Toward General Linguistic Intelligence

Dani Yogatama

DeepMind

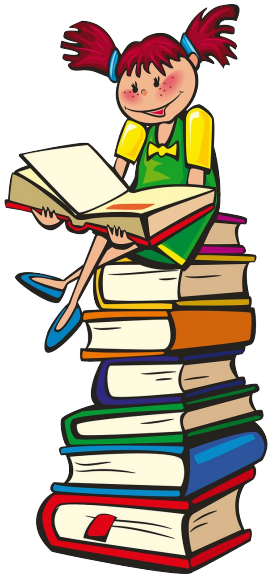
Language and Intelligence

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

AHOJ
HELLO
สวัสดีครับ
สะบายดี
Здравейте
Добар
Салам
SALUTON
ZDRAVO
BONJOUR
SVEIKI
HOLA
CHAO
TERVE
SZIA
HUJAMBO
MERHABA
NAMASTE
こんにちは
Γεια
ᎣᏏᏲ
ᏣᎳᎩ
ᏗᏂᏲᏍᏗᏱ
ᏗᏂᏲᏍᏗᏱᏗᏂᏲᏍᏗᏱ
ᏗᏂᏲᏍᏗᏱᏗᏂᏲᏍᏗᏱᏗᏂᏲᏍᏗᏱ

Language and Intelligence

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



Language and Intelligence

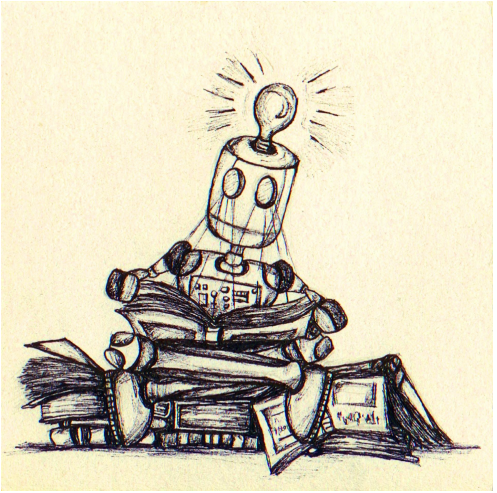
The most effective form of communication to transmit information and knowledge to others.

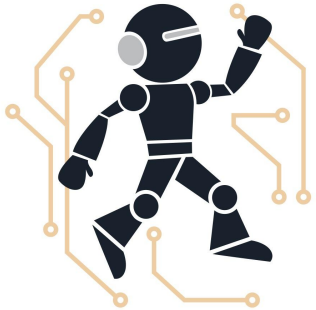
Language and Intelligence

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

General Linguistic Intelligence

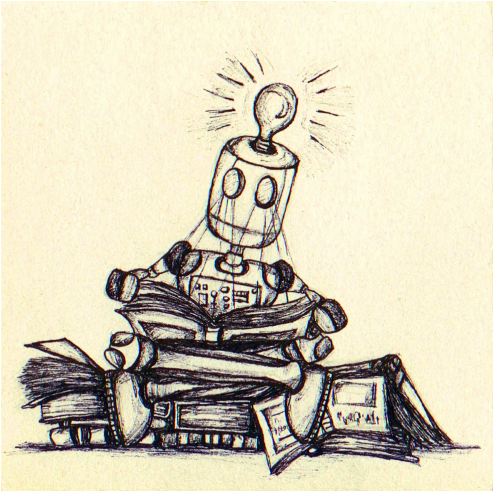
The ability to acquire, store, and reuse knowledge (about a language's lexicon, syntax, semantics, and pragmatic conventions) from textual data to **adapt to new tasks quickly without forgetting old ones**.

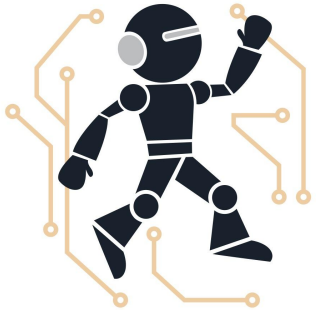




General Linguistic Intelligence

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**.



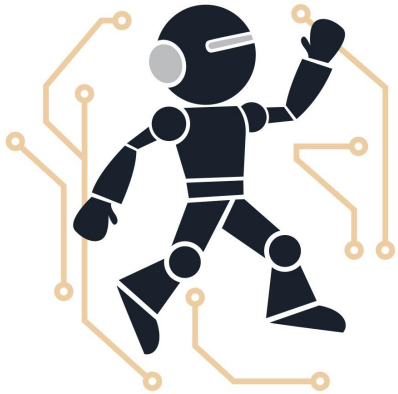


Challenges: Human Learning vs. Machine Learning

- ``Large'' datasets → sample efficient.
- Mostly task agnostic.
- Generalizable to new tasks.

Challenges: Human Learning vs. Machine Learning

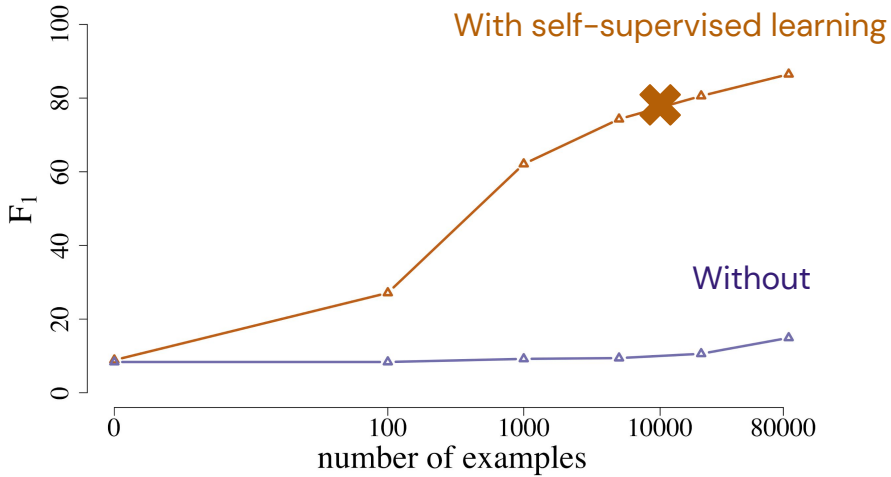
- ``Large'' datasets → sample efficient.
- Mostly task agnostic.
- Generalizable to new tasks.



- Large datasets (self-supervised learning) → large datasets (supervised fine tuning).
- Problem specific.
- Forget previous tasks given a new task.

The State of Natural Language Processing

- Current models still require many in-domain training examples.



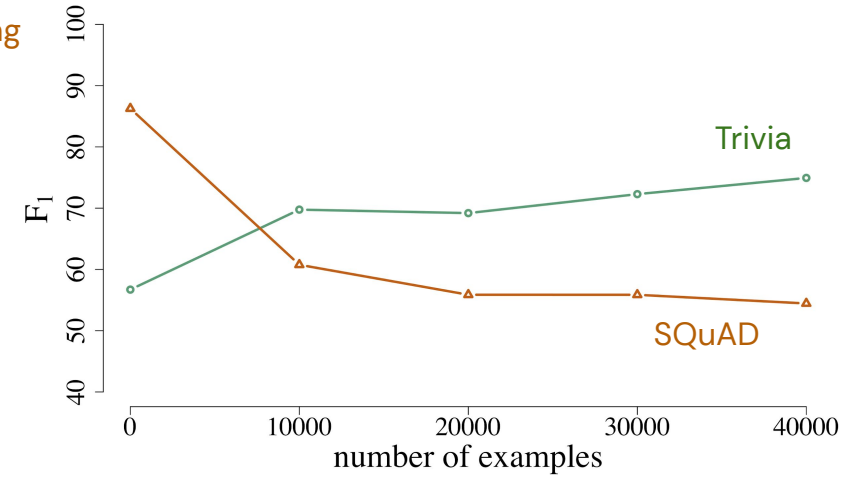
BERT model: Devlin et al. 2019

SQuAD dataset: Rajpurkar et al., 2016

Yogatama et al., arXiv 2019

The State of Natural Language Processing

- Current models still require many in-domain training examples.
- They overfit to a specific dataset (task) and often forget.



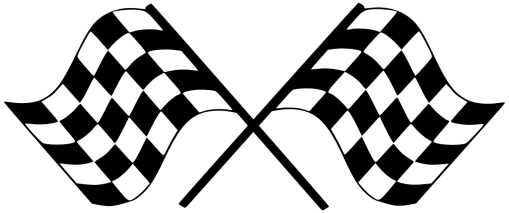
BERT model: Devlin et al. 2019

SQuAD dataset: Rajpurkar et al., 2016

Trivia dataset: Joshi et al., 2017

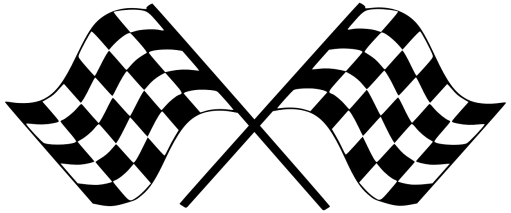
Yogatama et al., arXiv 2019

Research Areas



A universal language model that continually and efficiently learns to perform multiple complex tasks in many languages.

Research Areas



A universal language model that continually and efficiently learns to perform multiple complex tasks in many languages.

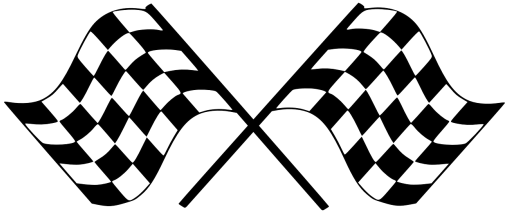
Memory

Yogatama and Mann; AISTATS 2014

Yogatama et al., ICLR 2018

de Masson d'Autume; NeurIPS 2019

Research Areas



A universal language model that continually and efficiently learns to perform multiple complex tasks in many languages.

Memory

Yogatama and Mann; AISTATS 2014

Yogatama et al., ICLR 2018

de Masson d'Autume; NeurIPS 2019

Representation Learning

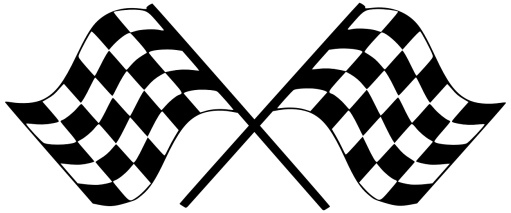
Yogatama and Smith; ACL 2014

Yogatama and Smith; ICML 2015

Yogatama et al., ICLR 2017

Maillard et al., JNLE 2019

Research Areas



A universal language model that continually and efficiently learns to perform multiple complex tasks in many languages.

Memory

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

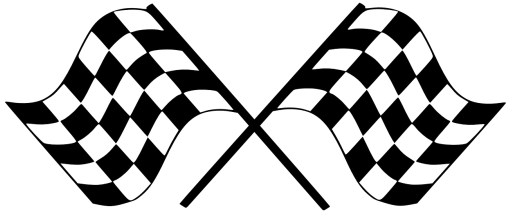
Representation Learning

Yogatama and Smith; ACL 2014
Yogatama and Smith; ICML 2015
Yogatama et al., ICLR 2017
Maillard et al., JNLE 2019

Generative Models

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

Research Areas



A universal language model that continually and efficiently learns to perform multiple complex tasks in many languages.

Memory

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

Representation Learning

Yogatama and Smith; ACL 2014
Yogatama and Smith; ICML 2015
Yogatama et al., ICLR 2017
Maillard et al., JNLE 2019

Generative Models

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

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.

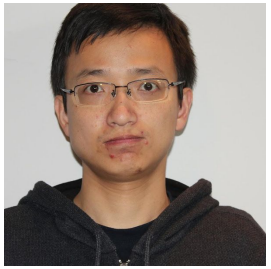
  Kong et al., ICLR 2020

Episodic Memory in Lifelong Language Learning

de Masson d'Autume et al., NeurIPS 2019

Cyprien

Sebastian



Lingpeng

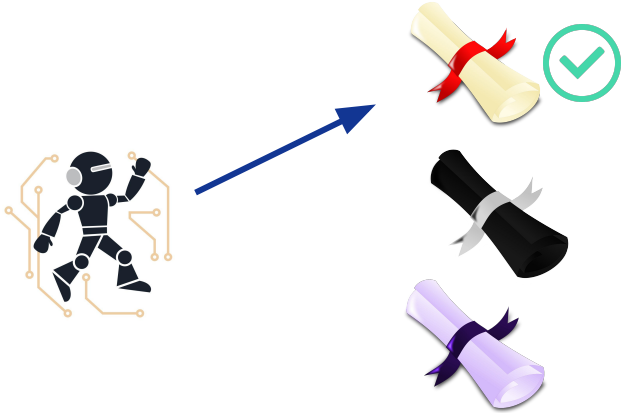
Dani

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 (McCloseky and Cohen, 1989; Ratcliff, 1990).

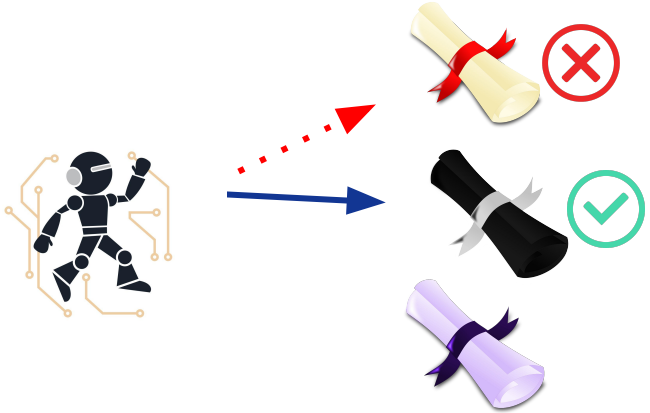
Background

- 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).



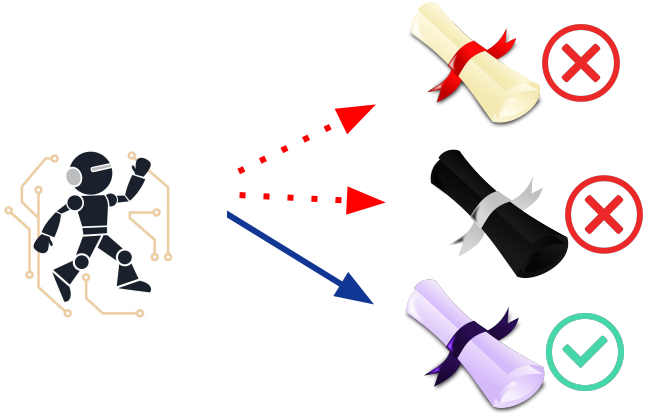
Background

- 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).



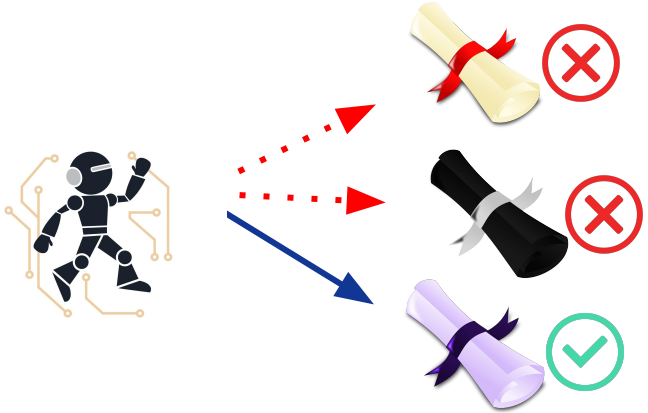
Background

- 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

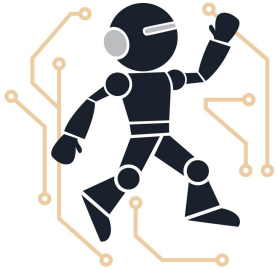
- A model should be able to reuse knowledge from related tasks to learn a new task faster.
- Current models not only fail to do this, they catastrophically forget previously learned tasks (McCloseky and Cohen, 1989; Ratcliff, 1990).



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?

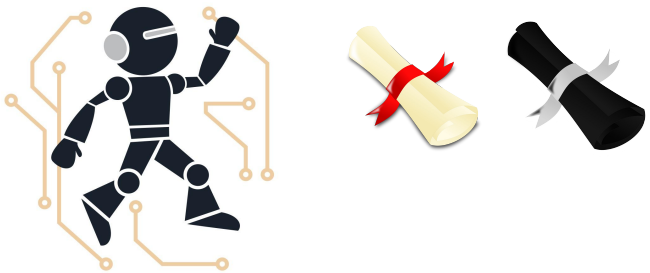
Problem Setup

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

Problem Setup

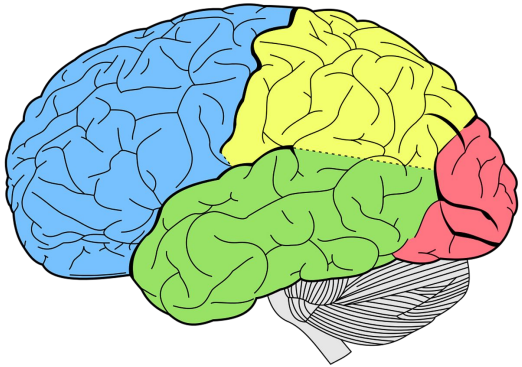
Training

Test

Never-Ending Language Learning

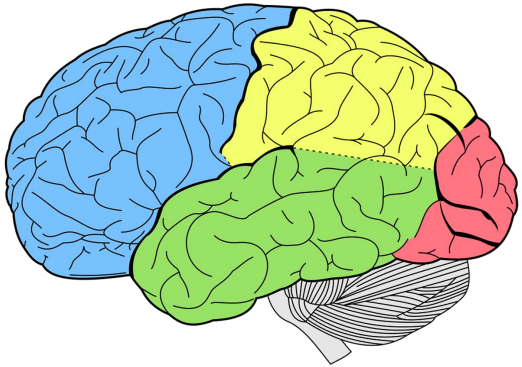
Mitchell et al., 2015

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

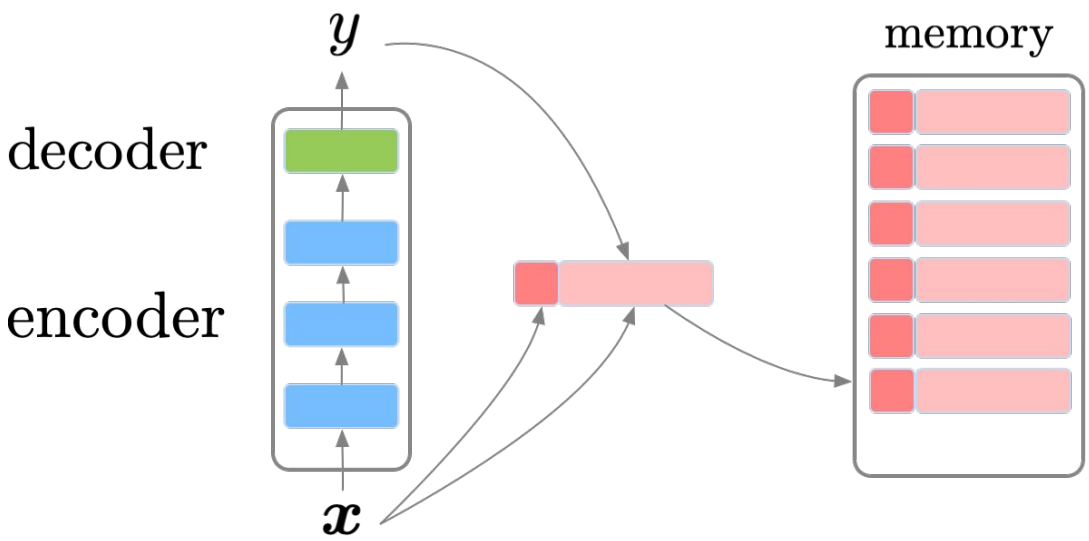


Episodic memory is a type of long-term memory of events and experiences. It is often associated with a module that stores training examples in neural networks..

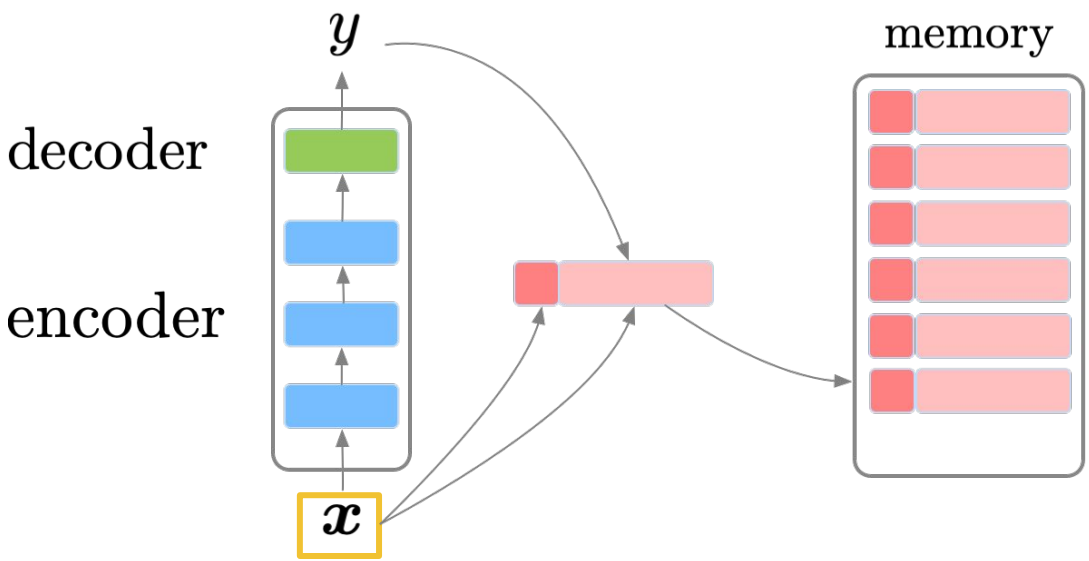
Contrast this with a short-term (working) memory in e.g., LSTMs (Hochreiter and Schmidhuber, 1997) , stack-augmented networks (Joulin and Mikolov, 2015; Grefenstette et al., 2015), or DNCs (Graves et al., 2016).

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

Question Answering Model



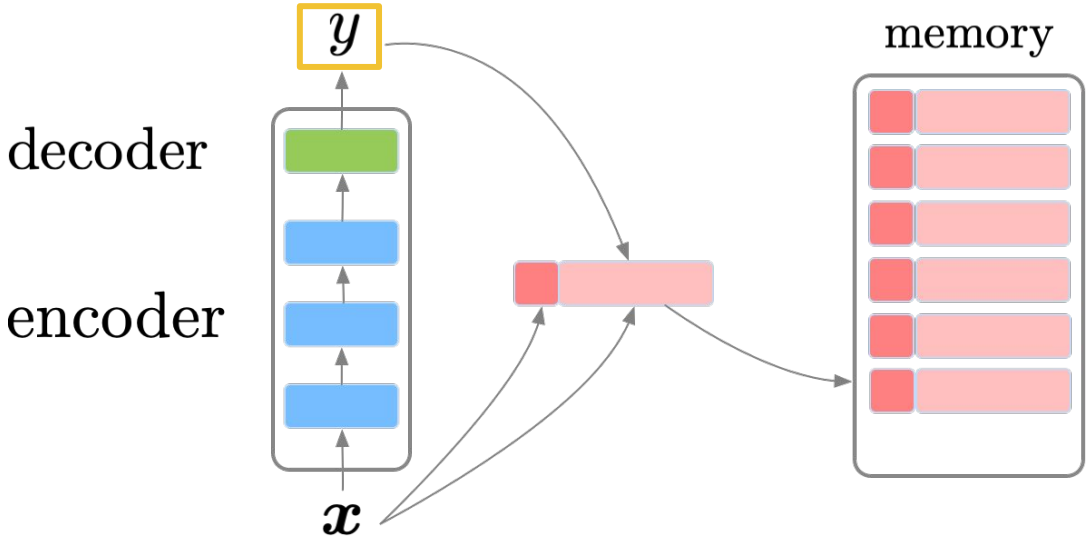
Question Answering Model



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

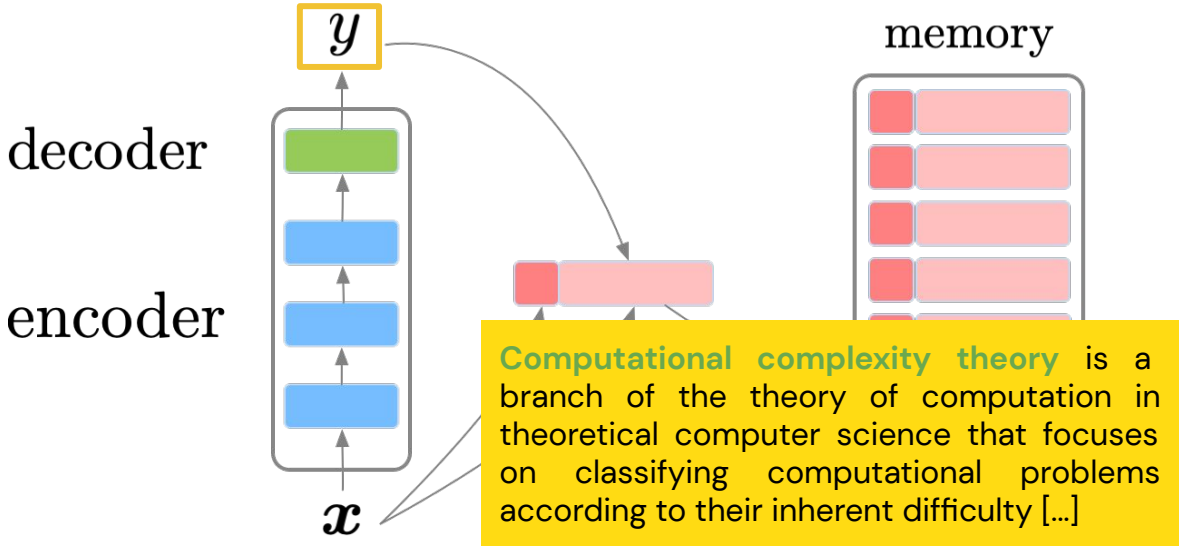
Question Answering Model

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



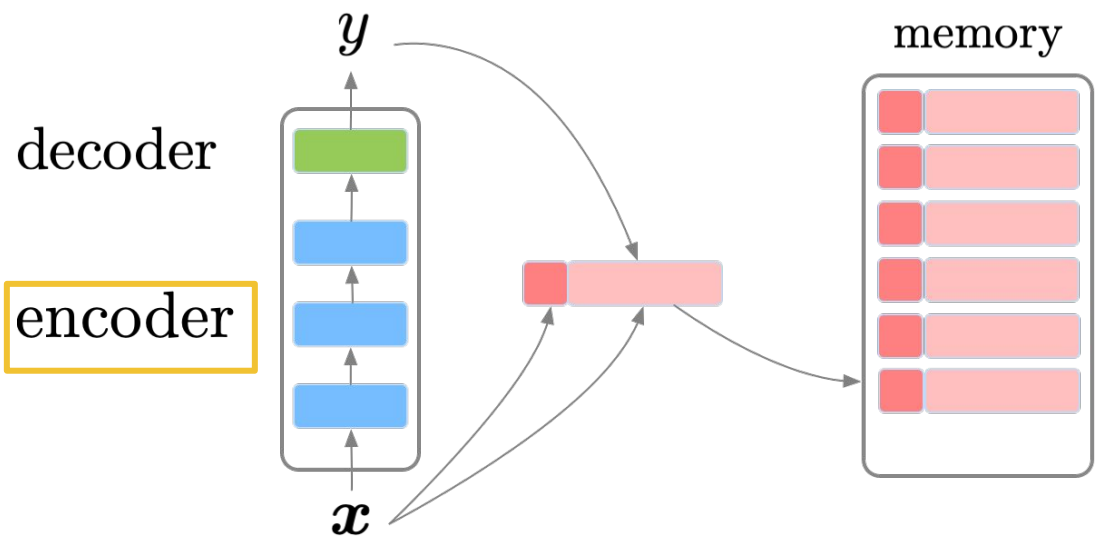
Question Answering Model

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



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

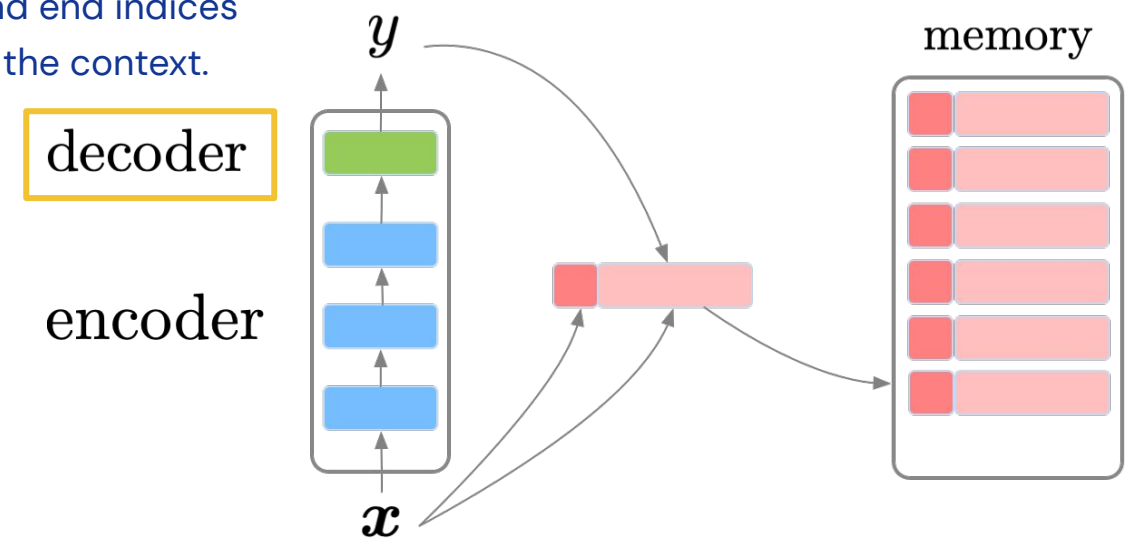
Question Answering Model



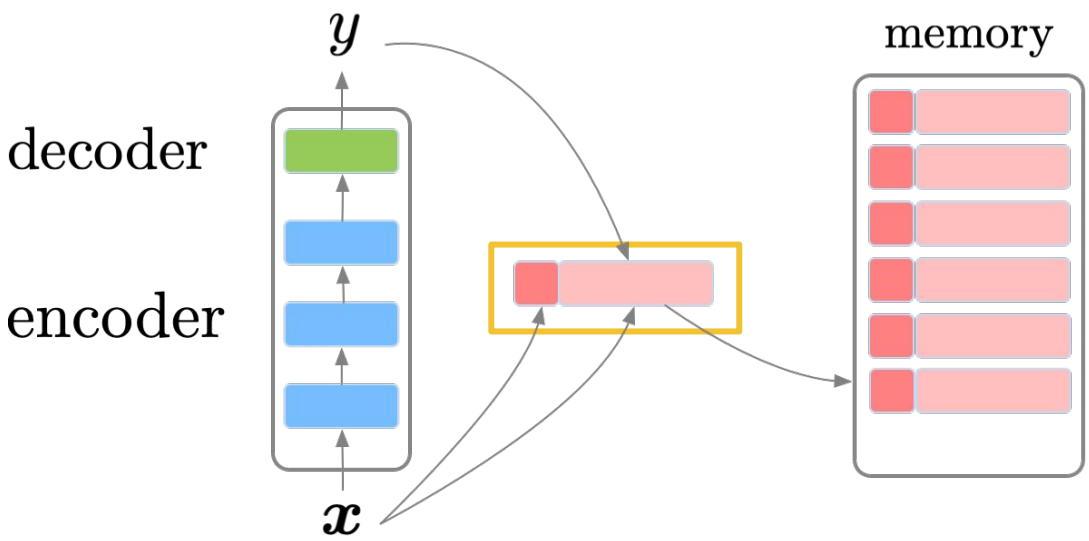
Encoder: a large neural network, e.g., ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), XLNet (Yang et al., 2019).

Question Answering Model

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

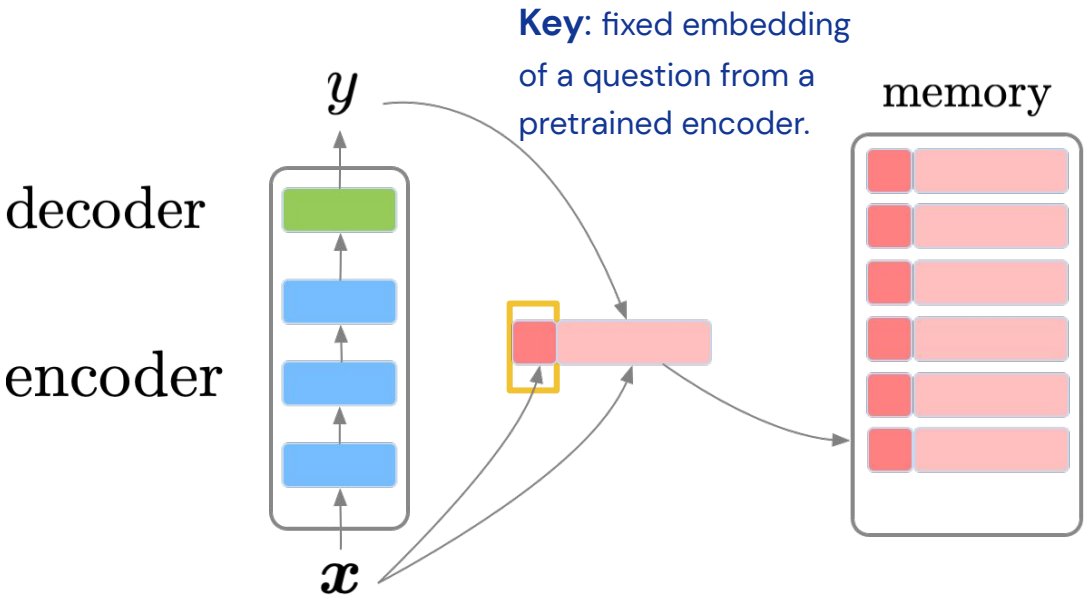


Question Answering Model

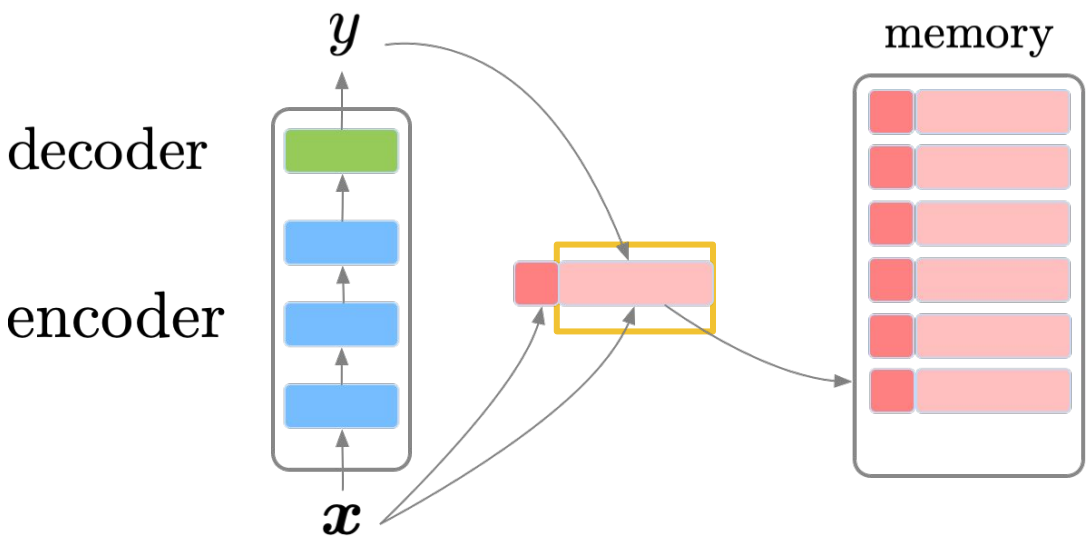


Memory entry: a key, value pair

Question Answering Model



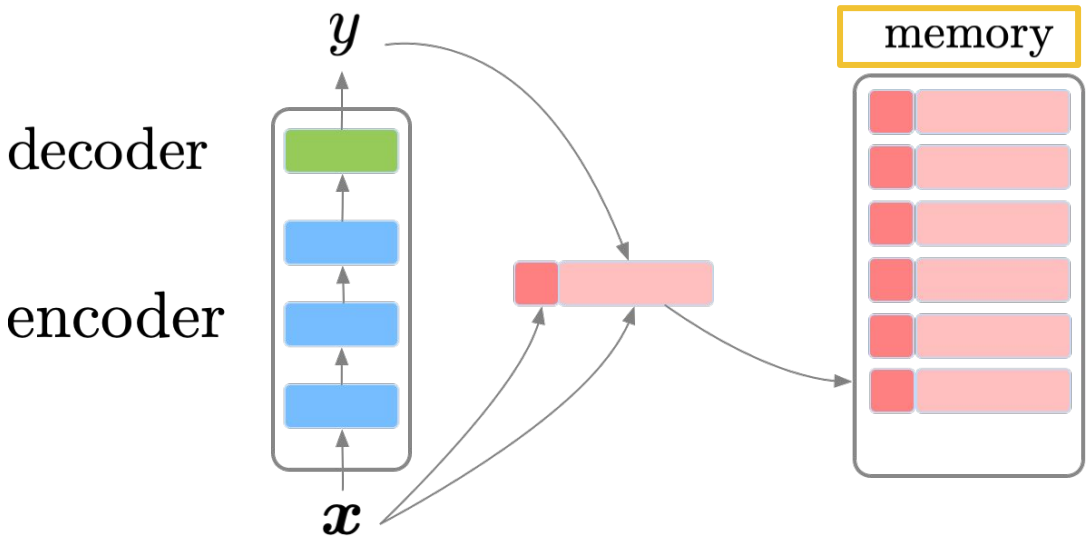
Question Answering Model



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

Question Answering Model

Memory: stores memory entries

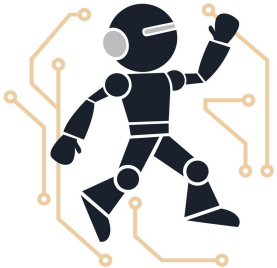


Training

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

Related to memory consolidation in human learning.

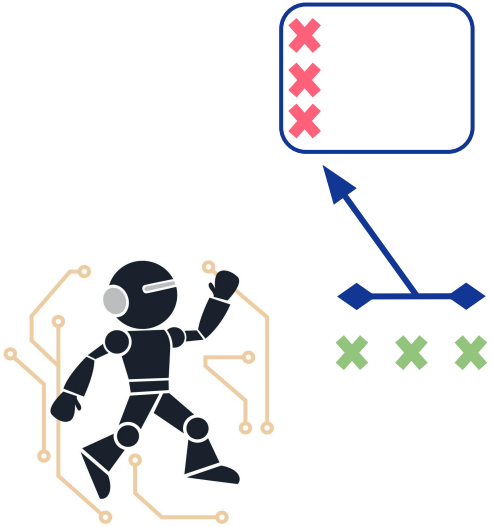
$$\mathcal{L} = \log p(\boldsymbol{y} \mid \boldsymbol{x}; \mathbf{W})$$



Training

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

Related to memory consolidation in human learning.

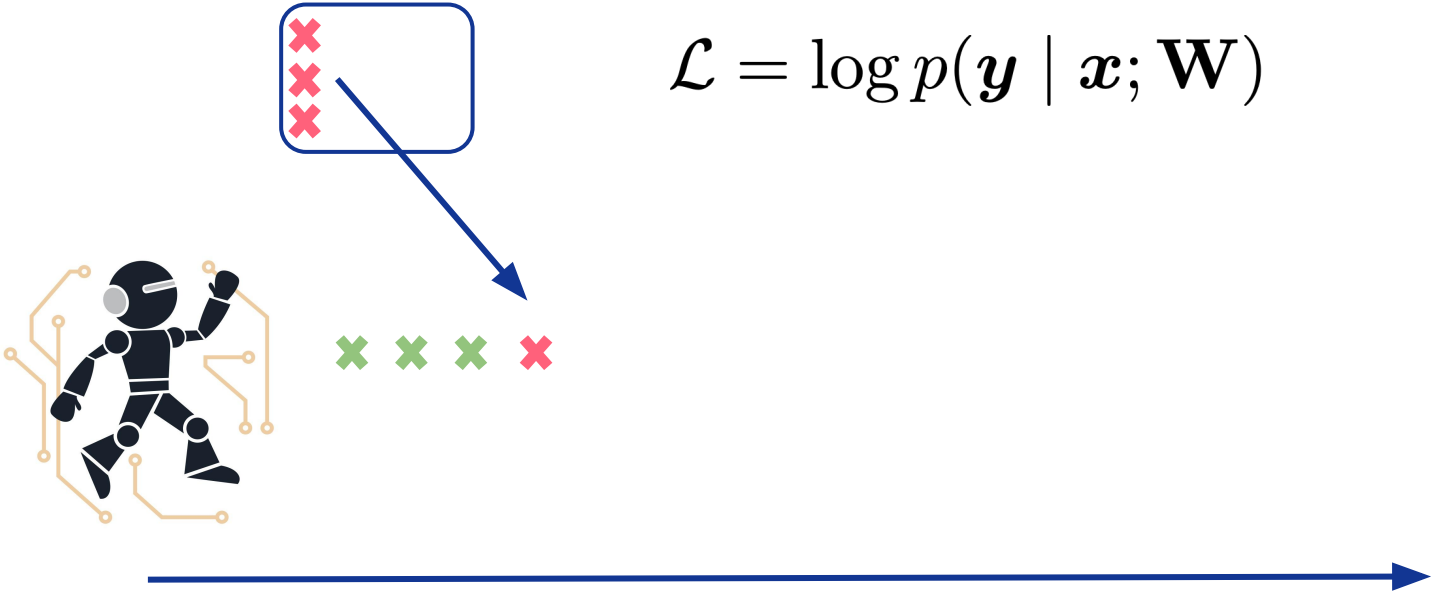


$$\mathcal{L} = \log p(\boldsymbol{y} \mid \boldsymbol{x}; \mathbf{W})$$

Training

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

Related to memory consolidation in human learning.

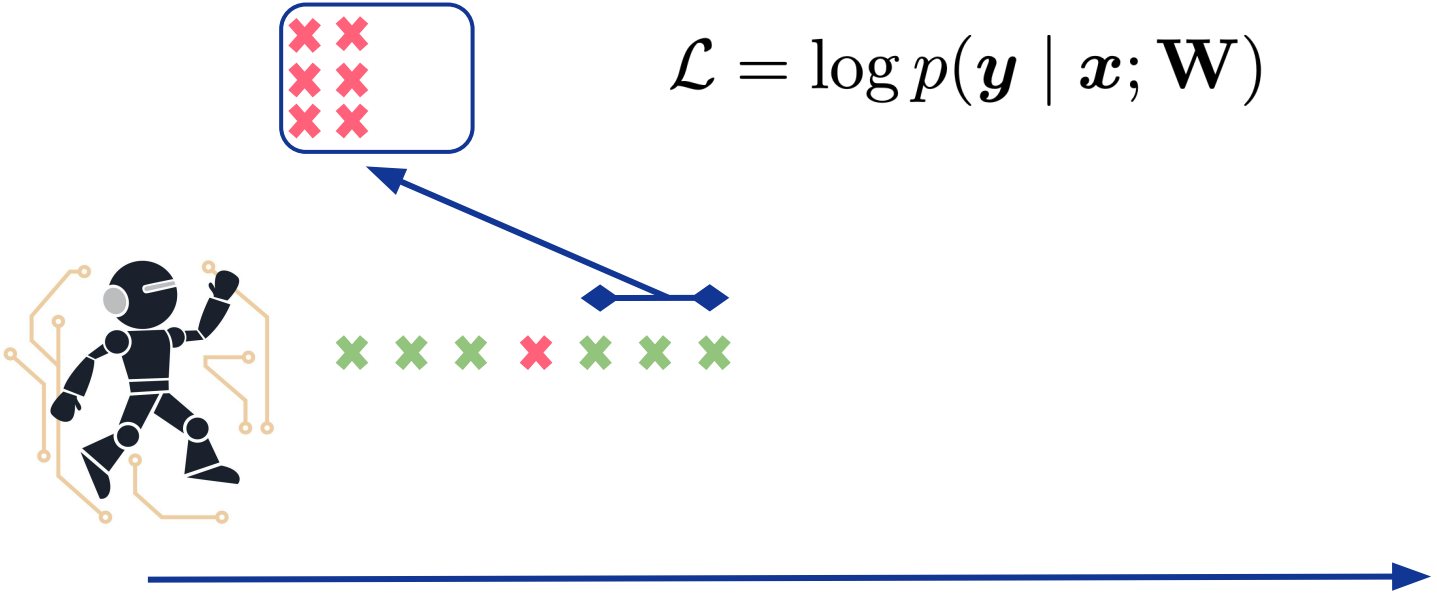


$$\mathcal{L} = \log p(\boldsymbol{y} \mid \boldsymbol{x}; \mathbf{W})$$

Training

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

Related to memory consolidation in human learning.

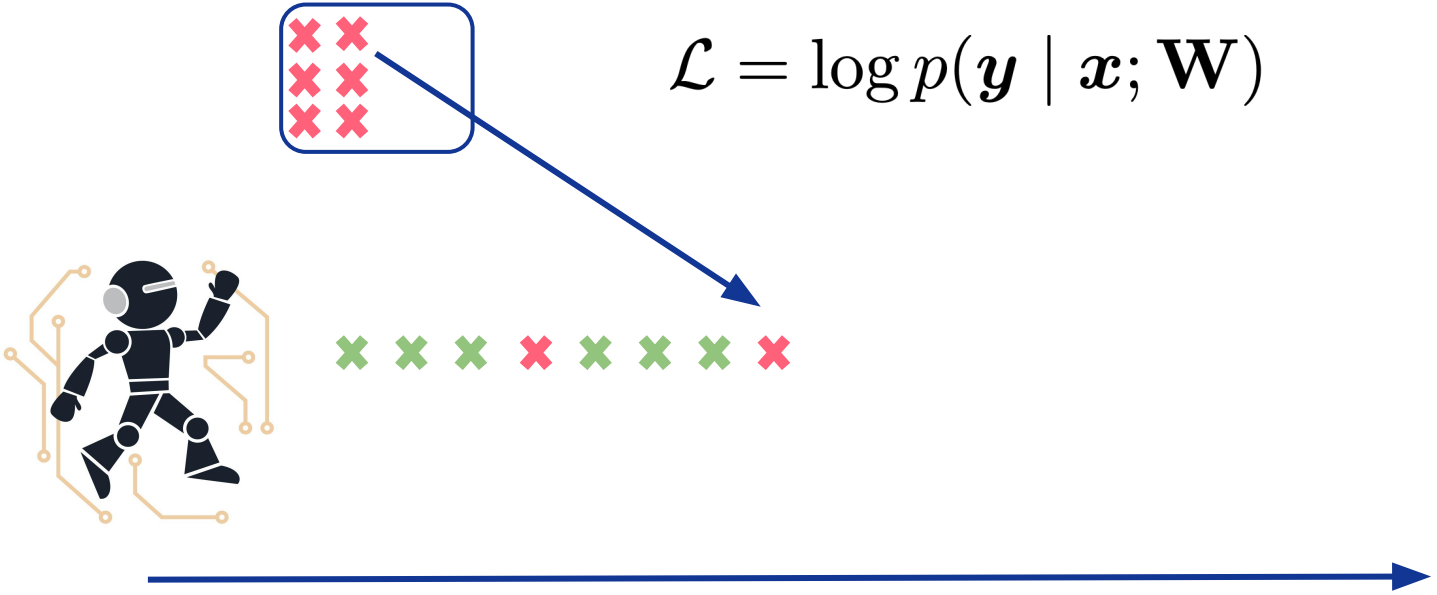


$$\mathcal{L} = \log p(\boldsymbol{y} \mid \boldsymbol{x}; \mathbf{W})$$

Training

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

Related to memory consolidation in human learning.

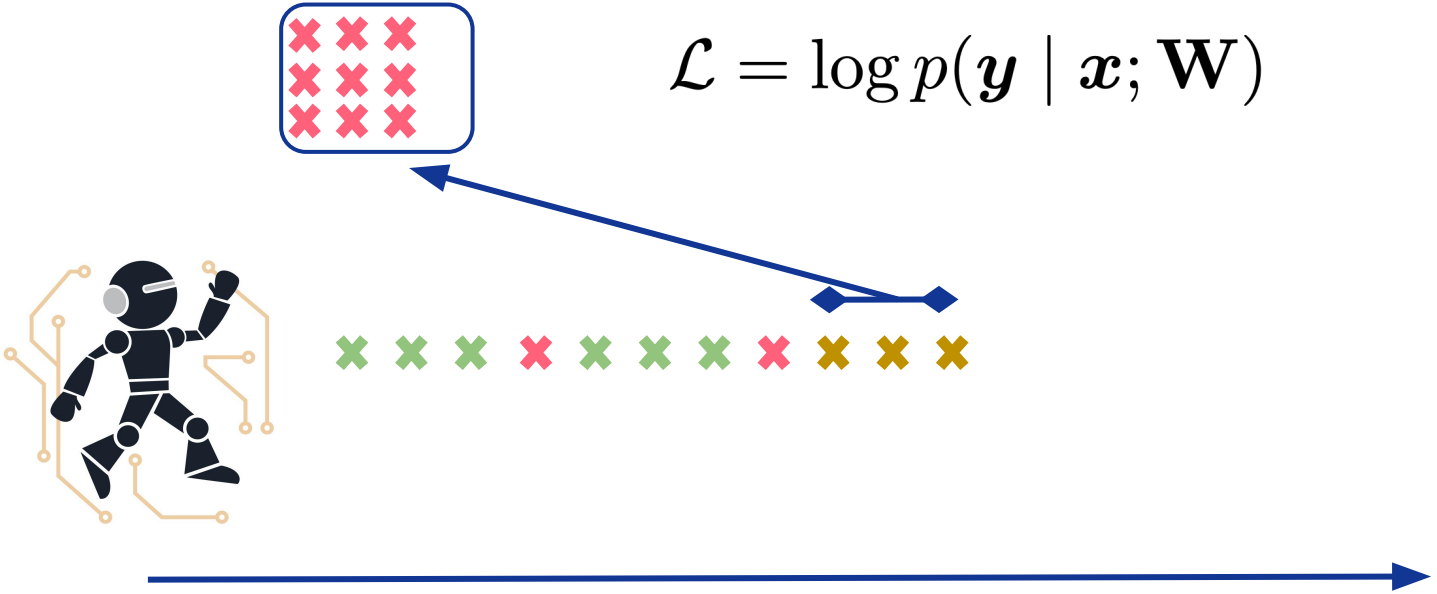


$$\mathcal{L} = \log p(\boldsymbol{y} \mid \boldsymbol{x}; \mathbf{W})$$

Training

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

Related to memory consolidation in human learning.

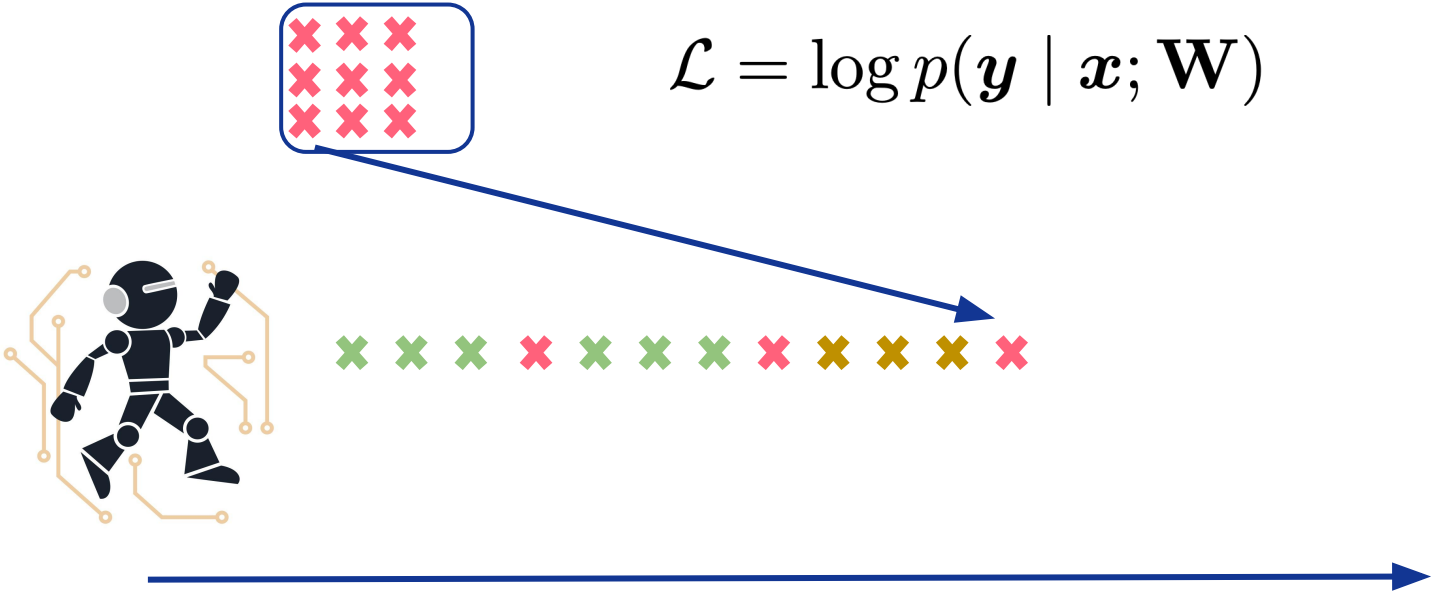


$$\mathcal{L} = \log p(\boldsymbol{y} \mid \boldsymbol{x}; \mathbf{W})$$

Training

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

Related to memory consolidation in human learning.

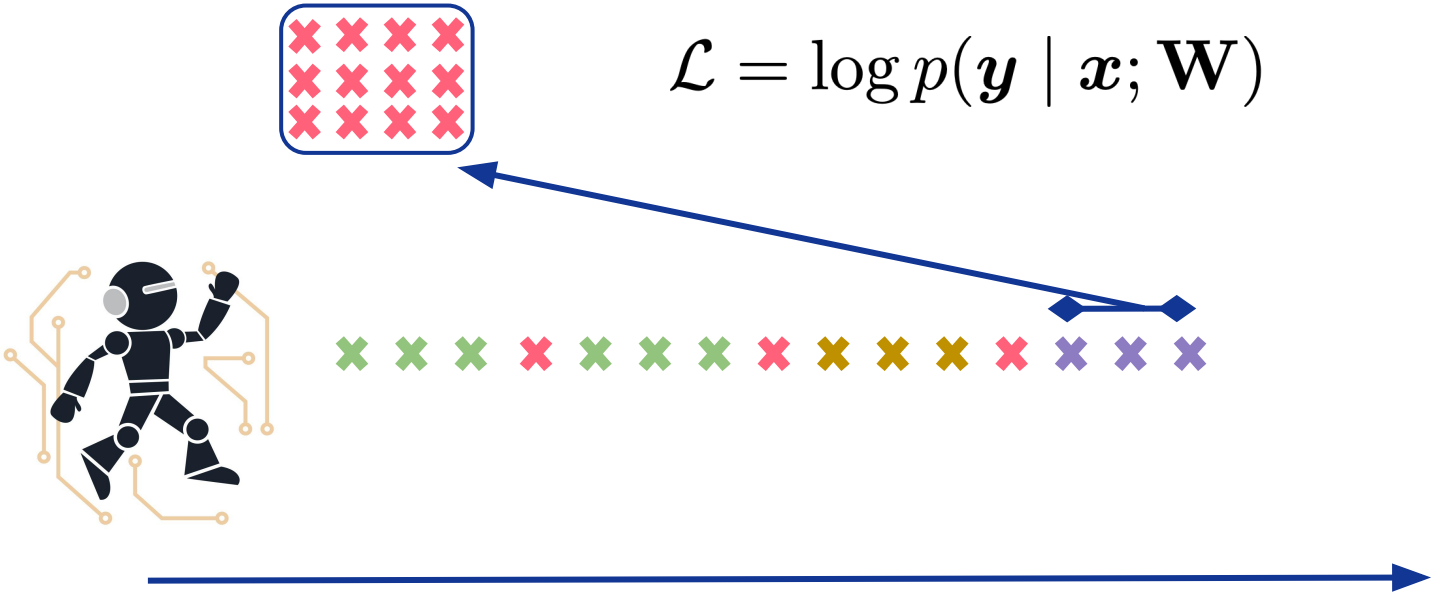


$$\mathcal{L} = \log p(\boldsymbol{y} \mid \boldsymbol{x}; \mathbf{W})$$

Training

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

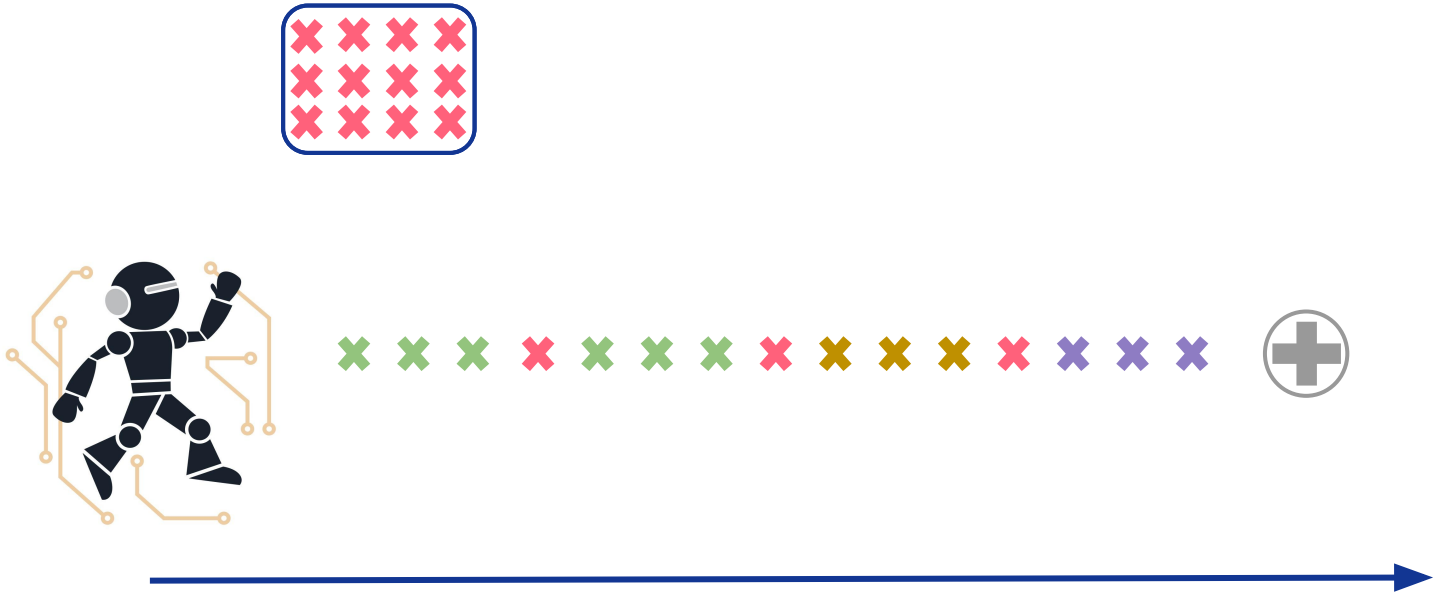
Related to memory consolidation in human learning.



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

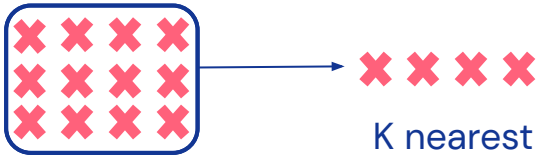
Inference (Prediction)

Local adaptation similar to MbPA (Sprechmann et al., 2018).

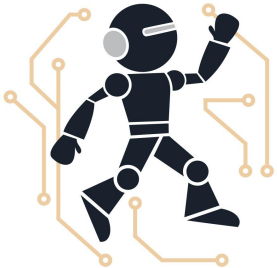


Inference (Prediction)

Local adaptation similar to MbPA (Sprechmann et al., 2018).



K nearest neighbors retrieval



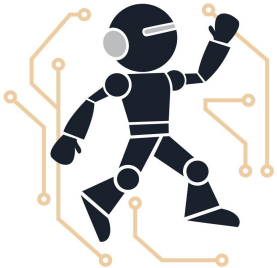
Query: in what country is normandy located
(17.48) in what area of france is calais located
(20.37) in what country is st john s located
(22.76) in what country is spoleto located
(23.12) in what part of africa is congo located
(23.83) on what island is palermo located

Inference (Prediction)

Local adaptation similar to MbPA (Sprechmann et al., 2018).

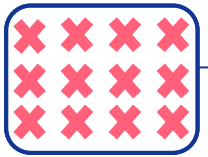
K nearest neighbors retrieval

Make a copy of itself



Inference (Prediction)

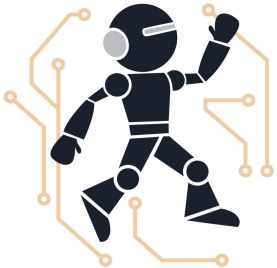
Local adaptation similar to MbPA (Sprechmann et al., 2018).



K nearest neighbors retrieval

Make a copy of itself

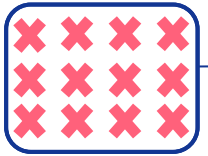
Retrain itself on the retrieved examples



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

Inference (Prediction)

Local adaptation similar to MbPA (Sprechmann et al., 2018).

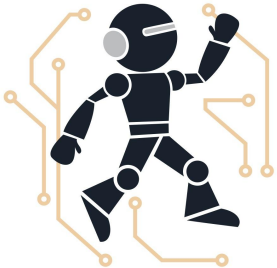


K nearest neighbors retrieval

Make a copy of itself

Retrain itself on the retrieved examples

Use the retrained copy to predict



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	Multitask (upper bound)
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.

10%	50%	100%
61.5	61.6	62.0

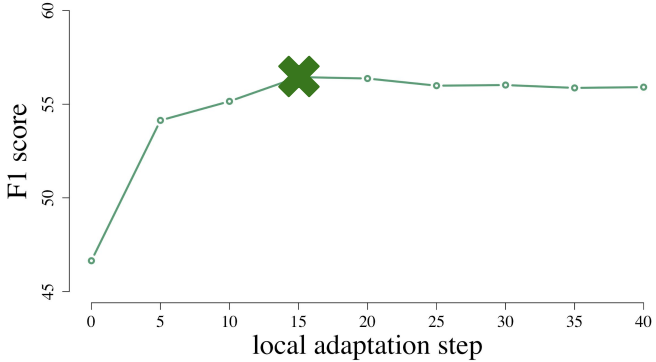
Takeaways and Limitations

- 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.

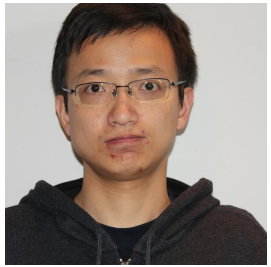
10%	50%	100%
61.5	61.6	62.0

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



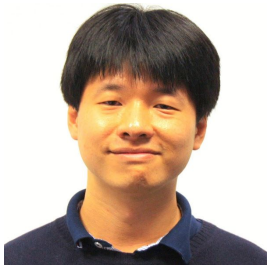
A Mutual Information Maximization Perspective of Language Representation Learning

Kong et al., ICLR 2020

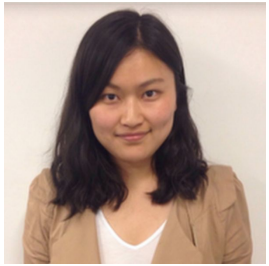


Lingpeng

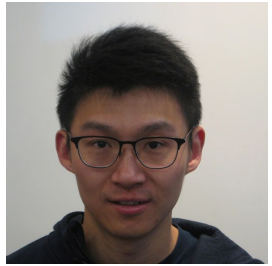
Cyprien



Wang



Lei

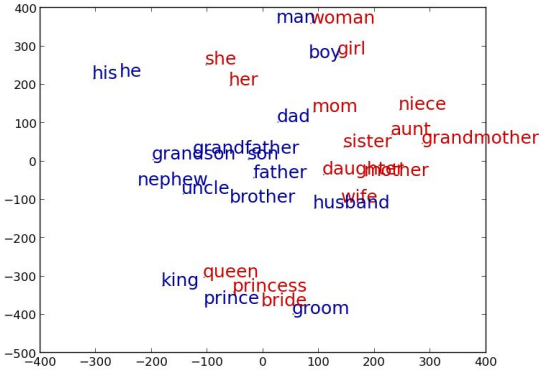


Zihang

Dani

Background

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



Skip gram, Mikolov et al., 2013.

GloVe, Pennington et al., 2014.

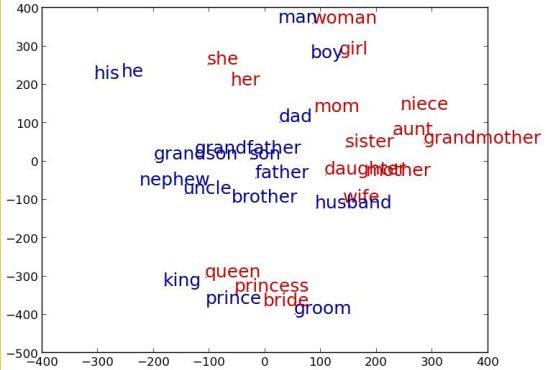
ELMo, Peters et al., 2018.

BERT, Devlin et al., 2019.

XLNet, Yang et al., 2019.

Background

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



Hypothesis: these methods are different instances of one framework.

Skip gram, Mikolov et al., 2013.

GloVe, Pennington et al., 2014.

ELMo, Peters et al., 2018.

BERT, Devlin et al., 2019.

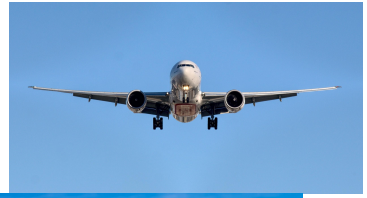
XLNet, Yang et al., 2019.

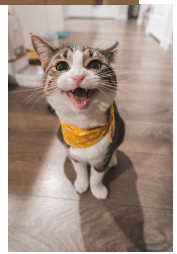
Contrastive Learning with InfoNCE

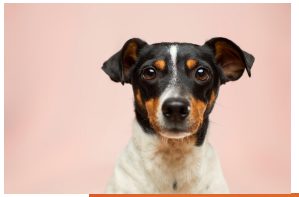
Main assumption: representations should capture similarity (Arora et al., 2019).

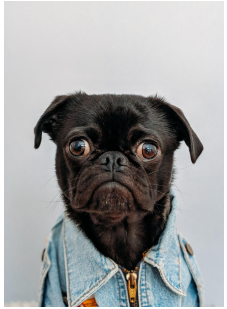
Contrastive Learning with InfoNCE

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









Contrastive Learning with InfoNCE

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).

$$I(A, B) \geq \mathbb{E}_{p(A,B)} \left[\mathbb{E}_{q(\tilde{\mathcal{B}})} \left[\log \frac{\exp f_{\boldsymbol{\theta}}(a, b)}{\sum_{\tilde{b} \in \tilde{\mathcal{B}}} \exp f_{\boldsymbol{\theta}}(a, \tilde{b})} \right] \right]$$

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).

$$I(A, B) \geq \mathbb{E}_{p(A,B)} \left[\mathbb{E}_{q(\tilde{\mathcal{B}})} \left[\log \frac{\boxed{\exp f_{\boldsymbol{\theta}}(a, b)}}{\sum_{\tilde{b} \in \tilde{\mathcal{B}}} \exp f_{\boldsymbol{\theta}}(a, \tilde{b})} \right] \right]$$

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).

$$I(A, B) \geq \mathbb{E}_{p(A,B)} \left[\mathbb{E}_{q(\tilde{\mathcal{B}})} \left[\log \frac{\exp f_{\boldsymbol{\theta}}(a, b)}{\boxed{\sum_{\tilde{b} \in \tilde{\mathcal{B}}} \exp f_{\boldsymbol{\theta}}(a, \tilde{b})}} \right] \right]$$

InfoNCE objective

Logeswaran and Lee, 2018

van den Oord, et al., 2019

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

Contrastive Learning with InfoNCE

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

$$I(A, B) \geq \mathbb{E}_{p(A,B)} \left[\mathbb{E}_{q(\tilde{\mathcal{B}})} \left[\log \frac{\exp f_{\boldsymbol{\theta}}(a, b)}{\sum_{\tilde{b} \in \tilde{\mathcal{B}}} \exp f_{\boldsymbol{\theta}}(a, \tilde{b})} \right] \right]$$

Carnegie Mellon University is located in Pittsburgh

Contrastive Learning with InfoNCE

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

$$I(A, B) \geq \mathbb{E}_{p(A,B)} \left[\mathbb{E}_{q(\tilde{\mathcal{B}})} \left[\log \frac{\exp \boxed{f_{\boldsymbol{\theta}}(a, b)}}{\sum_{\tilde{b} \in \tilde{\mathcal{B}}} \exp f_{\boldsymbol{\theta}}(a, \tilde{b})} \right] \right]$$

a b

Carnegie Mellon University is located in Pittsburgh

Contrastive Learning with InfoNCE

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

$$I(A,B) \geq \mathbb{E}_{p(A,B)}\left[\mathbb{E}_{q(\tilde{\mathcal{B}})}\left[\log \frac{\exp \boxed{f_{\boldsymbol{\theta}}(a,b)}}{\sum_{\tilde{b}\in\tilde{\mathcal{B}}} \exp f_{\boldsymbol{\theta}}(a,\tilde{b})}\right]\right]$$

a b a

Carnegie Mellon University is located in Pittsburgh

Contrastive Learning with InfoNCE

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

$$I(A, B) \geq \mathbb{E}_{p(A,B)} \left[\mathbb{E}_{q(\tilde{\mathcal{B}})} \left[\log \frac{\exp \boxed{f_{\boldsymbol{\theta}}(a, b)}}{\sum_{\tilde{b} \in \tilde{\mathcal{B}}} \exp f_{\boldsymbol{\theta}}(a, \tilde{b})} \right] \right]$$

a b

Carnegie Mellon University is located in Pittsburgh

Contrastive Learning with InfoNCE

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

$$I(A,B) \geq \mathbb{E}_{p(A,B)}\left[\mathbb{E}_{q(\tilde{\mathcal{B}})}\left[\log \frac{\exp \boxed{f_{\boldsymbol{\theta}}(a,b)}}{\sum_{\tilde{b}\in\tilde{\mathcal{B}}} \exp f_{\boldsymbol{\theta}}(a,\tilde{b})}\right]\right]$$

a b

Carnegie Mellon University is located in Pittsburgh

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

Contrastive Learning with InfoNCE

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

$$I(A,B) \geq \mathbb{E}_{p(A,B)}\left[\mathbb{E}_{q(\tilde{\mathcal{B}})}\left[\log \frac{\exp f_{\boldsymbol{\theta}}(a,b)}{\boxed{\sum_{\tilde{b}\in\tilde{\mathcal{B}}} \exp f_{\boldsymbol{\theta}}(a,\tilde{b})}}\right]\right]$$

{Pittsburgh, Boston, London, dog, cat}

a b

Carnegie Mellon University is located in Pittsburgh

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

Contrastive Learning with InfoNCE

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

$$I(A,B) \geq \mathbb{E}_{p(A,B)}\left[\boxed{\mathbb{E}_{q(\tilde{\mathcal{B}})}}\left[\log \frac{\exp f_{\boldsymbol{\theta}}(a,b)}{\sum_{\tilde{b}\in\tilde{\mathcal{B}}} \exp f_{\boldsymbol{\theta}}(a,\tilde{b})}\right]\right]$$

a b

Carnegie Mellon University is located in Pittsburgh

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

Contrastive Learning with InfoNCE

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

$$I(A,B) \geq \boxed{\mathbb{E}_{p(A,B)}} \left[\mathbb{E}_{q(\tilde{\mathcal{B}})} \left[\log \frac{\exp f_{\boldsymbol{\theta}}(a,b)}{\sum_{\tilde{b} \in \tilde{\mathcal{B}}} \exp f_{\boldsymbol{\theta}}(a,\tilde{b})} \right] \right]$$

a b

$\boxed{\text{Carnegie Mellon University}}$ is located in $\boxed{\text{Pittsburgh}}$

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

Contrastive Learning with InfoNCE

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

$$I(A, B) \geq \mathbb{E}_{p(A,B)} \left[\mathbb{E}_{q(\tilde{\mathcal{B}})} \left[\log \frac{\exp f_{\boldsymbol{\theta}}(a, b)}{\sum_{\tilde{b} \in \tilde{\mathcal{B}}} \exp f_{\boldsymbol{\theta}}(a, \tilde{b})} \right] \right]$$

a b

Carnegie Mellon University is located in Pittsburgh

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

Contrastive Learning with InfoNCE

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

$$I(A, B) \geq \mathbb{E}_{p(A,B)} \left[\mathbb{E}_{q(\tilde{\mathcal{B}})} \left[\log \frac{\exp f_{\boldsymbol{\theta}}(a, b)}{\sum_{\tilde{b} \in \tilde{\mathcal{B}}} \exp f_{\boldsymbol{\theta}}(a, \tilde{b})} \right] \right]$$

Harvard University is located in Cambridge

Contrastive Learning with InfoNCE

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

$$I(A, B) \geq \mathbb{E}_{p(A,B)} \left[\mathbb{E}_{q(\tilde{\mathcal{B}})} \left[\log \frac{\exp \boxed{f_{\boldsymbol{\theta}}(a, b)}}{\sum_{\tilde{b} \in \tilde{\mathcal{B}}} \exp f_{\boldsymbol{\theta}}(a, \tilde{b})} \right] \right]$$

a b

Harvard University is located in Cambridge

Contrastive Learning with InfoNCE

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

$$I(A,B) \geq \mathbb{E}_{p(A,B)}\left[\mathbb{E}_{q(\tilde{\mathcal{B}})}\left[\log \frac{\exp \boxed{f_{\boldsymbol{\theta}}(a,b)}}{\sum_{\tilde{b}\in\tilde{\mathcal{B}}} \exp f_{\boldsymbol{\theta}}(a,\tilde{b})}\right]\right]$$

a b a

Harvard University is located in Cambridge

Contrastive Learning with InfoNCE

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

$$I(A, B) \geq \mathbb{E}_{p(A,B)} \left[\mathbb{E}_{q(\tilde{\mathcal{B}})} \left[\log \frac{\exp \boxed{f_{\boldsymbol{\theta}}(a, b)}}{\sum_{\tilde{b} \in \tilde{\mathcal{B}}} \exp f_{\boldsymbol{\theta}}(a, \tilde{b})} \right] \right]$$

a b

Harvard University is located in Cambridge

Contrastive Learning with InfoNCE

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

$$I(A,B) \geq \mathbb{E}_{p(A,B)}\left[\mathbb{E}_{q(\tilde{\mathcal{B}})}\left[\log \frac{\exp \boxed{f_{\boldsymbol{\theta}}(a,b)}}{\sum_{\tilde{b}\in\tilde{\mathcal{B}}} \exp f_{\boldsymbol{\theta}}(a,\tilde{b})}\right]\right]$$

a b

Harvard University is located in Cambridge

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

Contrastive Learning with InfoNCE

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

$$I(A,B) \geq \mathbb{E}_{p(A,B)}\left[\mathbb{E}_{q(\tilde{\mathcal{B}})}\left[\log \frac{\exp f_{\boldsymbol{\theta}}(a,b)}{\boxed{\sum_{\tilde{b}\in\tilde{\mathcal{B}}} \exp f_{\boldsymbol{\theta}}(a,\tilde{b})}}\right]\right]$$

{Cambridge, Pittsburgh, London, dog, cat}

a b

Harvard University is located in Cambridge

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

Contrastive Learning with InfoNCE

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

$$I(A,B) \geq \mathbb{E}_{p(A,B)}\left[\boxed{\mathbb{E}_{q(\tilde{\mathcal{B}})}}\left[\log \frac{\exp f_{\boldsymbol{\theta}}(a,b)}{\sum_{\tilde{b}\in\tilde{\mathcal{B}}} \exp f_{\boldsymbol{\theta}}(a,\tilde{b})}\right]\right]$$

a b

Harvard University is located in Cambridge

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

Contrastive Learning with InfoNCE

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

$$I(A,B) \geq \boxed{\mathbb{E}_{p(A,B)}}\left[\mathbb{E}_{q(\tilde{\mathcal{B}})}\left[\log \frac{\exp f_{\boldsymbol{\theta}}(a,b)}{\sum_{\tilde{b}\in\tilde{\mathcal{B}}} \exp f_{\boldsymbol{\theta}}(a,\tilde{b})}\right]\right]$$

a b

Harvard University is located in Cambridge

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

Contrastive Learning with InfoNCE

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

$$I(A,B) \geq \mathbb{E}_{p(A,B)}\left[\mathbb{E}_{q(\tilde{\mathcal{B}})}\left[\log \frac{\exp f_{\boldsymbol{\theta}}(a,b)}{\sum_{\tilde{b}\in\tilde{\mathcal{B}}} \exp f_{\boldsymbol{\theta}}(a,\tilde{b})}\right]\right]$$

a b

Harvard University is located in Cambridge

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

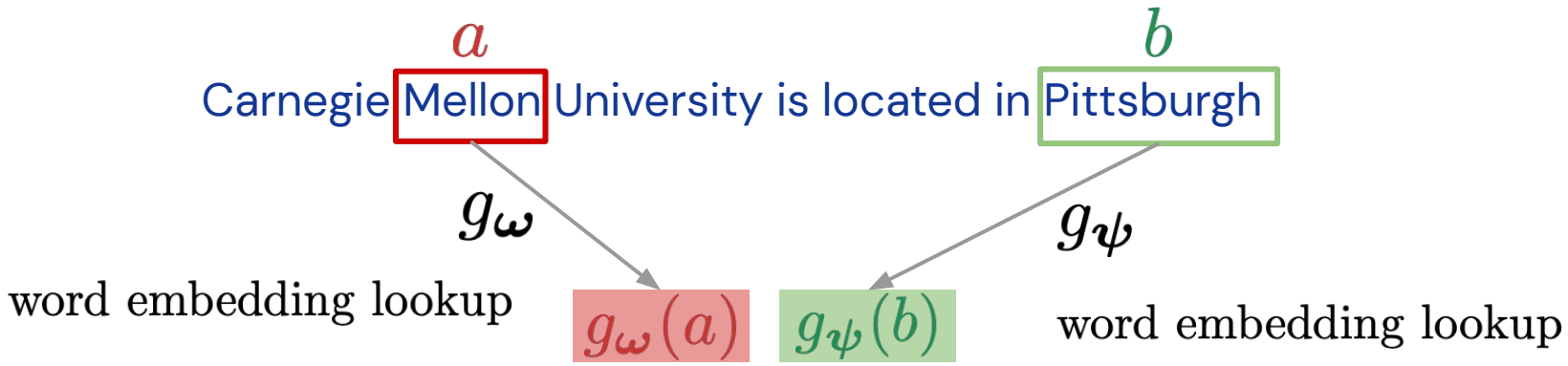
Skip-gram

Mikolov et al., 2013

Carnegie Mellon University is located in Pittsburgh

Skip-gram

Mikolov et al., 2013



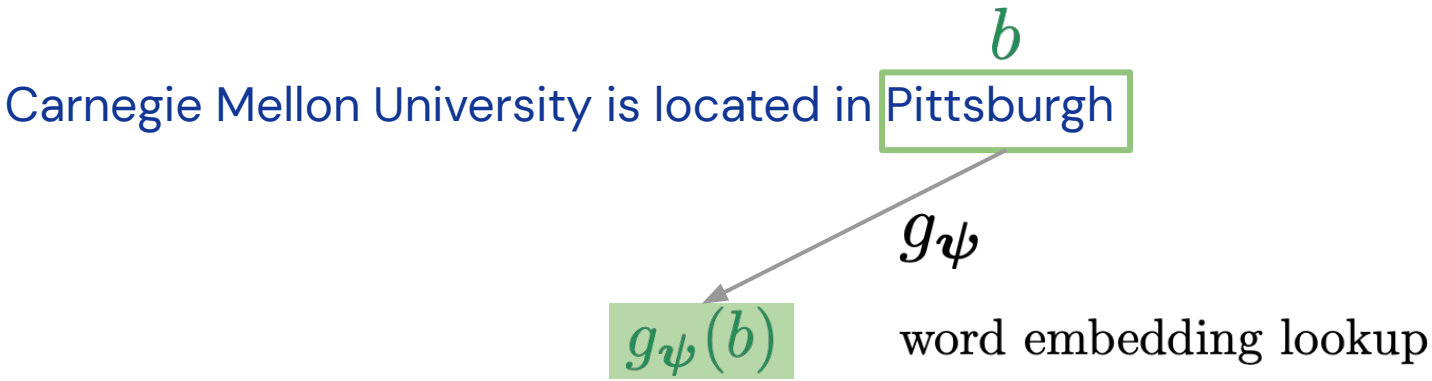
BERT

Devlin et al., 2019

Carnegie Mellon University is located in Pittsburgh

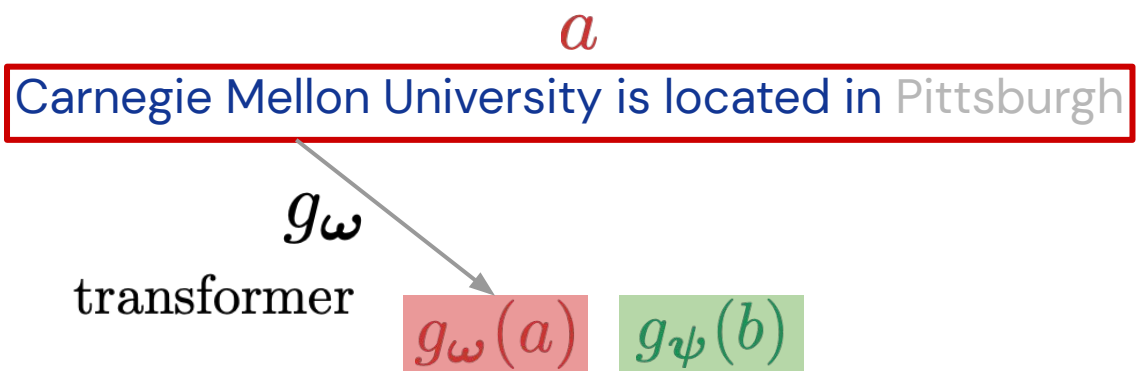
BERT

Devlin et al., 2019



BERT

Devlin et al., 2019



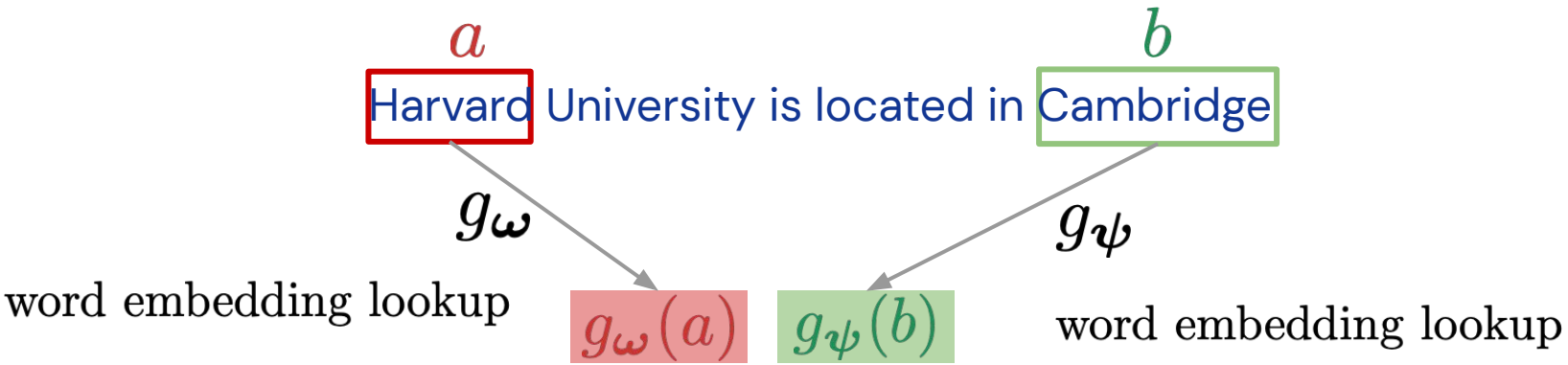
Skip-gram

Mikolov et al., 2013

Harvard University is located in Cambridge

Skip-gram

Mikolov et al., 2013



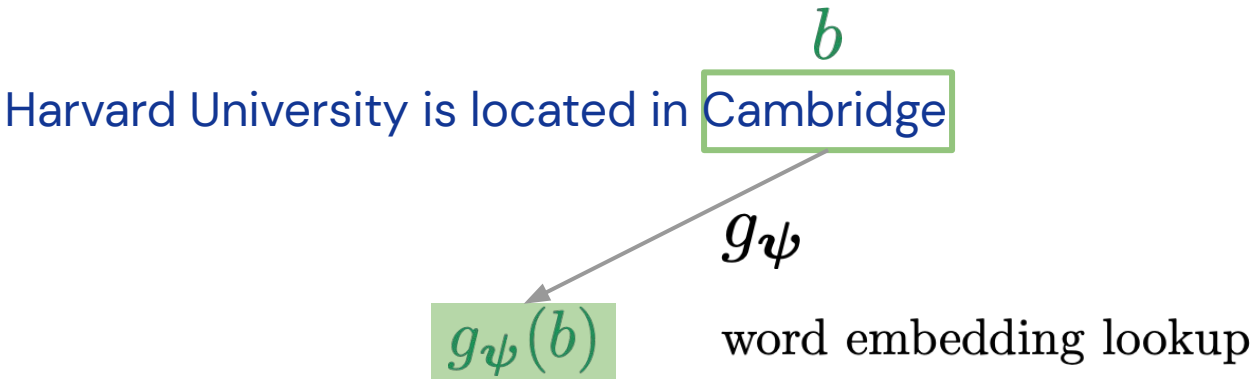
BERT

Devlin et al., 2019

Harvard University is located in Cambridge

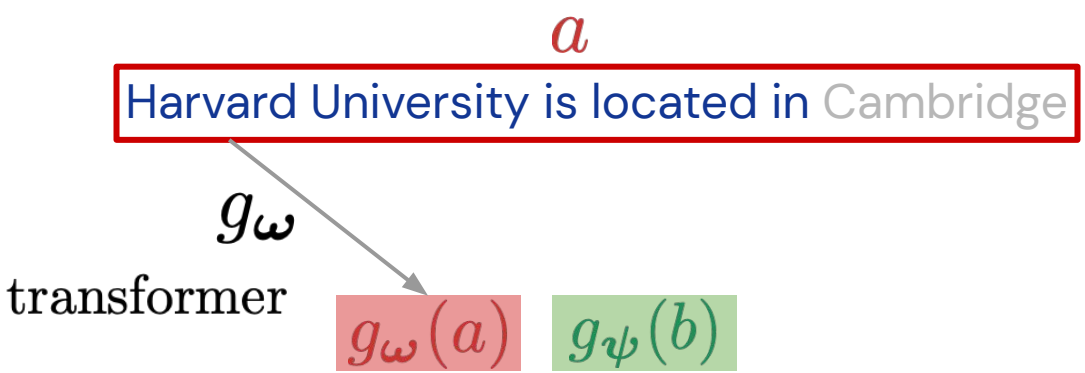
BERT

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-MLM**	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-MLM**	context	word	transformer	lookup
Yang et al., 2019	**XLNet**	context	word	TXL++	lookup

- Connections to representation learning 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-MLM**	context	word	transformer	lookup
Yang et al., 2019	**XLNet**	context	word	TXL++	lookup

- Connections to representation learning methods used in other domains (vision, speech).

- A better understanding on how to construct new self-supervised tasks.

Model

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

Model

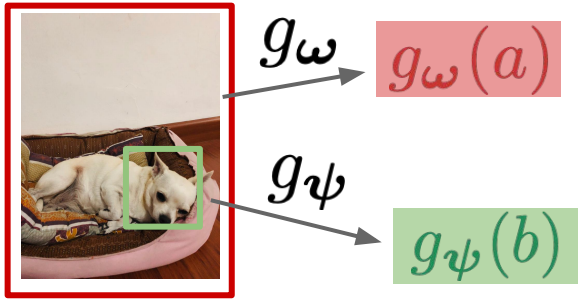
Deep InfoMax (DIM; Hjelm et al., 2019)

g_ω

$g_\omega(a)$

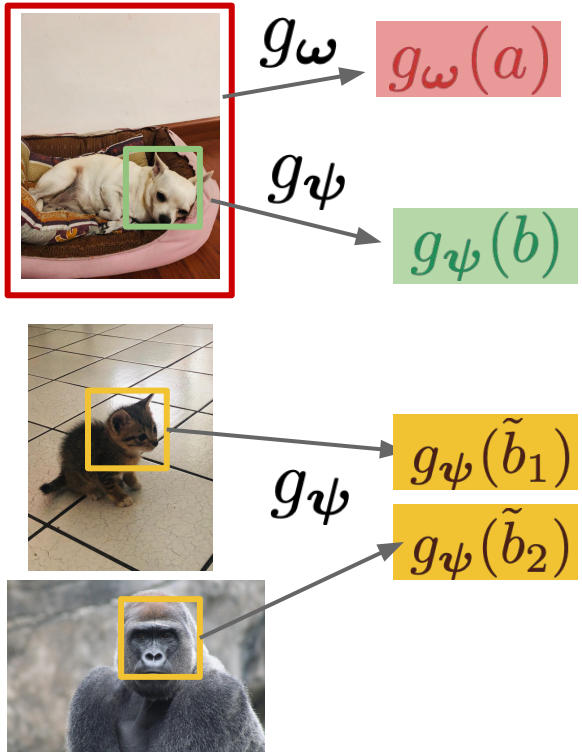
Model

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



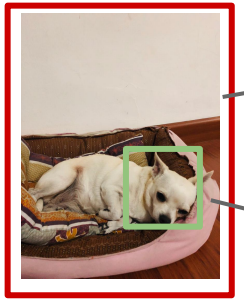
Model

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



Model

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

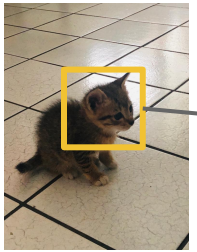


$g_{\boldsymbol{\omega}}$

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

$g_{\boldsymbol{\psi}}$

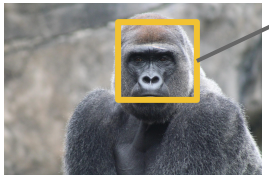
$g_{\boldsymbol{\psi}}(b)$



$g_{\boldsymbol{\psi}}$

$g_{\boldsymbol{\psi}}(\tilde{b}_1)$

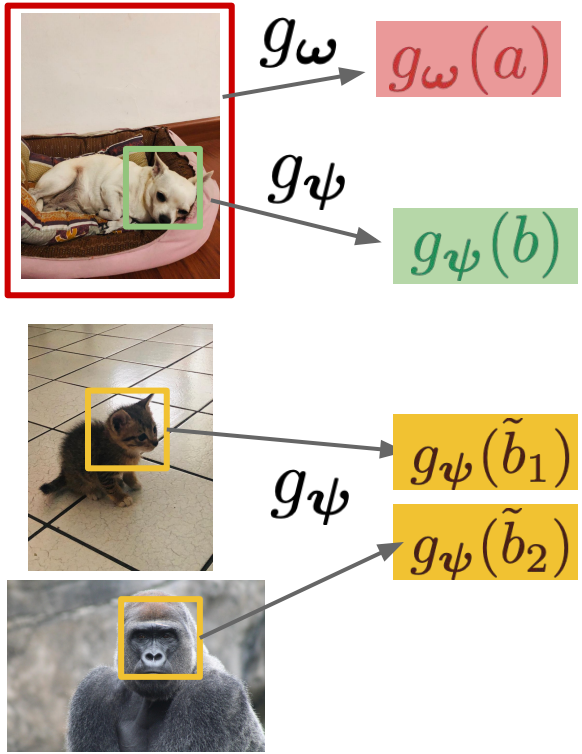
$g_{\boldsymbol{\psi}}(\tilde{b}_2)$



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

Model

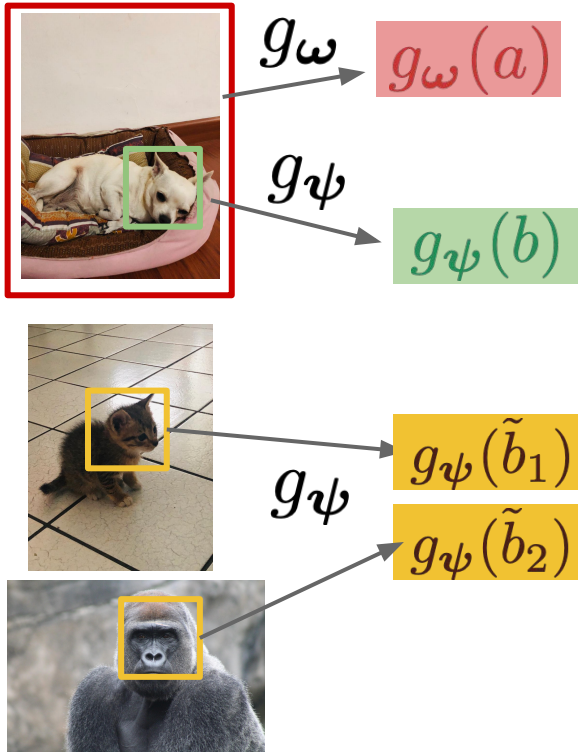
Deep InfoMax (DIM; Hjelm et al., 2019)

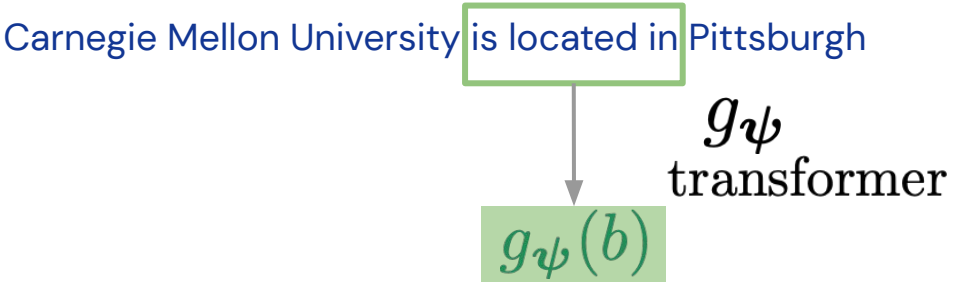


Carnegie Mellon University is located in Pittsburgh

Model

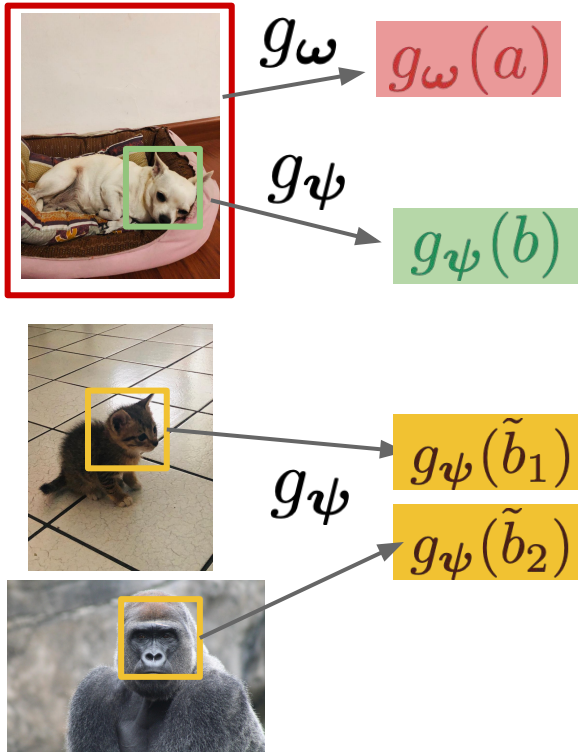
Deep InfoMax (DIM; Hjelm et al., 2019)

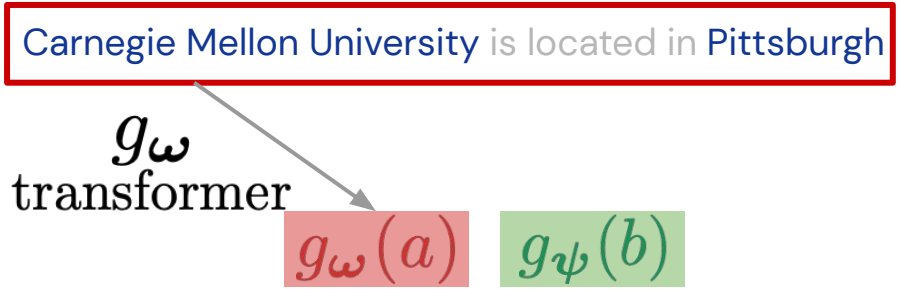




Model

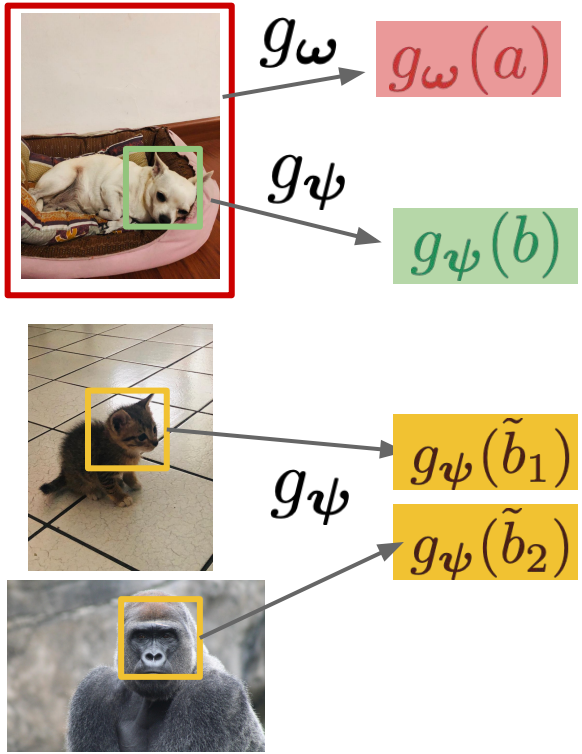
Deep InfoMax (DIM; Hjelm et al., 2019)





Model

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



Carnegie Mellon University is located in Pittsburgh

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

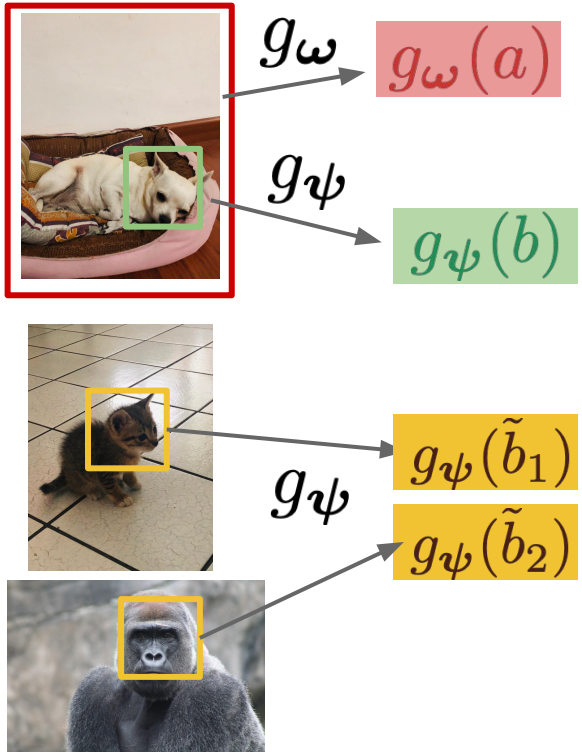
Starcraft II is a fun game

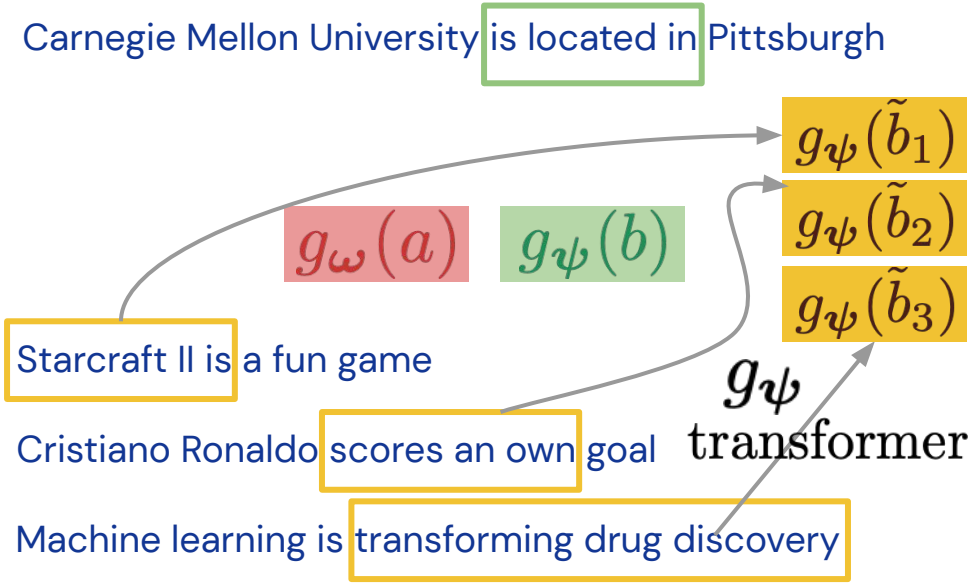
Cristiano Ronaldo scores an own goal

Machine learning is transforming drug discovery

Model

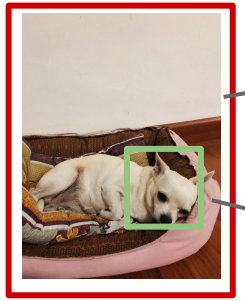
Deep InfoMax (DIM; Hjelm et al., 2019)





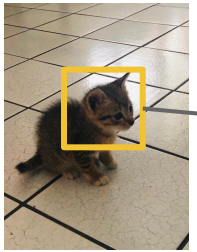
Model

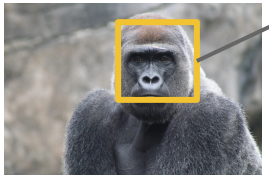
Deep InfoMax (DIM; Hjelm et al., 2019)



g_ω → $g_\omega(a)$

g_ψ → $g_\psi(b)$





g_ψ → $g_\psi(\tilde{b}_1)$, $g_\psi(\tilde{b}_2)$

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

Carnegie Mellon University is located in Pittsburgh

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

$g_\psi(\tilde{b}_1)$

$g_\psi(\tilde{b}_2)$

$g_\psi(\tilde{b}_3)$

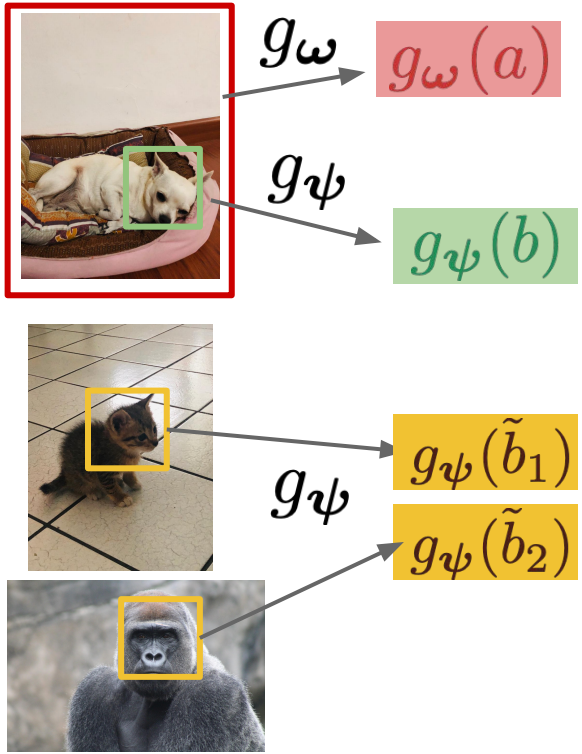
Starcraft II is a fun game

Cristiano Ronaldo scores an own goal

Machine learning is transforming drug discovery

Model

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



Carnegie Mellon University is located in Pittsburgh

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

$g_\psi(\tilde{b}_1)$
$g_\psi(\tilde{b}_2)$
$g_\psi(\tilde{b}_3)$

Starcraft II is a fun game

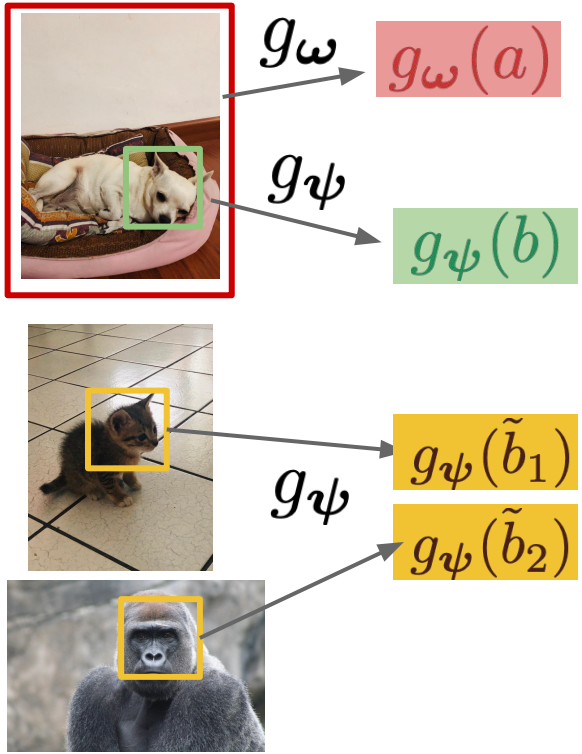
Cristiano Ronaldo scores an own goal

Machine learning is transforming drug discovery

$$\mathcal{I}_{\text{INFOWORD}} = \lambda_{\text{MLM}}\mathcal{I}_{\text{MLM}} + \lambda_{\text{DIM}}\mathcal{I}_{\text{DIM}}$$

Model

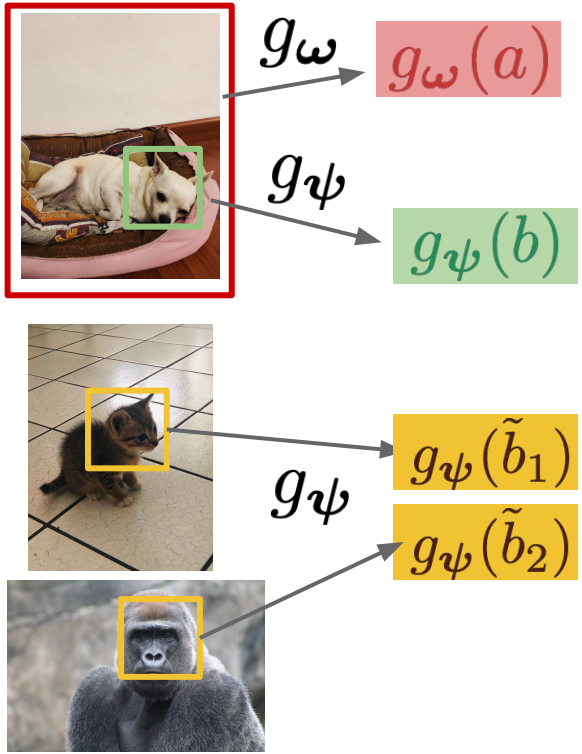
Deep InfoMax (DIM; Hjelm et al., 2019)

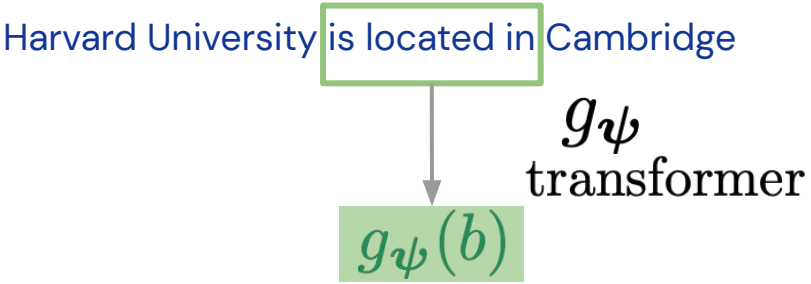


Harvard University is located in Cambridge

Model

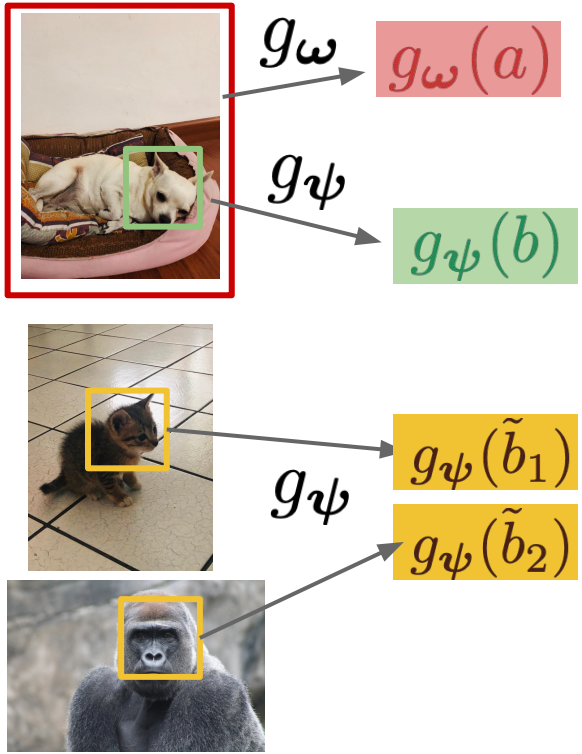
Deep InfoMax (DIM; Hjelm et al., 2019)

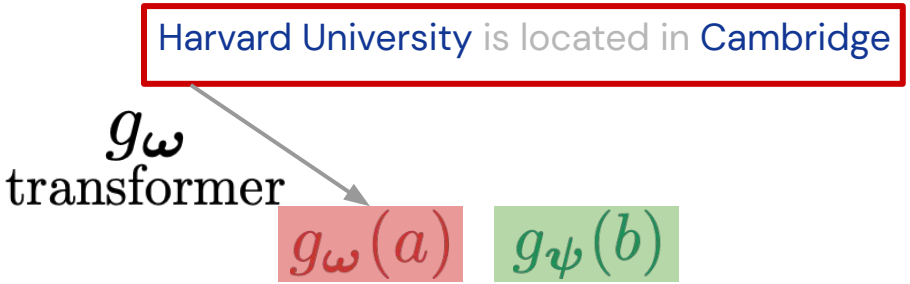




Model

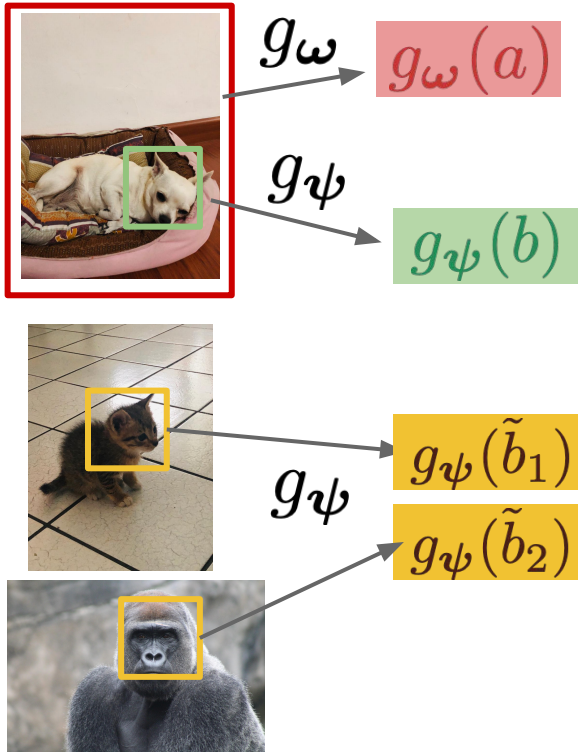
Deep InfoMax (DIM; Hjelm et al., 2019)





Model

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



Harvard University is located in Cambridge

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

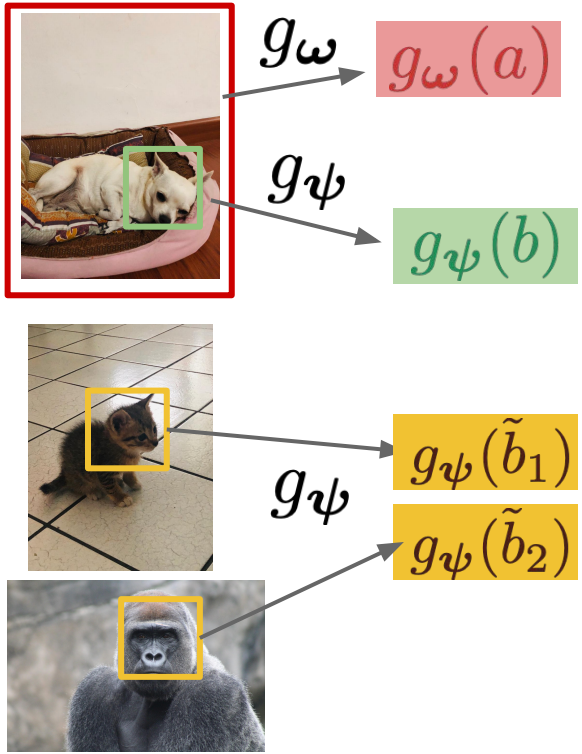
Starcraft II is a fun game

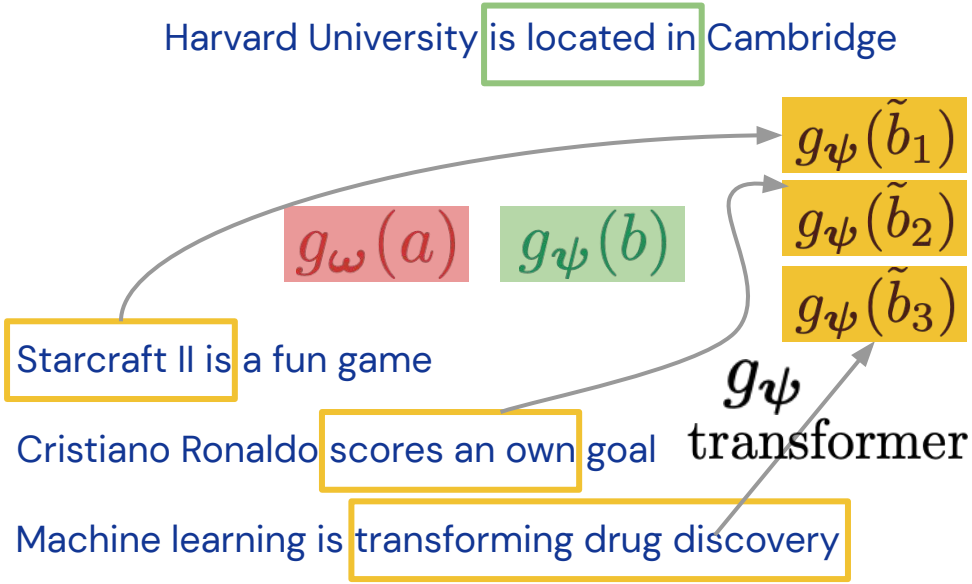
Cristiano Ronaldo scores an own goal

Machine learning is transforming drug discovery

Model

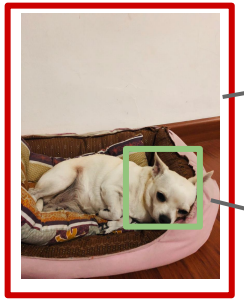
Deep InfoMax (DIM; Hjelm et al., 2019)





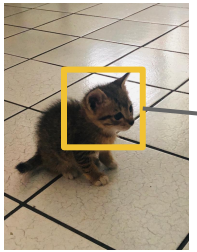
Model

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

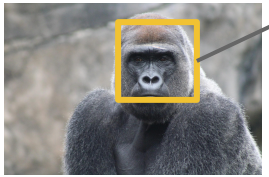


g_ω → $g_\omega(a)$

g_ψ → $g_\psi(b)$



g_ψ → $g_\psi(\tilde{b}_1)$



→ $g_\psi(\tilde{b}_2)$

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

Harvard University is located in Cambridge

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

$g_\psi(\tilde{b}_1)$

$g_\psi(\tilde{b}_2)$

$g_\psi(\tilde{b}_3)$

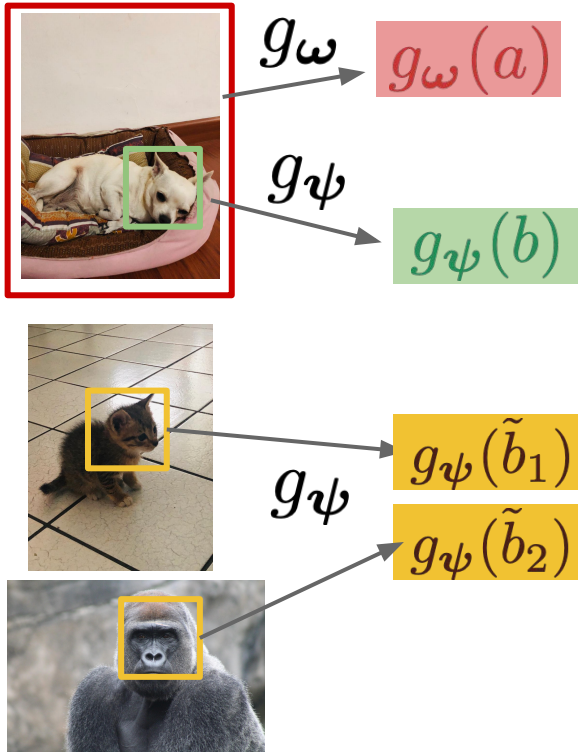
Starcraft II is a fun game

Cristiano Ronaldo scores an own goal

Machine learning is transforming drug discovery

Model

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



Harvard University is located in Cambridge

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

$g_\psi(\tilde{b}_1)$
$g_\psi(\tilde{b}_2)$
$g_\psi(\tilde{b}_3)$

Starcraft II is a fun game

Cristiano Ronaldo scores an own goal

Machine learning is transforming drug discovery

$$\mathcal{I}_{\text{INFOWORD}} = \lambda_{\text{MLM}}\mathcal{I}_{\text{MLM}} + \lambda_{\text{DIM}}\mathcal{I}_{\text{DIM}}$$

Experiments

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

		F1	Exact Match
Small Model	BERT	90.9	84.4
	Ours	**91.4**	**84.7**
Large Model	BERT	92.7	86.6
	Ours	**93.1**	**87.3**

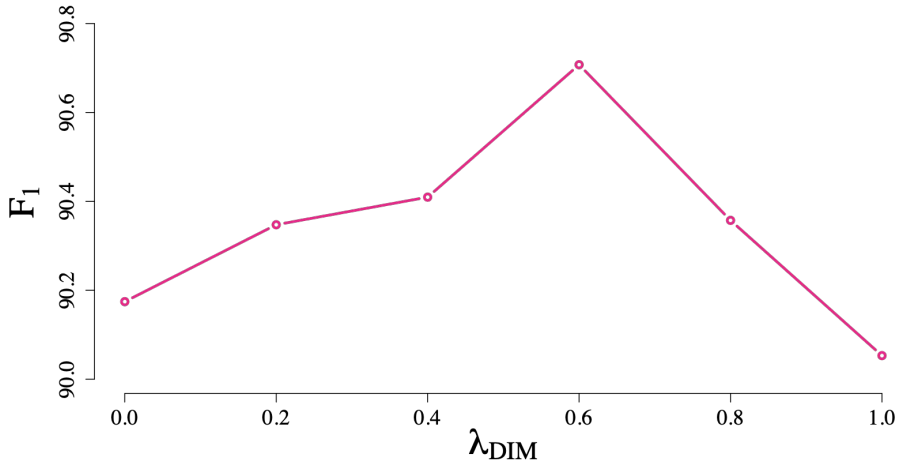
F1 and exact match scores (0–100), higher is better

BERT: Devlin et al., 2019

Experiments

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

		F1	Exact Match
Small Model	BERT	90.9	84.4
	Ours	**91.4**	**84.7**
Large Model	BERT	92.7	86.6
	Ours	**93.1**	**87.3**



$$\mathcal{I}_{\text{InfoWord}} = \lambda_{\text{MLM}}\mathcal{I}_{\text{MLM}} + \lambda_{\text{DIM}}\mathcal{I}_{\text{DIM}}$$

F1 and exact match scores (0–100), higher is better

BERT: Devlin et al., 2019

Takeaways and Limitations

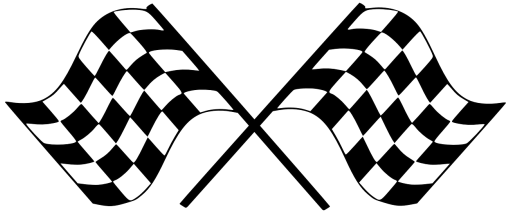
- Progress in language representation learning has largely been driven by advances in model architectures.

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

Takeaways and Limitations

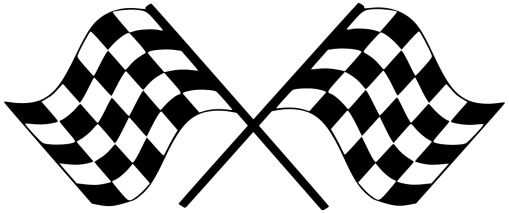
- Progress in language representation learning has largely been driven by advances in model architectures.
- It is possible to transfer ideas across domains when designing self-supervised tasks.
- All variants of existing models, fail to incorporate **global context** (they rely on local views).

Future Directions



A universal language model that continually and efficiently learns to perform multiple complex tasks in many languages.

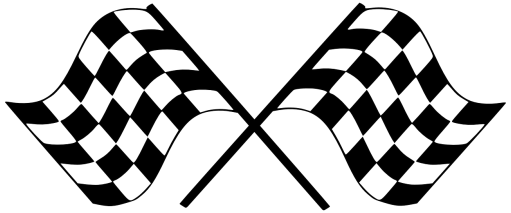
Future Directions



A universal language model that continually and efficiently learns to perform multiple complex tasks in many languages.

Generative Models

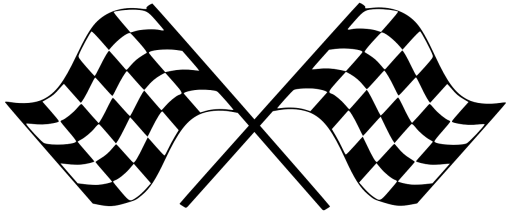
Future Directions



A universal language model that continually and efficiently learns to perform multiple complex tasks in many languages.

Generative Models

Future Directions

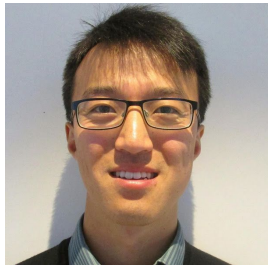


A universal language model that continually and efficiently learns to perform multiple complex tasks in many languages.

Generative Models

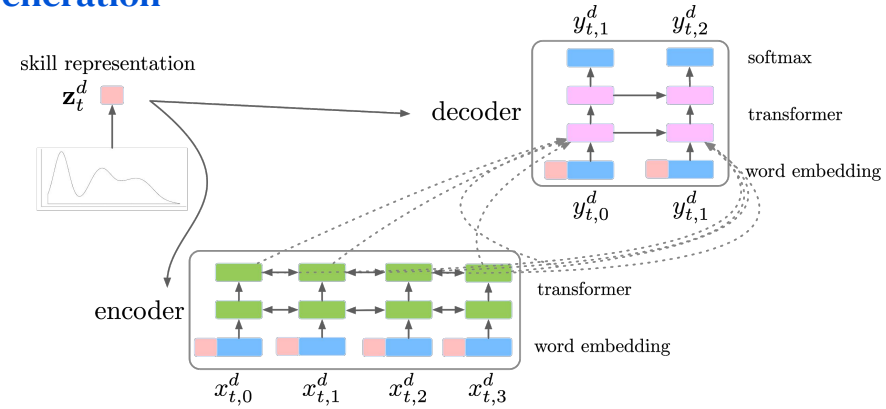
Modelling Latent Skills for Multitask Language Generation

Cao and Yogatama, arXiv 2020

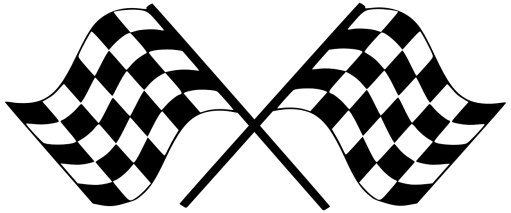


Kris

Dani



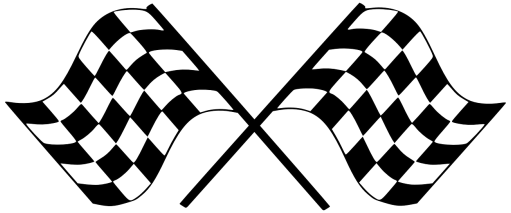
Future Directions



Memory

A universal language model that continually and efficiently learns to perform multiple complex tasks in many languages.

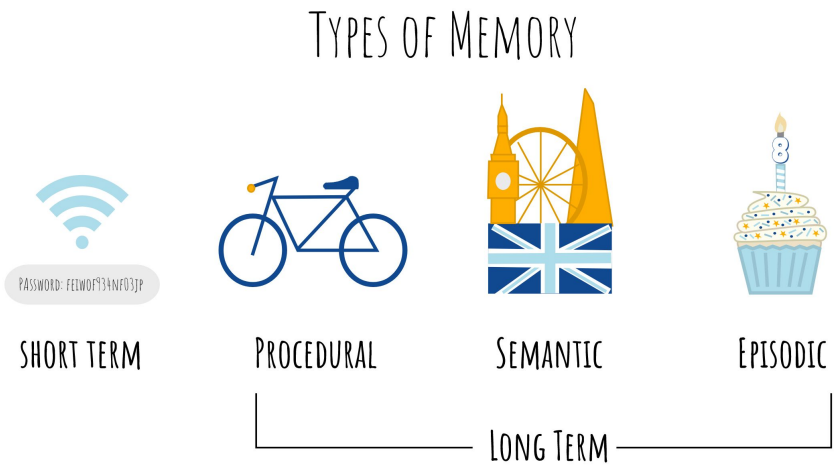
Future Directions



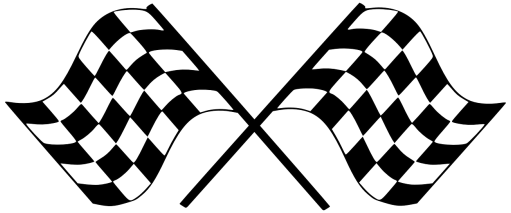
A universal language model that continually and efficiently learns to perform multiple complex tasks in many languages.

Memory

Image from https://kwiklearning.com/kwik-tips/the-memory-lanes-in-your-brain-how-we-make-memories/



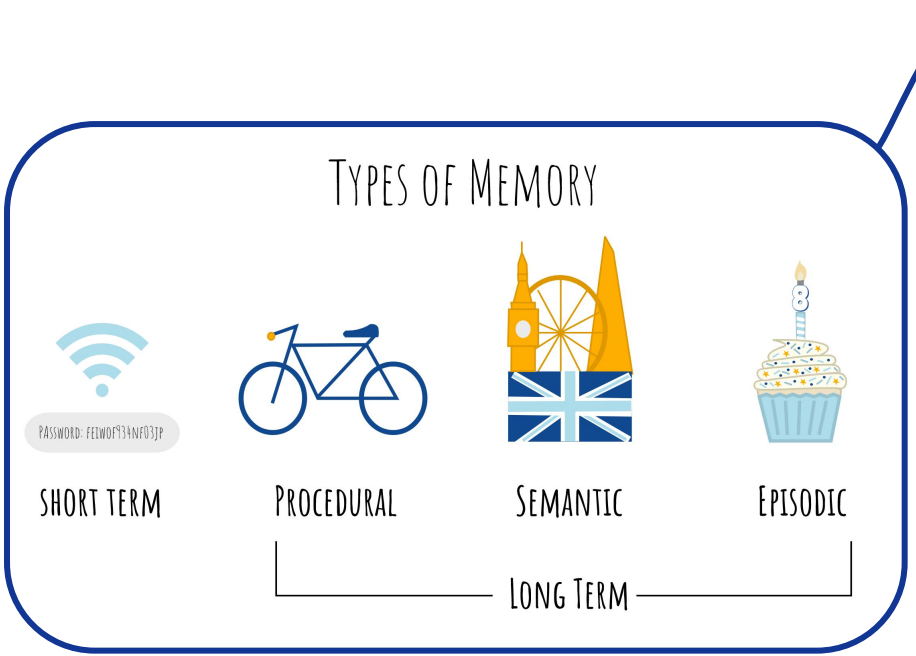
Future Directions



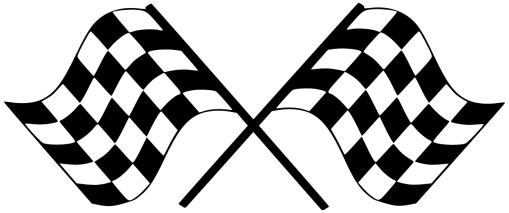
A universal language model that continually and efficiently learns to perform multiple complex tasks in many languages.

Memory

Image from https://kwiklearning.com/kwik-tips/the-memory-lanes-in-your-brain-how-we-make-memories/



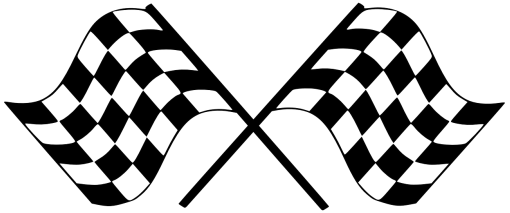
Future Directions



A universal language model that continually and efficiently learns to perform multiple complex tasks in many languages.

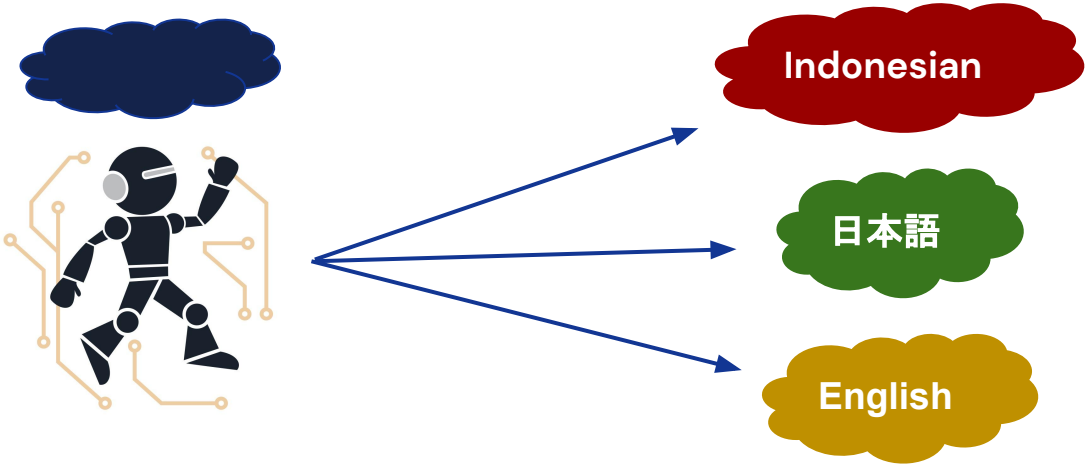
Representation Learning

Future Directions

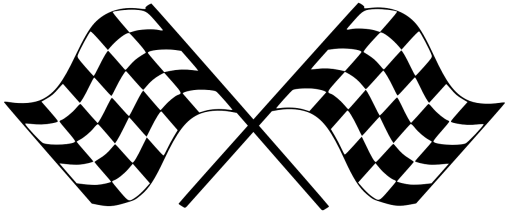


A universal language model that continually and efficiently learns to perform multiple complex tasks in many languages.

Representation Learning



Future Directions



A universal language model that continually and efficiently learns to perform multiple complex tasks in many languages.

Representation Learning

On the Crosslingual Transferability of Monolingual Representations

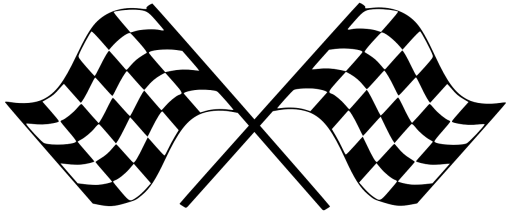
Artetxe et al., arXiv 2019

Mikel

Sebastian

Dani

Future Directions



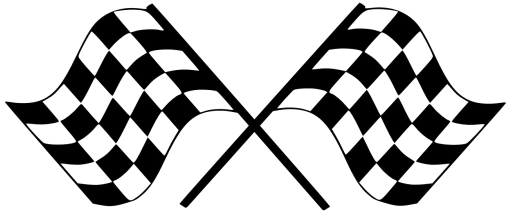
A universal language model that continually and efficiently learns to perform multiple complex tasks in many languages.

Memory

Representation Learning

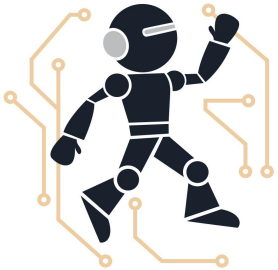
Generative Models

Future Directions

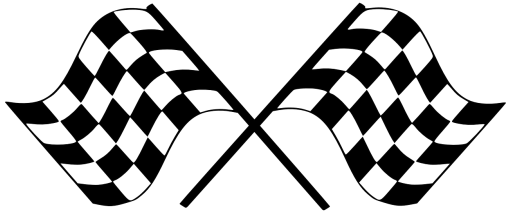


A universal language model that continually and efficiently learns to perform multiple complex tasks in many languages.

Memory	Representation Learning	Generative Models

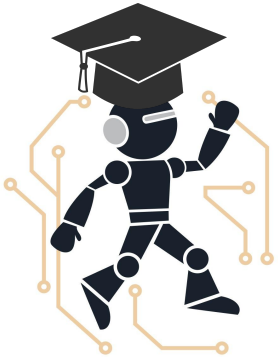


Future Directions



A universal language model that continually and efficiently learns to perform multiple complex tasks in many languages.

Memory | Representation Learning | Generative Models



tack շնորհակալություն Danke
ありがとうございました Salamat
grazie **Thank you** multumesc
धन्यवाद நன்றி
Terima kasih Dankie 감사합니다 Merci
Спасибо شكرا لكم σας ευχαριστώ
teşekkür ederim 谢谢 cảm ơn bạn

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

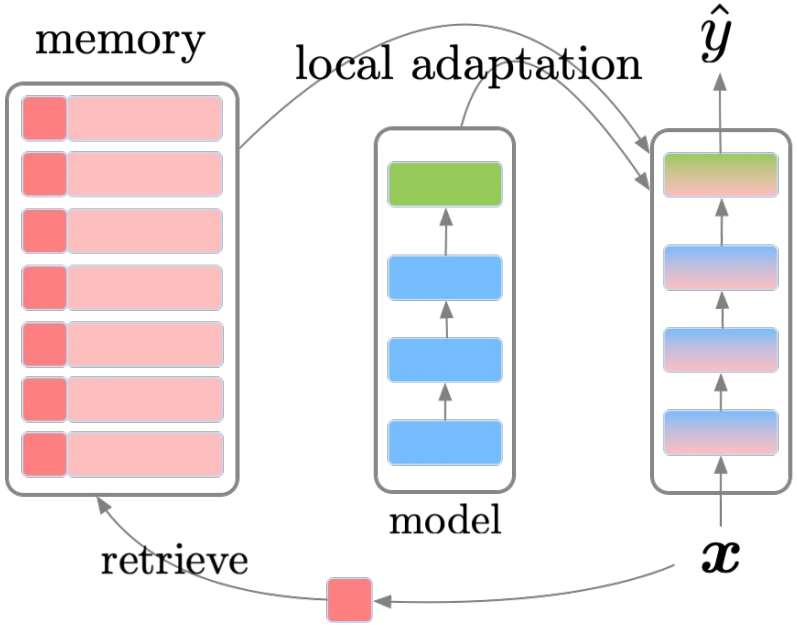
DeepMind

Future Directions

- The future is generative (models).
  - Elegantly deal with multiple tasks (McCann et al., 2018; Radford et al., 2019).
  - Approach their asymptotic error faster (Yogatama et al., arXiv 2017).
  - Crucial to imagination and planning (Weber et al., 2017).

Inference (Prediction)

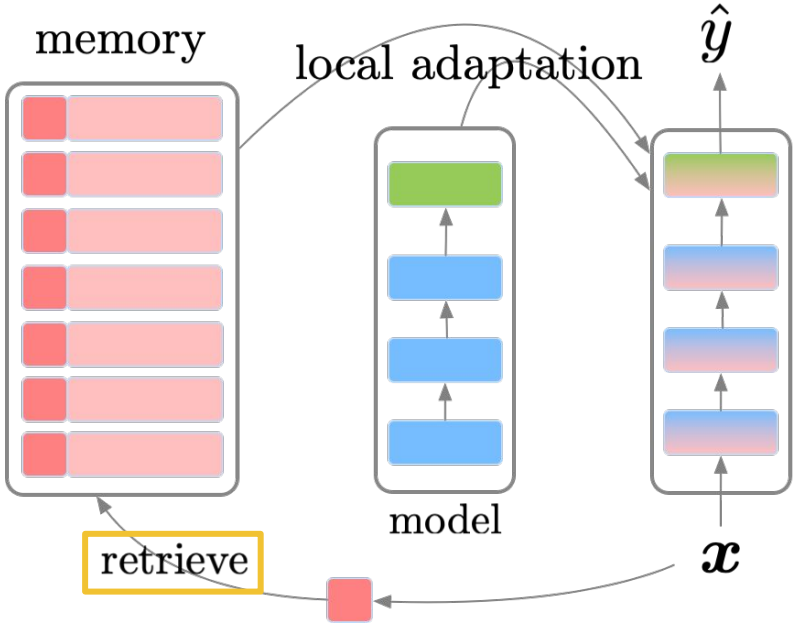
Local adaptation similar to MbPA (Sprechmann et al., 2018).



Inference (Prediction)

Local adaptation similar to MbPA (Sprechmann et al., 2018).

Query: in what country is normandy located
(17.48) in what area of france is calais located
(20.37) in what country is st john s located
(22.76) in what country is spoleto located
(23.12) in what part of africa is congo located
(23.83) on what island is palermo located

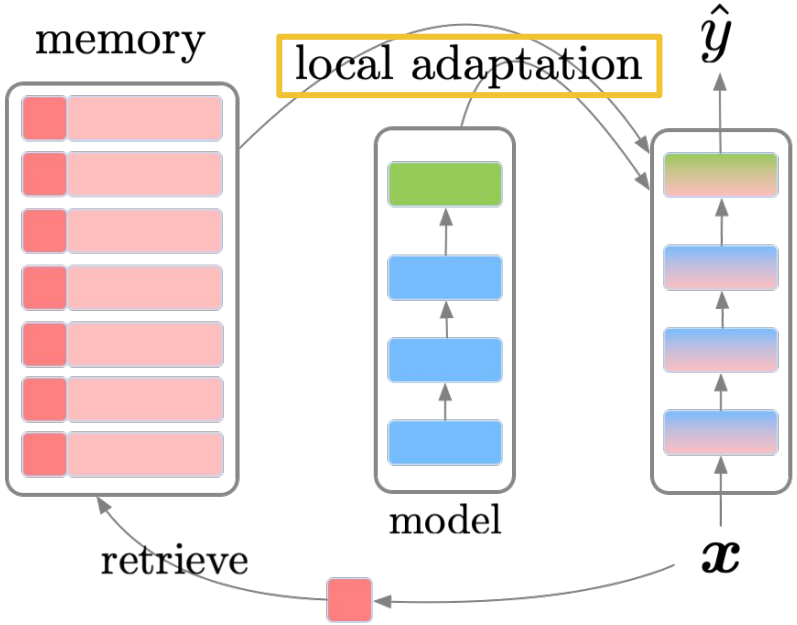


Inference (Prediction)

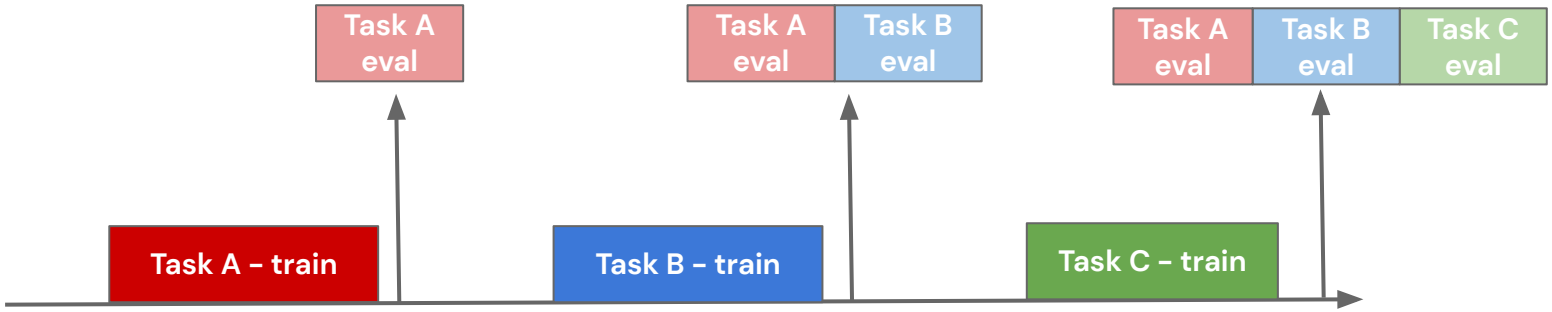
Local adaptation similar to MbPA (Sprechmann et al., 2018).

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

Query: in what country is normandy located
(17.48) in what area of france is calais located
(20.37) in what country is st john s located
(22.76) in what country is spoleto located
(23.12) in what part of africa is congo located
(23.83) on what island is palermo located



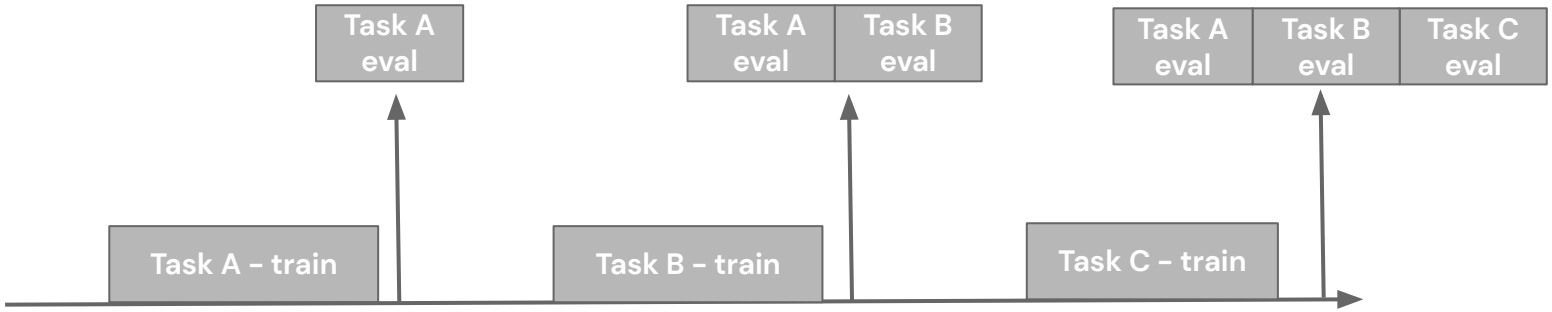
Problem Setup



Standard setup, models know which task an example belongs to.

Never-Ending Language Learning, Mitchell et al., 2015

Problem Setup



Our more difficult and realistic setup.

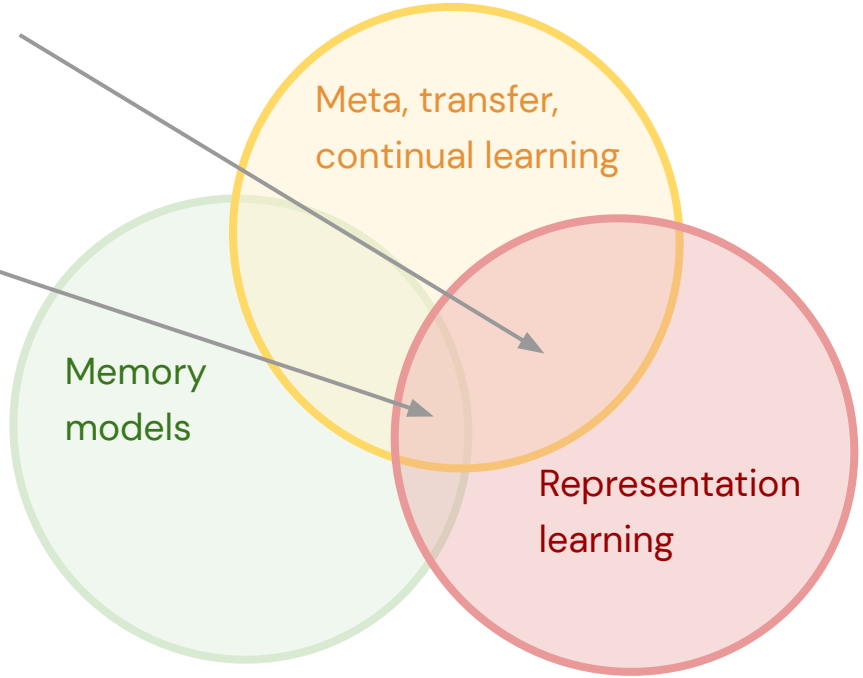
This Talk

- Episodic memory in lifelong language learning.

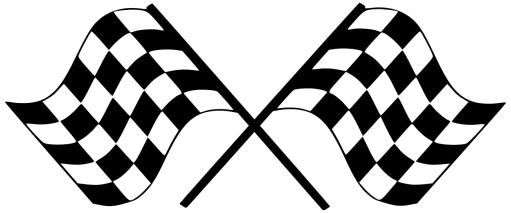
  de Masson d'Autume et al., NeurIPS 2019

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

  Kong et al., ICLR 2020



Research Areas

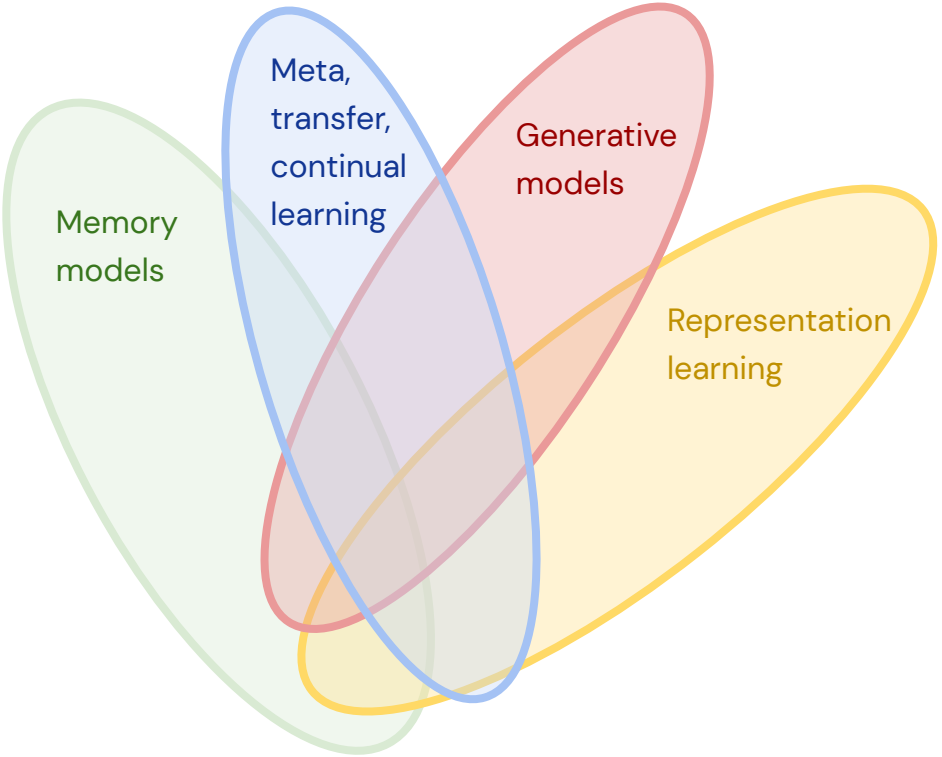
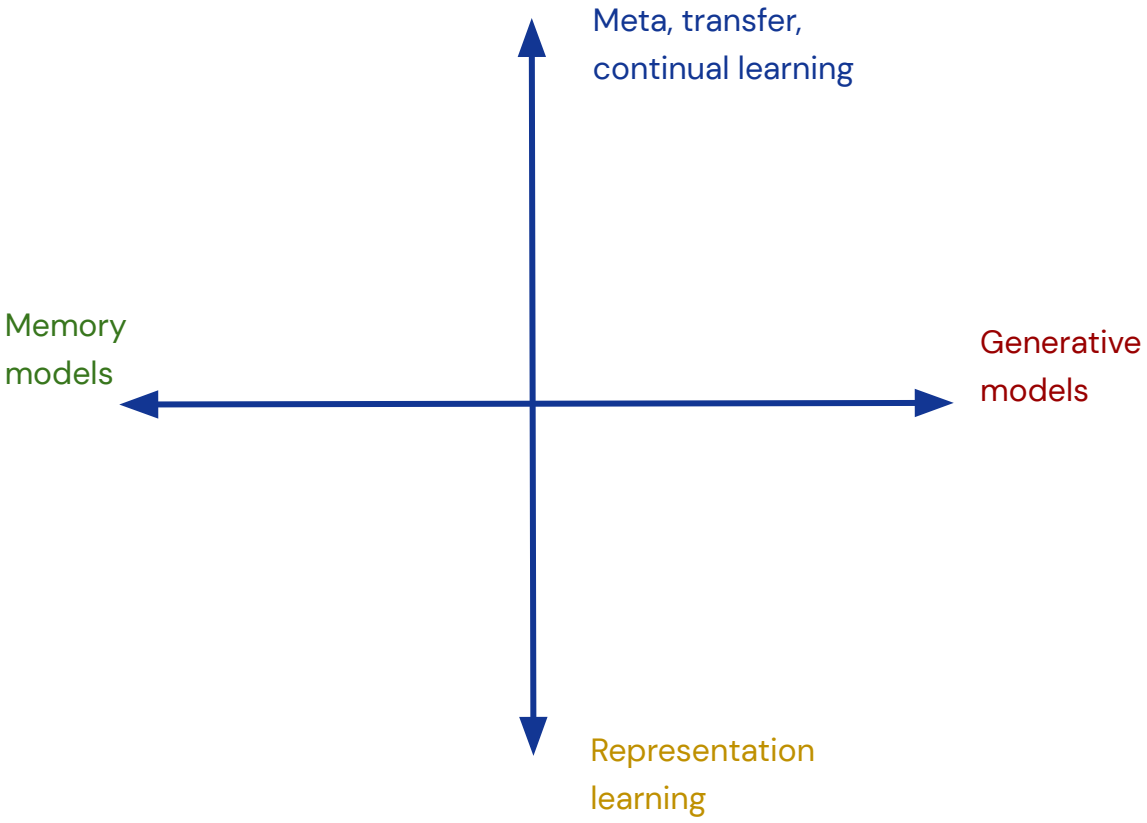


A universal language model that continually learns to perform multiple tasks in many languages.

- Semi-parametric models/memory-augmented neural networks.
  Yogatama and Mann; AISTATS 2014; Yogatama et al., ICLR 2018; de Masson d'Autume; NeurIPS 2019
- Architectural advances and structural biases for self-supervised representation learning.
  Yogatama and Smith; ACL 2014; ICML 2015; Yogatama et al., ICLR 2017; Maillard et al., JNLE 2019
- Generative language models.
  Yogatama et al., TACL 2014; Yogatama et al., arXiv 2017; Kong et al., ICLR 2018.

Reasoning (Dhingra et al., 2020), interactions with other modalities (Baltrusaitis et al., 2019), robustness (Huang et al., EMNLP 2019), fairness (Manzini et al., 2019), and others

Future Directions



Frontiers of NLP