Toward General Linguistic Intelligence

Dani Yogatama

DeepMind

Why Language?

- Language is a primary medium through which we acquire new skills and knowledge (+visual perception)
  - Reading the web, listening to instructions from other people, etc.

- Language is the most effective form of communication to transmit knowledge to others.
  - Writing papers, conversing with others, etc.

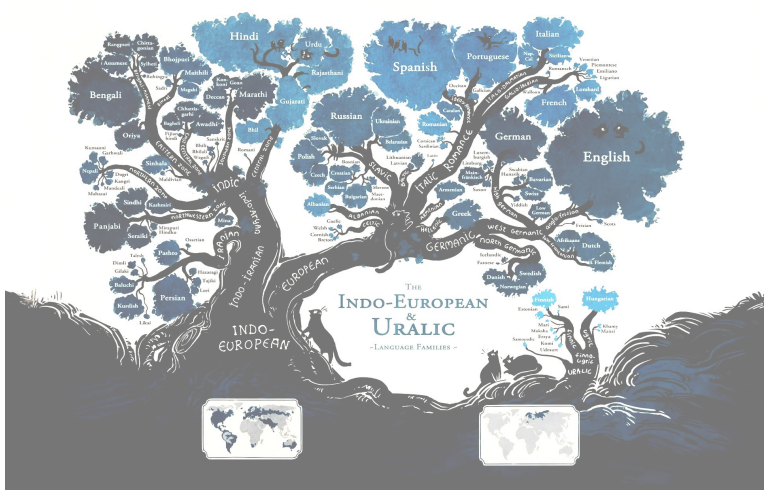


Image from: https://www.theguardian.com/education/gallery/2015/jan/23/a-language-family-tree-in-pictures

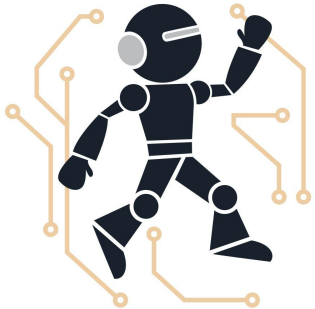
Why Language?

- Language is a primary medium through which we acquire new skills and knowledge (+visual perception)
  - Reading the web, listening to instructions from other people, etc.

- Language is the most effective form of communication to transmit knowledge to others.
  - Writing papers, conversing with others, etc.

Language is key to human intelligence and is going to be important for artificial general intelligence to emerge.

Human Learning vs. Machine Learning?



Mostly task agnostic

``Large'' datasets → sample efficient

Generalizable to new tasks

Problem specific

Massive datasets → massive datasets

Forget previous tasks given a new task

General Linguistic Intelligence

- The ability to reuse previously acquired knowledge about a language's lexicon, syntax, semantics, and pragmatic conventions to **adapt to new tasks quickly without forgetting old ones**.

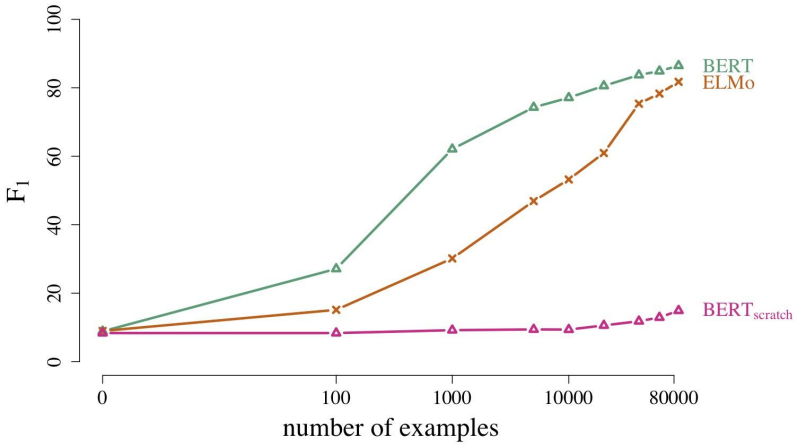
- Imagine an agent that keeps getting better by continuously reading everything on the web and is able to answer questions (and perform other tasks) in all languages, including those that require it to reason and make new inferences (e.g., which method is good for problem X?).

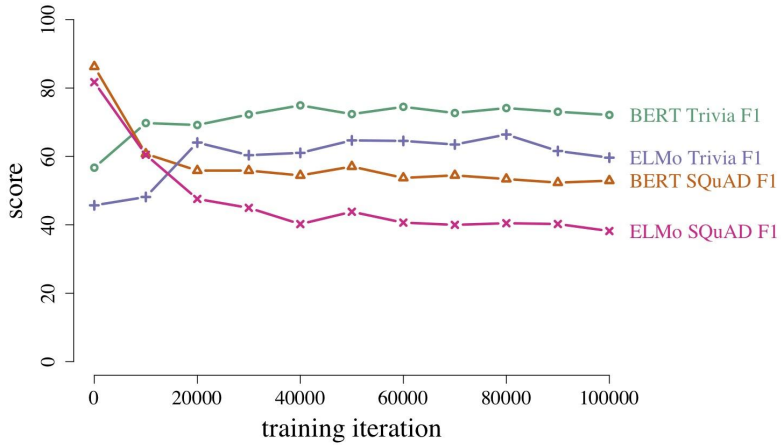
The State of Natural Language Processing

- Advances in deep learning techniques have considerably improved performance on many important downstream tasks (question answering, machine translation, sentiment analysis, etc.).
- In particular, self-supervised representation learning has resulted in state-of-the-art models that require fewer labeled training examples on the task of interest.

The State of Natural Language Processing

- Advances in deep learning techniques have considerably improved performance on many important downstream tasks (question answering, machine translation, sentiment analysis, etc.).
- In particular, self-supervised representation learning has resulted in state-of-the-art models that require fewer labeled training examples on the task of interest.
- They still require many training examples (in the order of tens/hundreds of thousands) and only work well for a specific purpose (overfit to a dataset as opposed to solving the task). **Yogatama et al., arXiv 2019**





Research Areas

- State-of-the-art LMs are large parametric deep models with task-specific components, which assume stationary data distribution and are trained in a discriminative fashion on many examples (i.e., fine tuning).

Research Areas

- State-of-the-art LMs are large parametric deep models with task-specific components, which assume stationary data distribution and are trained in a discriminative fashion on many examples (i.e., fine tuning).

- Self-supervised learning has shown impressive progress in reducing sample complexity.
  - Architectural and structural biases (Yogatama and Smith, ICML 2014; Yogatama and Smith, ACL 2014; Yogatama et al., ICLR 2017; Maillard et al., JNLE 2019).
- Semi-parametric models (memory-augmented neural networks).
  - Allow a model to store and reuse previously acquired knowledge effectively (Yogatama and Mann; AISTATS 2014; Yogatama et al., ICLR 2018; de Masson d'Autume; NeurIPS 2019).
- Generative models as an alternative to discriminative training.
  - Able do any linguistic task, more sample efficient (Yogatama et al., TACL 2014; Yogatama et al., arXiv 2017; Kong et al., ICLR 2018).

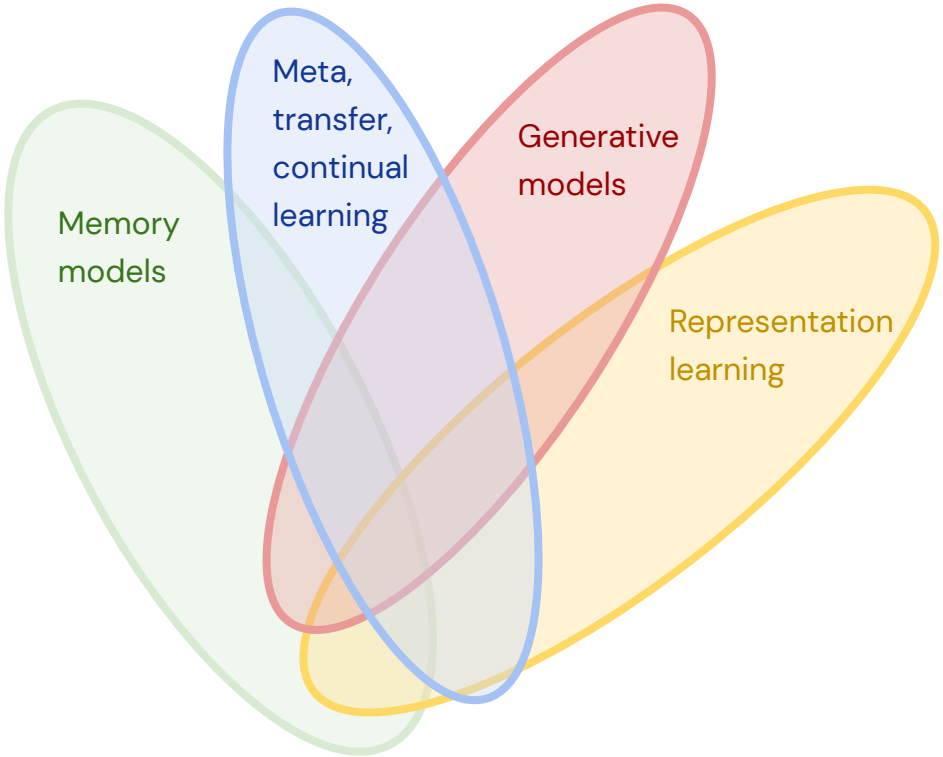
Research Areas

- State-of-the-art LMs are large parametric deep models with task-specific components, which assume stationary data distribution and are trained in a discriminative fashion on many examples (i.e., fine tuning).

- Self-supervised learning has shown impressive progress in reducing sample complexity.
  - Architectural and structural biases (Yogatama and Smith, ICML 2014; Yogatama and Smith, ACL 2014; Yogatama et al., ICLR 2017; Maillard et al., JNLE 2019).
- Semi-parametric models (memory-augmented neural networks).
  - Allow a model to store and reuse previously acquired knowledge effectively (Yogatama and Mann; AISTATS 2014; Yogatama et al., ICLR 2018; de Masson d'Autume; NeurIPS 2019).
- Generative models as an alternative to discriminative training.
  - Able do any linguistic task, more sample efficient (Yogatama et al., TACL 2014; Yogatama et al., arXiv 2017; Kong et al., ICLR 2018).

- Other abilities: commonsense reasoning, interactions with other modalities, robustness to adversaries, etc.

Research Areas



This Talk

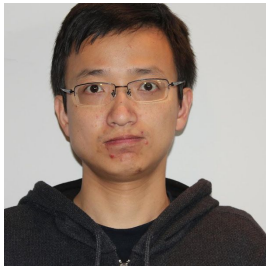
- Episodic memory in lifelong language learning (de Masson d'Autume et al., NeurIPS 2019).
- A framework for self-supervised language representation learning methods (Kong et al., ICLR 2020).

Episodic Memory in Lifelong Language Learning

de Masson d'Autume et al., NeurIPS 2019

Cyprien

Sebastian

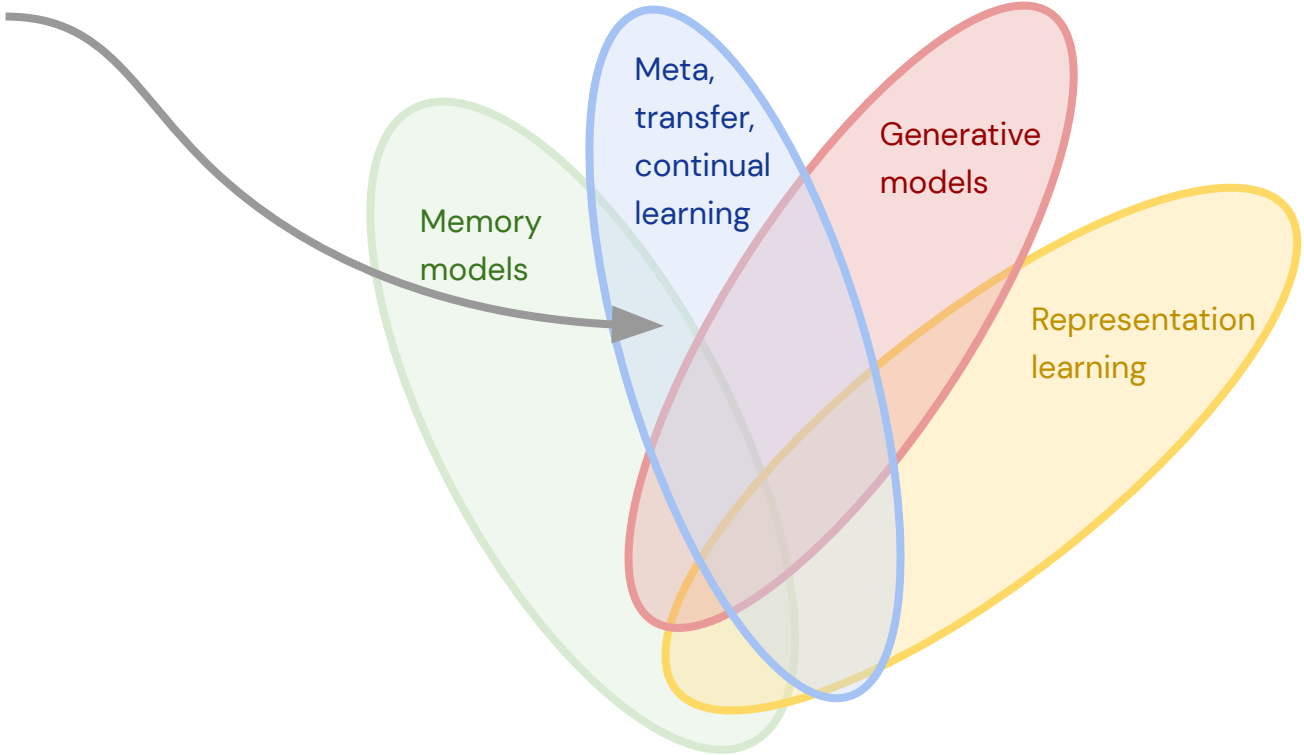


Lingpeng

Dani

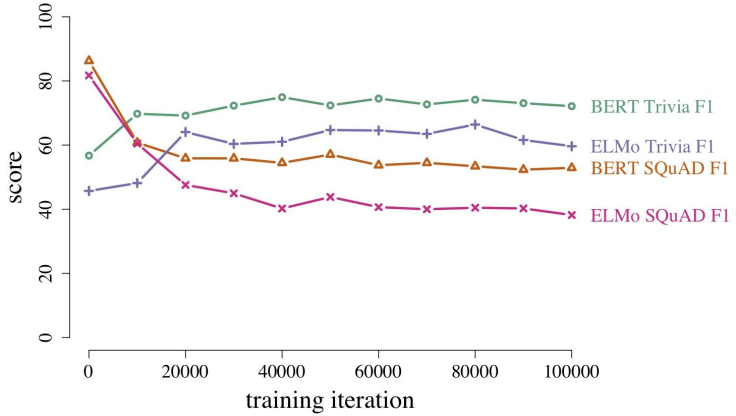
Episodic Memory in Lifelong Language Learning

de Masson d'Autume et al., NeurIPS 2019



Background

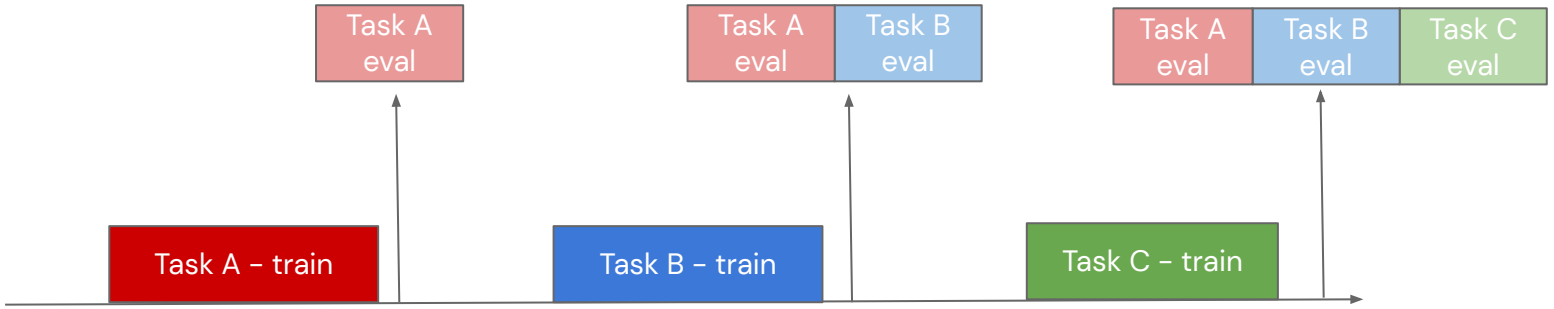
- Current models suffer from **catastrophic forgetting**: when learning a new task, they tend to forget previously learned tasks.



- See Yogatama et al., EMNLP 2011 and Yogatama et al., TACL 2014 for our earlier work in this area.
- **Hypothesis**: augmenting a language model with an episodic memory module mitigates catastrophic forgetting.

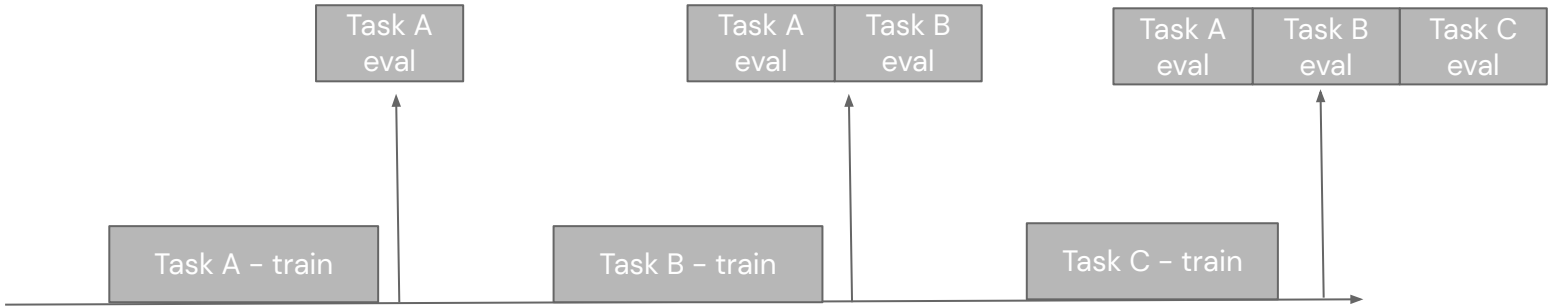
Problem Setup

- A new, realistic, lifelong learning setting: the model is exposed to several different datasets presented sequentially in one pass without dataset identifiers.



Problem Setup

- A new, realistic, lifelong learning setting: the model is exposed to several different datasets presented sequentially in one pass without dataset identifiers.



- In this work, for simplicity, we consider Task A, B, and C to be different datasets from with the same formulation (e.g., question answering, text classification).
- In this talk, I will use question answering as an example (eventually every task can be formulated as question answering, but I will talk about that later).

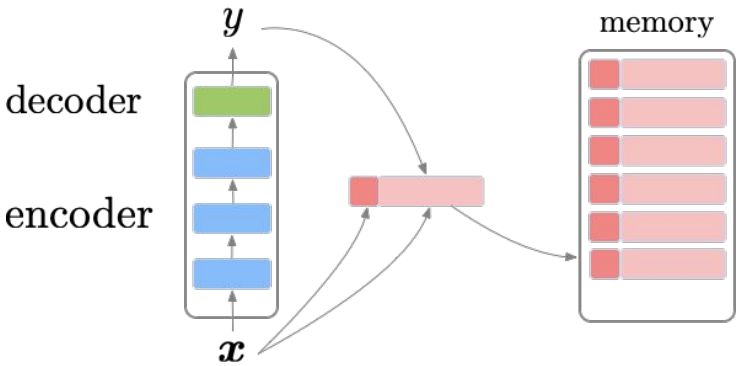
On Memory-Augmented Neural Networks

- Episodic memory is a long-term memory of events. In humans, it encodes individual experiences of the world. In a neural network model, it is typically associated with a memory module that stores previously seen (training) examples.

- Contrast this with a short-term (working) memory presents in e.g., an LSTM (Hochreiter and Schmidhuber, 1997) , stack-augmented neural network (Joulin and Mikolov, 2015; Grefenstette et al., 2015), or differentiable neural computers (Graves et al., 2016). See **Yogatama et al., ICLR 2018** for an overview and comparisons of how this type of memory works in language models.

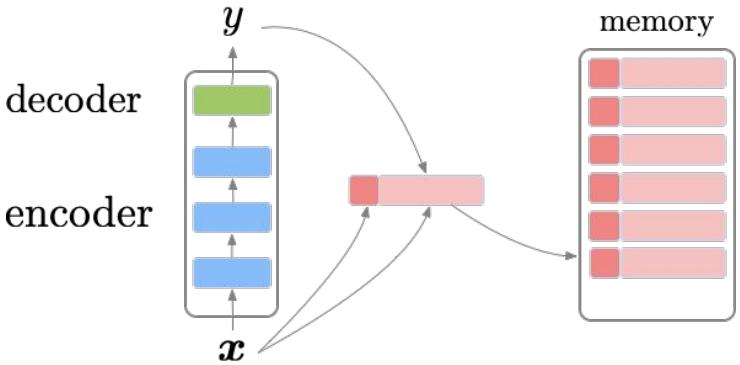
Models

- A context-question encoder; e.g., ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), etc.
- An answer decoder, i.e., a span decoder that predicts two indices (start and end indices).
- A key-value episodic memory module.
  - Key: embedding of a question, computed using a pretrained fixed encoder (e.g., BERT).
  - Value: context, question, and answer in textual forms (strings).



Models

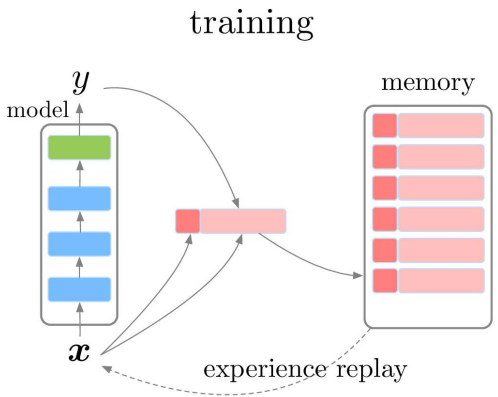
- A context-question encoder; e.g., ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), etc.
- An answer decoder, i.e., a span decoder that predicts two indices (start and end indices).
- A key-value episodic memory module.
  - Key: embedding of a question, computed using a pretrained fixed encoder (e.g., BERT).
  - Value: context, question, and answer in textual forms (strings).



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

Models

- During training, we perform *sparse* experience replay where we retrain on randomly sampled examples from the memory at a 1% rate (similar to memory consolidation in human learning).

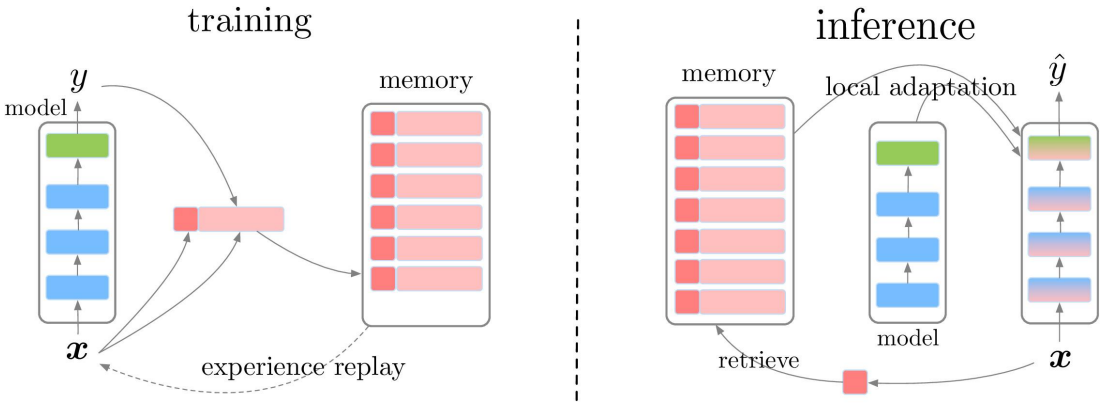


Models

- During training, we perform *sparse* experience replay where we retrain on randomly sampled examples from the memory at a 1% rate (similar to memory consolidation in human learning).

- During inference (prediction), we perform local adaptation similar to MbPA (Sprechmann et al., 2018).

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



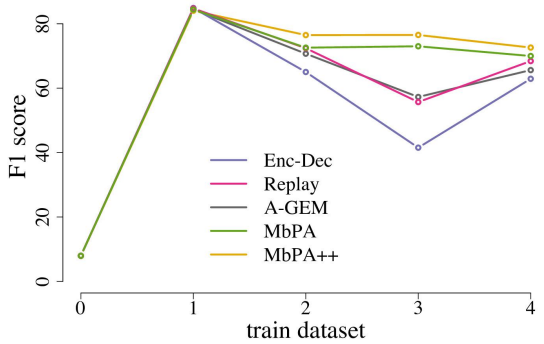
Experiments

- Four question answering datasets where:
  - The contexts come from **different domains** (e.g., Wikipedia articles, web pages).
  - The questions are posed in **different styles** (e.g., information seeking dialog, trivia questions).

Experiments

- Four question answering datasets where:
  - The contexts come from **different domains** (e.g., Wikipedia articles, web pages).
  - The questions are posed in **different styles** (e.g., information seeking dialog, trivia questions).

	Enc-Dec	A-GEM	MbPA	MbPA++	Multitask (upper bound)
QA	53.1	56.2	60.3	**62.4**	67.8
TextCat	18.4	66.9	68.8	**70.6**	73.6



Summary and Limitations

- Episodic memory mitigates catastrophic forgetting and allows the model to deal with changes in data distribution.

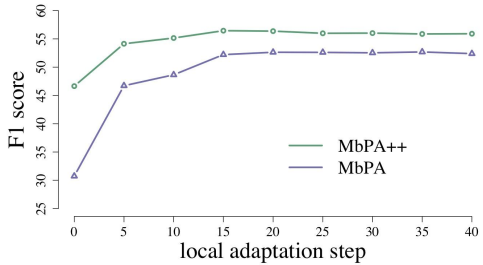
Summary and Limitations

- Episodic memory mitigates catastrophic forgetting and allows the model to deal with changes in data distribution.

- The current memory module has a linear **space complexity** in the number of examples seen.
  - We have some evidence that only a small subset of examples needs to be stored, but choosing them is difficult.

10%	50%	100%
67.6	70.3	70.6

  - **Constant** capacity is desired: compression, forgetting, learning what to remember.
- Local adaptation at inference time is computationally expensive.

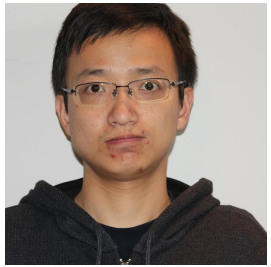


This Talk

- 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).

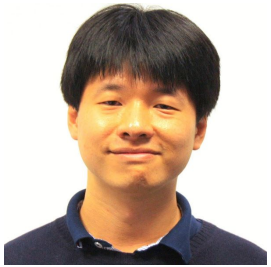
A Mutual Information Maximization Perspective of Language Representation Learning

Kong et al., ICLR 2020

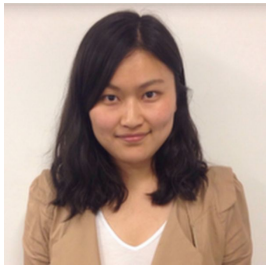


Lingpeng

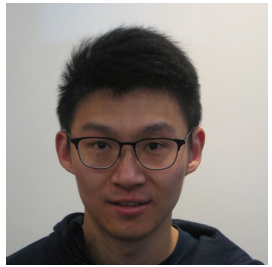
Cyprien



Wang



Lei

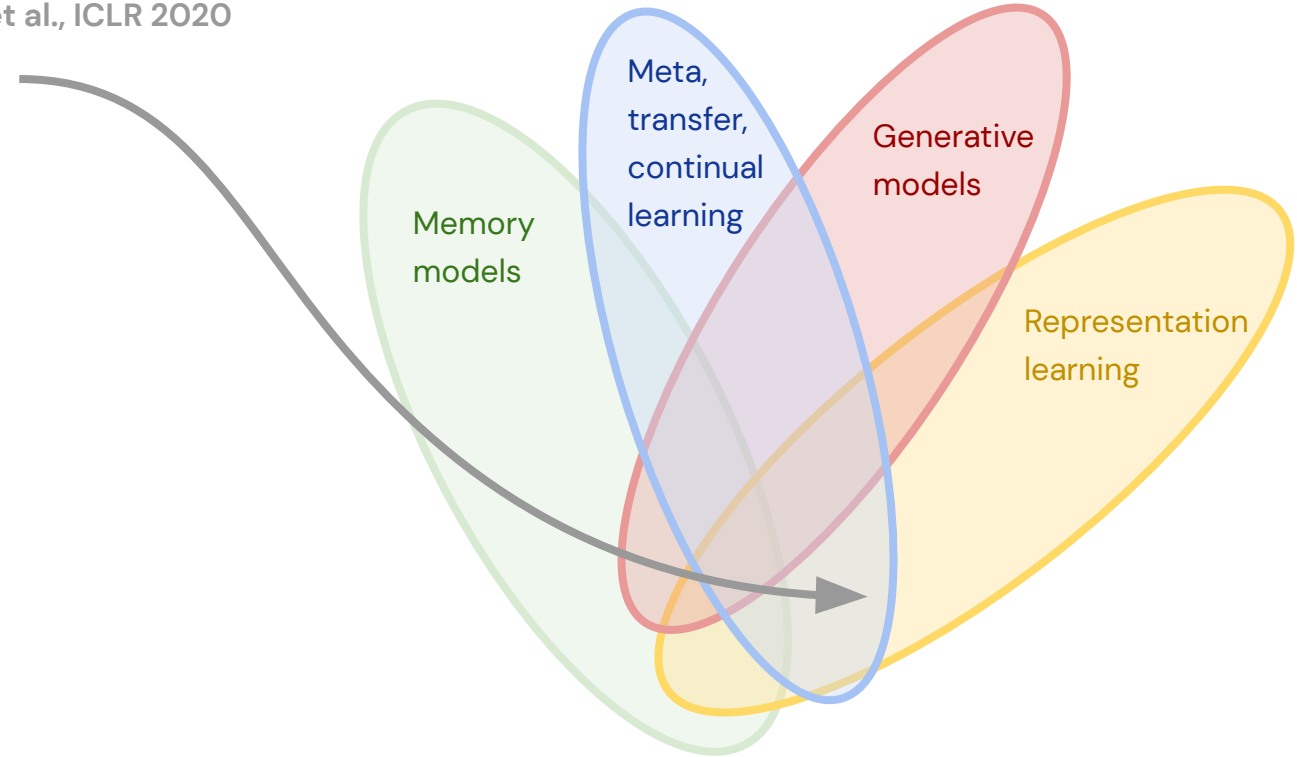


Zihang

Dani

A Mutual Information Maximization Perspective of Language Representation Learning

Kong et al., ICLR 2020



Background

- Unsupervised or self-supervised training has been shown to produce better models for downstream tasks.
- The field has progressed from using (optimized) bag-of-words (Yogatama et al., EMNLP 2015) to word embeddings (skip gram, Mikolov et al., 2013; GloVe, Pennington et al., 2014) to contextual embeddings (BERT, Devlin et al., 2019; XLNet, Yang et al., 2019).

- Key to sample efficiency, which is a necessary property of general linguistic intelligence.

- **Hypothesis**: both classical and modern word representation learning methods are trained to maximize an objective function that is a lower bound on the mutual information between different parts of a sentence.

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

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

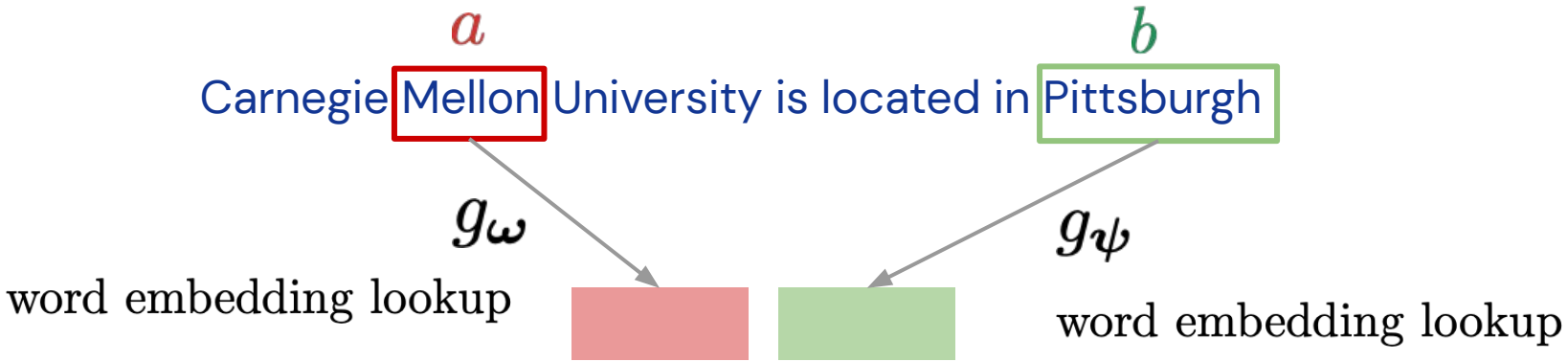
Skip-gram

Mikolov et al., 2013

Carnegie Mellon University is located in Pittsburgh

Skip-gram

Mikolov et al., 2013



BERT

Devlin et al., 2019

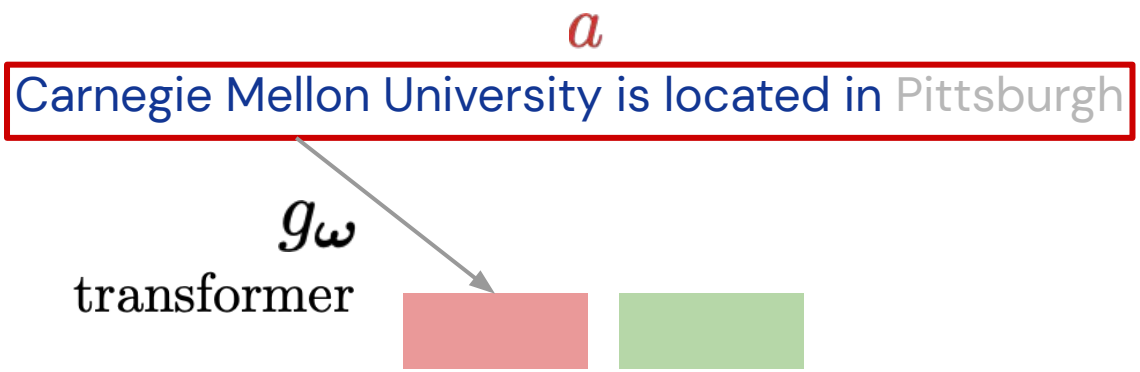
Carnegie Mellon University is located in Pittsburgh

BERT

Devlin et al., 2019

BERT

Devlin et al., 2019



Why is this interesting?

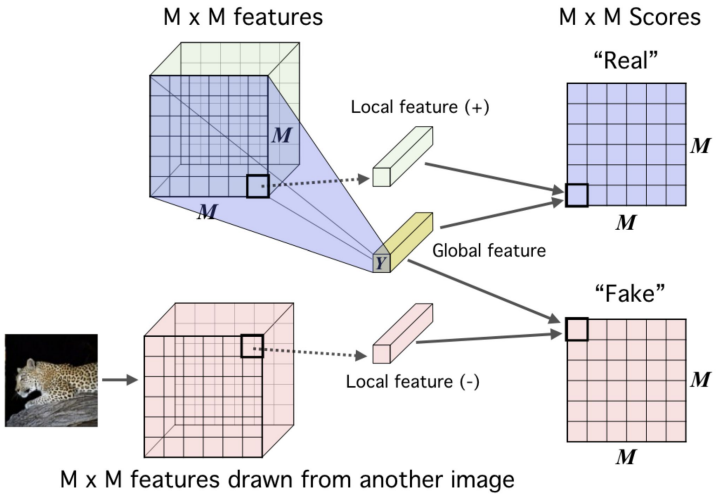
- A framework that unifies classical word embedding methods and modern contextual embeddings and connect them to self-supervised representation learning methods used in other domains (vision, speech).

	a	b	$p(a, b)$	g_ω	g_ψ
Skip-gram	word	word	word and its context	lookup	lookup
BERT-MLM	context	word	masked tokens probability	transformer	lookup
XLNet	context	word	factorization permutation	TXL++	lookup
InfoWord-DIM	context	n-grams	Sentence and its n-grams	transformer	N/A

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

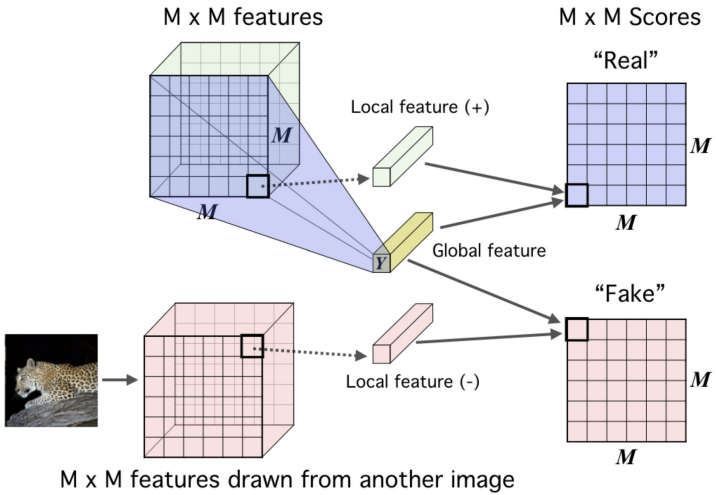
InfoWord

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



InfoWord

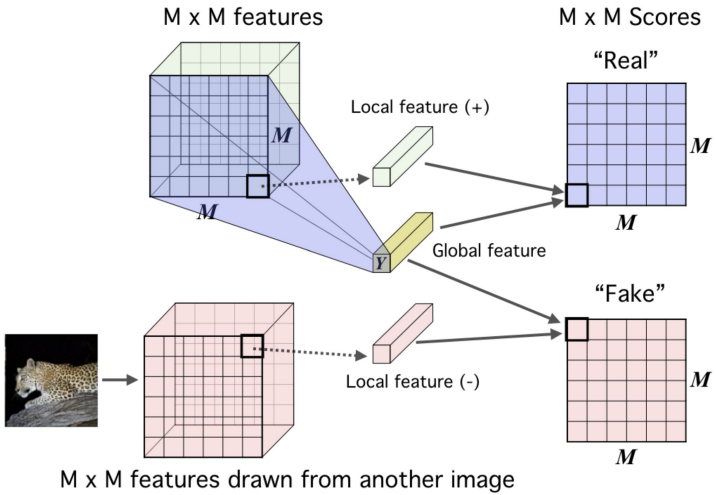
Deep InfoMax (DIM; Hjelm et al., 2019)



Carnegie Mellon University is located in Pittsburgh

InfoWord

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



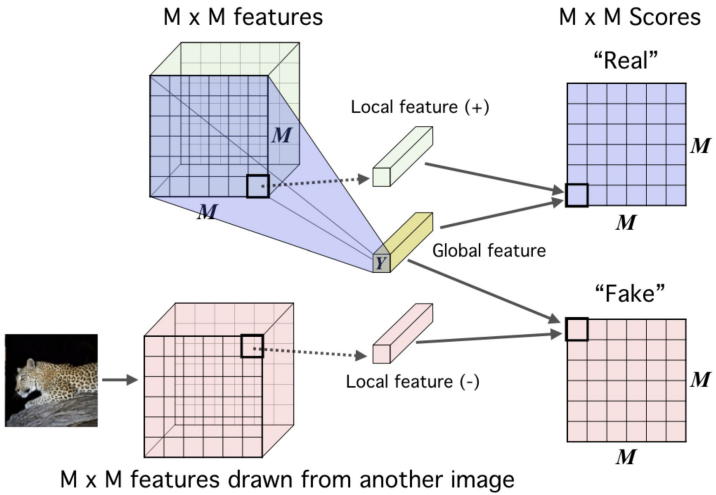
Carnegie Mellon University is located in Pittsburgh

g_ψ

transformer

InfoWord

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



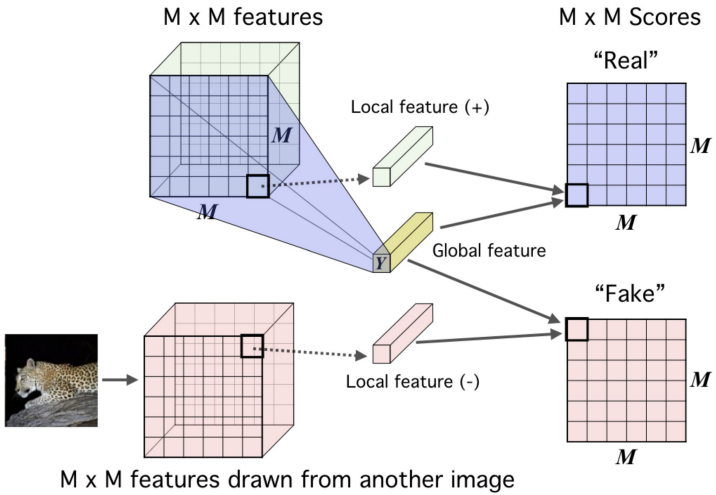
Carnegie Mellon University is located in Pittsburgh

g_ω

transformer

InfoWord

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



Carnegie Mellon University is located in Pittsburgh

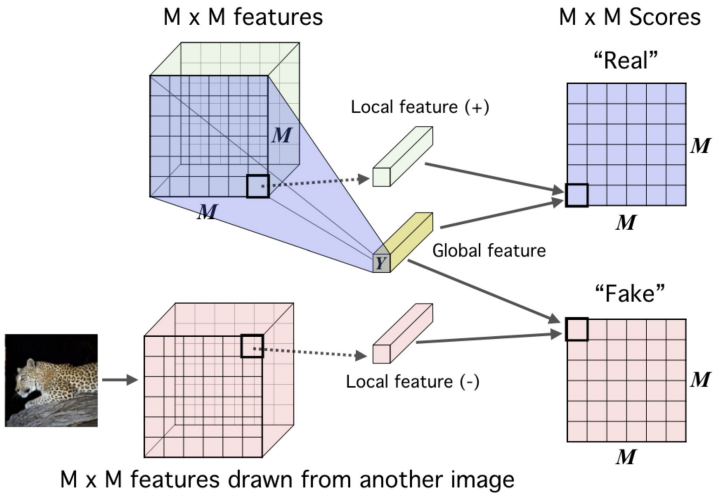
Starcraft II is a boring game

Cristiano Ronaldo scores an own goal

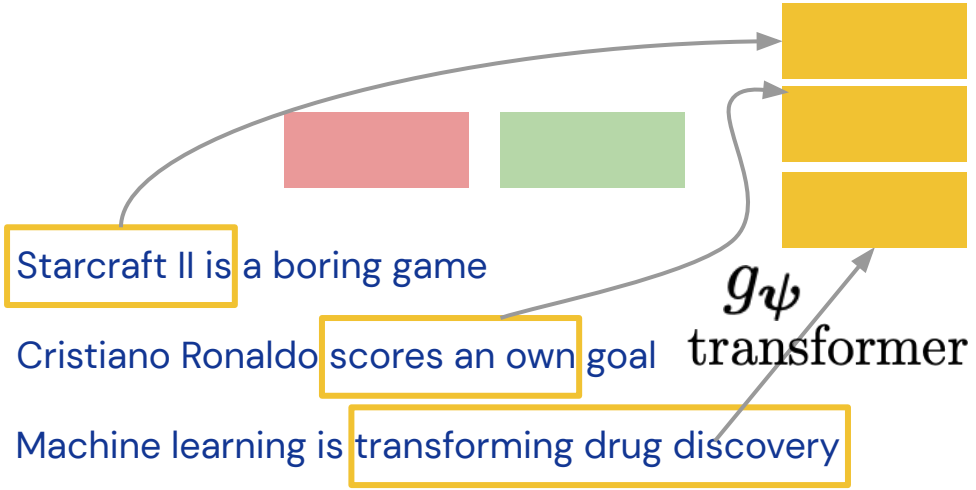
Machine learning is transforming drug discovery

InfoWord

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

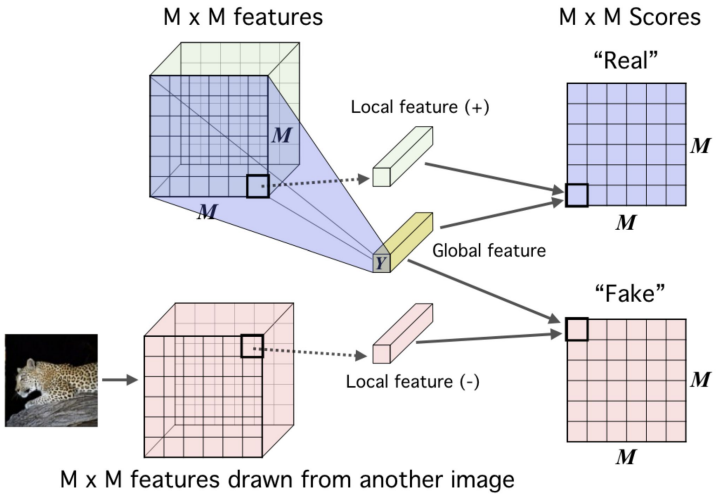


Carnegie Mellon University is located in Pittsburgh



InfoWord

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



Carnegie Mellon University is located in Pittsburgh

Starcraft II is a boring game

Cristiano Ronaldo scores an own goal

Machine learning is transforming drug discovery

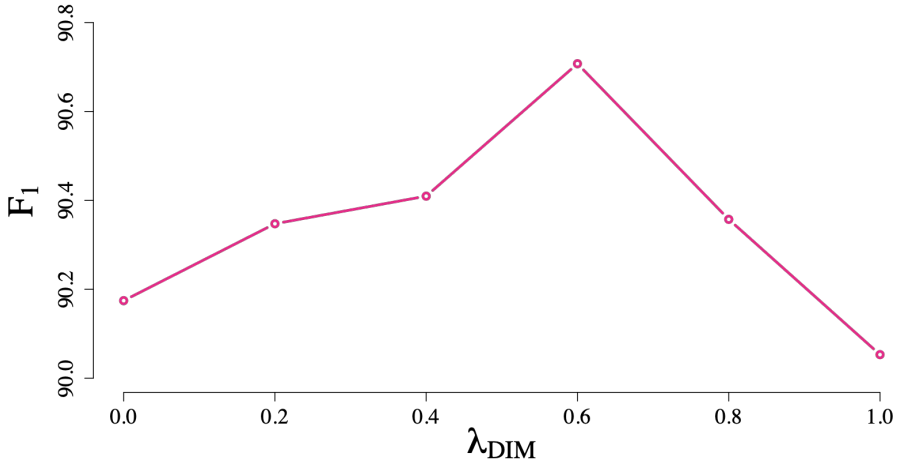
$$\mathcal{I}_{\text{DIM}} = \mathbb{E}_{p(\hat{\boldsymbol{x}}_{i:j}, \boldsymbol{x}_{i:j})} \left[\underbrace{g_{\boldsymbol{\omega}}(\hat{\boldsymbol{x}}_{i:j})}_{a}^{\top} \underbrace{g_{\boldsymbol{\omega}}(\boldsymbol{x}_{i:j})}_{b} - \log \sum_{\tilde{\boldsymbol{x}}_{i:j} \in \tilde{\mathcal{S}}}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

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

Experiments

- Question answering on SQuAD 1.1 (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**



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

Summary and Limitations

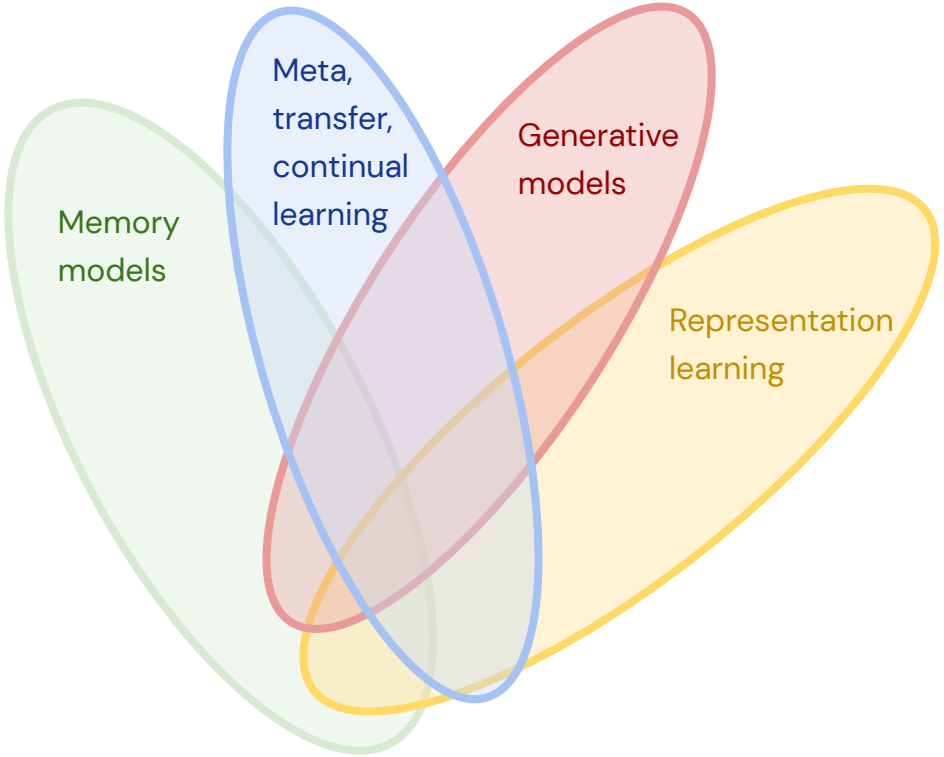
- Progress in self-supervised language representation learning has largely been driven by advances in model architectures.
- A framework that unifies classical word embedding methods and modern contextual embeddings and connects them to representation learning methods used in other domains.
- The framework facilitates transfer of ideas across domains when designing self-supervised tasks.

Summary and Limitations

- Progress in self-supervised language representation learning has largely been driven by advances in model architectures.
- A framework that unifies classical word embedding methods and modern contextual embeddings and connects them to representation learning methods used in other domains.
- The framework facilitates transfer of 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

- Language-independent representations.
  - Humans acquire language with a mechanism that is independent of the language.
  - Learning to compose words into sentences (Yogatama et al., ICLR 2017; Maillard et al., JNLE 2019)
  - Humans learn abstractions that generalize to other languages.

→ Provides an insight into learning generalizable representations.

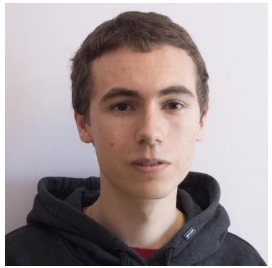
Future Directions

- Language-independent representations.
  - Humans acquire language with a mechanism that is independent of the language.
  - Learning to compose words into sentences (Yogatama et al., ICLR 2017; Maillard et al., JNLE 2019)
  - Humans learn abstractions that generalize to other languages.

→ Provides an insight into learning generalizable representations.

On the Crosslingual Transferability of Monolingual Representations

Artetxe et al., arXiv 2019



Mikel

Sebastian

Dani

	EN	ES	TR	AR	Average (11 langs.)
Multilingual Training	82.7	74.3	42.9	52.3	54.5
Monolingual Training + Transfer	83.9	61.3	51.2	61.0	66.8

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; Kong et al., ICLR 2018).
  - Crucial to imagination and planning (Weber et al., 2017).

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

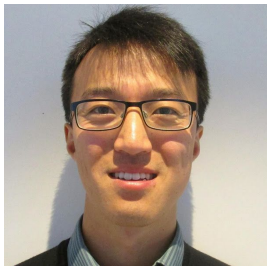
Future Directions

- 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; Kong et al., ICLR 2018).
  - Crucial to imagination and planning (Weber et al., 2017).

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

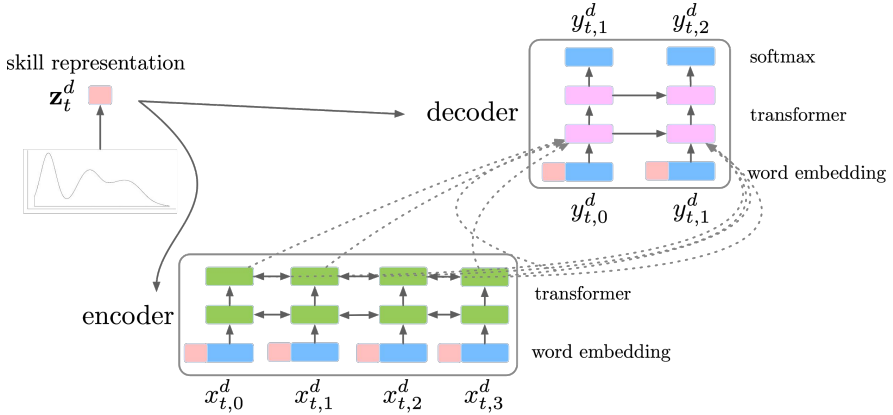
Modelling Latent Skills for Multitask Language Generation

Cao and Yogatama, arXiv 2020



Kris

Dani



Future Directions

- Semi-parametric models---a combination of a big parametric neural network with non-parametric components such as memory modules (cache, stacks, memory blocks).
  - Human memory has specialized systems for different functions.
  - Implicit representation of memory as parameters (weights) suffers from forgetting.
  - The separation of compute and storage is a necessary architectural bias.
  - Important for multihop and commonsense reasoning.

Learning what to remember and forget, how to consolidate (compress) past experiences to facilitate lifelong learning.

Thank you!

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

DeepMind

Frontiers of NLP