+

CHULA ΣNGINEERING
Foundation toward Innovation
COMPUTER

NLP

Lifelong Learning

2110572: Natural Language Processing Systems

Kasidis Kanwatchara & Thanapapas Horsuwan

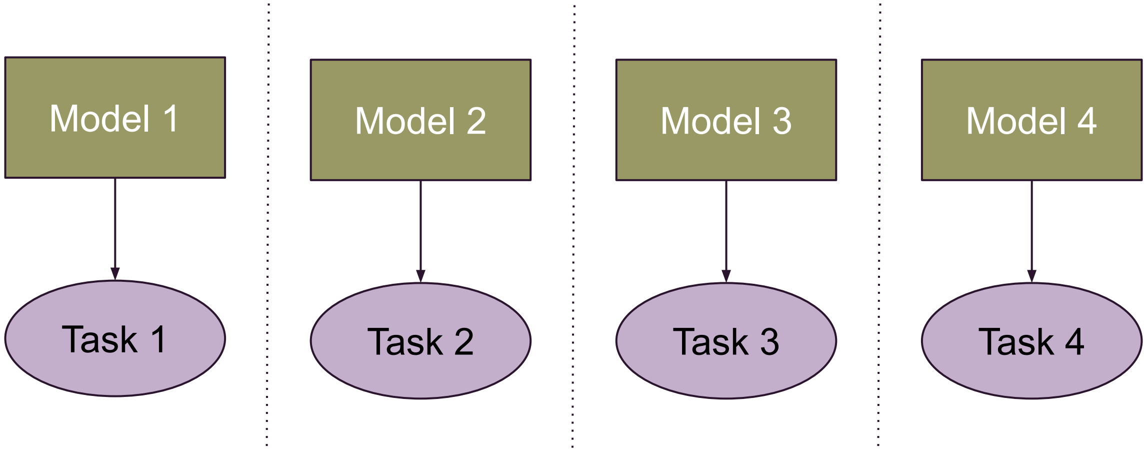
Outline

- Introduction
- Approaches to lifelong learning
  - Architectural based
  - Regularization based
  - Data based
- Lifelong language learning
  - LAMOL
  - MbPA++
- Benchmark
- Closing remarks

Introduction

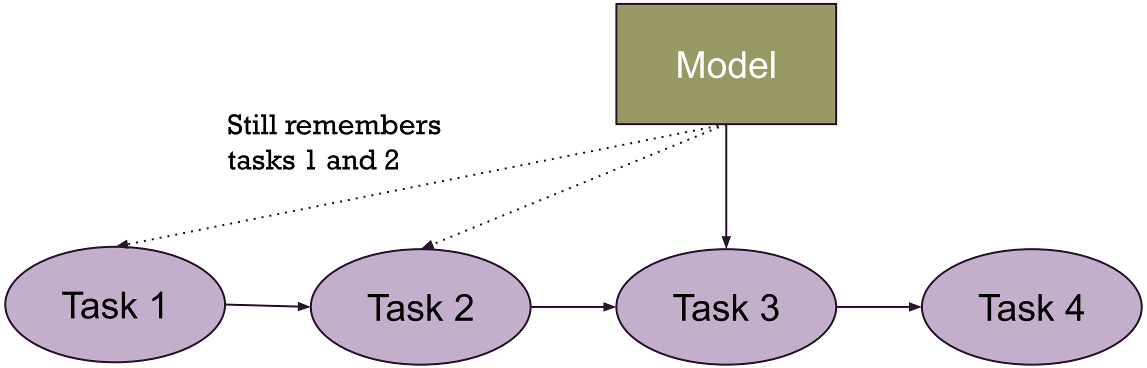
Introduction

Today, the machine learning models we train are highly specialized on a single task.
Nevertheless, they cannot do anything else that they were not trained to do.
This is called "Isolated learning".



Introduction (cont.)

If we want to create a true AI, we will need it to be able to learn like humans do; task by task, without forgetting what it has learned so far. This learning paradigm is called "Sequential Learning".



Introduction (cont.)

However, if we just subject the models we have today to sequential learning, they will not be able to retain any knowledge from the past due to **Catastrophic Forgetting (CF)** [1].

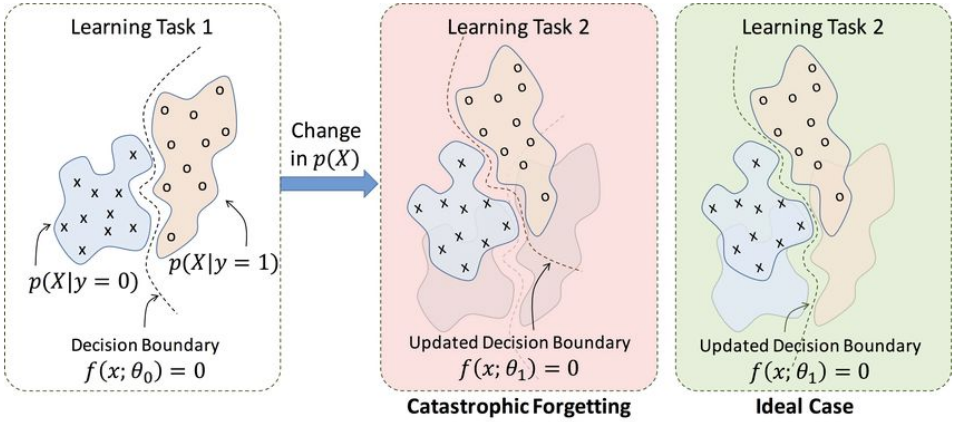
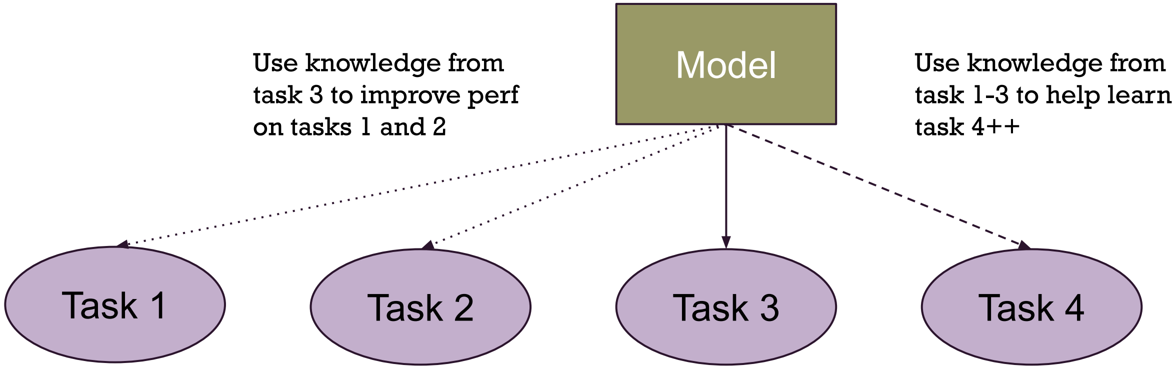


Figure from [2]

[1] Michael McCloskey, Neal J. Cohen. (1989). Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem.
[2] Kolouri, Soheil & Ketz, Nicholas & Zou, Xinyun & Krichmar, Jeff & Pilly, Praveen. (2019). Attention-Based Structural-Plasticity.

Lifelong Learning (Continual Learning)

A long-standing research field that focuses on solving the problem of CF.



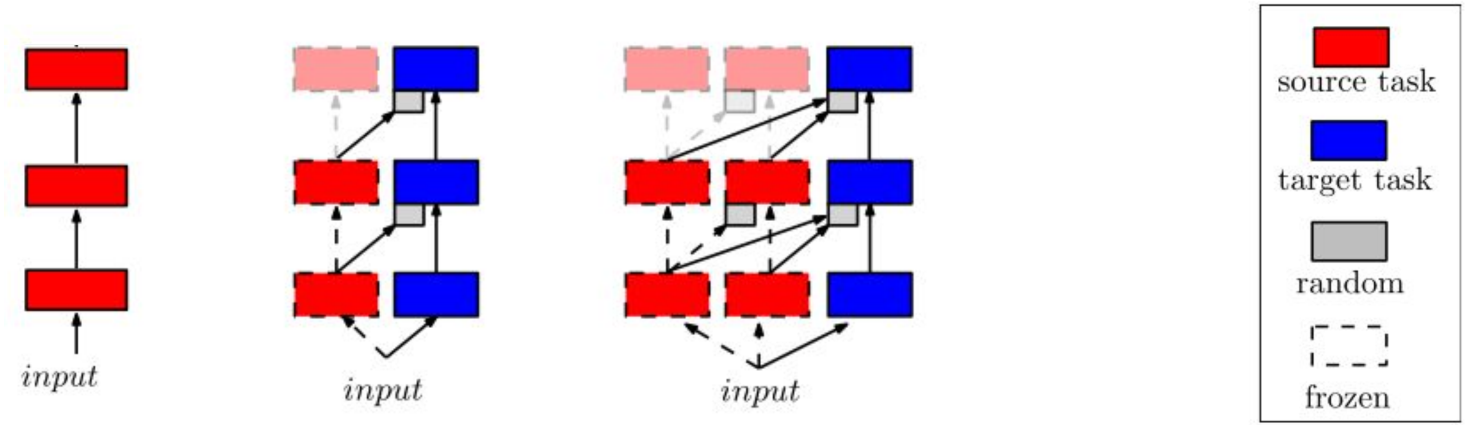
Approaches to lifelong learning

Approaches to Lifelong Learning

Architectural-based	Regularization-based	Data-based
Introducing task-specific parameters	Add a regularization term that aids knowledge consolidation	Keep some samples in memory

1. Architectural-based

Progressive Neural Network (PNN) - Add a new “column” when encountering a new task and keep the original network frozen



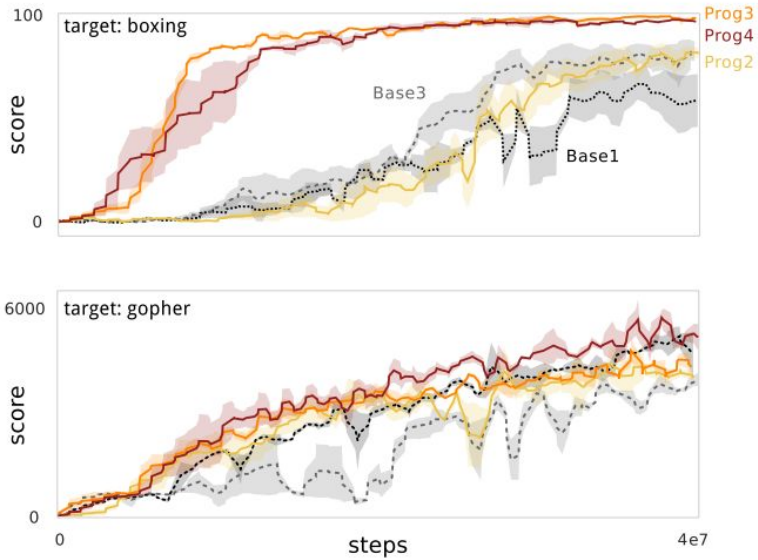
Task 1 Task 2 Task 3

Progressive Neural Networks. CoRR abs/1606.04671 (2016)

1. Architectural-based (cont.)

In reinforcement learning experiments, PNNs show signs of **positive transfer**, i.e., using past experiences to help learn new tasks.

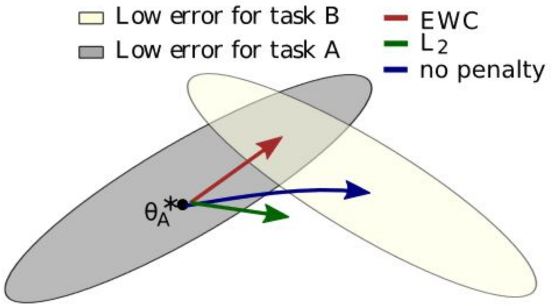
However, the number of parameters grows linearly as new tasks keep coming, making it not practical.

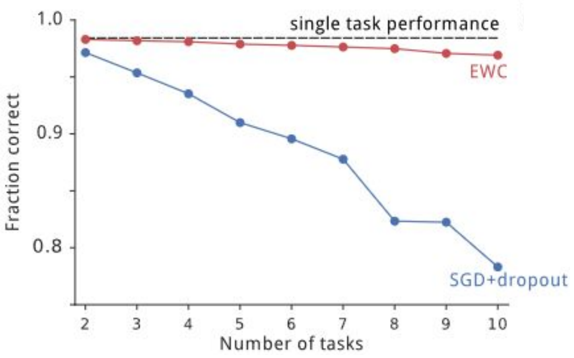


Progressive Neural Networks. Rusu et al., 2016

2. Regularization-based

Elastic Weight Consolidation (EWC) - uses a regularization term to prevent the model weights from shifting from the old model too much, thus preventing catastrophic forgetting.





The downside is that since the model capacity is fixed, eventually, the model will not be able to learn anything new due to the regularization.

Overcoming catastrophic forgetting in neural networks. Kirkpatrick et al., 2017.

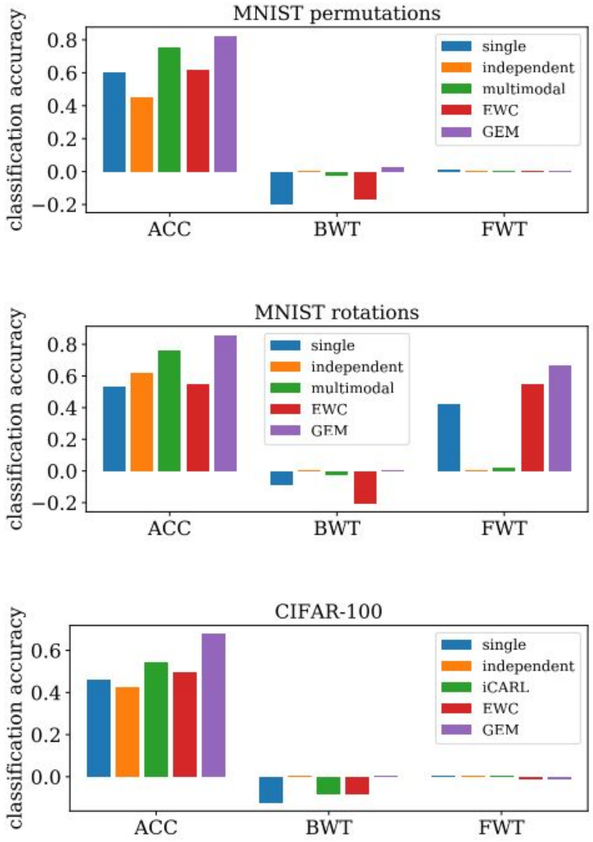
3. Data-based (Rehearsal-based)

Gradient Episodic Memory (GEM) - keeps a small subset of data from previous tasks to prevent loss on these exemplars from increasing when trained on new tasks.

This can be achieved by projecting the gradient so that it satisfies the following equality constraint:

$$\begin{aligned} \text{minimize}_\theta \quad & \ell(f_\theta(x,t),y) \\ \text{subject to} \quad & \ell(f_\theta,\mathcal{M}_k) \leq \ell(f_\theta^{t-1},\mathcal{M}_k) \text{ for all } k < t, \end{aligned}$$

where t refers to a task descriptor, M_k is the exemplars in the buffer and f_θ is the network.



Gradient episodic memory for continual learning. Lopez-Paz et al., 2017.

Lifelong language learning

Lifelong Language Learning (LLL)

A sub-field of LL that focuses on NLP tasks. The amount of research word in this field is rather scant compared to other fields.

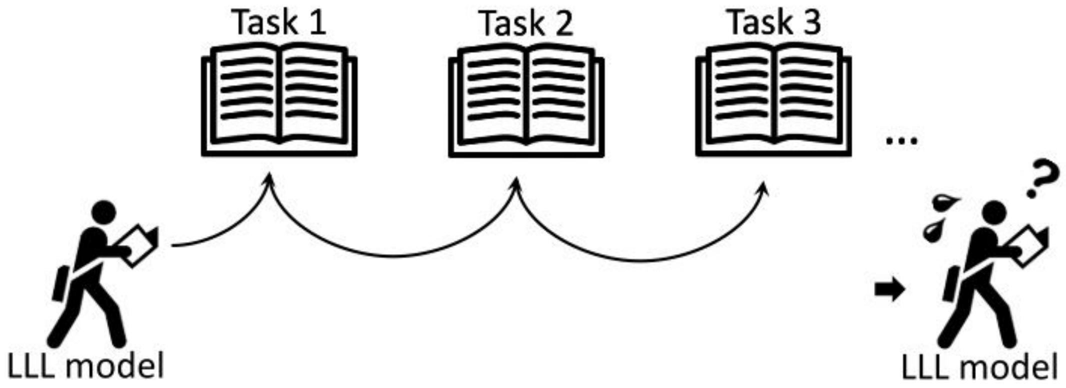
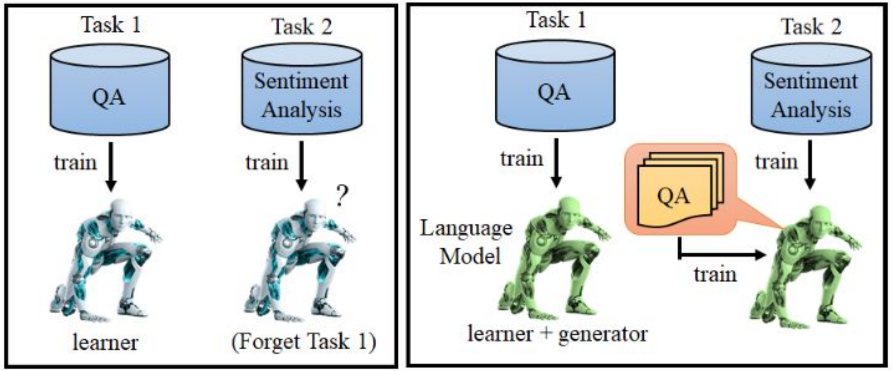


Image from : Lifelong Language Knowledge Distillation. Chuang et al., 2020.

LAMOL

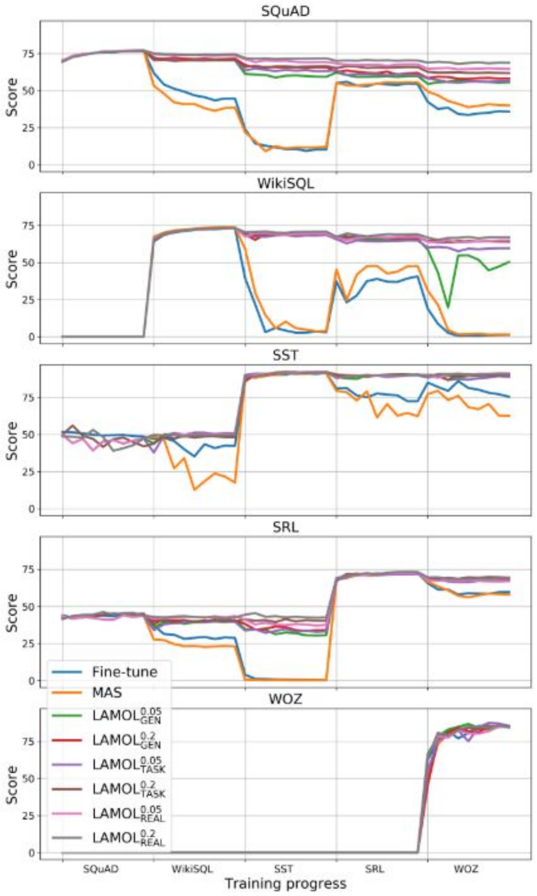
LAMOL

- Uses a single GPT2 model to solve all NLP tasks in a unified format.
- Before training on a new task, LAMOL generates pseudo samples to augment the training data of the new task.
- Results show that LAMOL can effectively prevent CF.



Sequential fine-tuning

LAMOL



LAMOL: LAnguage MOdeling for Lifelong Language Learning. Sun et al., 2020

LAMOL (cont.)

To be able to solve multiple NLP tasks using a single architecture, LAMOL uses the decaNLP formatting that frames all NLP tasks into the QA task.

Given a context and a question, the GPT model just has to generate the correct answer.

Examples

Question	Context	Answer
What is a major importance of Southern California in relation to California and the US?	...Southern California is a major economic center for the state of California and the US....	major economic center
What is the translation from English to German?	Most of the planet is ocean water.	Der Großteil der Erde ist Meerwasser
What is the summary?	Harry Potter star Daniel Radcliffe gains access to a reported £320 million fortune...	Harry Potter star Daniel Radcliffe gets £320M fortune...
Hypothesis: Product and geography are what make cream skimming work. Entailment, neutral, or contradiction?	Premise: Conceptually cream skimming has two basic dimensions – product and geography.	Entailment
Is this sentence positive or negative?	A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.	positive

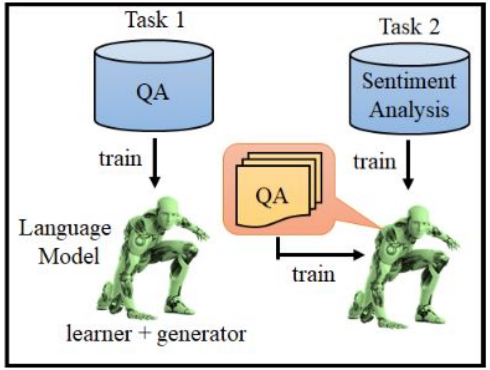
LAMOL: LAnguage MOdeling for Lifelong Language Learning. Sun et al., 2020
The Natural Language Decathlon: Multitask Learning as Question Answering. McCann et al., 2018

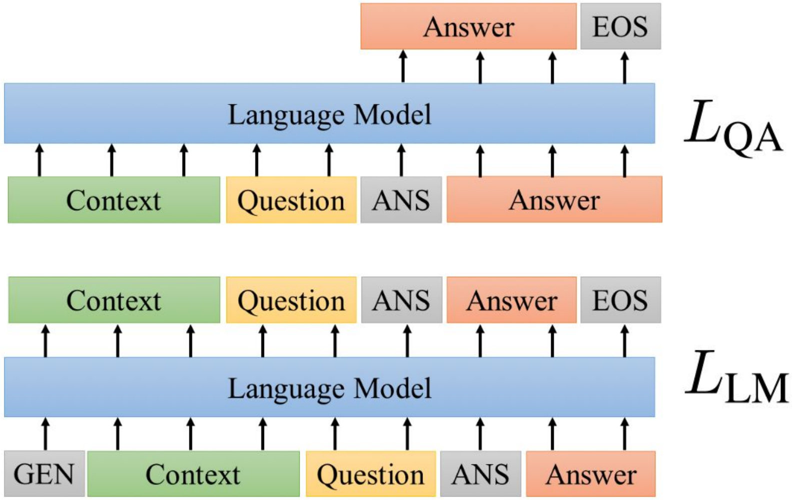
LAMOL (cont.)

Recall that LAMOL generates **pseudo samples** to use in conjunction with the new task data.

To improve the quality of the pseudo sample generation, LAMOL trains on an auxiliary LM loss.

$$L = L_{\mathrm{QA}} + \lambda L_{\mathrm{LM}}$$





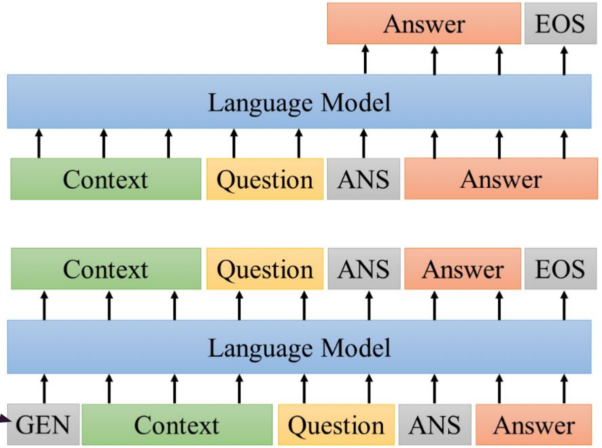
LAMOL: LAnguage MOdeling for Lifelong Language Learning. Sun et al., 2020

LAMOL (cont.)

Since LAMOL generates a limited amount (20% of the size of the new task data) of pseudo samples, we would like to balance the amount of pseudo samples from each old task.

To this extent, a **task-specific token** is added for every new task so that during the pseudo sample generation, we can inform the model from which task we want the data from.

e.g. we input __movie__ as the task-specific token, replacing GEN, and the model knows that we want a pseudo sample from the IMDB task.



LAMOL: LAnguage MOdeling for Lifelong Language Learning. Sun et al., 2020

LAMOL (cont.)

By training on the mixture of pseudo samples and the new data, LAMOL can effectively prevent CF.

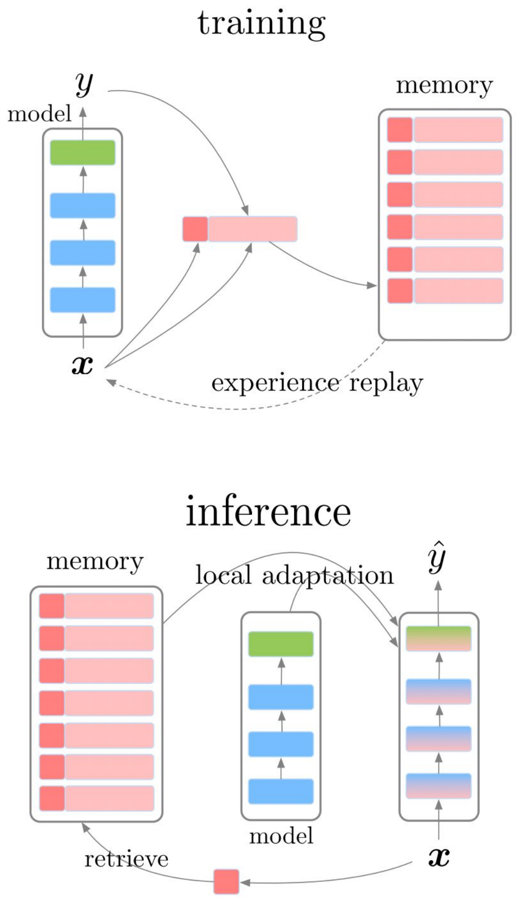
	Methods	SST SRL WOZ	SST WOZ SRL	SRL SST WOZ	SRL WOZ SST	WOZ SST SRL	WOZ SRL SST	Average	Std
	Fine-tuned	50.2	24.7	62.9	31.3	32.8	33.9	39.3	12
Reg-based	EWC	50.6	48.4	64.7	35.5	43.9	39.0	47.0	8.7
Reg-based	MAS	36.5	45.3	56.6	31.0	49.7	30.8	41.6	8.9
Data-based	GEM	50.4	29.8	63.3	32.6	44.1	36.3	42.8	11
	$\text{LAMOL}^{0}_{\text{GEN}}$	46.5	36.6	56.6	38.6	44.9	45.2	44.8	6.0
	$\text{LAMOL}^{0.05}_{\text{GEN}}$	79.6	78.9	73.1	73.7	68.6	75.7	74.9	3.4
	$\text{LAMOL}^{0.2}_{\text{GEN}}$	80.0	80.7	79.6	78.7	78.4	80.5	**79.7**	0.8
	$\text{LAMOL}^{0}_{\text{TASK}}$	41.0	33.5	50.1	41.9	49.3	41.5	42.9	5.2
	$\text{LAMOL}^{0.05}_{\text{TASK}}$	77.3	76.9	78.1	74.7	73.4	75.8	76.0	1.5
	$\text{LAMOL}^{0.2}_{\text{TASK}}$	79.4	79.9	80.1	78.7	79.8	79.0	79.5	**0.5**
	$\text{LAMOL}^{0.05}_{\text{REAL}}$	81.0	78.9	80.1	80.9	77.7	78.0	79.4	1.2
	$\text{LAMOL}^{0.02}_{\text{REAL}}$	81.8	80.6	81.6	81.2	80.4	80.5	81.0	0.5
Upper bound	Multitasked				81.5				

LAMOL: LAnguage MOdeling for Lifelong Language Learning. Sun et al., 2020

MbPA++

MbPA++

- Uses a BERT model with **a memory buffer** that randomly stores examples encountered during training.
- Do **experience replay** during training at a regular interval.
- During testing, chooses examples similar to the testing sample and do **local adaptation** for a fixed number of step before testing on the sample.



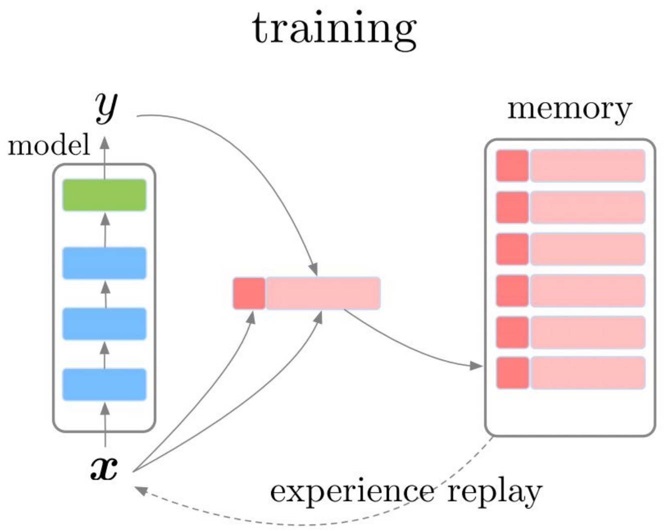
Episodic Memory in Lifelong Language Learning. d'Autume et al., 2019

MbPA++

During training, MbPA++ randomly stores training samples in a memory.

For every 10,000 steps, 100 stored samples are randomly taken from the memory and used to train the model for 1 step.

This is called "**experience replay**" and is done to prevent CF.



Episodic Memory in Lifelong Language Learning. d'Autume et al., 2019

MbPA++

During inference, given a test example, the model would take *k* most similar examples and uses it to perform gradient based "**local adaptation**" (basically finetune) for 30 steps.

Note that for the next test example, the model parameter would revert back to before local adaptation.

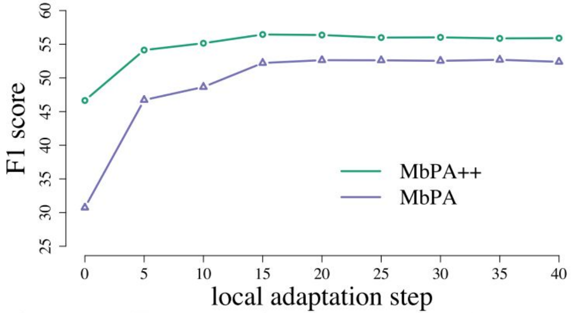


Figure 3: F_1 scores for MbPA**++** and MbPA as the # of local adaptation steps increases.

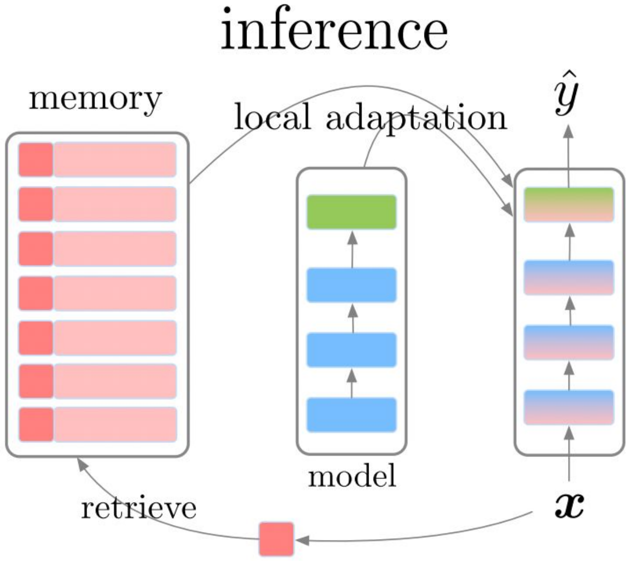


Table 3: Results for different # of retrieved examples K.

	8	16	32	64	128
class.	68.4	69.3	70.6	71.3	71.6
QA	60.2	60.8	62.0	-	-

Episodic Memory in Lifelong Language Learning. d'Autume et al., 2019

MbPA++

By using experience replay and local adaptation, MbPA++ achieves SOTA results (at that time).

	Sequential		No local adaptation	No ER	Not using k-nearest neighbour at local adaptation		
Order	**ENC-DEC**	**A-GEM**	**REPLAY**	**MBPA**	**MBPA$_{++}^{rand}$**	**MBPA++**	**MTL**
i	14.8	70.6	67.2	68.9	59.4	**70.8**	73.7
ii	27.8	65.9	64.7	68.9	58.7	**70.9**	73.2
iii	26.7	67.5	64.7	68.8	57.1	**70.2**	73.7
iv	4.5	63.6	44.6	68.7	57.4	**70.7**	73.7
class.-avg.	18.4	66.9	57.8	68.8	58.2	**70.6**	73.6

Episodic Memory in Lifelong Language Learning. d'Autume et al., 2019

+

Benchmark

Benchmark

In LLL, we can use any dataset and just train the model on a sequence of them. The two most popular task sequences are the following:

Task	Dataset	# Train	# Test	Metric
Question answering	SQuAD	87599	10570	nF1
Semantic parsing	WikiSQL	56355	15878	lfEM
Sentiment analysis	SST	6920	1821	EM
Semantic role labeling	QA-SRL	6414	2201	nF1
Goal-oriented dialogue	WOZ	2536	1646	dsEM
Text classification	AGNews Amazon DBPedia Yahoo Yelp	115000	7600	EM

LAMOL: LAnguage MOdeling for Lifelong Language Learning. Sun et al., 2020

Benchmark

The de facto benchmark is training on multiple permutations of the same task sequence and then test the model on all learned tasks and report the average.

Average the score of 5 tasks

Text classification We use the following text classification dataset orders for comparing our results with (d'Autume et al., 2019):

i. Yelp→AGNews→DBPedia→Amazon→Yahoo
ii. DBPedia→Yahoo→AGNews→Amazon→Yelp
iii. Yelp→Yahoo→Amazon→DBpedia→AGNews
iv. AGNews→Yelp→Amazon→Yahoo→DBpedia

Order	Enc-Dec	Online EWC	A-GEM†	Replay	MbPA++†	MbPA++ (Our Impl.)	Meta-MbPA	MTL	MTL (1%)	LAMOL‡
					Text Classification					
i.	35.5	43.8	70.7	63.4	70.8	75.3	**77.9**	-	-	76.7
ii.	44.8	49.8	65.9	73.0	70.9	74.6	**76.7**	-	-	77.2
iii.	42.4	59.5	67.5	65.8	70.2	75.6	**77.3**	-	-	76.1
iv.	28.6	52.0	63.6	74.0	70.7	75.5	**77.6**	-	-	76.1
Average	37.8	51.3	66.9	69.1	70.6	75.3	**77.3**	78.9	50.4	76.5

Efficient Meta Lifelong-Learning with Limited Memory. Wang et al., 2020

Benchmark

Then the scores are usually **averaged over multiple permutations** to show the robustness of the method across different data ordering.

Multitask learning (MTL) is considered to be the upper bound since MTL sees all the data at the same time thus there is no CF.

Order	Enc-Dec	Online EWC	A-GEM†	Replay	MbPA++†	MbPA++ (Our Impl.)	Meta-MbPA (1%)	MTL	MTL (1%)	LAMOL‡
					Text Classification					
i.	35.5	43.8	70.7	63.4	70.8	75.3	**77.9**	-	-	76.7
ii.	44.8	49.8	65.9	73.0	70.9	74.6	**76.7**	-	-	77.2
iii.	42.4	59.5	67.5	65.8	70.2	75.6	**77.3**	-	-	76.1
iv.	28.6	52.0	63.6	74.0	70.7	75.5	**77.6**	-	-	76.1
Average	37.8	51.3	66.9	69.1	70.6	75.3	**77.3**	78.9	50.4	76.5

Efficient Meta Lifelong-Learning with Limited Memory. Wang et al., 2020

Our research in LLL – R-LAMOL

Rational LAMOL: A Rationale-based Lifelong Learning Framework

Kasidis Kanwatchara, Thanapapas Horsuwan, Piyawat Lertvittayakumjorn, Boonserm Kijsirikul, Peerapon Vateekul

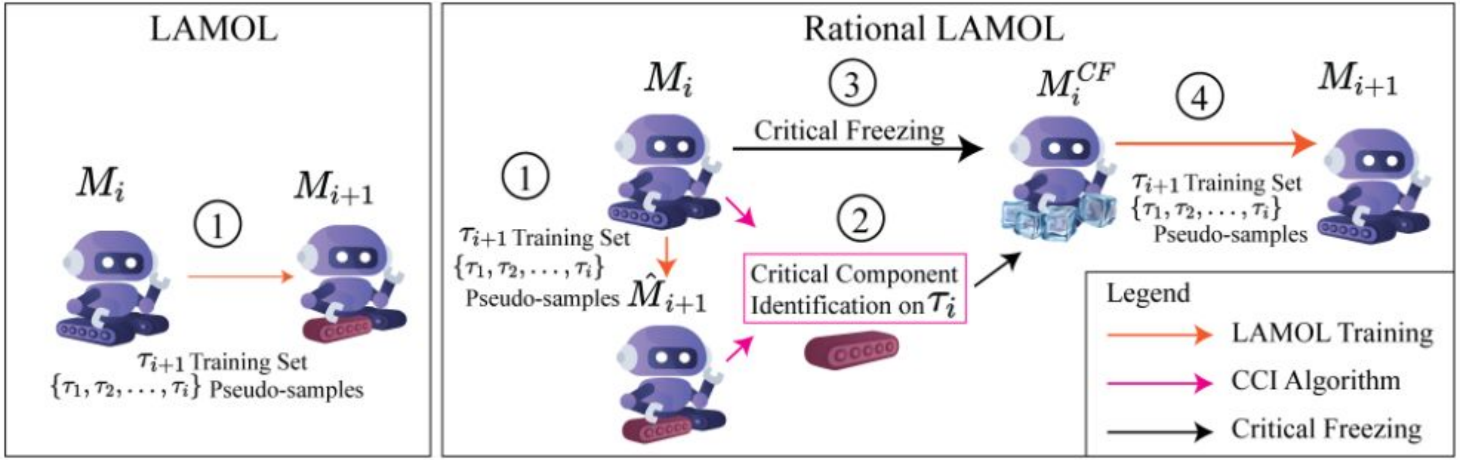
Abstract

Lifelong learning (LL) aims to train a neural network on a stream of tasks while retaining knowledge from previous tasks. However, many prior attempts in NLP still suffer from the catastrophic forgetting issue, where the model completely forgets what it just learned in the previous tasks. In this paper, we introduce Rational LAMOL, a novel end-to-end LL framework for language models. In order to alleviate catastrophic forgetting, Rational LAMOL enhances LAMOL, a recent LL model, by applying critical freezing guided by human rationales. When the human rationales are not available, we propose exploiting unsupervised generated rationales as substitutions. In the experiment, we tested Rational LAMOL on permutations of three datasets from the ERASER benchmark. The results show that our proposed framework outperformed vanilla LAMOL on most permutations. Furthermore, unsupervised rationale generation was able to consistently improve the overall LL performance from the baseline without relying on human-annotated rationales.

Anthology ID: 2021.acl-long.229

Volume: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)

Our research in LLL – R-LAMOL (cont.)



Methods	BMS	BSM	MBS	MSB	SBM	SMB	Average	Std.
LAMOL	57.39	55.98	65.89	66.71	67.63	60.08	62.28	5.09
Partial Brute Force $_{\text{block}}$	62.97	64.05	66.73	67.75	65.22	69.05	65.96	2.30
Rational LAMOL $_{\text{block}}$	62.49	59.55	66.09	68.04	68.55	59.94	64.11	4.57
Rational LAMOL $_{\text{head}}$	64.35	61.70	65.22	67.76	56.59	60.62	62.71	3.93
Gen-Rational LAMOL $_{\text{block}}$	66.82	59.97	66.38	65.11	66.94	64.49	64.95	2.63
Gen-Rational LAMOL $_{\text{head}}$	67.35	57.36	66.51	63.85	63.98	65.52	64.10	3.57
Multitask	67.32							

Our research in LLL – DoubleLM

December 01 2022

Enhancing Lifelong Language Learning by Improving Pseudo-Sample Generation

In Special Collection: CogNet

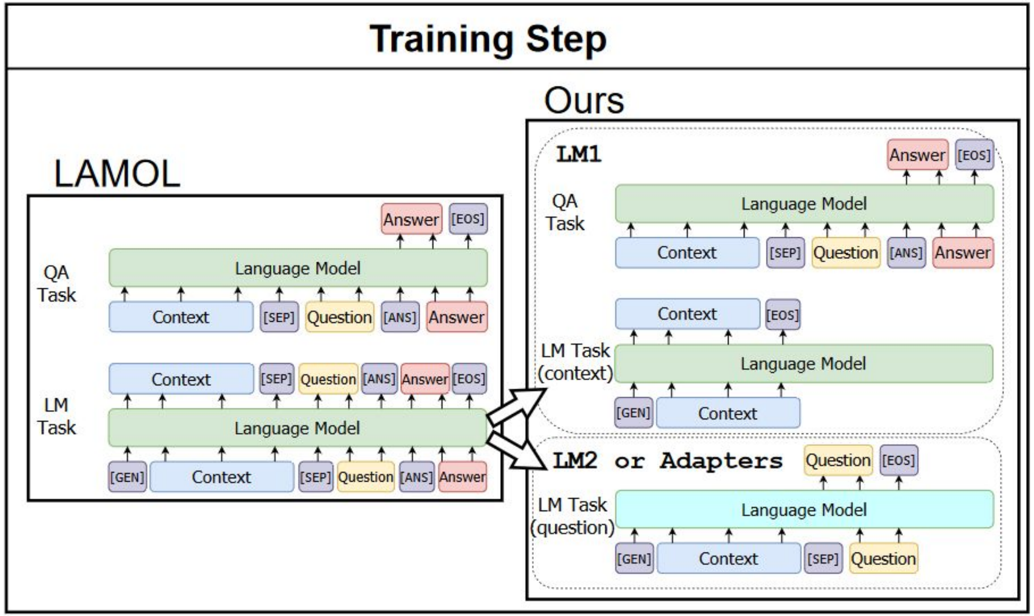
Kasidis Kanwatchara, Thanapapas Horsuwan, Piyawat Lertvittayakumjorn, Boonserm Kijsirikul, Peerapon Vateekul

> Author and Article Information

Computational Linguistics (2022) 48 (4): 819–848.

https://doi.org/10.1162/coli_a_00449 **Article history**

Our research in LLL – DoubleLM (cont.)



Methods	FBTMS	SMTBF	Average
LAMOL	57.01	44.32	50.67
LLKD	42.73	47.04	44.89
LM+Adapter	65.51	62.18	63.85
LM+Adapter+RT	66.03	67.74	66.88
LAMOL_{real}	70.95	71.83	71.39
Multitask		68.89	

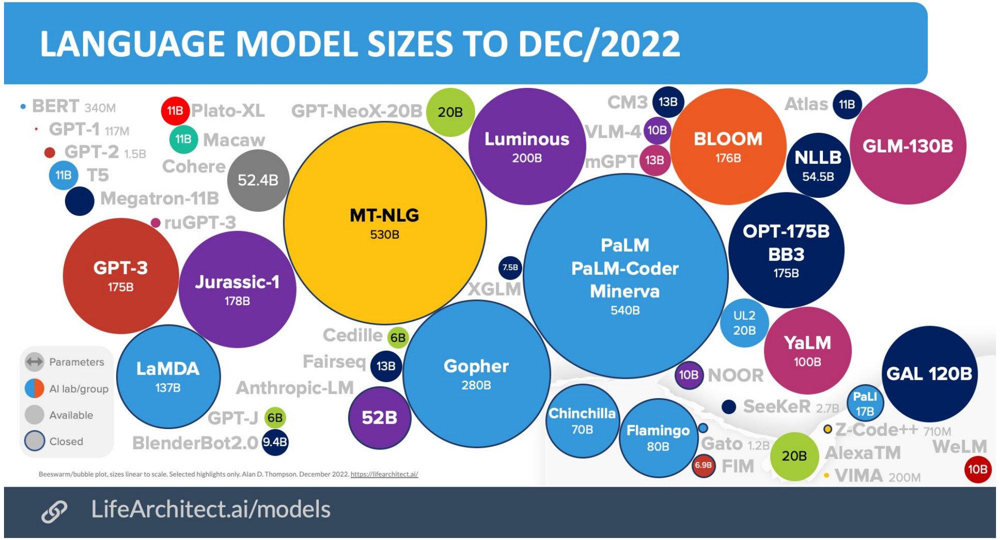
+

Closing remarks

Is LLL a deprecated field?

With the advent of LLMs and their ability to do zero-/few-shot learning, the future of LLL does not seem so bright anymore.

By doing in-context learning or prompt tuning, we can induce the desired behaviour without having to take a gradient step on the model. This also means that there will be no catastrophic forgetting.



Is LLL a deprecated field?

Still LLMs are not yet perfect.

Modern LLMs are limited by the data they were trained on. What if, after training a model, there are facts that we later found out to be false?

LLL might still be relevant in the future but it probably will be about adding/modifying knowledge into the model.

Yann LeCun
13 February

MetaAI's chief scientist

My unwavering opinion on current (auto-regressive) LLMs
1. They are useful as writing aids.
2. They are "reactive" & don't plan nor reason.
3. They make stuff up or retrieve stuff approximately.
4. That can be mitigated but not fixed by human feedback.
5. Better systems will come.
6. Current LLMs should be used as writing aids, not much more.
7. Marrying them with tools such as search engines is highly non trivial.
8. There *will* be better systems that are factual, non toxic, and controllable. They just won't be auto-regressive LLMs.
9.I have been consistent with the above while defending Galactica as a scientific writing aid.
10. Warning folks that AR-LLMs make stuff up and should not be used to get factual advice.
11. Warning that only a small superficial portion of human knowledge can ever be captured by LLMs.
12. Being clear that better system will be appearing, but they will be based on different principles. They will not be auto-regressive LLMs.
13. Why do LLMs appear much better at generating code than generating general text?
Because, unlike the real world, the universe that a program manipulates (the state of the variables) is limited, discrete, deterministic, and fully observable.
The real world is none of that.
14. Unlike what the most acerbic critics of Galactica have claimed
- LLMs *are* being used as writing aids.
- They *will not* destroy the fabric of society by causing the mindless masses to believe their made-up nonsense.
- People will use them for what they are helpful with.

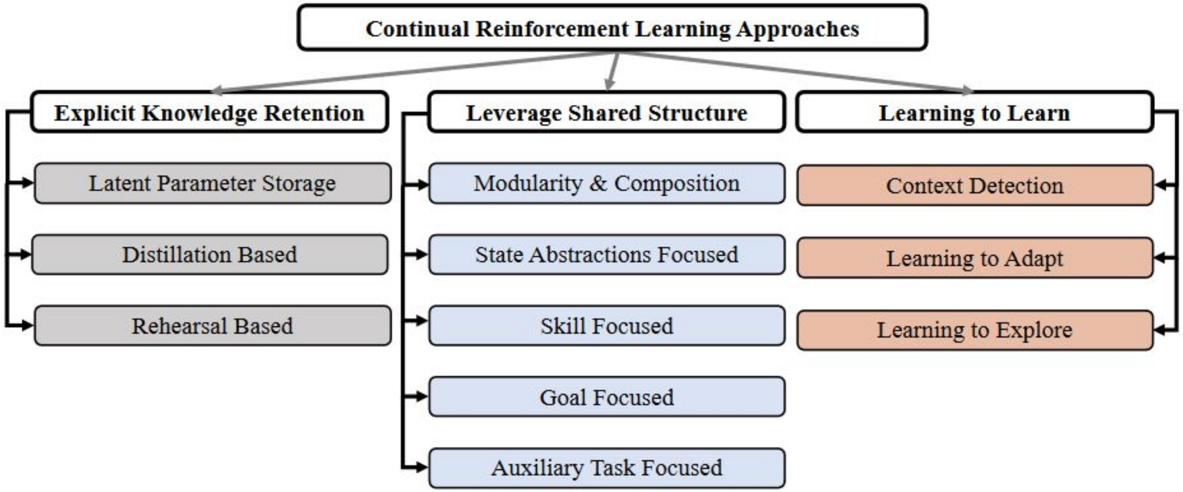
Models referred to as "GPT 3.5"

GPT-3.5 series is a series of models that was trained on a blend of text and code from before Q4 2021. The following models are in the GPT-3.5 series:

1. `code-davinci-002` is a base model, so good for pure code-completion tasks
2. `text-davinci-002` is an InstructGPT model based on `code-davinci-002`
3. `text-davinci-003` is an improvement on `text-davinci-002`
4. `gpt-3.5-turbo-0301` is an improvement on `text-davinci-003`, optimized for chat

But not in RL.

For an RL agent to be useful in real life, it must learn to adapt its knowledge to use in new non-stationary environment.



Towards Continual Reinforcement Learning: A Review and Perspectives. Khetarpal et al., 2022