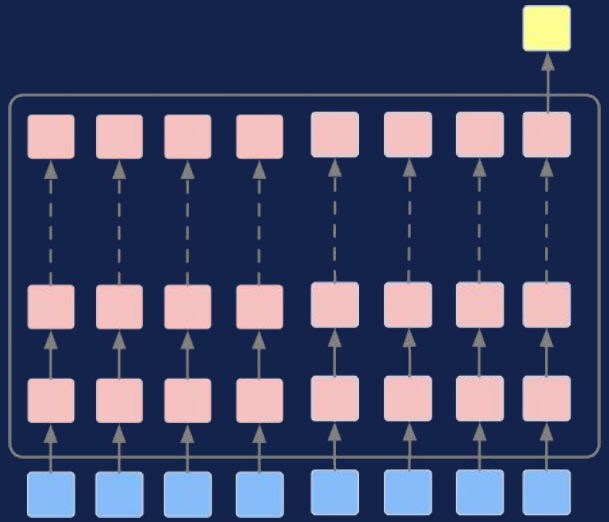


Storage (Memory)

Episodic memory in lifelong language learning.

de Masson d'Autume et al., NeurIPS 2019

This Talk



Computation module

Storage (Memory)

Adaptive semiparametric language models.

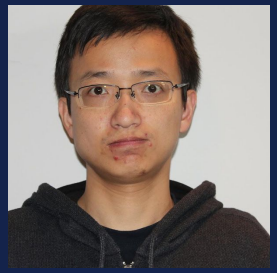
Yogatama et al., in review

Episodic Memory in Lifelong Language Learning

de Masson d'Autume et al., NeurIPS 2019

Cyprien

Sebastian

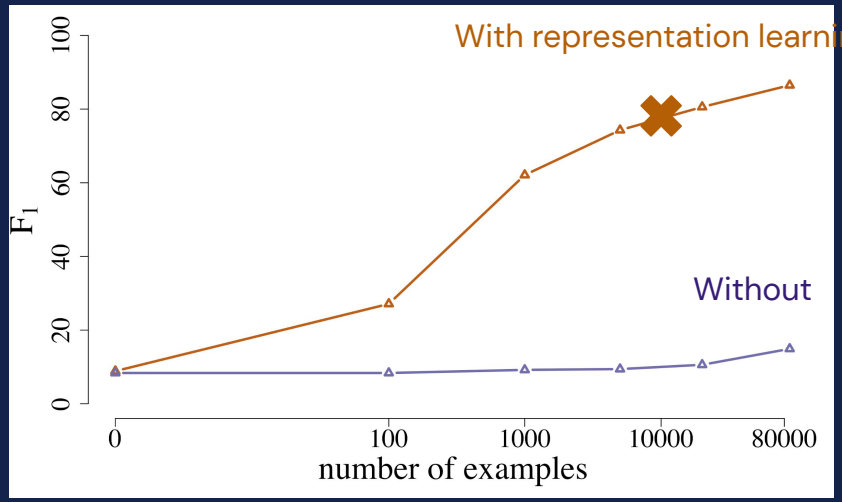


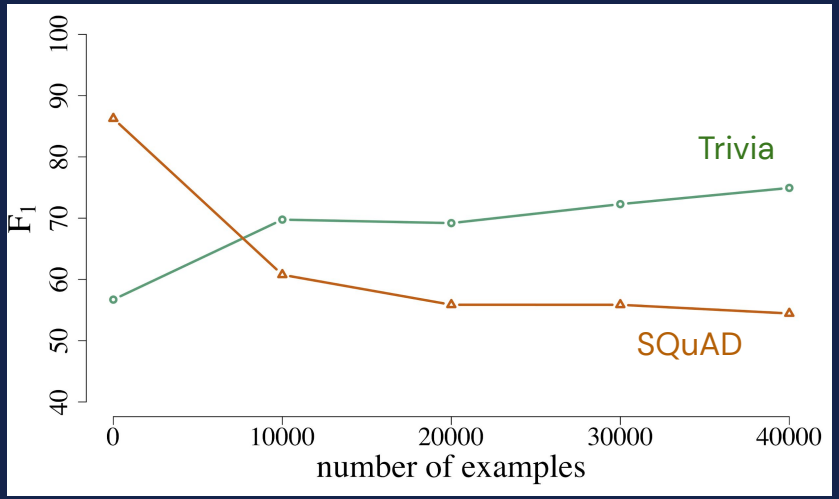
Lingpeng

Dani

Background

- Great progress, but current models 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

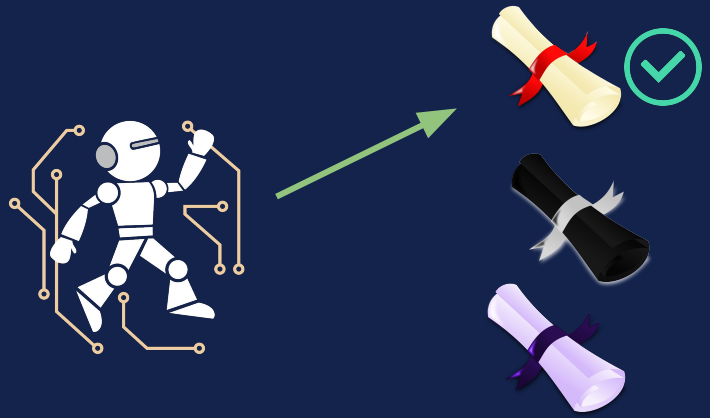
QA dataset 2: Trivia, Joshi et al., 2017

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).

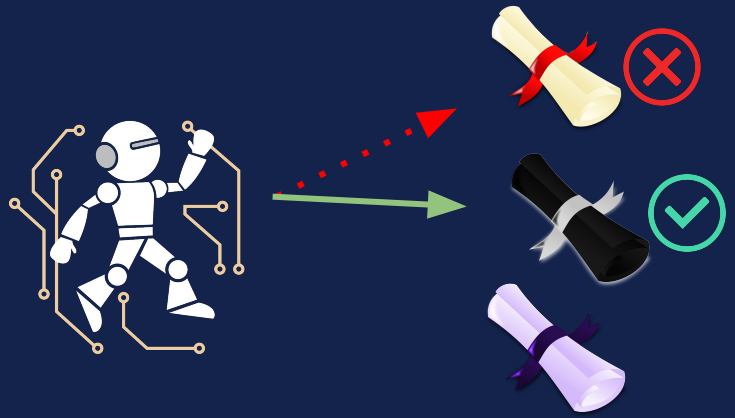
Background

- 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).



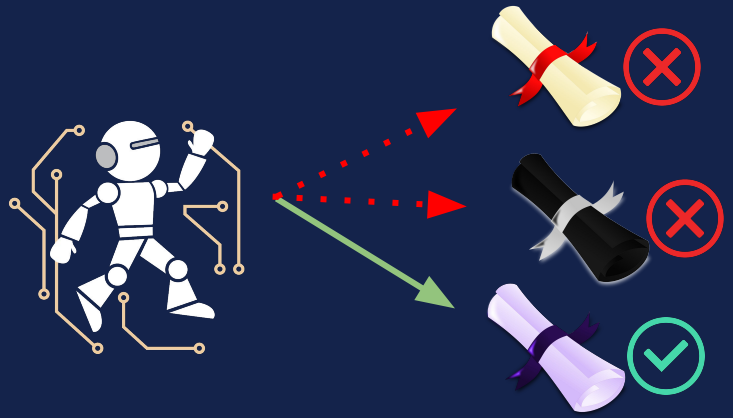
Background

- 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).



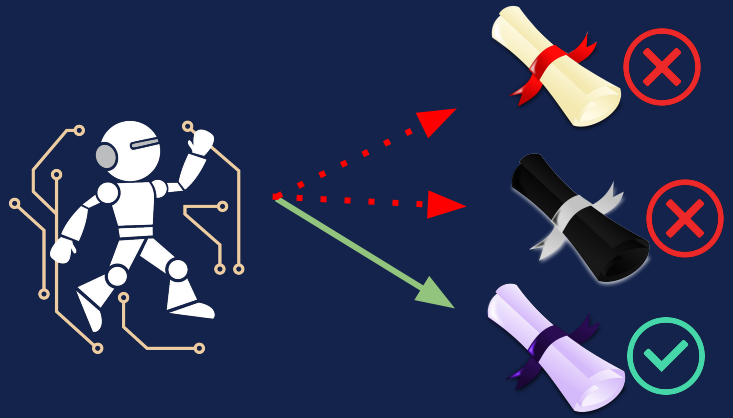
Background

- 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.

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

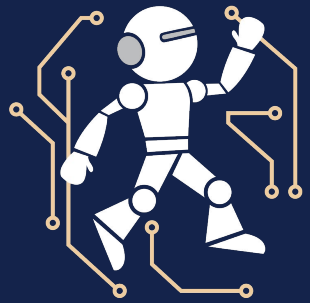
SQuAD: Rajpurkar et al., 2016

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?

Training

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.

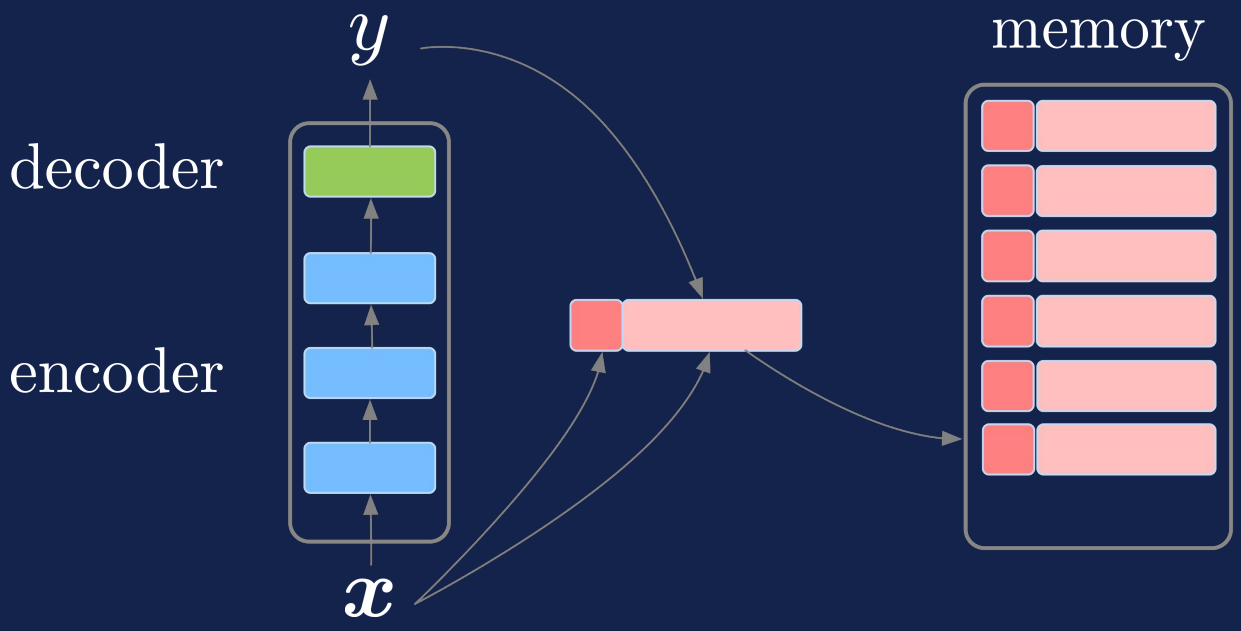
Why did people feel that way?

Problem Setup

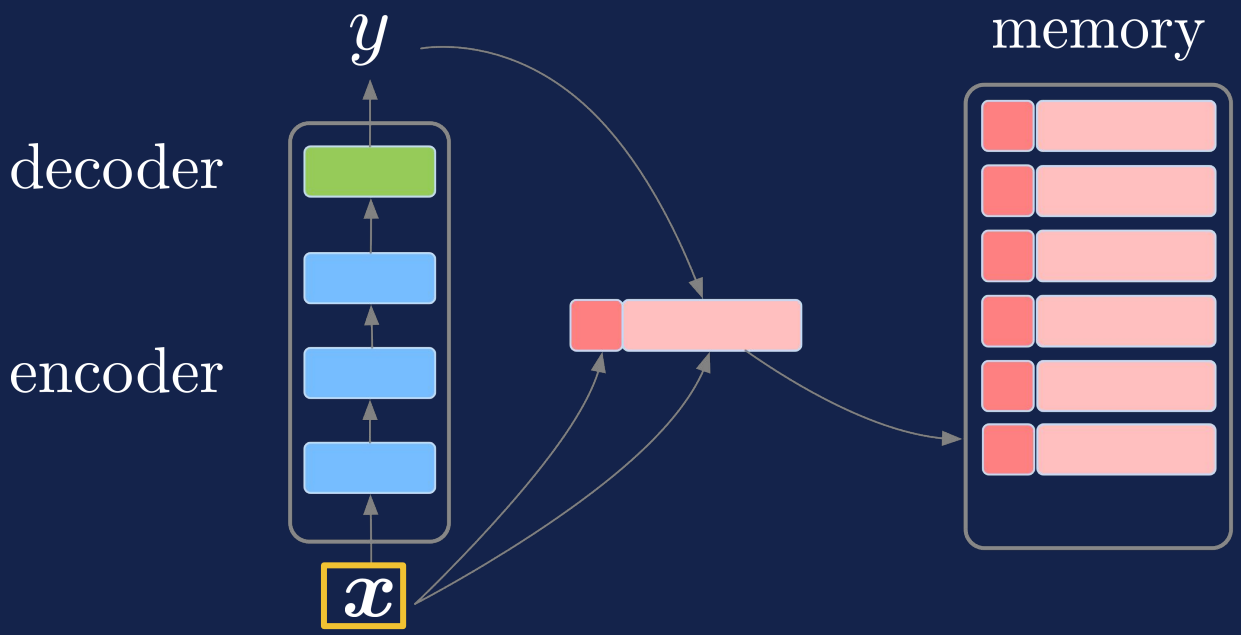
Training

Test

Question Answering Model



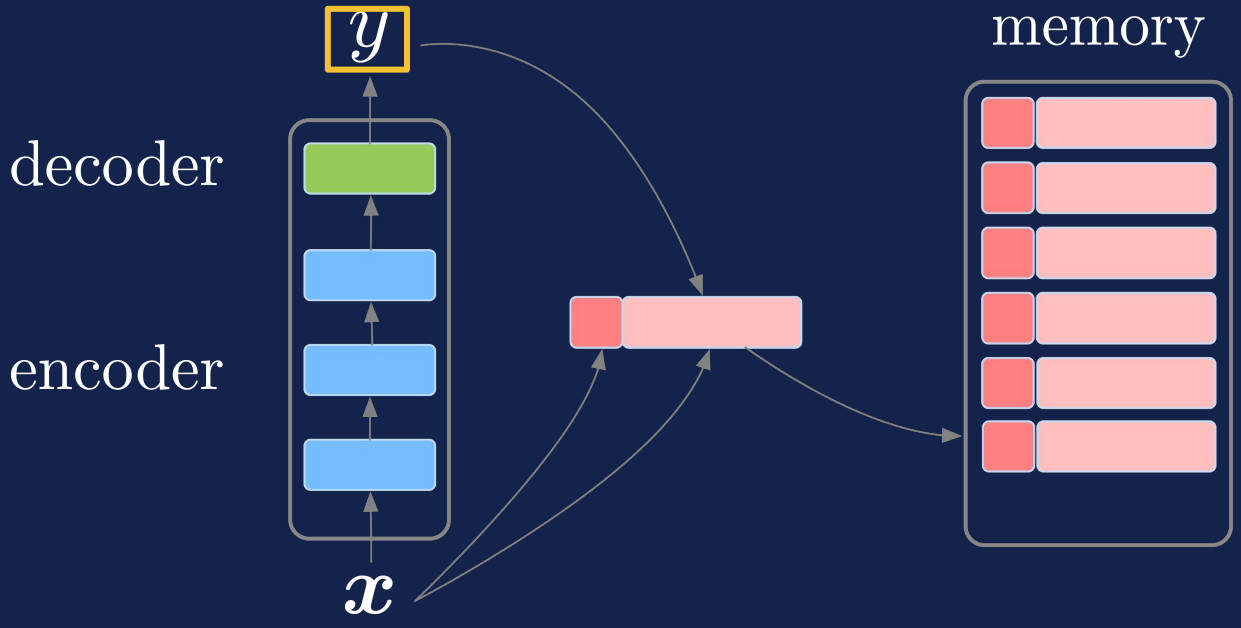
Question Answering Model



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

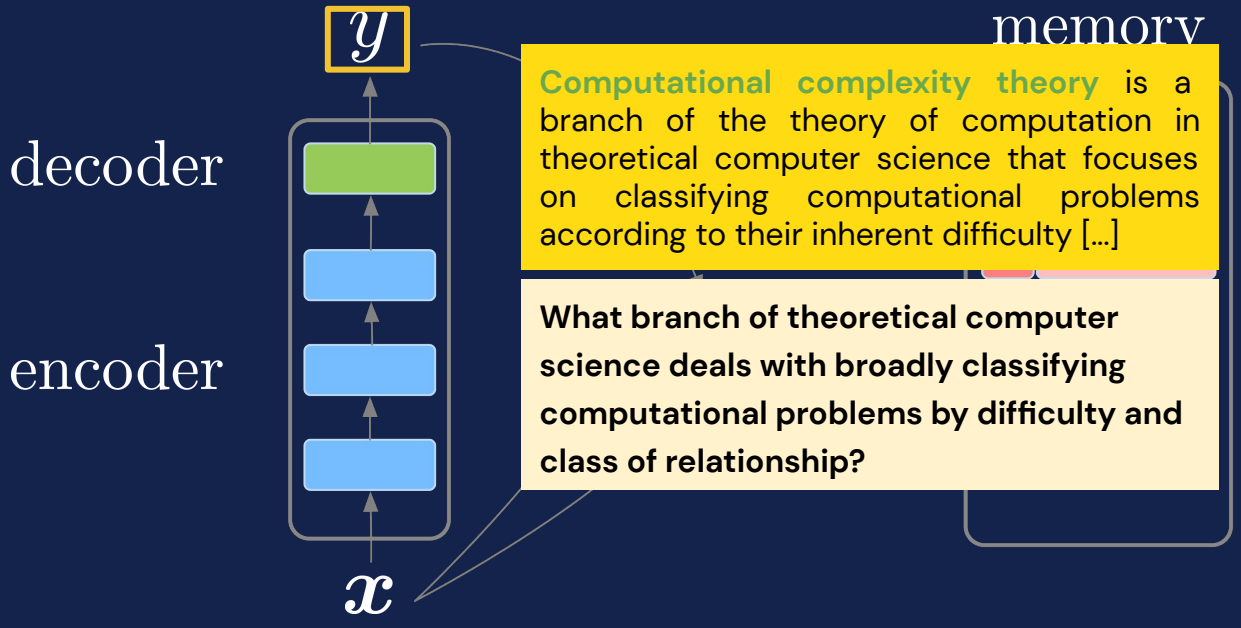
Question Answering Model

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

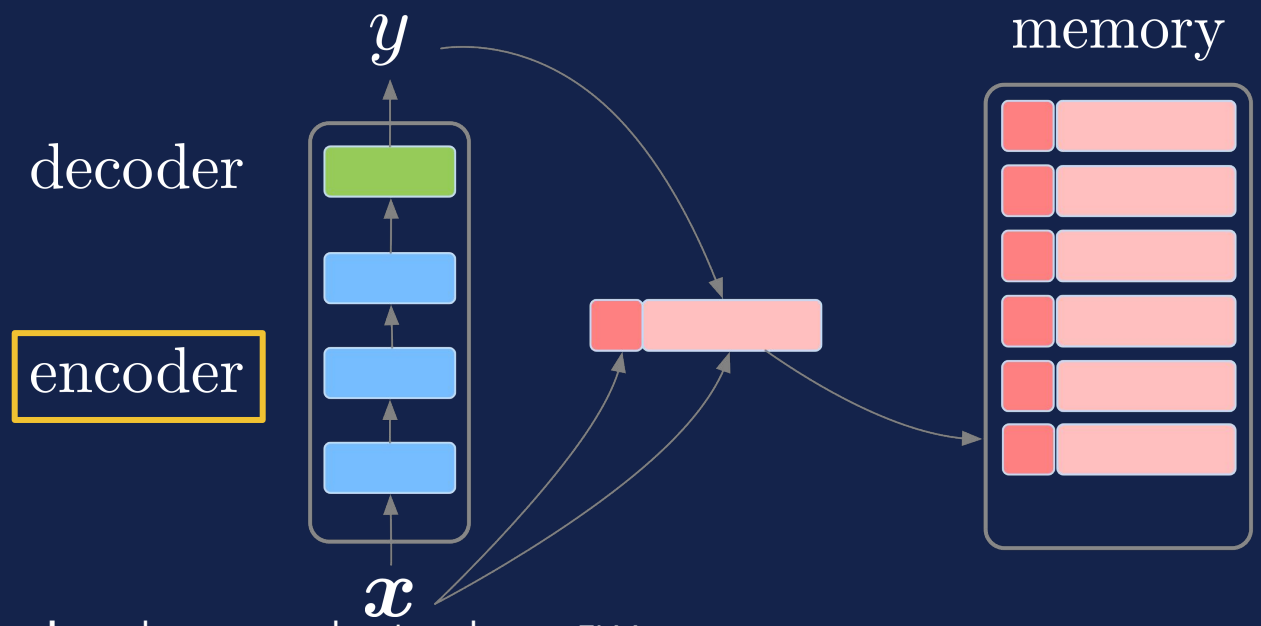


Question Answering Model

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



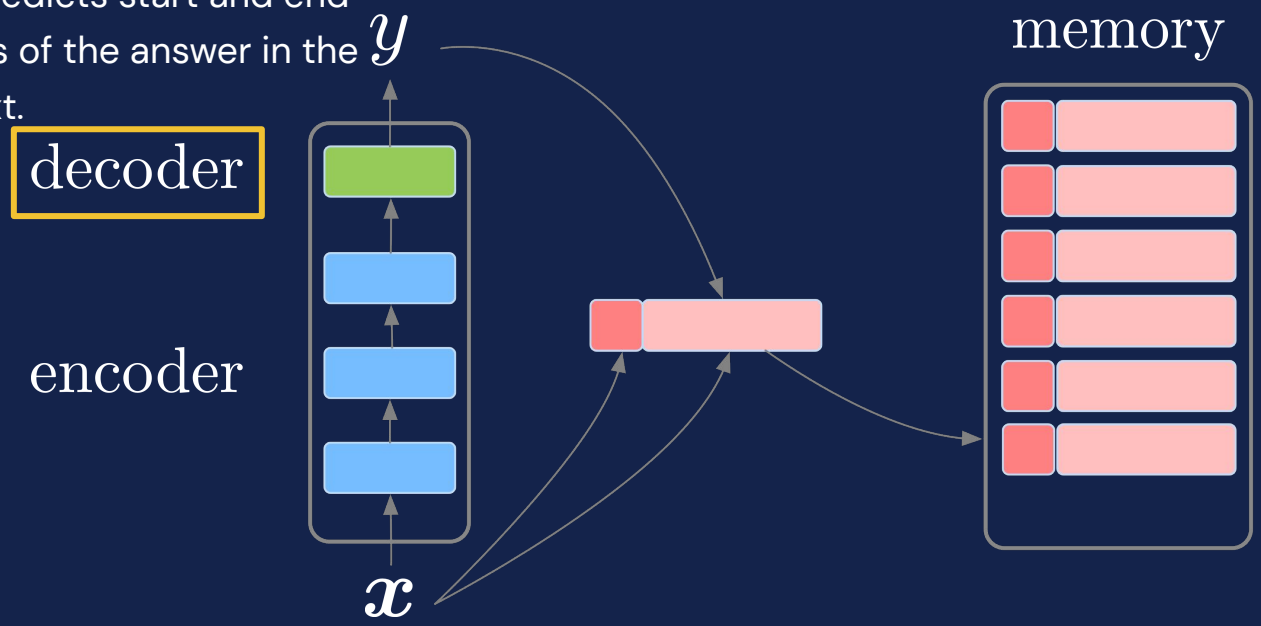
Question Answering Model



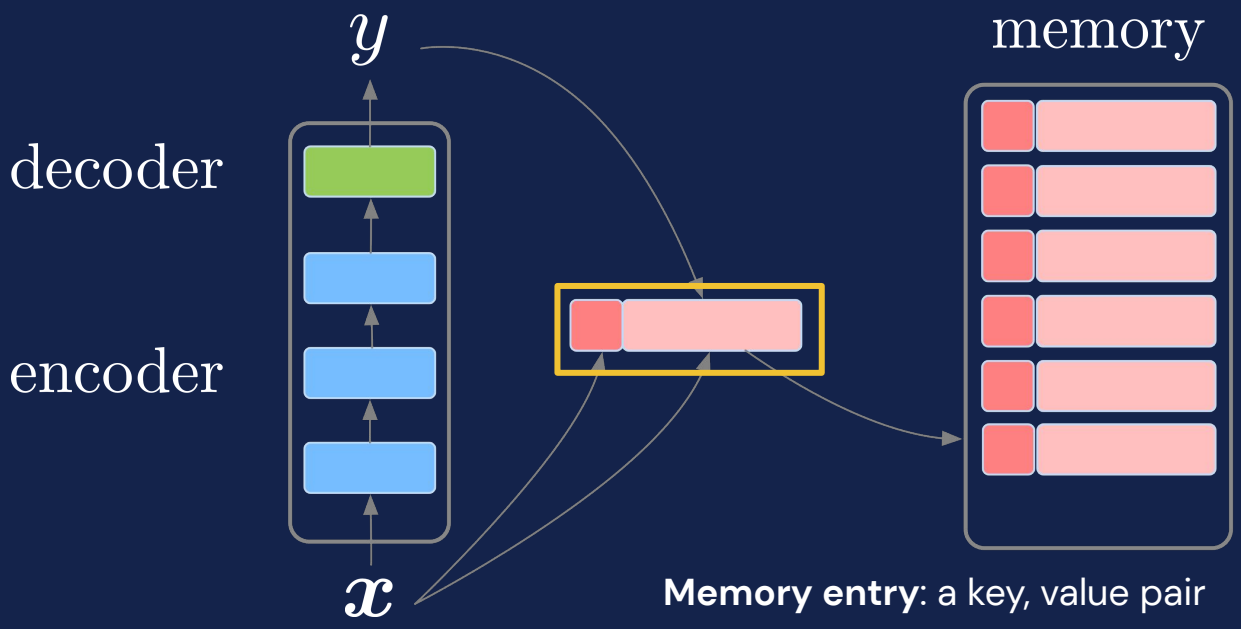
Encoder: a large neural network, e.g., ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), XLNet (Yang et al., 2019).

Question Answering Model

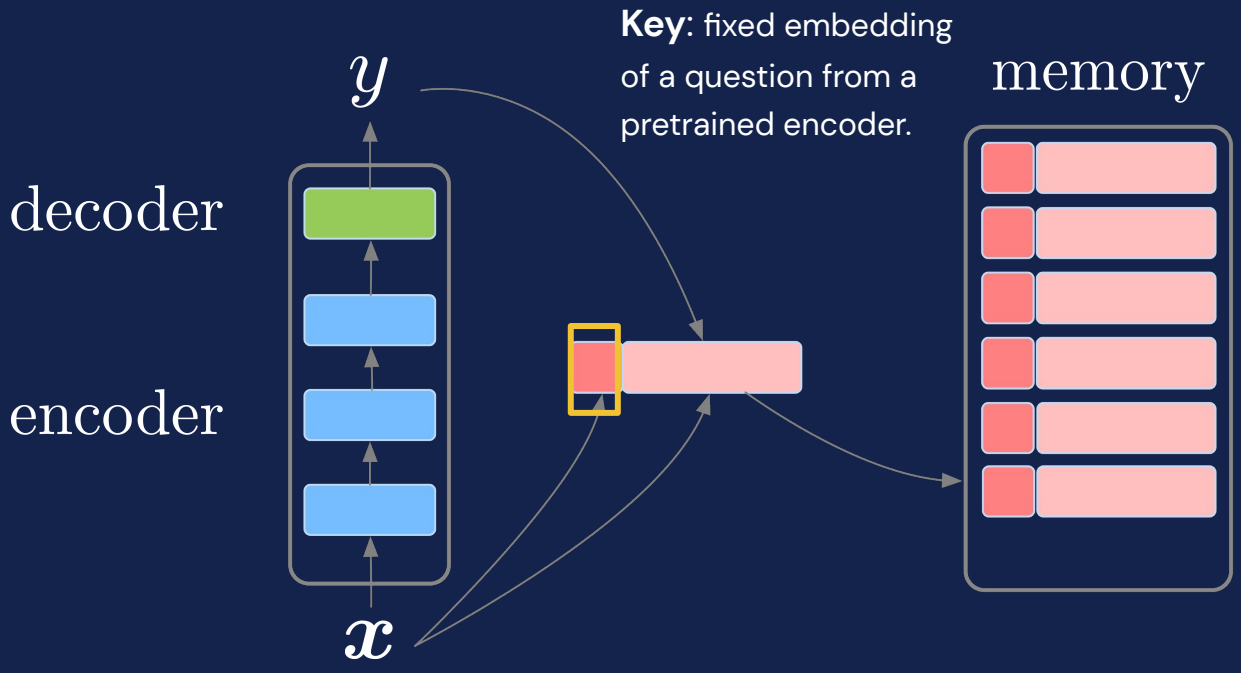
Decoder: a linear function that predicts start and end indices of the answer in the context.



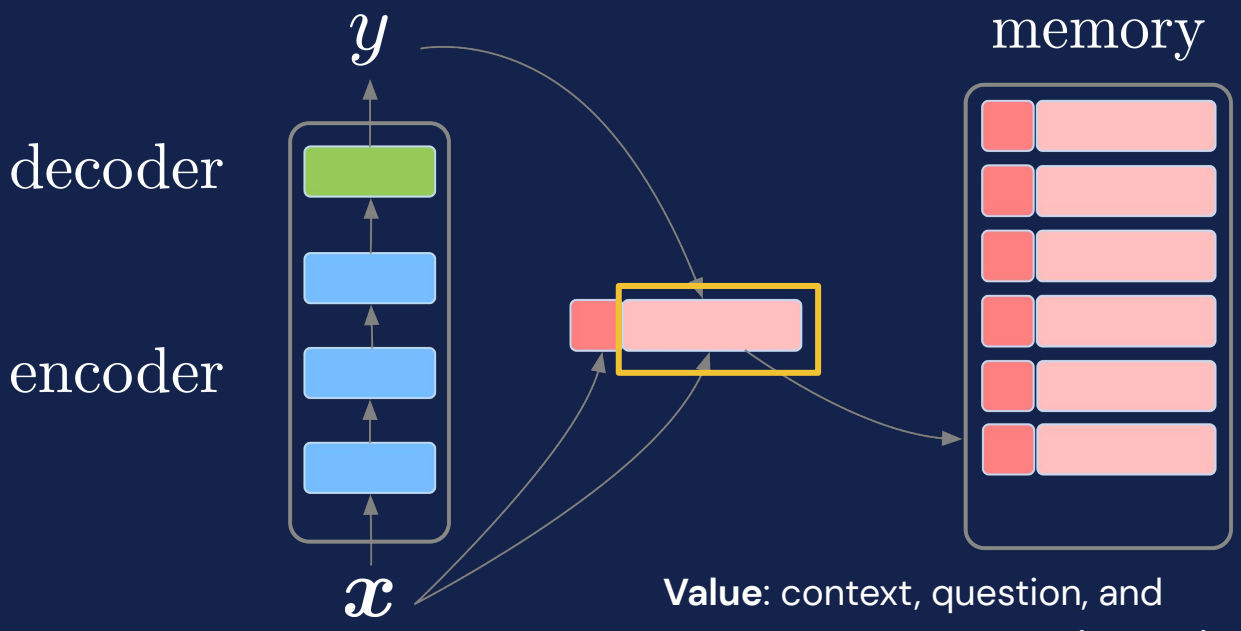
Question Answering Model



Question Answering Model



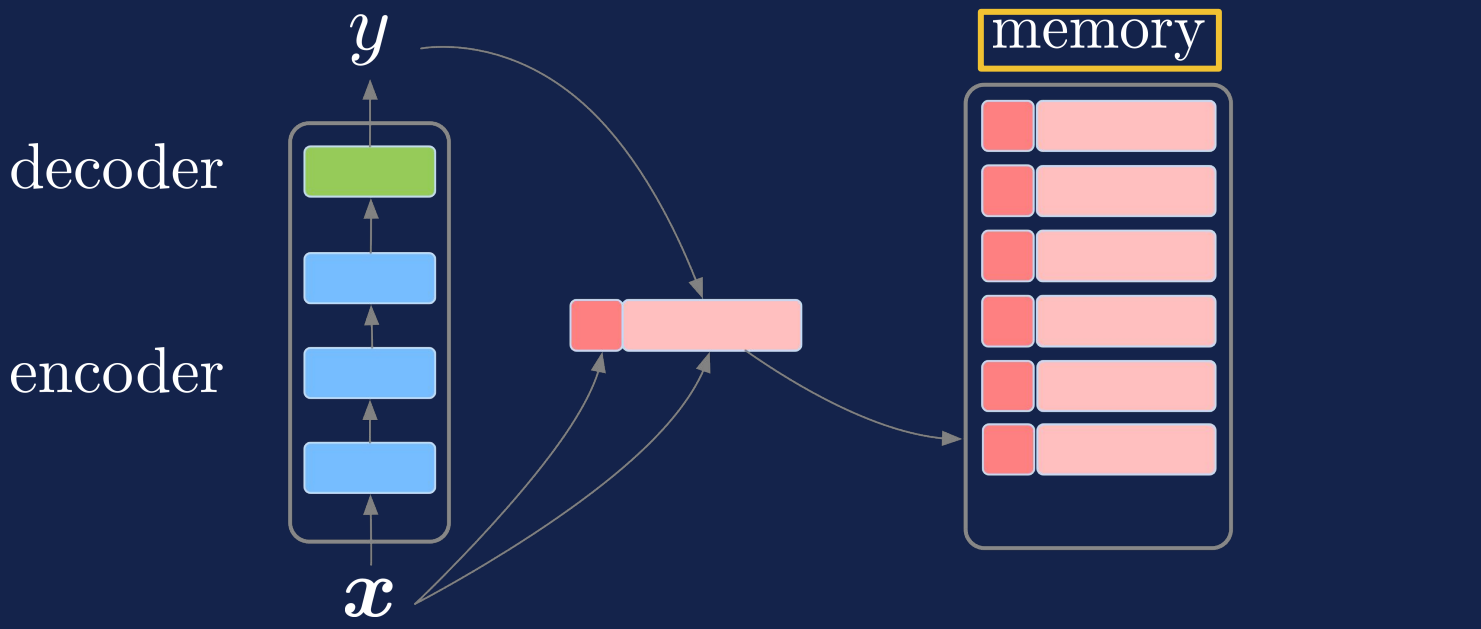
Question Answering Model



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

Question Answering Model

Memory: stores memory entries

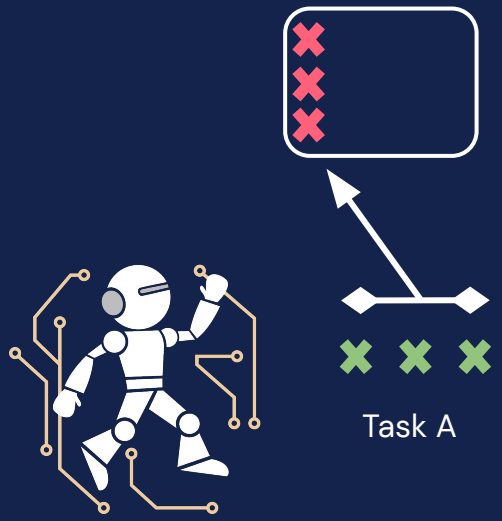


Training

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

Task A

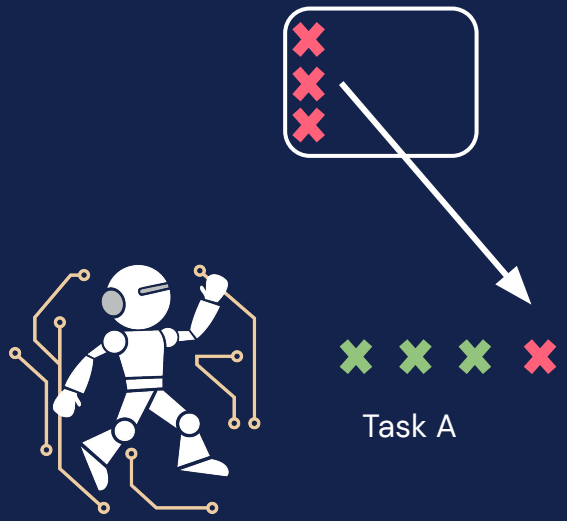
Training



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

Training

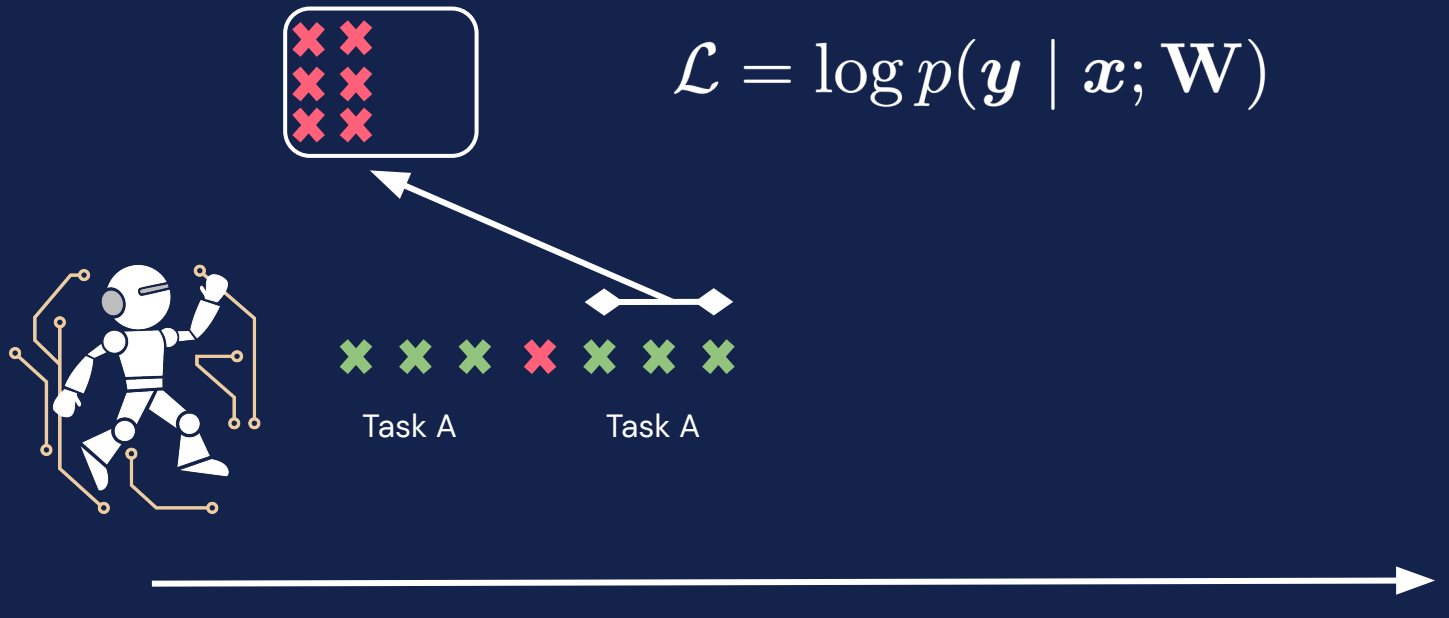
***Sparse* experience replay**: retrain on randomly sampled examples from the memory at a 1% rate.



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

Training

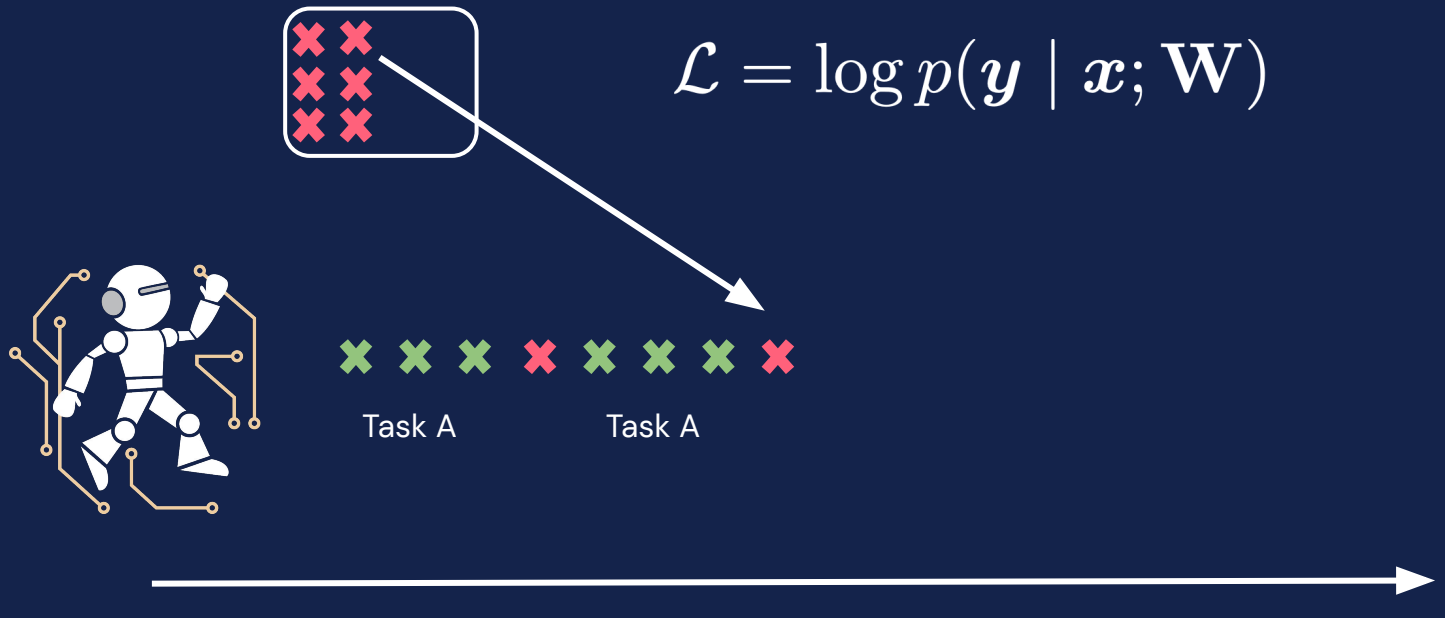
***Sparse* experience replay**: retrain on randomly sampled examples from the memory at a 1% rate.



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

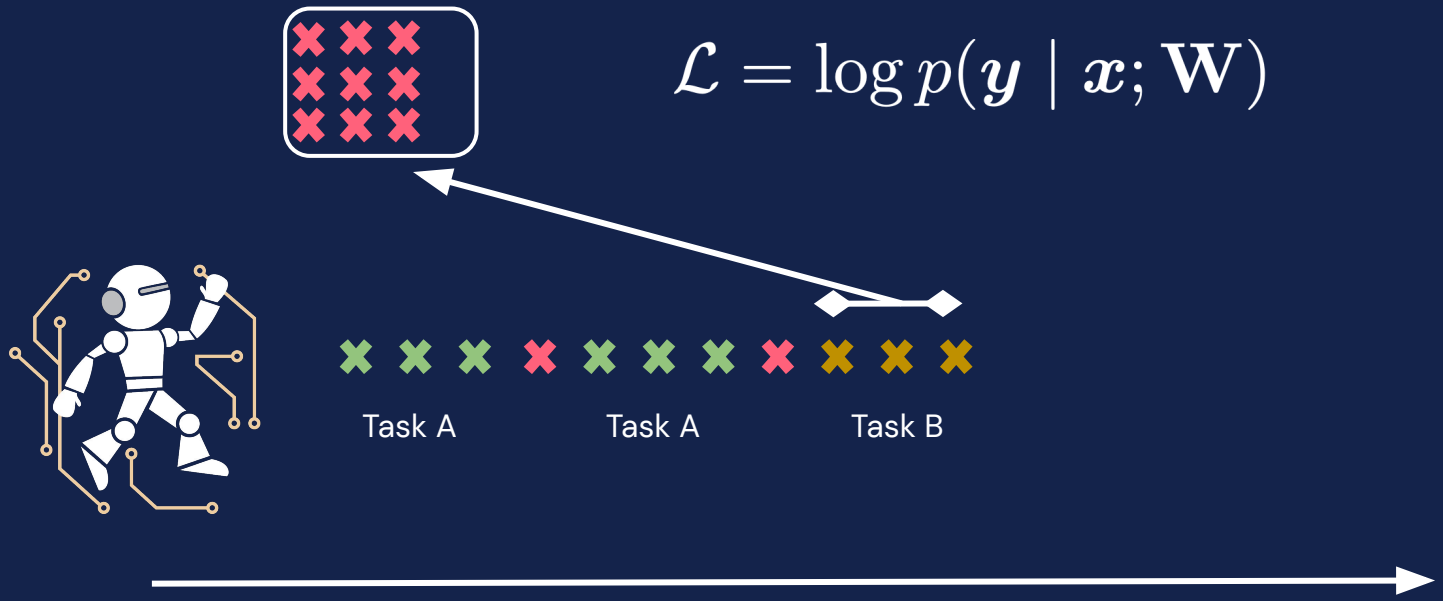
Training

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



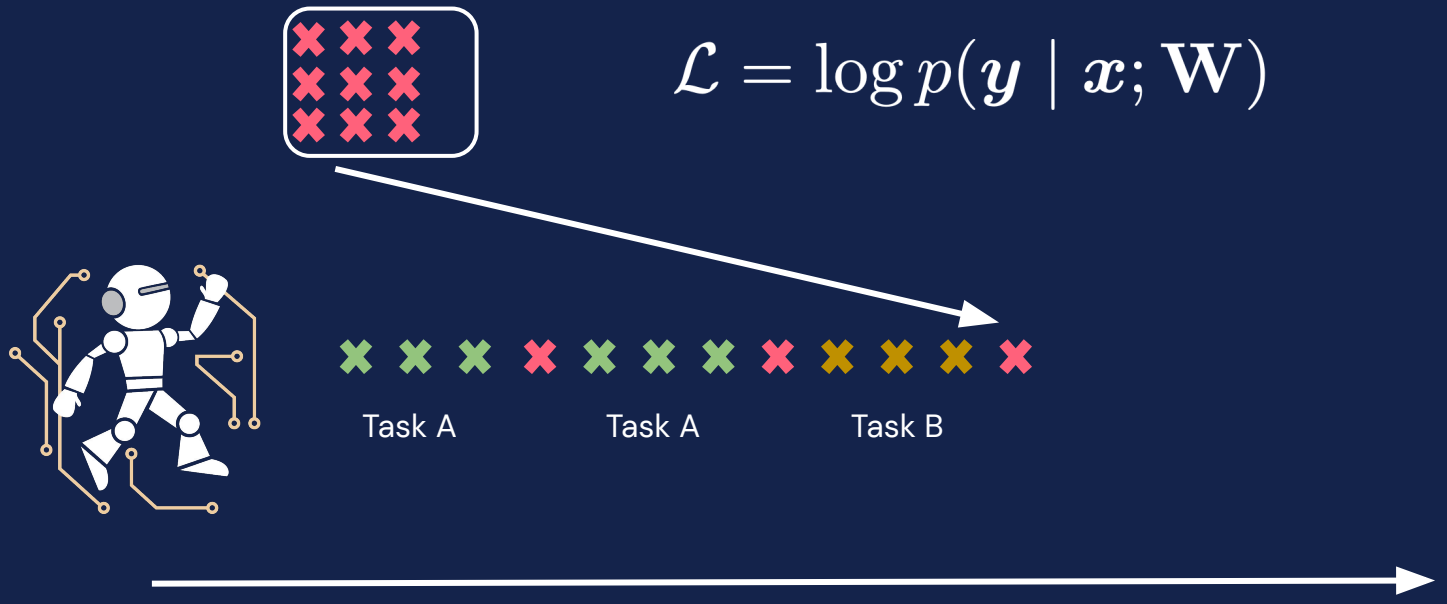
Training

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



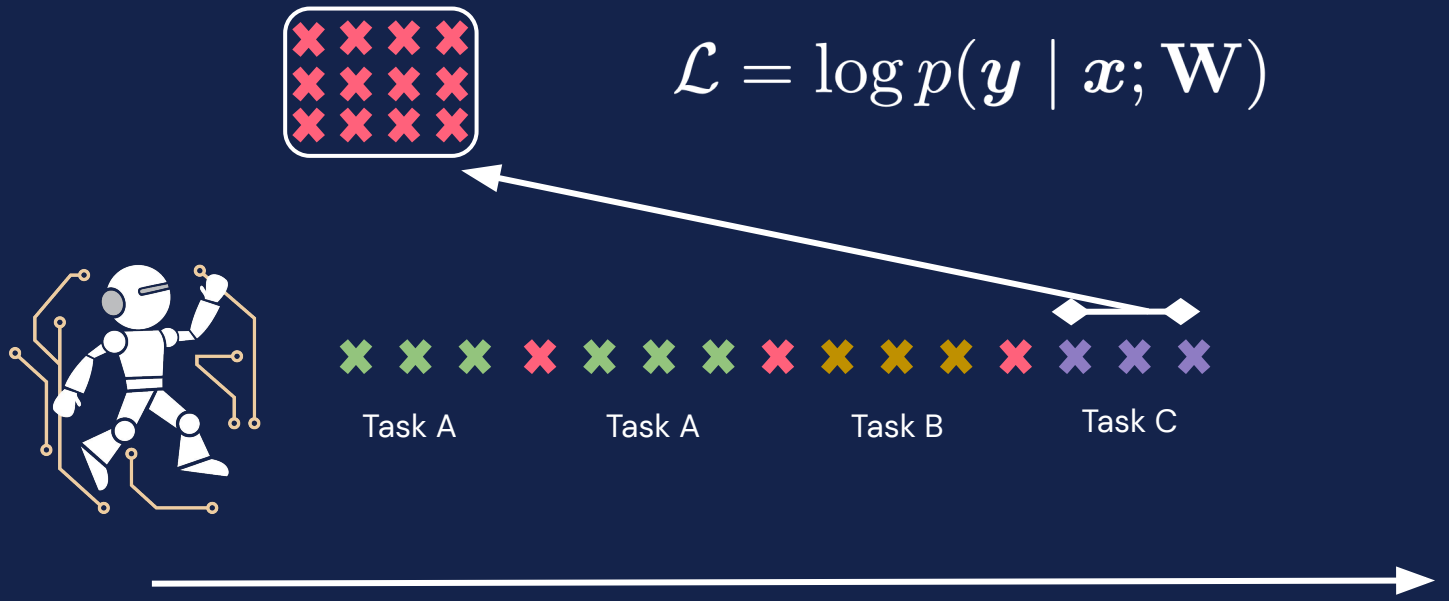
Training

***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)

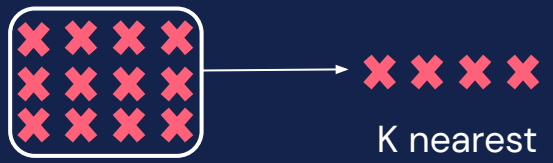
Local adaptation (Sprechmann et al., 2018).

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).

Inference (Prediction)

Local adaptation (Sprechmann et al., 2018).

K nearest neighbors retrieval

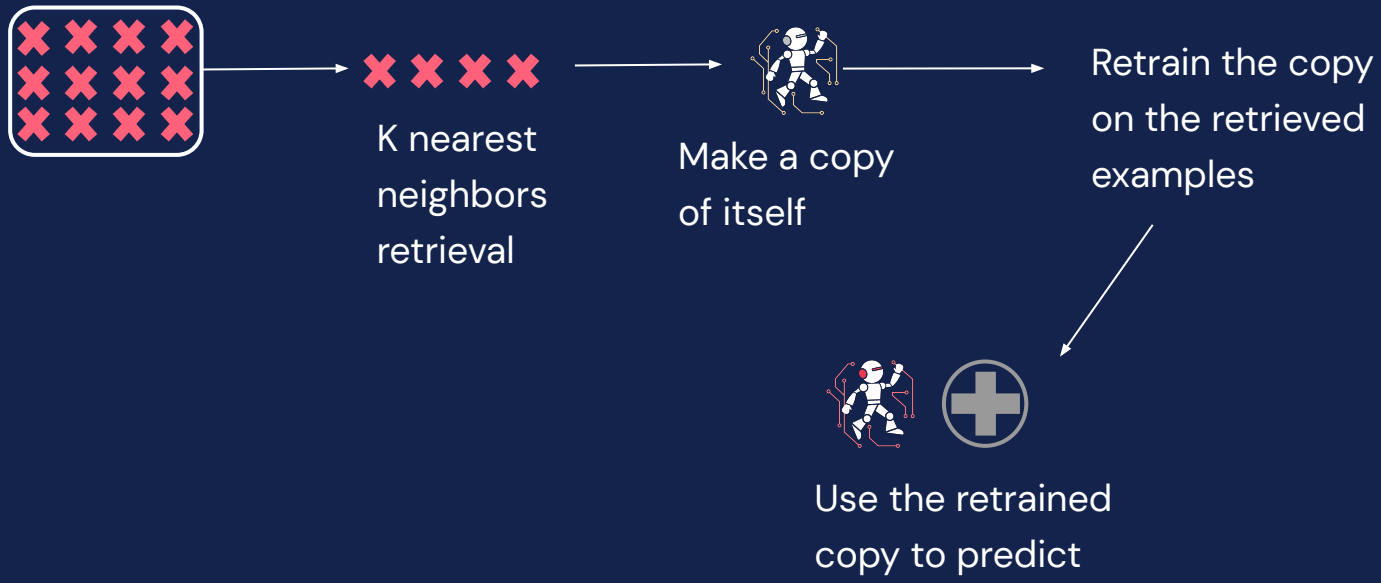
Make a copy of itself

Retrain the copy on the retrieved examples

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

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	MTL
QA	53.1	56.2	60.3	**62.4**	67.8

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

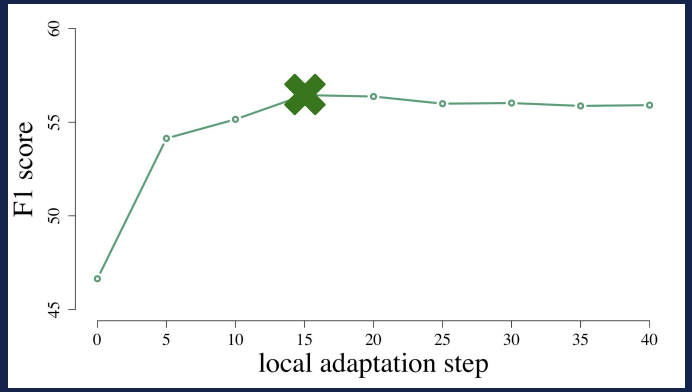
Takeaways and Limitations

- 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

- Local adaptation at inference time is **computationally expensive**.

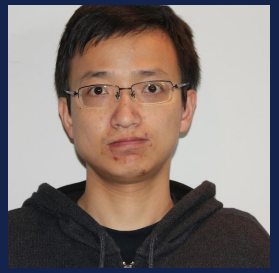


Adaptive Semiparametric Language Models

Yogatama et al., in review

Dani

Cyprien



Lingpeng

Background

- Existing memory-augmented language models are designed for one memory type.

- In language models, many tokens can be predicted from local context.

Background

- Existing memory-augmented language models are designed for one memory type.
- In language models, many tokens can be predicted from local context.
- Goal: an integrated language model architecture that can adaptively decide when to use local context, short-term, and/or long-term memory.

Background

- Existing memory-augmented language models are designed for one memory type.
- In language models, many tokens can be predicted from local context.
- Goal: an integrated language model architecture that can adaptively decide when to use local context, short-term, and/or long-term memory.

Hypothesis: encouraging each component to focus on a specific function results in a better language model.

Problem Setup

NYU Wikipedia

New York University (NYU) is a private research university based in New York City. Founded in

Problem Setup

NYU Wikipedia

New York University (NYU) is a private research university based in New York City. Founded in **1831**

Problem Setup

NYU Wikipedia

New York University (NYU) is a private research university based in New York City. Founded in 1831 **by**

Problem Setup

NYU Wikipedia

New York University (NYU) is a private research university based in New York City. Founded in 1831 by **Albert**

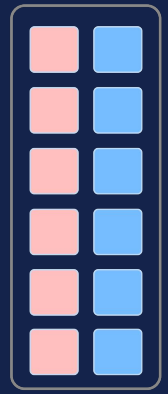
Problem Setup

NYU Wikipedia

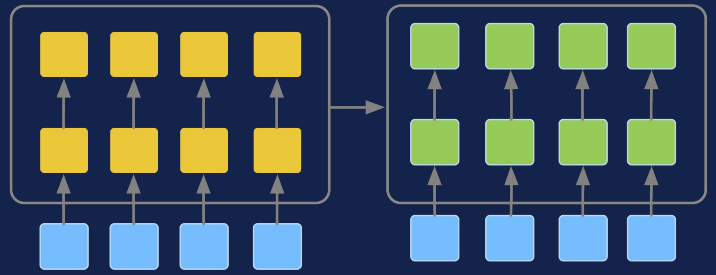
New York University (NYU) is a private research university based in New York City. Founded in 1831 by Albert **Gallatin**

Language Model

long-term memory

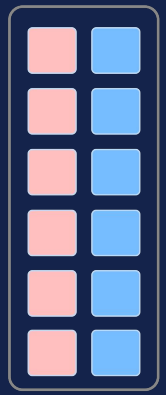


short-term memory



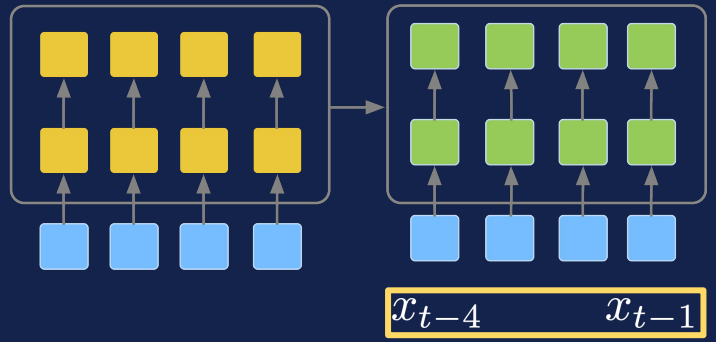
Language Model

long-term memory



New York University (NYU) is a private research university based in New York City. Founded in 1831 by

short-term memory

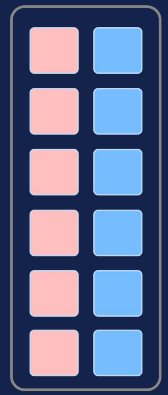


encoder

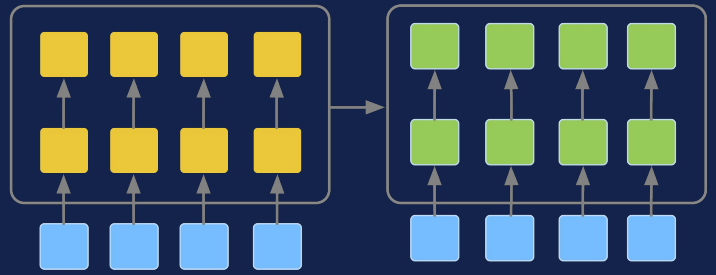
Input: a sequence of context tokens.

Language Model

long-term memory



short-term memory

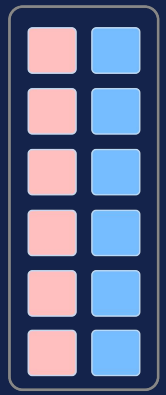


Encoder: transformer

(Vaswani et al., 2017)

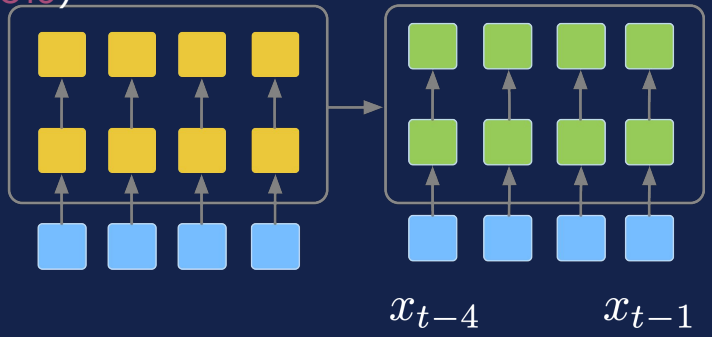
Language Model

long-term memory



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

short-term memory



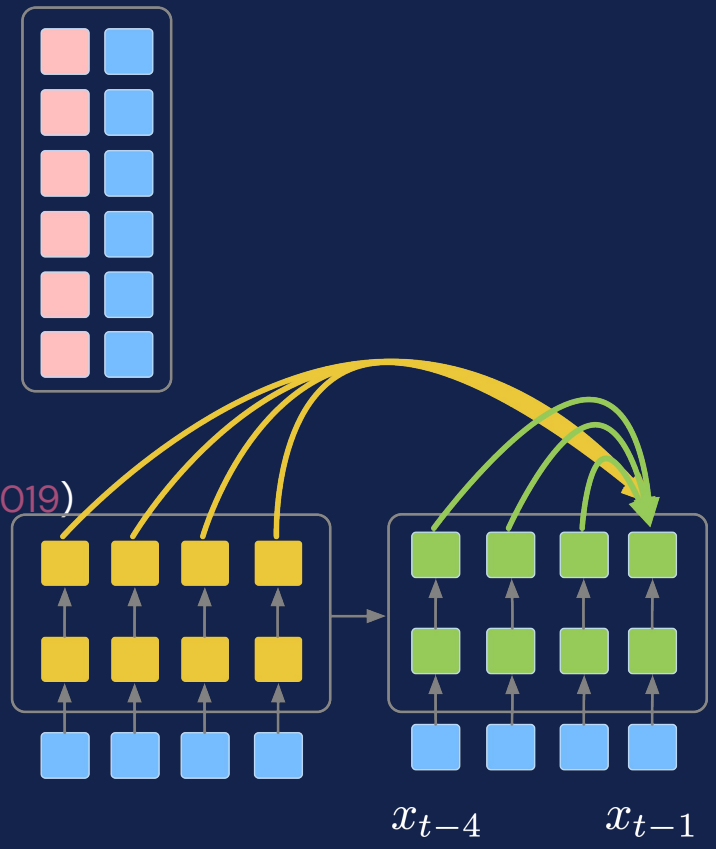
encoder

Language Model

long-term memory

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

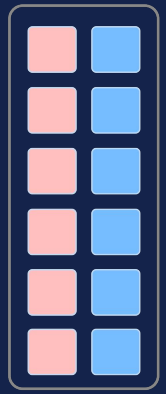
short-term memory



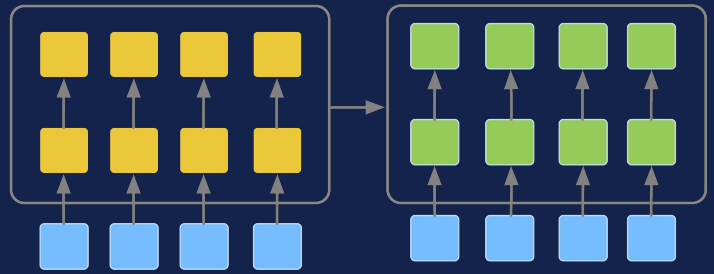
Language Model

Long-term memory:
key-value database

long-term memory



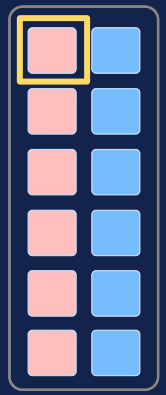
short-term memory



encoder

Language Model

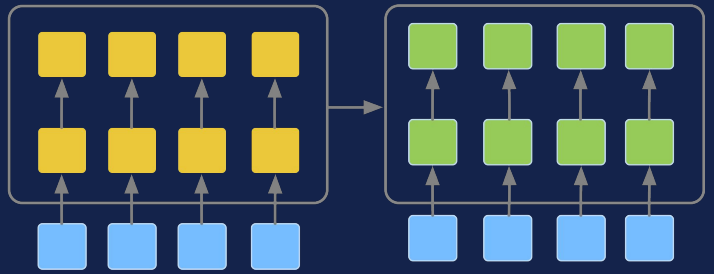
long-term memory



Key: compressed long-term context

Abraham Alfonse Albert Gallatin, born the Gallatin (January 29, 1761 - August 12, 1849) was an American

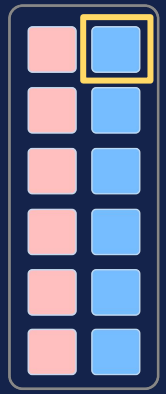
short-term memory



encoder

Language Model

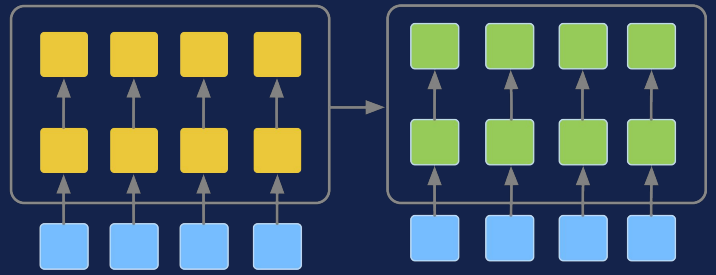
long-term memory



politician

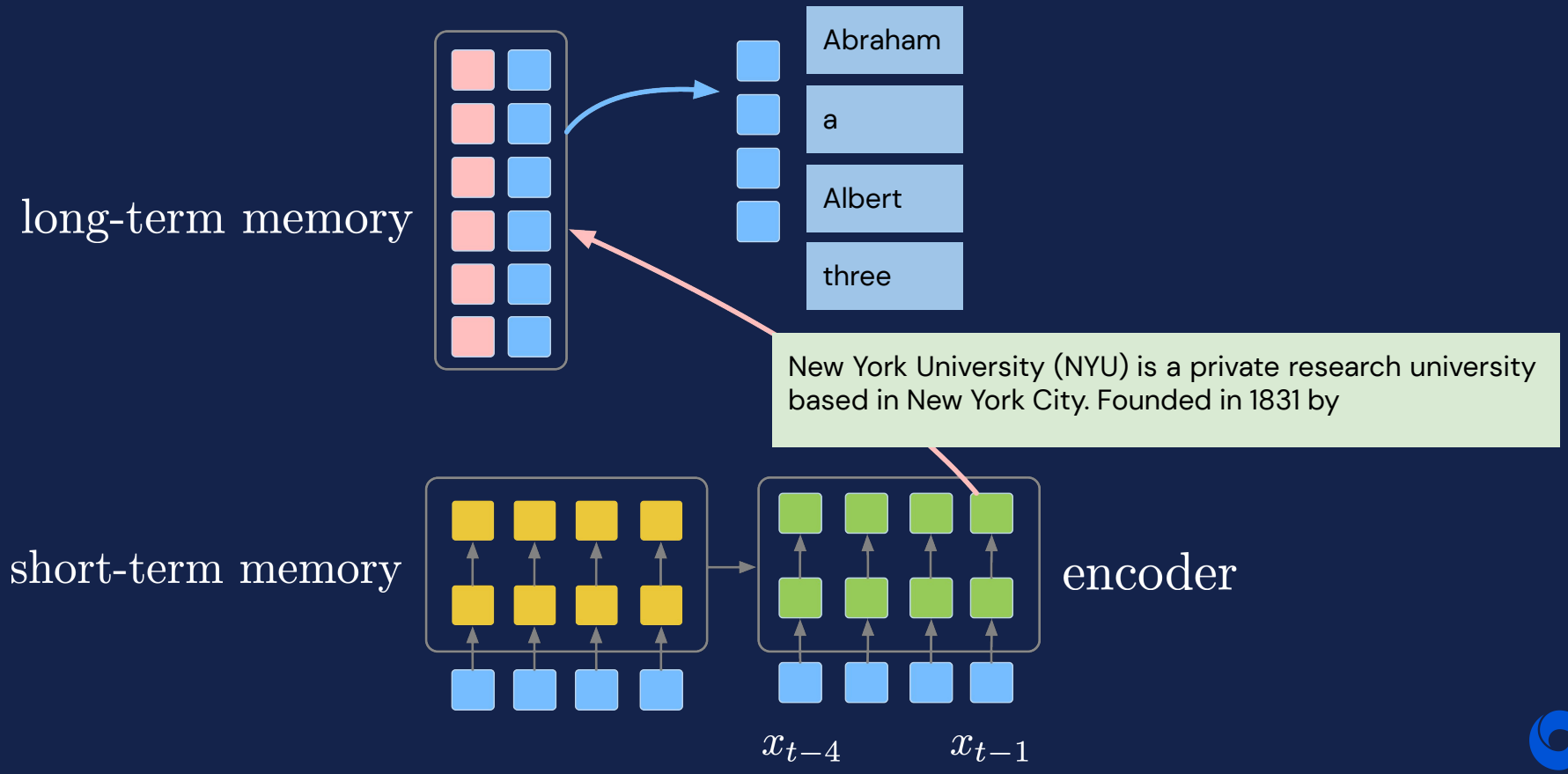
Value: output token for the respective context

short-term memory



encoder

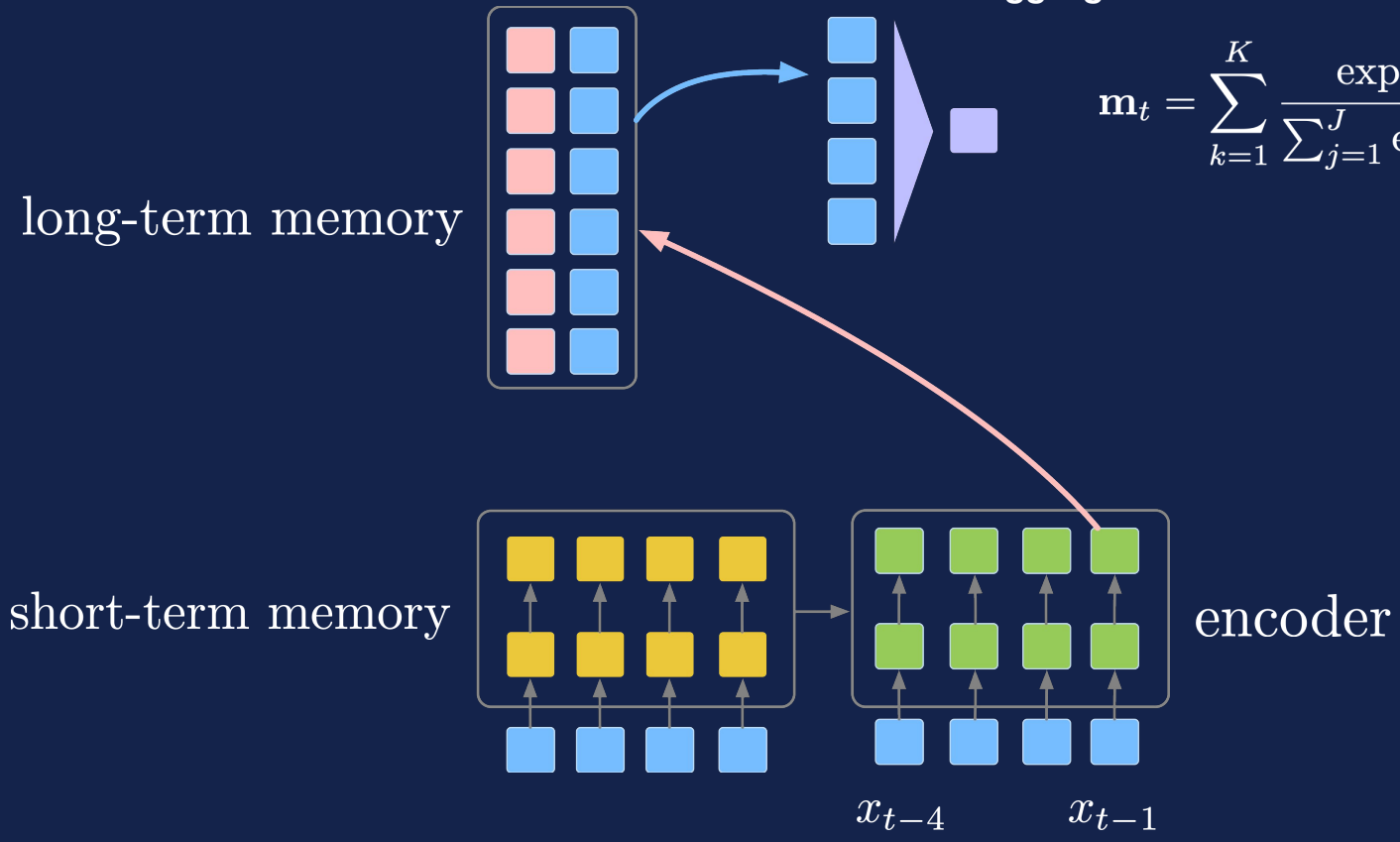
Language Model



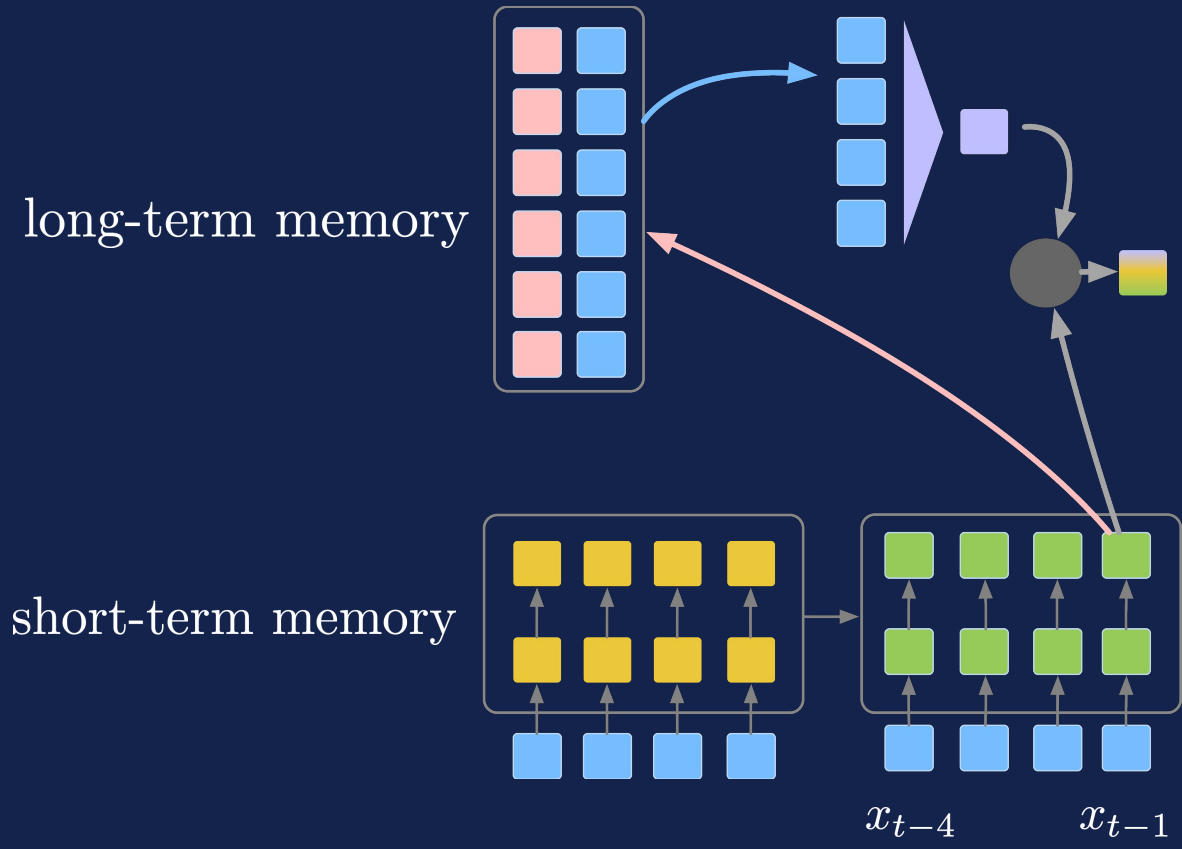
Language Model

Aggregator: attention over retrieved values

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



Language Model

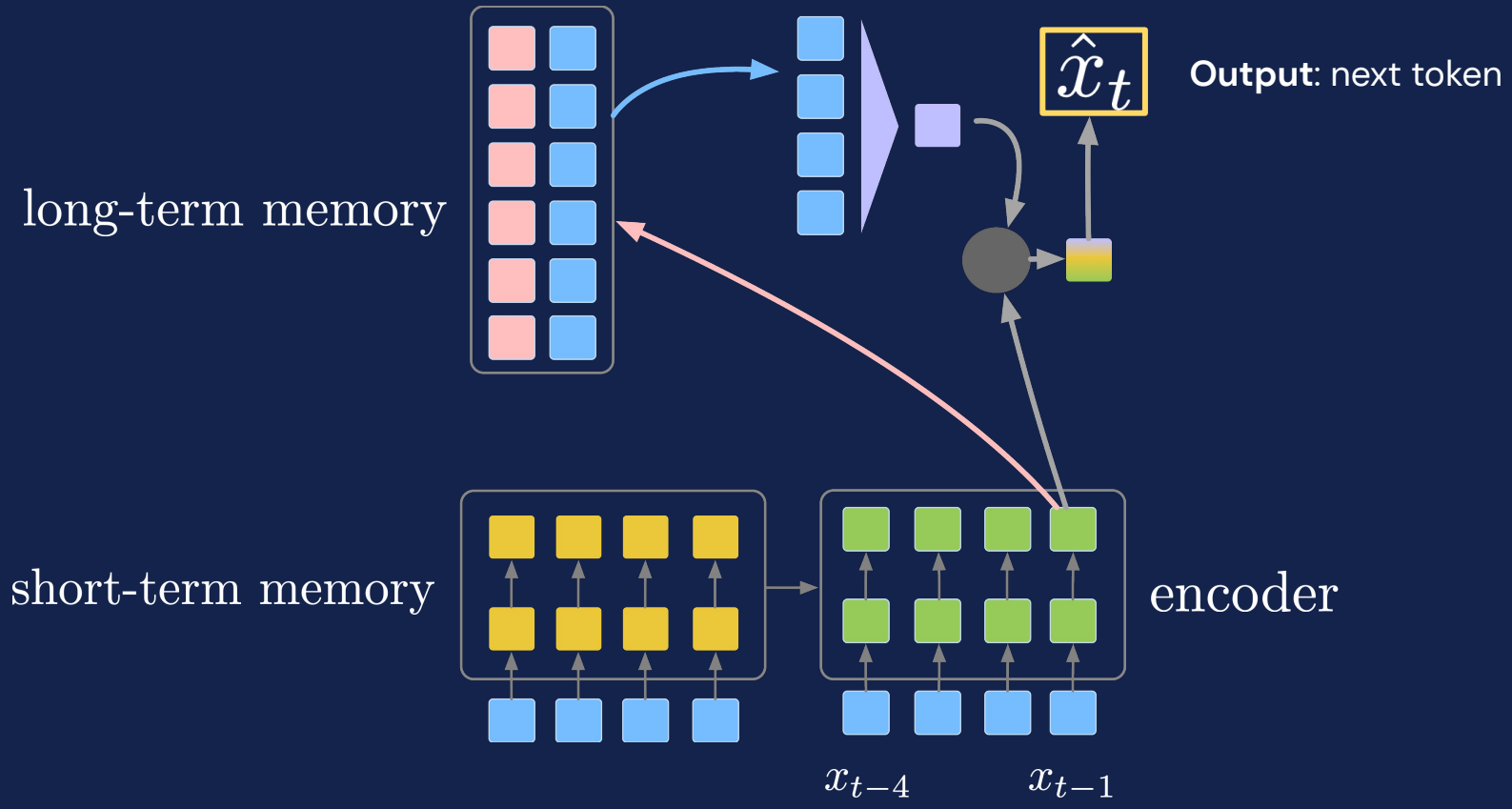


Gate: combine local context/short-term memory and long-term memory

$$\mathbf{g}_t = \sigma(\mathbf{w}_g^\top \mathbf{h}_{t-1}^R)$$

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

Language Model



Experiments

- Word-level language modeling.
  - WikiText-103: Merity et al., 2017.
  - 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

Perplexity (1-inf), lower is better

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

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

kNN-LM: Khandelwal et al., 2020

Experiments

BPC (0-inf), lower is better

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

Transformer: Vaswani et al., 2017

Transformer-XL: Dai et al., 2019

kNN-LM: Khandelwal et al., 2020

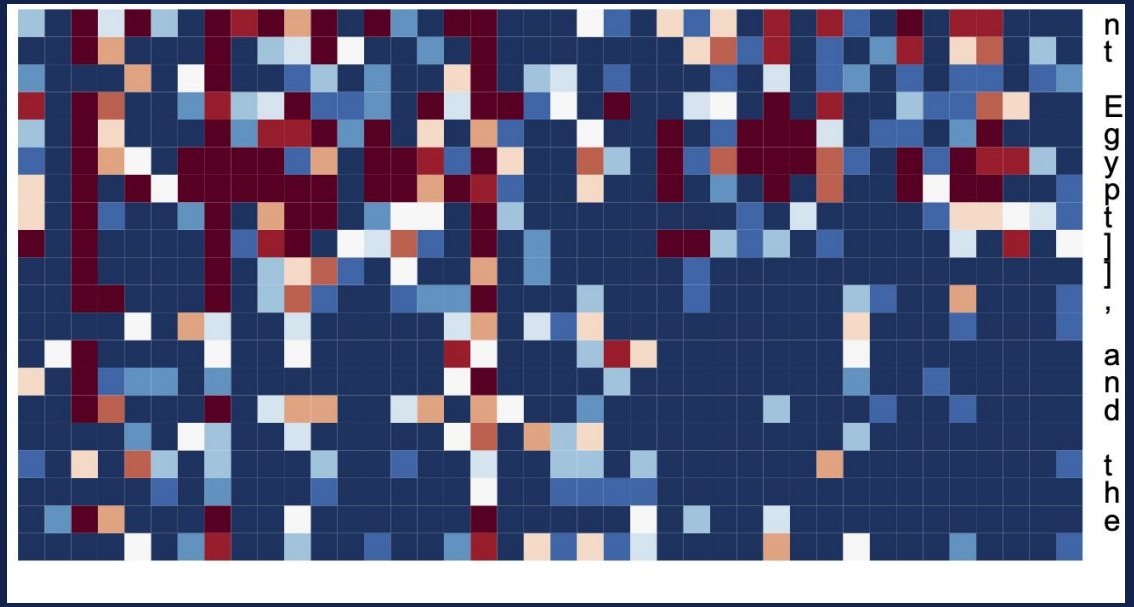
Analysis

What is in the long-term memory?

these	yellow	To	Agency	from	his	Stone	Stone	that
a	North	Korean	statement	on	KC	NA	said	last
is	year	\ n	The	by	Hay	Esc	;	Editing
month	.	\ n	Reporting	by	Joyce	Lee	;	Additional
reporting	by	Barbara	She	n	strom	;	New	;
reporting	by	David	Brun	n	strom	in	Washington	;
Editing	by	Dale	Sanders	and	Gareth	Coloney		
Editing	by	Chris	Reese	and	Peter	Coloney		

Analysis

How does the model combine/use information from different sources?



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.

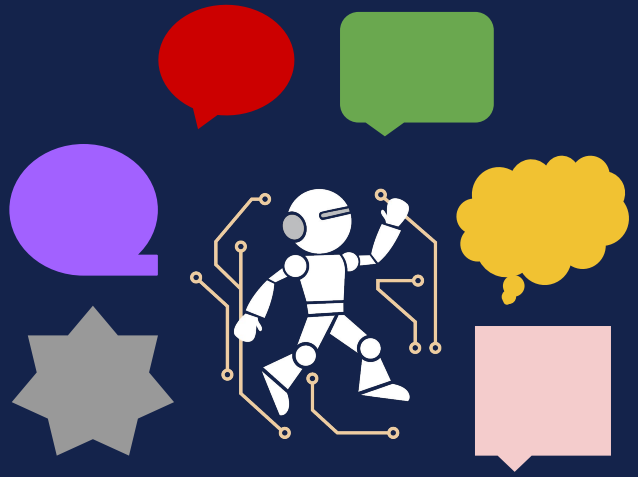
	CPUs	Hours
WikiText-103	1,000	6
WMT	9,000	18
enwik8	1,000	8

Future Directions

A language model that continually learns in an efficient way.

Future Directions

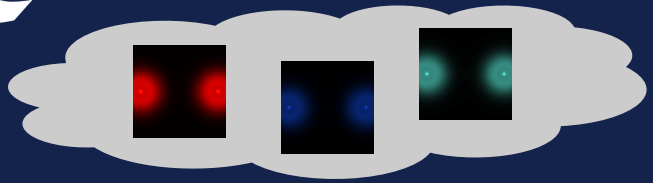
A language model that continually learns in an efficient way.

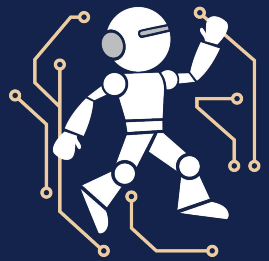


Integration of different memory types from various sources and modalities

Future Directions

A language model that continually learns in an efficient way.





Structured storage space that dynamically manages its complexity
(forgetting, compression/forming abstractions)

https://dyogatama.github.io

dyogatama@google.com

DeepMind