Toward General Linguistic Intelligence

Dani Yogatama

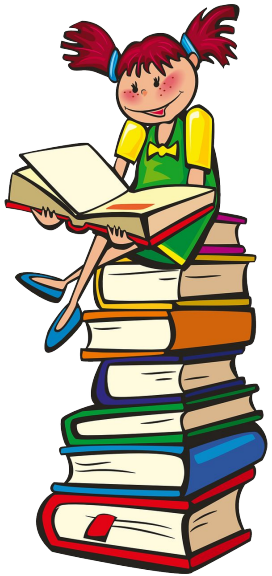
DeepMind

Language and Intelligence

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

Language and Intelligence

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



Language and Intelligence

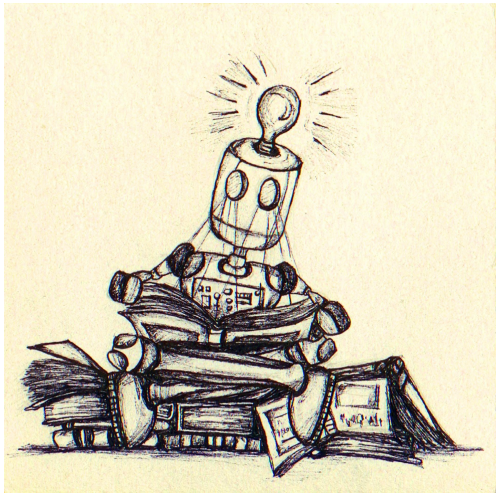
The most effective form of communication to transmit information and knowledge to others.

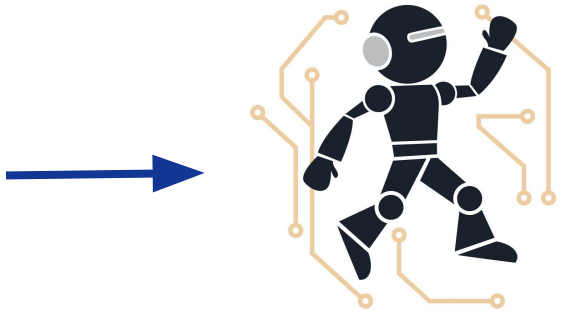
Language and Intelligence

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**.



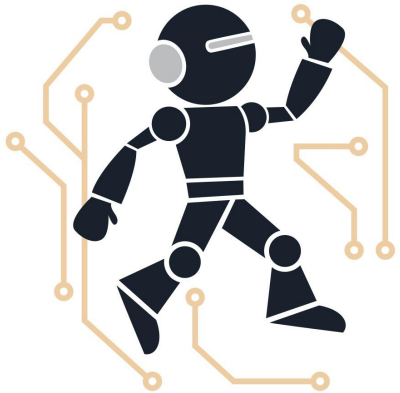


Challenges: Human Learning vs. Machine Learning

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

Challenges: Human Learning vs. Machine Learning

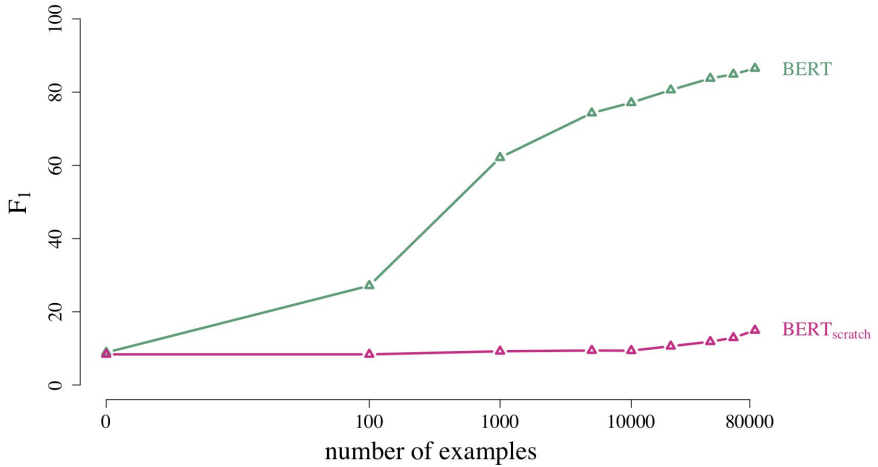
- ``Large'' datasets → sample efficient.
- Mostly task agnostic.
- Generalizable to new tasks.



- 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.



Yogatama et al., arXiv 2019

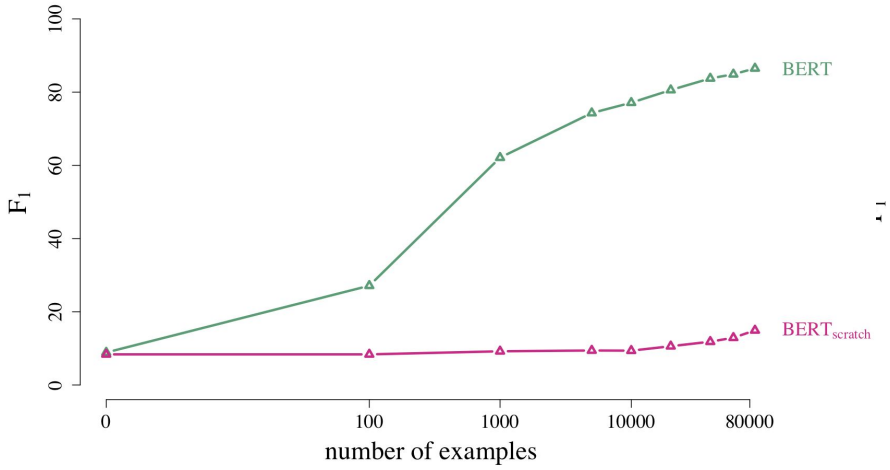
BERT model: Devlin et al. 2019

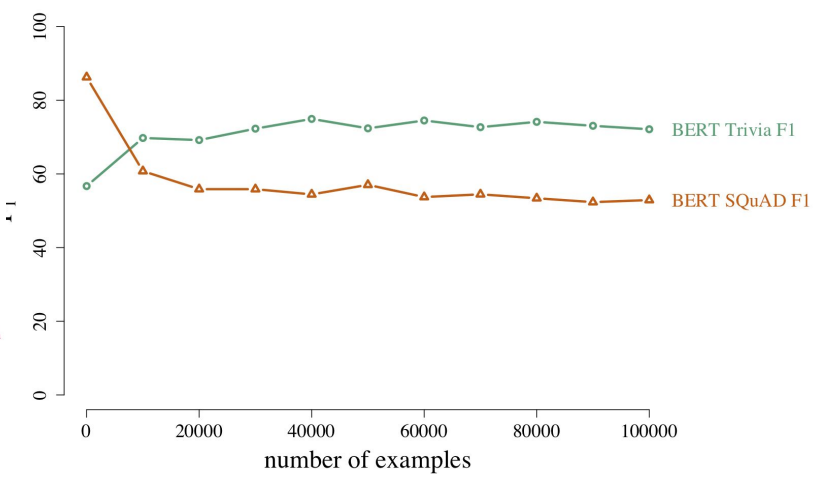
SQuAD dataset: Rajpurkar et al., 2016

Trivia dataset: Joshi et al., 2017

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.





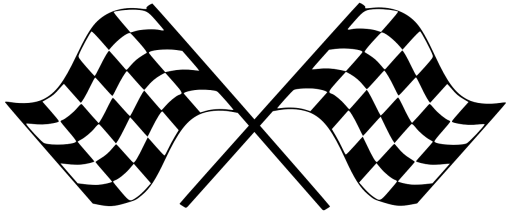
BERT model: Devlin et al. 2019

SQuAD dataset: Rajpurkar et al., 2016

Trivia dataset: Joshi et al., 2017

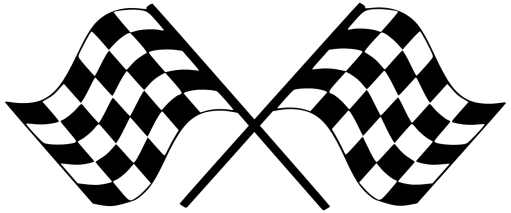
Yogatama et al., arXiv 2019

Research Areas



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

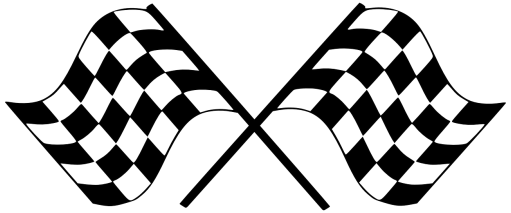
Research Areas



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.

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, 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

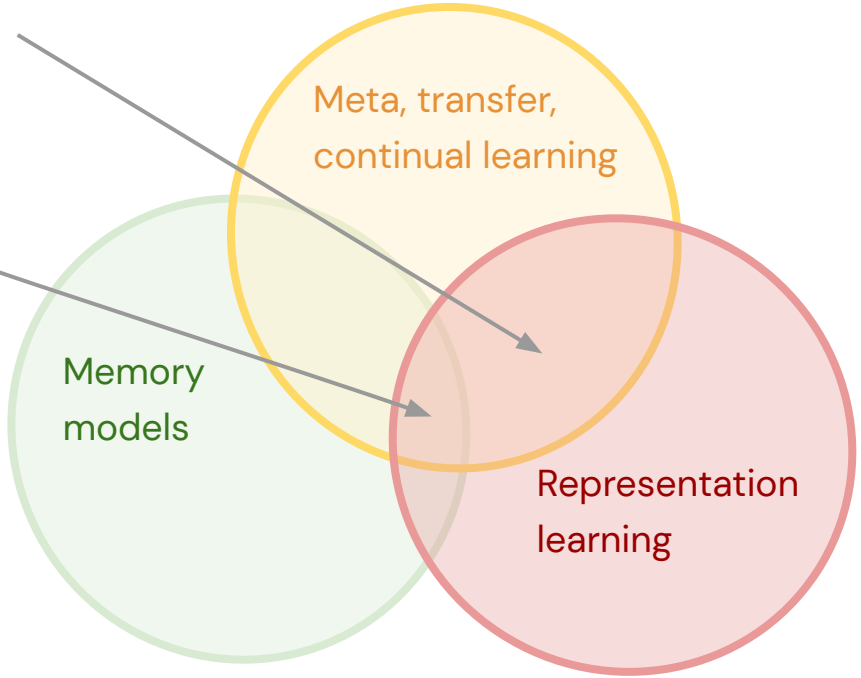
This Talk

- 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

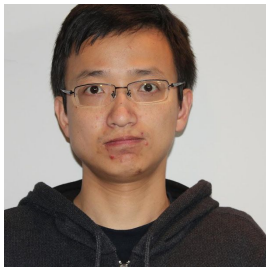


Episodic Memory in Lifelong Language Learning

de Masson d'Autume et al., NeurIPS 2019

Cyprien

Sebastian



Lingpeng

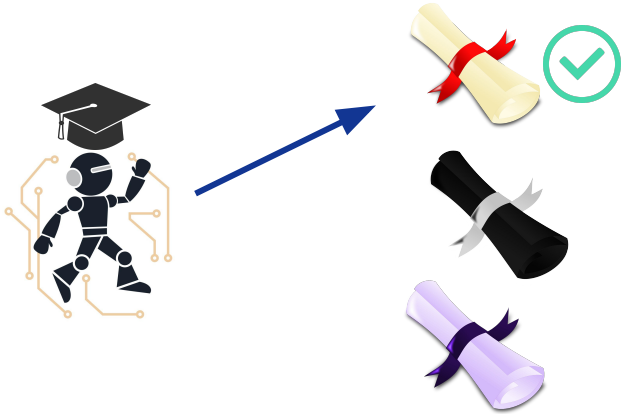
Dani

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).

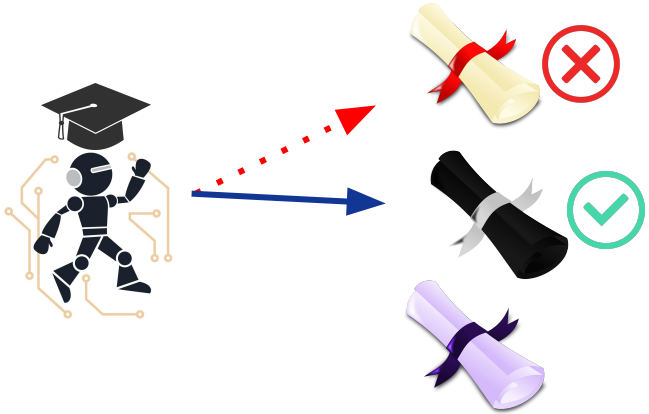
Background

- 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).



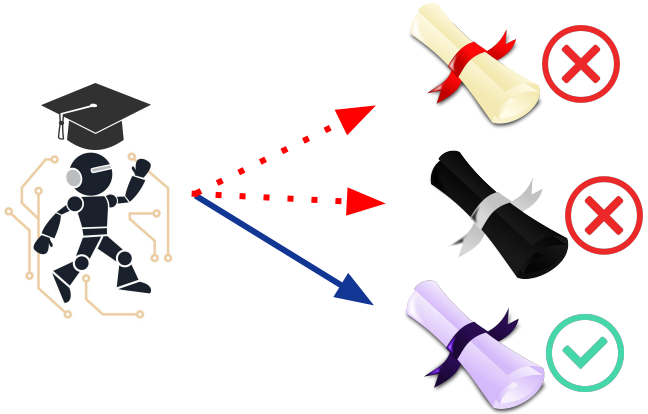
Background

- 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).



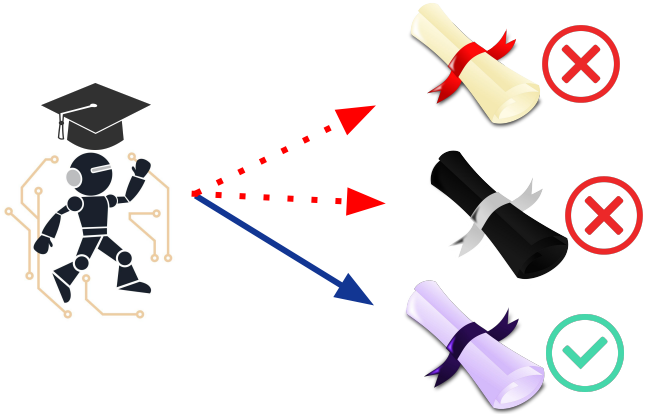
Background

- 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).



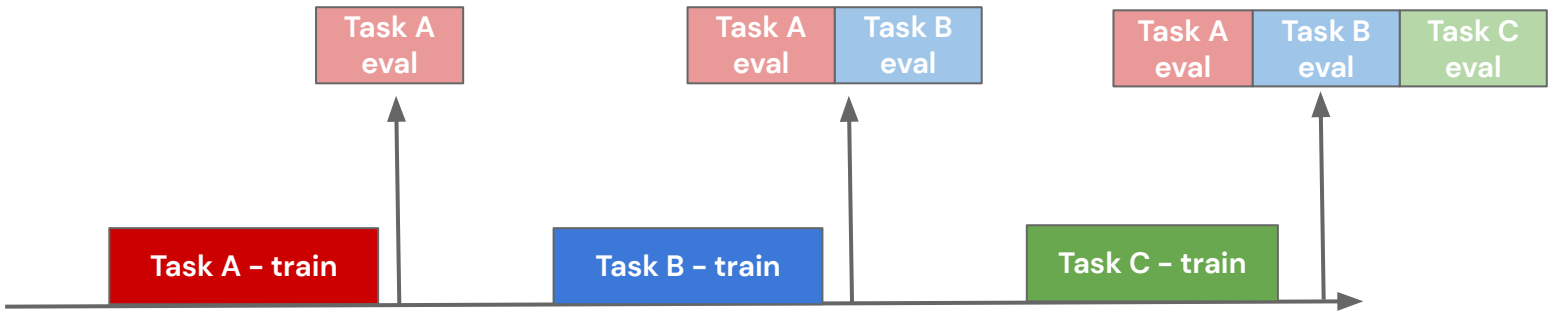
Background

- 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).



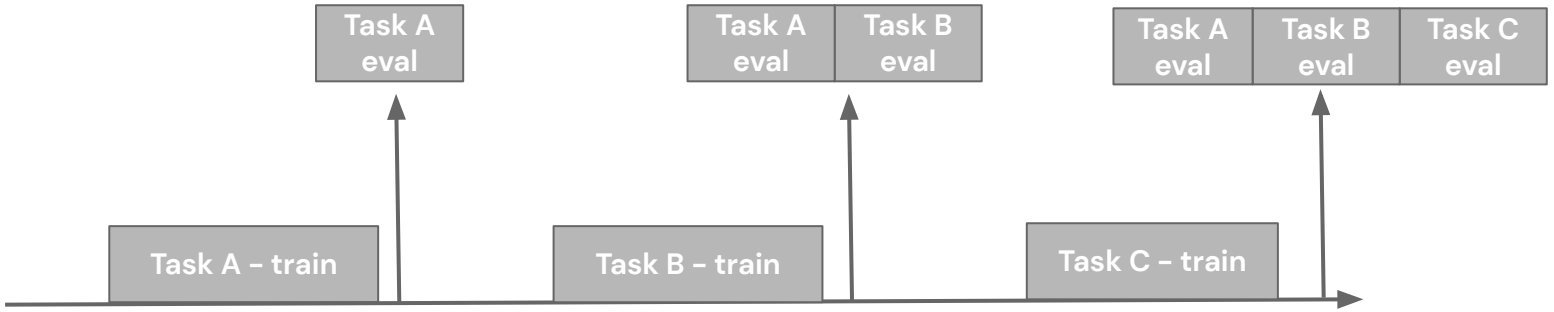
Hypothesis: episodic memory mitigates catastrophic forgetting in language learning.

Problem Setup



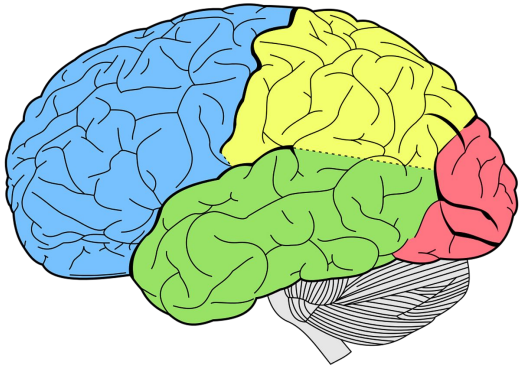
Standard setup, models know which task an example belongs to.

Problem Setup



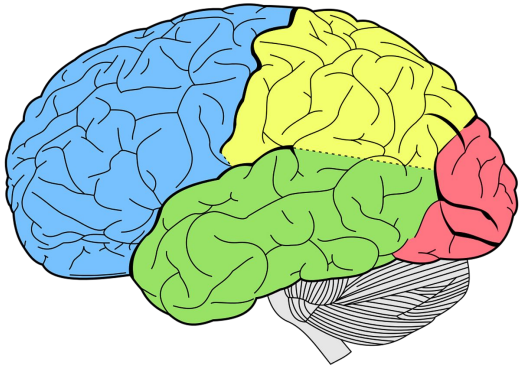
Our more difficult and realistic setup.

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

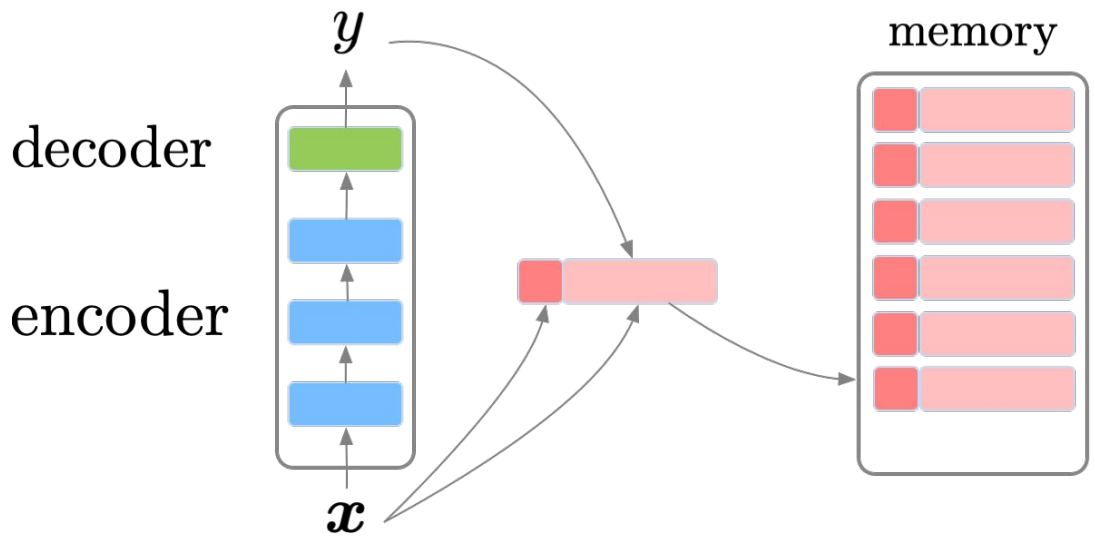


Episodic memory is a type of long-term memory of events and experiences. It is often associated with a module that stores training examples in neural networks..

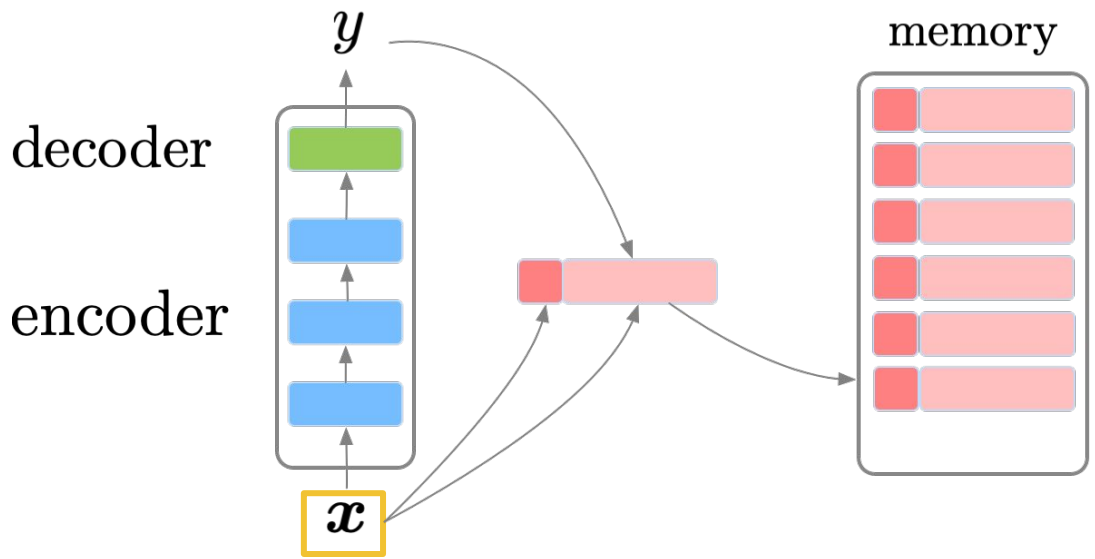
Contrast this with a short-term (working) memory in e.g., LSTMs (Hochreiter and Schmidhuber, 1997) , stack-augmented networks (Joulin and Mikolov, 2015; Grefenstette et al., 2015), or DNCs (Graves et al., 2016).

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

Question Answering Model



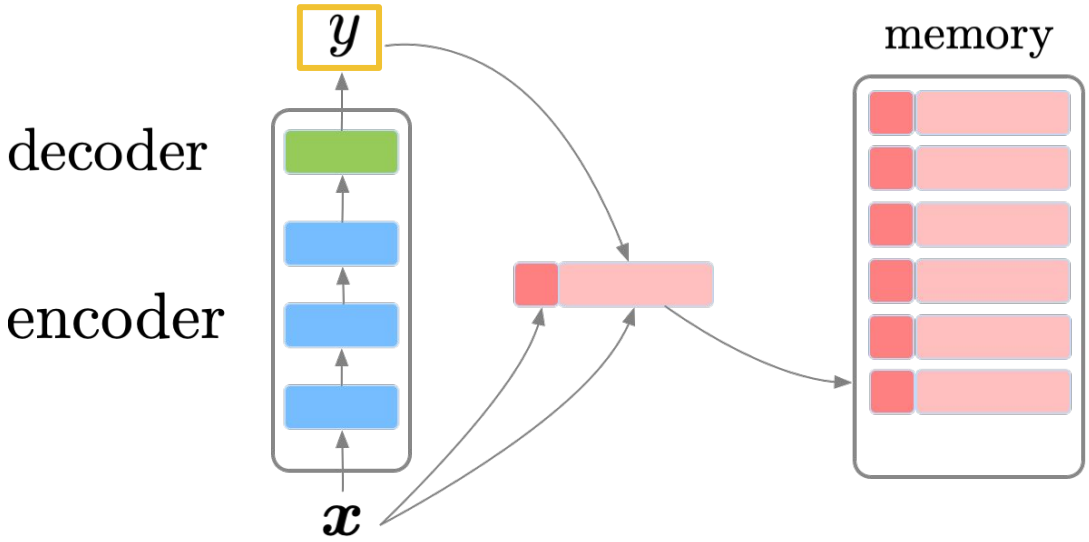
Question Answering Model



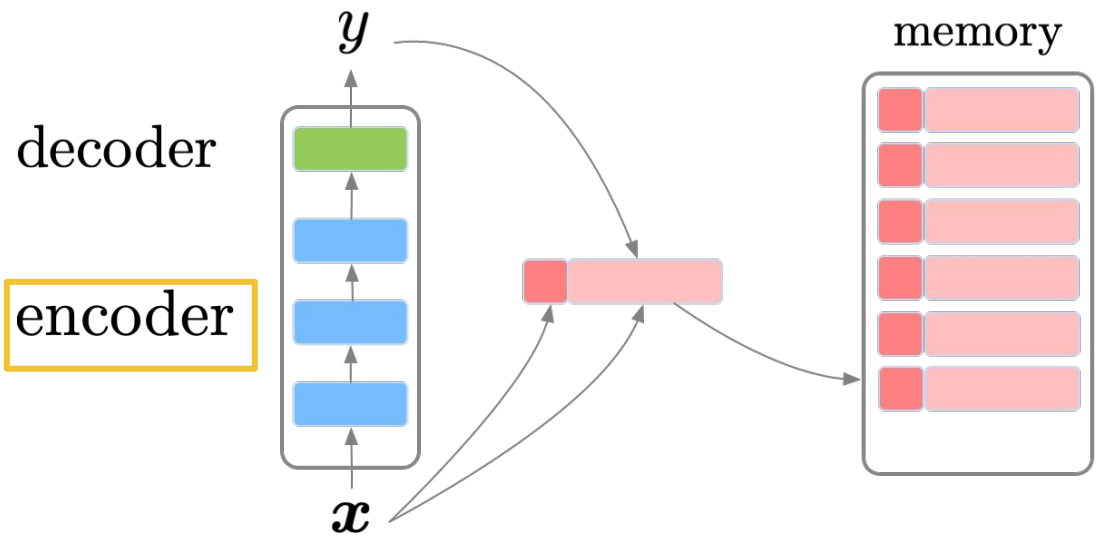
Input: a concatenation of context (e.g., a Wikipedia article) and question.

Question Answering Model

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



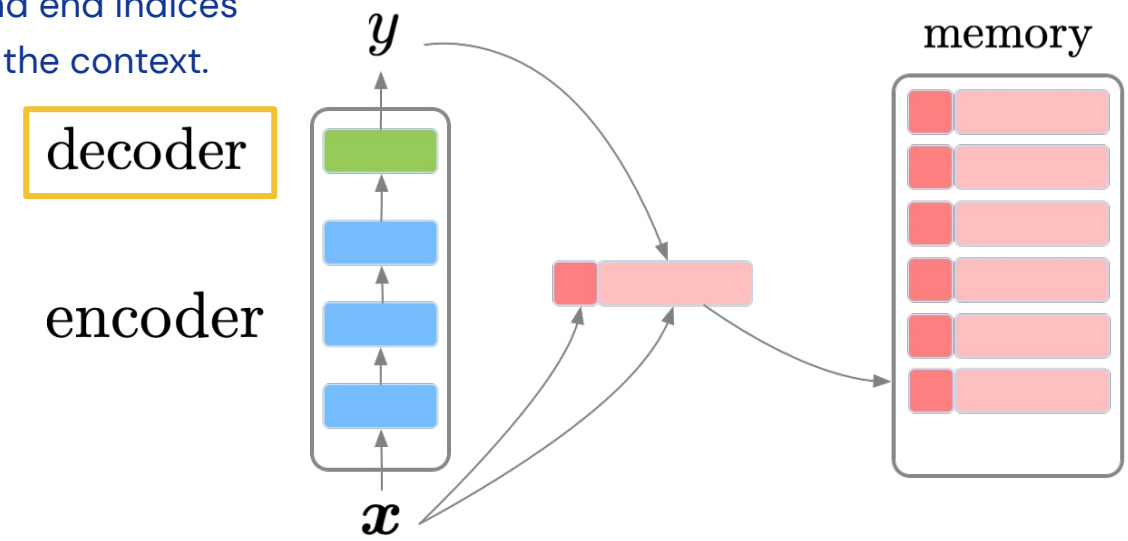
Question Answering Model



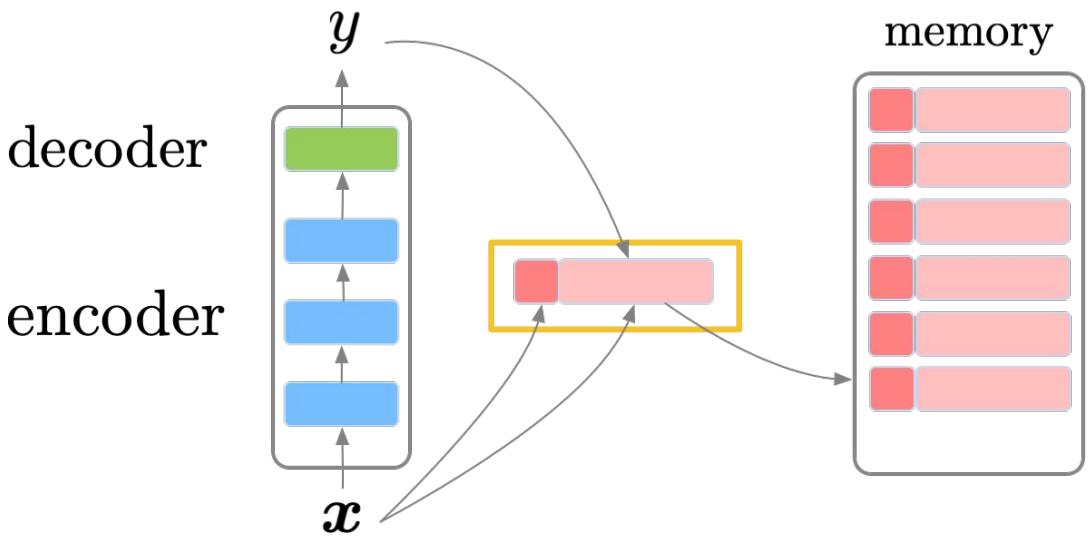
Encoder: a large neural network, e.g., ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), XLNet (Yang et al., 2019).

Question Answering Model

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

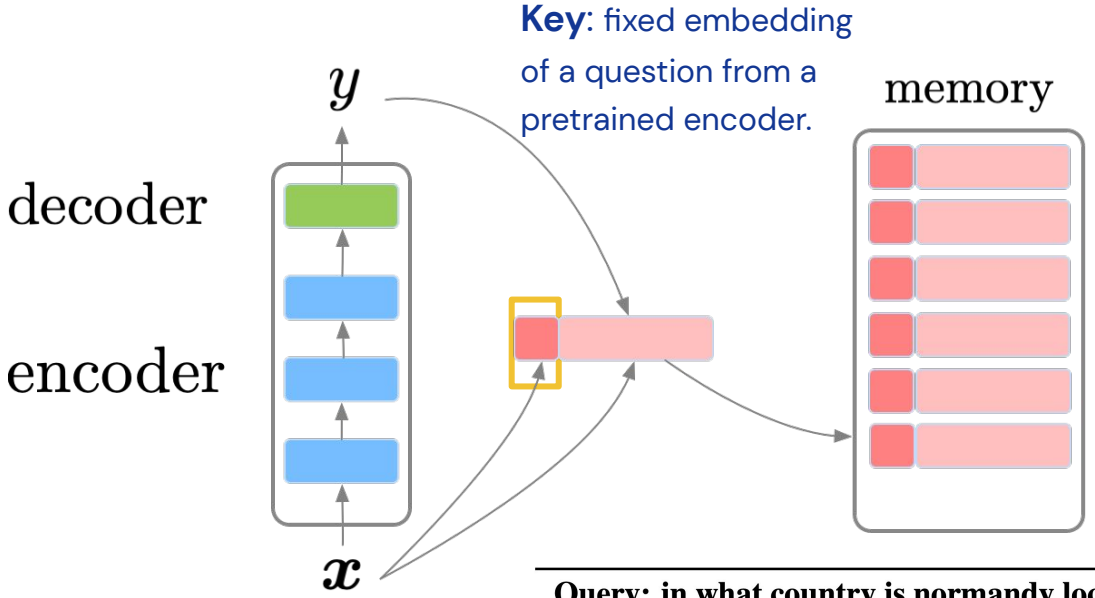


Question Answering Model



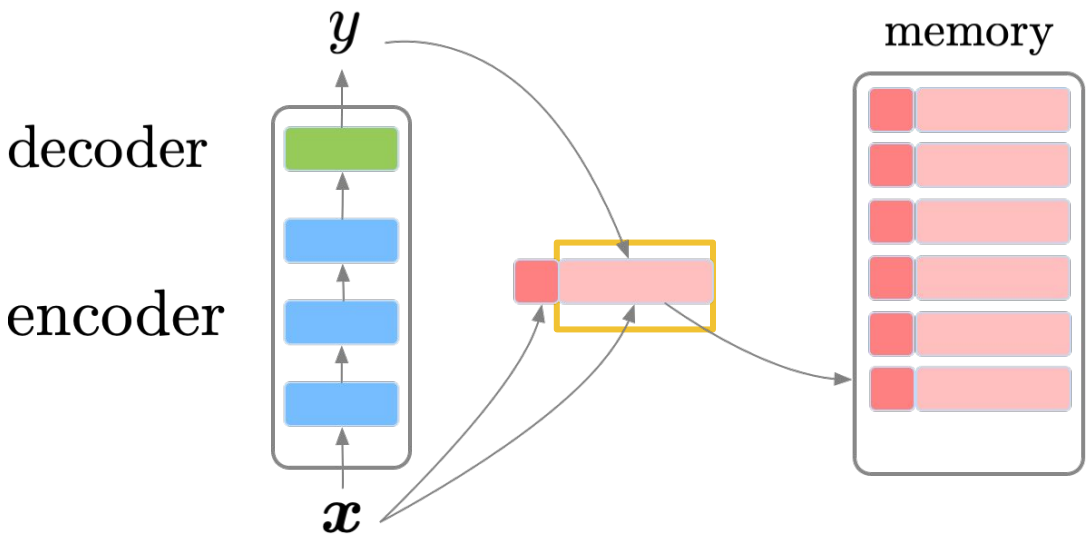
Memory entry: a key, value pair

Question Answering Model



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

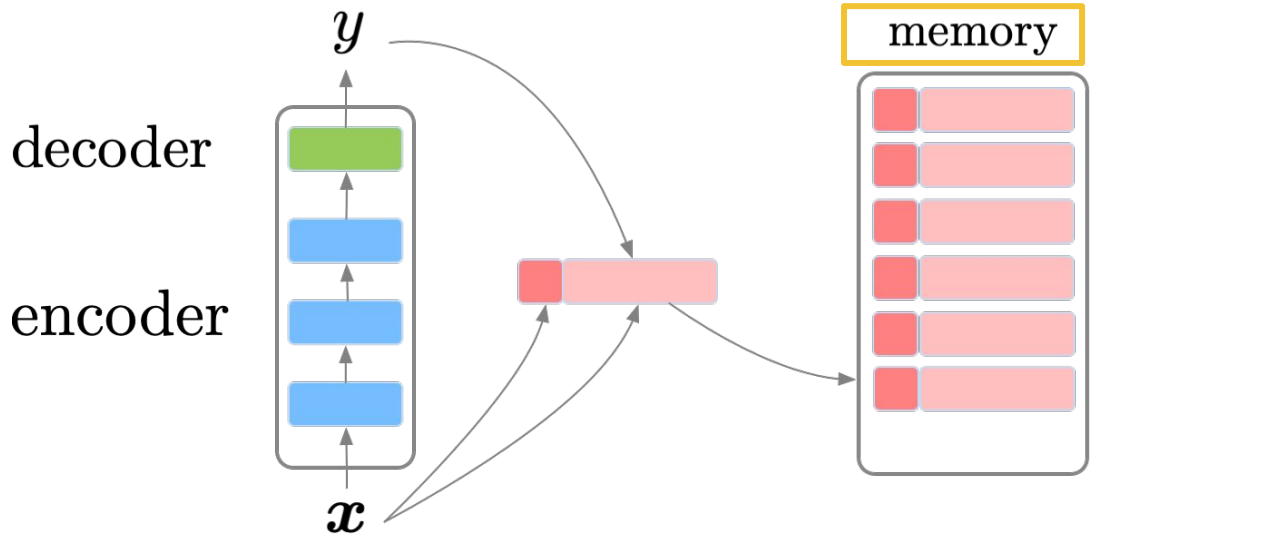
Question Answering Model



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

Question Answering Model

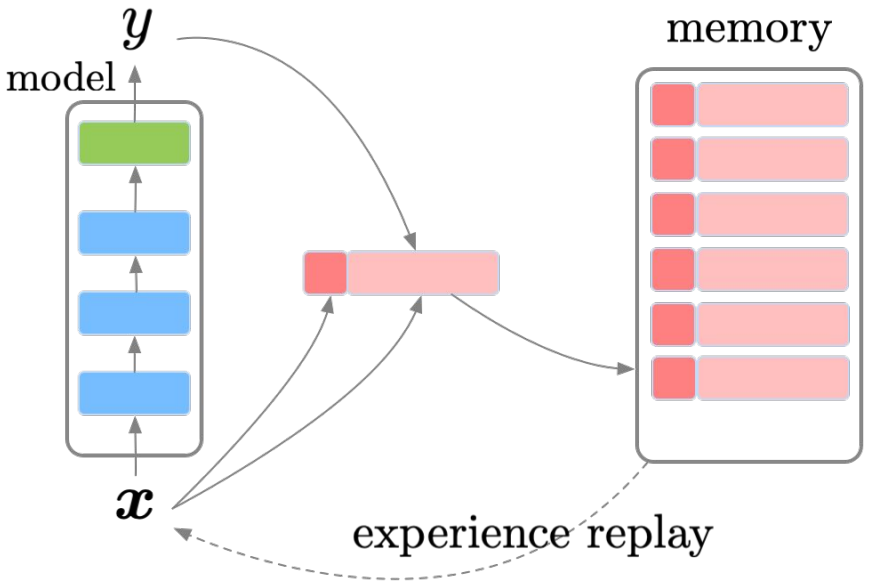
Memory: stores memory entries



Training

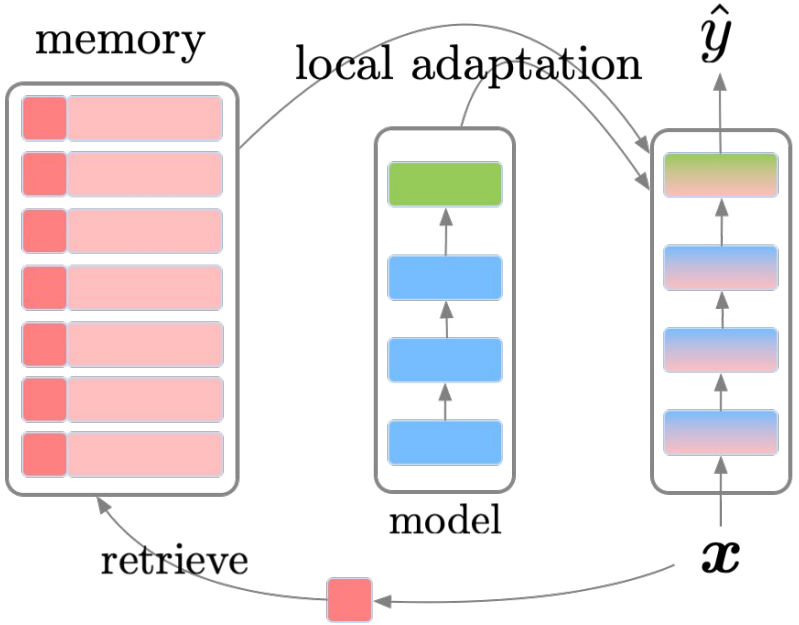
***Sparse* experience replay**: retrain on randomly sampled examples from the memory at a 1% rate.

Related to memory consolidation in human learning.



Inference (Prediction)

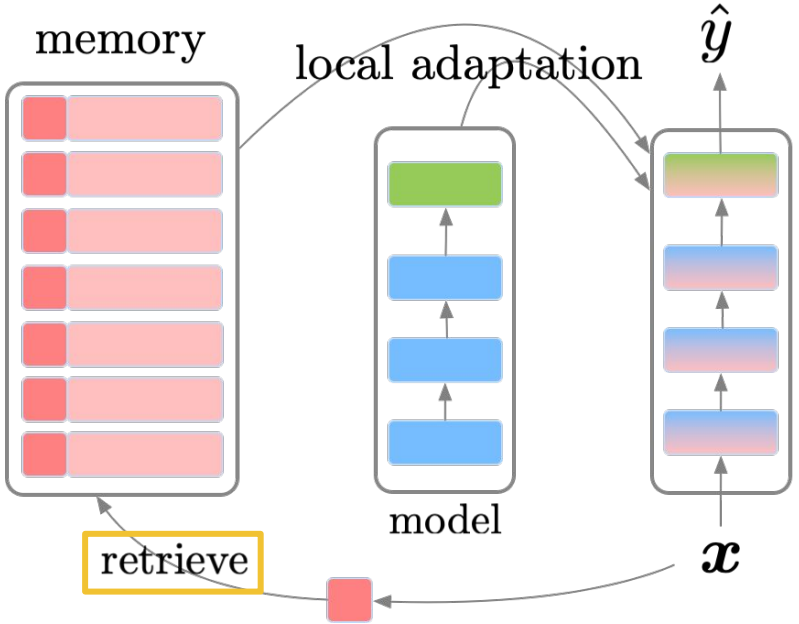
Local adaptation similar to MbPA (Sprechmann et al., 2018).



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

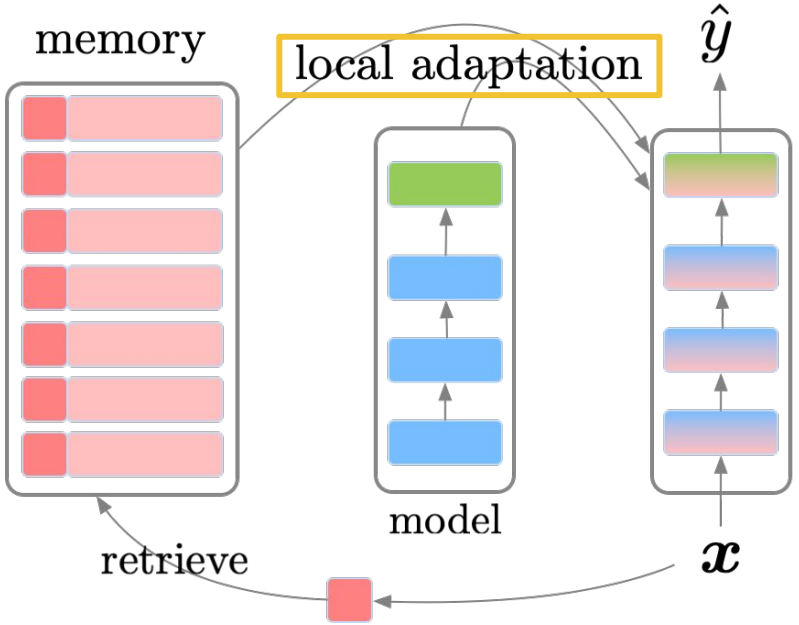


Inference (Prediction)

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



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	MbPA++	Multitask (upper bound)
QA	53.1	56.2	60.3	**62.4**	67.8

A-GEM: Chaudhry et al., 2019

MbPA: Chaudhry et al., 2019

Takeaways and Limitations

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

Takeaways and Limitations

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

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

10%	50%	100%
67.6	70.3	70.6

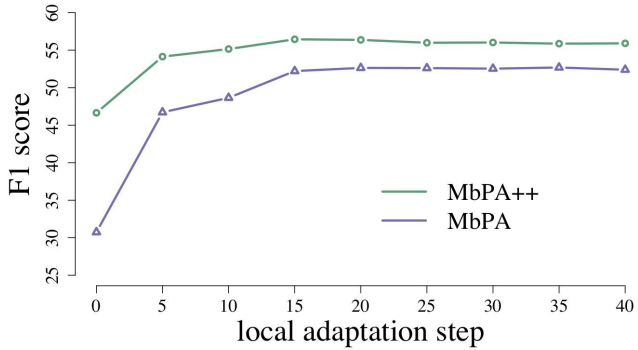
Takeaways and Limitations

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

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

10%	50%	100%
67.6	70.3	70.6

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



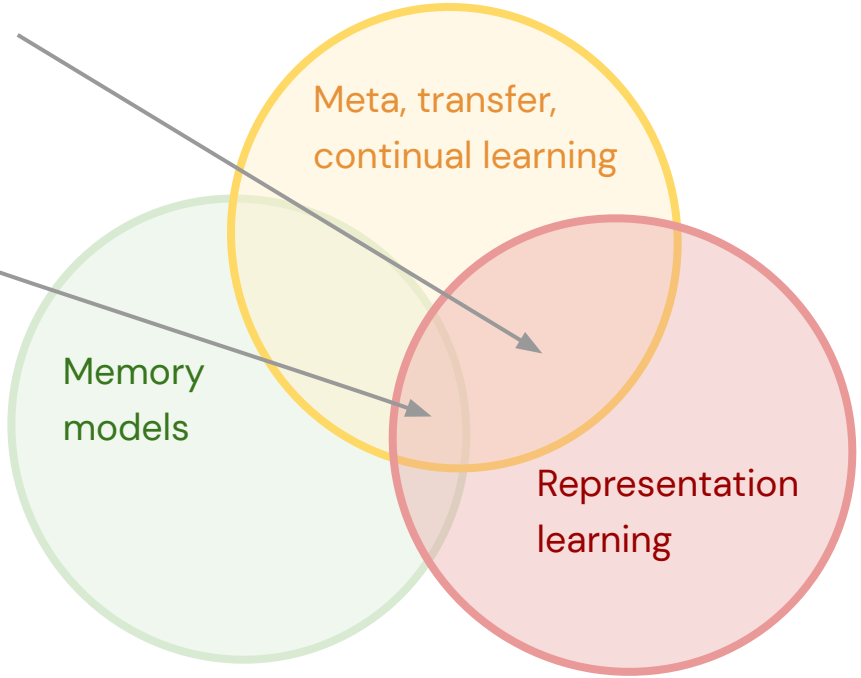
This Talk

- Episodic memory in lifelong language learning.

  de Masson d'Autume et al., NeurIPS 2019

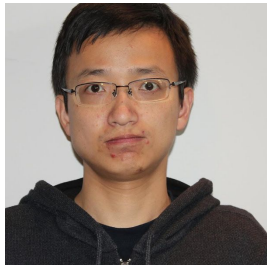
- A framework for self-supervised language representation learning methods.

  Kong et al., ICLR 2020

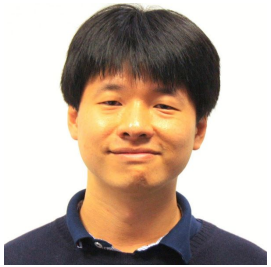


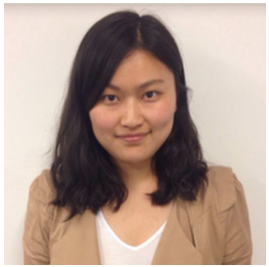
A Mutual Information Maximization Perspective of Language Representation Learning

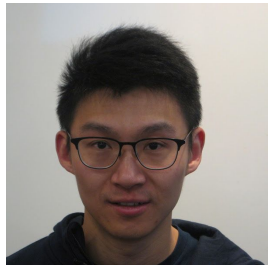
Kong et al., ICLR 2020


Lingpeng


Cyprien


Wang

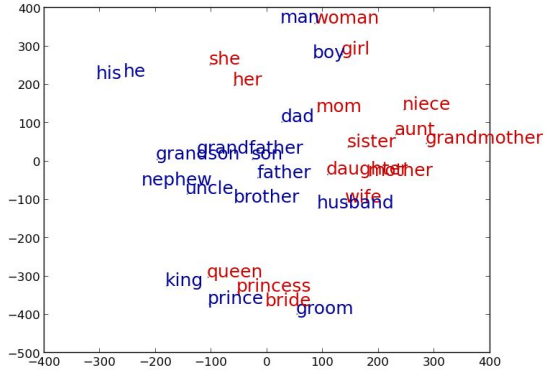

Lei


Zihang


Dani

Background

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



Skip gram, Mikolov et al., 2013.

GloVe, Pennington et al., 2014.

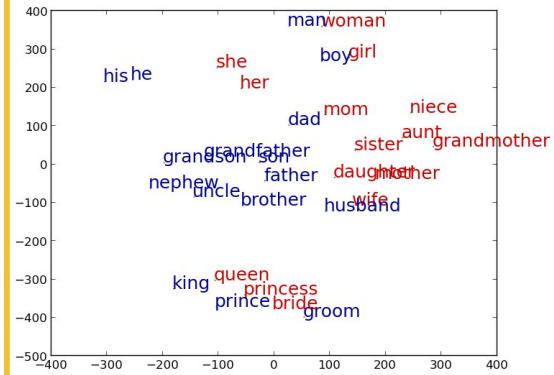
ELMo, Peters et al., 2018.

BERT, Devlin et al., 2019.

XLNet, Yang et al., 2019.

Background

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



Hypothesis: these methods are different instantiations 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.

Mutual Information and InfoNCE (Contrastive Learning)

InfoNCE (Logeswaran and Lee, 2018; van den Oord, et al., 2019) is a bound on mutual information.

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

Carnegie Mellon University is located in Pittsburgh

Mutual Information and InfoNCE (Contrastive Learning)

InfoNCE (Logeswaran and Lee, 2018; van den Oord, et al., 2019) is a bound on mutual information.

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

a b

Carnegie Mellon University is located in Pittsburgh

Mutual Information and InfoNCE (Contrastive Learning)

InfoNCE (Logeswaran and Lee, 2018; van den Oord, et al., 2019) is a bound on mutual information.

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

Mutual Information and InfoNCE (Contrastive Learning)

InfoNCE (Logeswaran and Lee, 2018; van den Oord, et al., 2019) is a bound on mutual information.

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

a b

Carnegie Mellon University is located in Pittsburgh

Mutual Information and InfoNCE (Contrastive Learning)

InfoNCE (Logeswaran and Lee, 2018; van den Oord, et al., 2019) is a bound on mutual information.

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

a b

Carnegie Mellon University is located in Pittsburgh

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

Mutual Information and InfoNCE (Contrastive Learning)

InfoNCE (Logeswaran and Lee, 2018; van den Oord, et al., 2019) is a bound on mutual information.

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

Pittsburgh + negative samples (other words) drawn from a proposal distribution

a b

Carnegie Mellon University is located in Pittsburgh

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

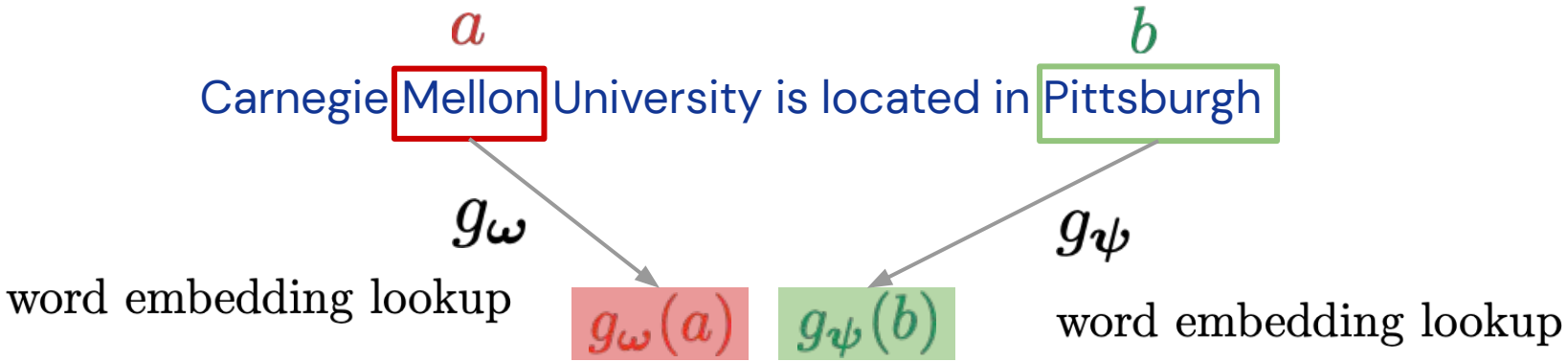
Skip-gram

Mikolov et al., 2013

Carnegie Mellon University is located in Pittsburgh

Skip-gram

Mikolov et al., 2013



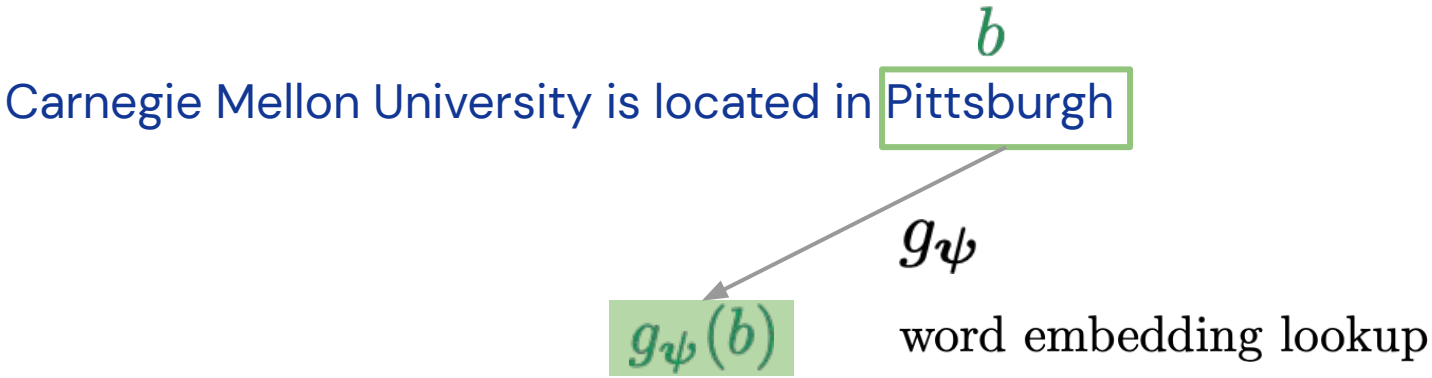
BERT

Devlin et al., 2019

Carnegie Mellon University is located in Pittsburgh

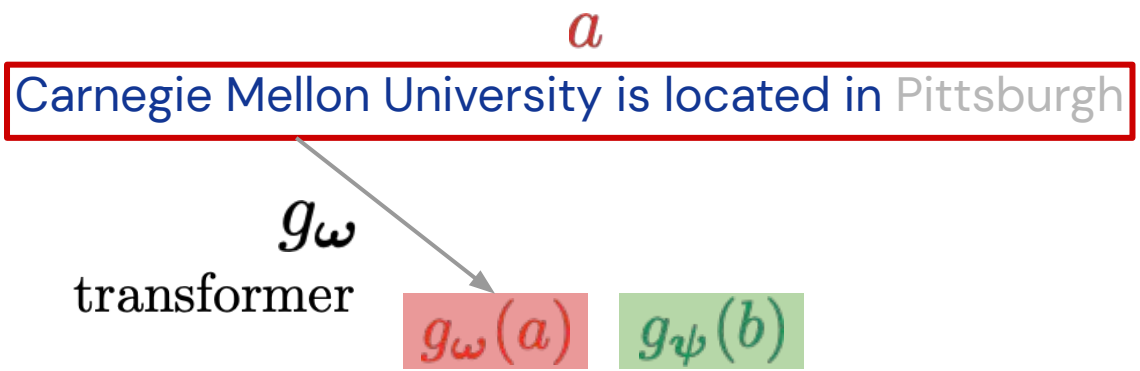
BERT

Devlin et al., 2019



BERT

Devlin et al., 2019



Why is this interesting?

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

	a	b	g_ω	g_ψ
Skip-gram	word	word	lookup	lookup
BERT-MLM	context	word	transformer	lookup
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
Skip-gram	word	word	lookup	lookup
BERT-MLM	context	word	transformer	lookup
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.

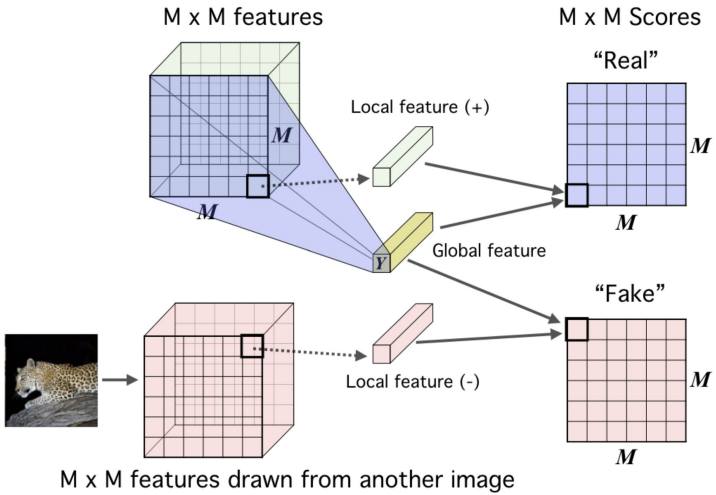
	a	b	g_ω	g_ψ
Skip-gram	word	word	lookup	lookup
BERT-MLM	context	word	transformer	lookup
XLNet	context	word	TXL++	lookup

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

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

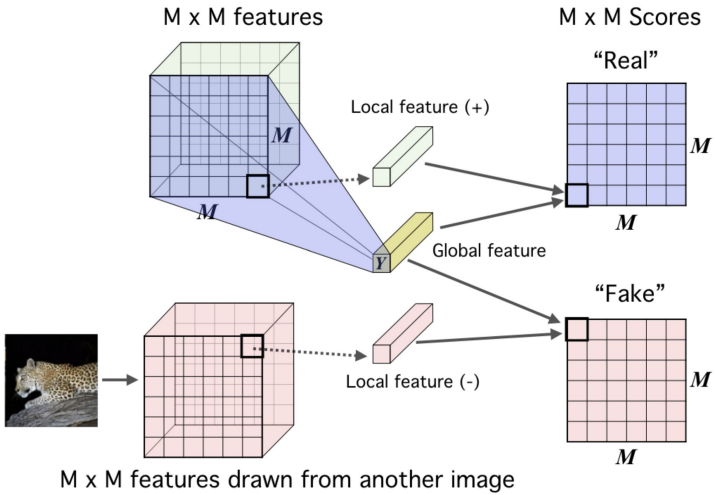
InfoWord

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



InfoWord

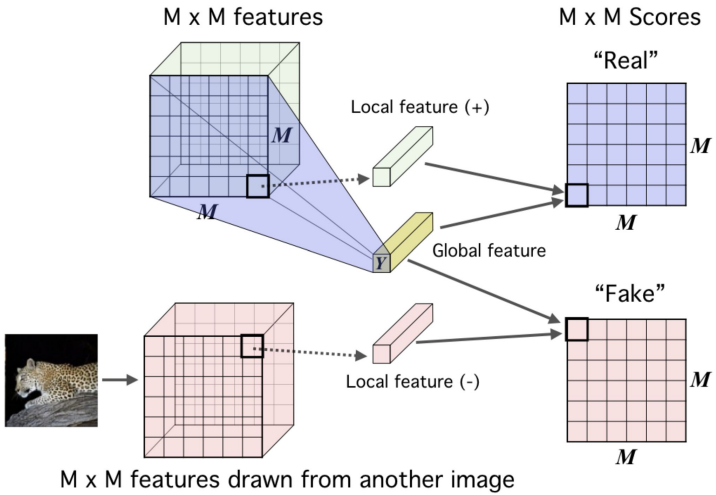
Deep InfoMax (DIM; Hjelm et al., 2019)



Carnegie Mellon University is located in Pittsburgh

InfoWord

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



Carnegie Mellon University is located in Pittsburgh

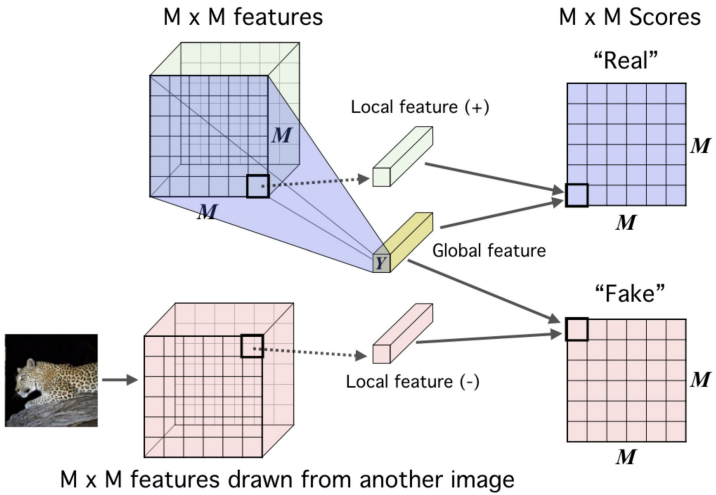
g_ψ

transformer

$g_\psi(b)$

InfoWord

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



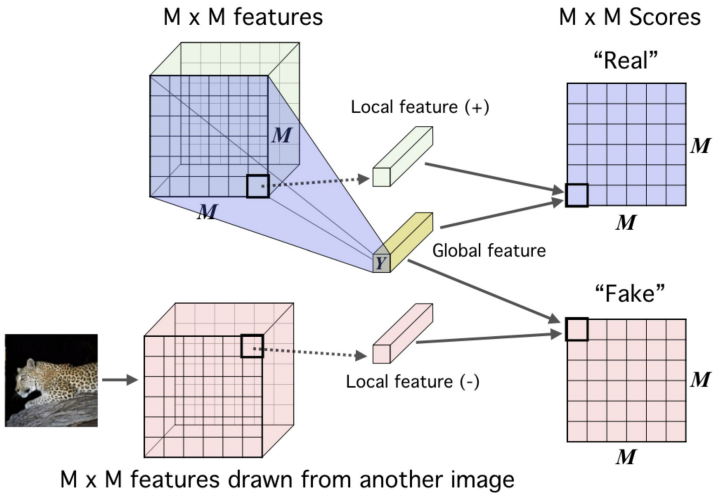
Carnegie Mellon University is located in Pittsburgh

g_ω
transformer

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

InfoWord

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



Carnegie Mellon University is located in Pittsburgh

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

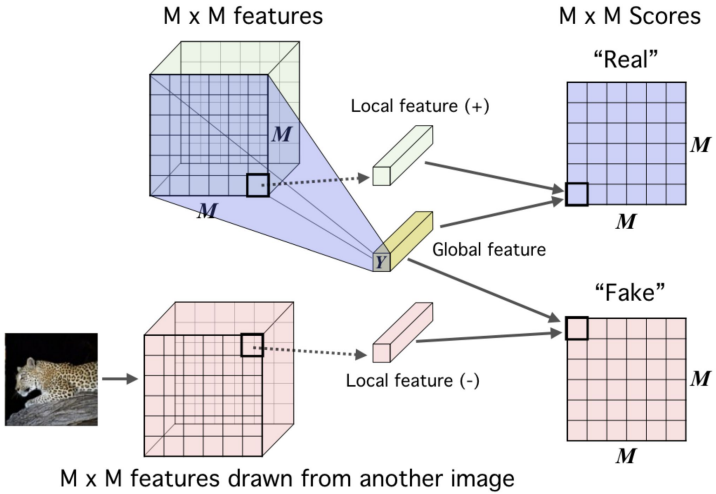
Starcraft II is a boring game

Cristiano Ronaldo scores an own goal

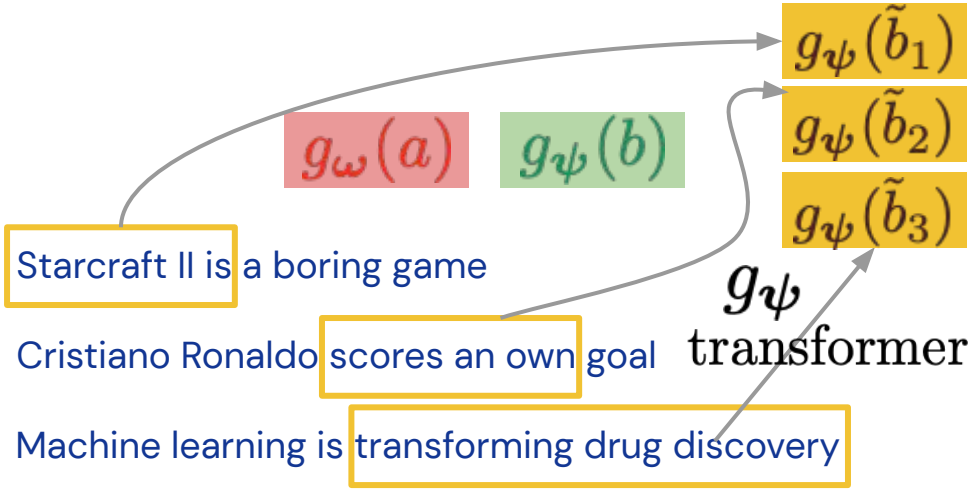
Machine learning is transforming drug discovery

InfoWord

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

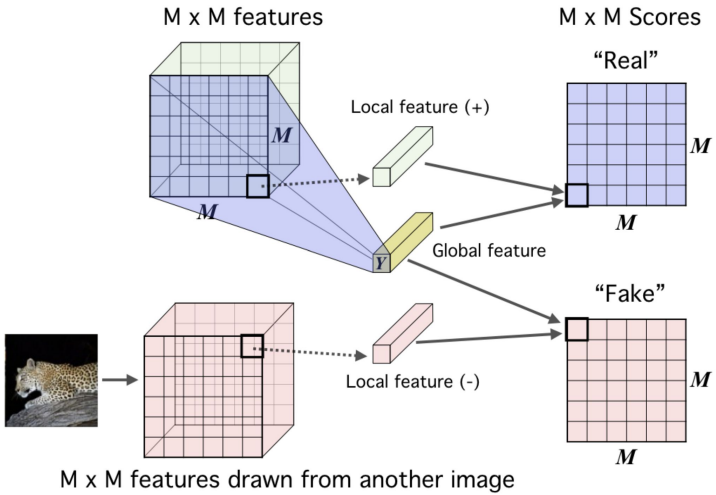


Carnegie Mellon University is located in Pittsburgh



InfoWord

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 boring game

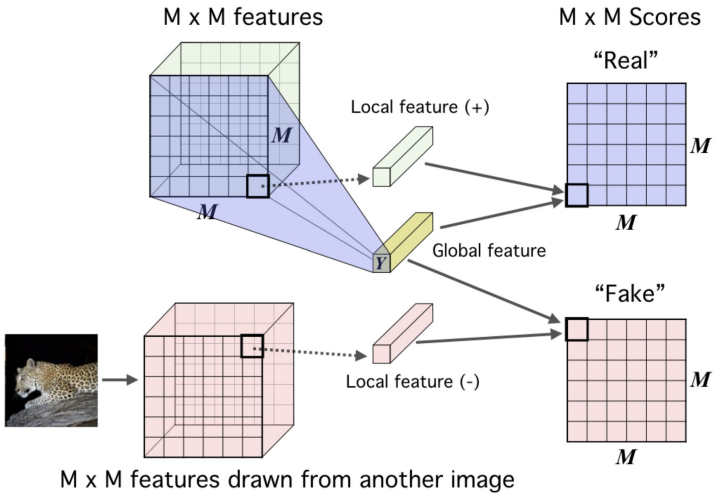
Cristiano Ronaldo scores an own goal

Machine learning is transforming drug discovery

$$\mathcal{I}_{\text{DIM}} = \mathbb{E}_{p(\hat{\boldsymbol{x}}_{i:j}, \boldsymbol{x}_{i:j})} \left[g_\omega(\hat{\boldsymbol{x}}_{i:j})^\top g_\omega(\boldsymbol{x}_{i:j}) - \log \sum_{\tilde{\boldsymbol{x}}_{i:j} \in \tilde{\mathcal{S}}} \exp(g_\omega(\hat{\boldsymbol{x}}_{i:j})^\top g_\omega(\tilde{\boldsymbol{x}}_{i:j})) \right]$$

InfoWord

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 boring 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
BASE	BERT	90.9	84.4
	InfoWord	**91.4**	**84.7**
LARGE	BERT	92.7	86.6
	InfoWord	**93.1**	**87.3**

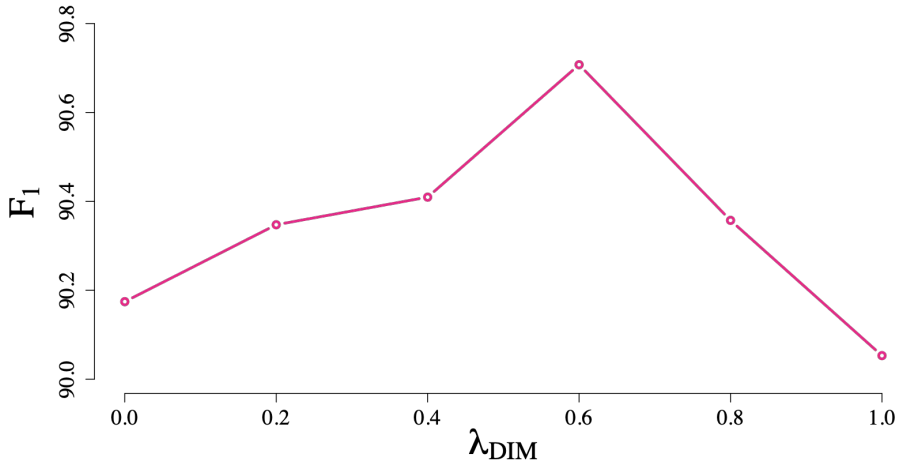
F1 and exact match scores (0-100), higher is better

Experiments

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

		F1	Exact Match
BASE	BERT	90.9	84.4
	InfoWord	**91.4**	**84.7**
LARGE	BERT	92.7	86.6
	InfoWord	**93.1**	**87.3**

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



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

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.

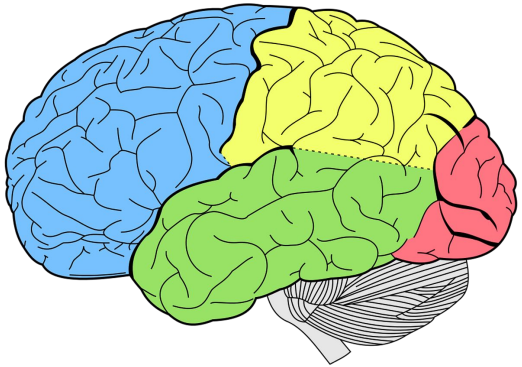
Takeaways and Limitations

- 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

Future Directions

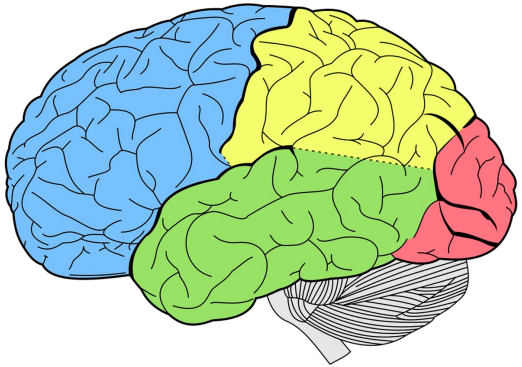
- Semi-parametric models: a combination of a big parametric neural network with non-parametric memory modules (cache, stacks, memory blocks).
  - Human memory has specialized systems for different functions.
  - Important for reasoning.



Future Directions

- Semi-parametric models: a combination of a big parametric neural network with non-parametric memory modules (cache, stacks, memory blocks).
  - Human memory has specialized systems for different functions.
  - Important for reasoning.

→ Learning what to remember and forget/compress, integration of modules.



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).

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).

→ Hierarchical models for efficient out-of-distribution generalization.

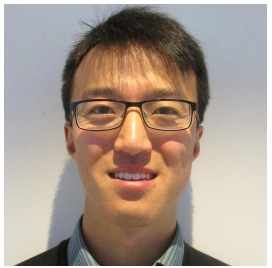
Future Directions

- 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).

→ Hierarchical models for efficient out-of-distribution generalization.

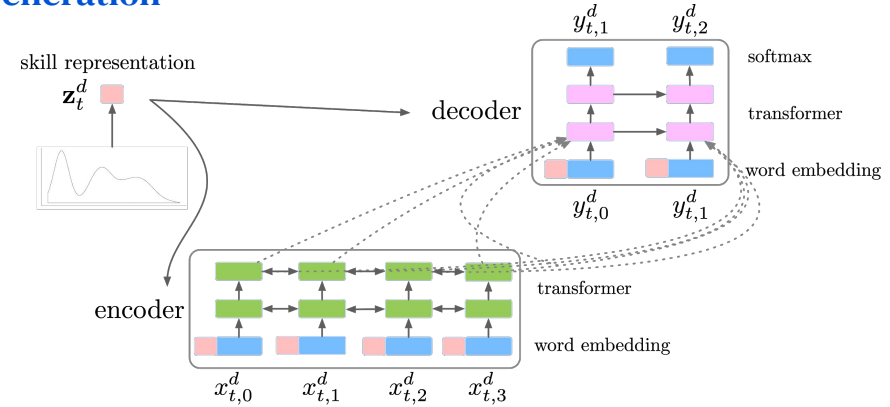
Modelling Latent Skills for Multitask Language Generation

Cao and Yogatama, arXiv 2020



Kris

Dani



Future Directions

- Language-independent representations.
  - Humans acquire language with a mechanism that is independent of the language.
  - Humans learn abstractions that generalize to other languages.

Future Directions

- Language-independent representations.
  - Humans acquire language with a mechanism that is independent of the language.
  - Humans learn abstractions that generalize to other languages.

→ Learning generalizable abstractions.

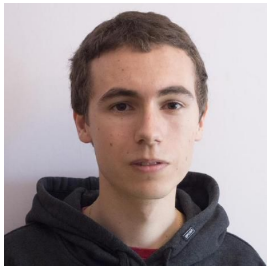
Future Directions

- Language-independent representations.
  - Humans acquire language with a mechanism that is independent of the language.
  - Humans learn abstractions that generalize to other languages.

→ Learning generalizable abstractions.

On the Crosslingual Transferability of Monolingual Representations

Artetxe et al., arXiv 2019



Mikel

Sebastian

Dani

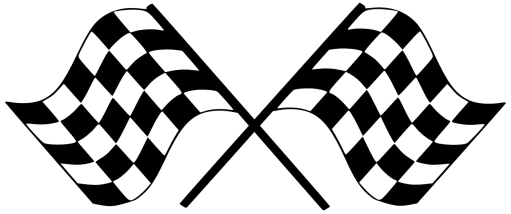
	Average (11 langs.)
Multilingual Training	65.7
Monolingual Training + Transfer	66.8

Thank you!

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

DeepMind

Research Areas

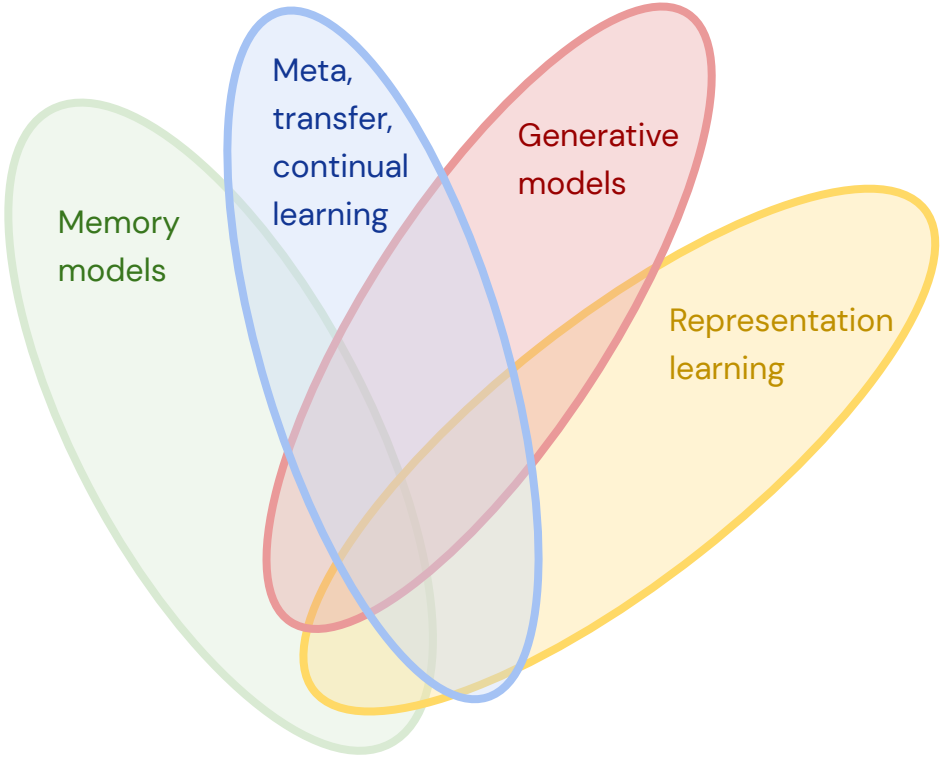


A universal language model that continually learns to perform multiple tasks in many languages.

- 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