Episodic Memory in Lifelong Language Learning

NIPS 19

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, Dani Yogatama

DeepMind

Xiachong Feng

Outline

- Author
- Background
- Task
- Model
- Experiment
- Result

Author

Cyprien de Masson d’Autume

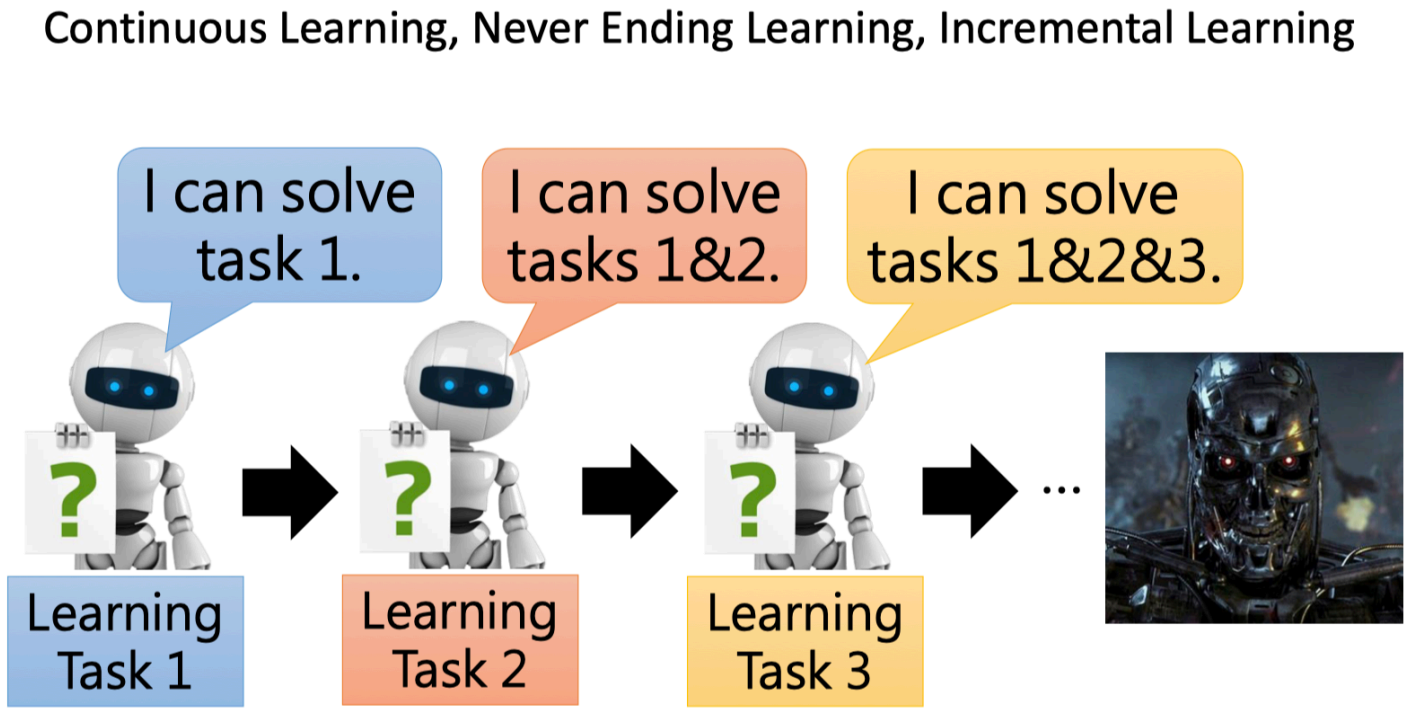
Sebastian Ruder
DeepMind

Lingpeng Kong
（孔令鹏）
DeepMind

Dani Yogatama
DeepMind

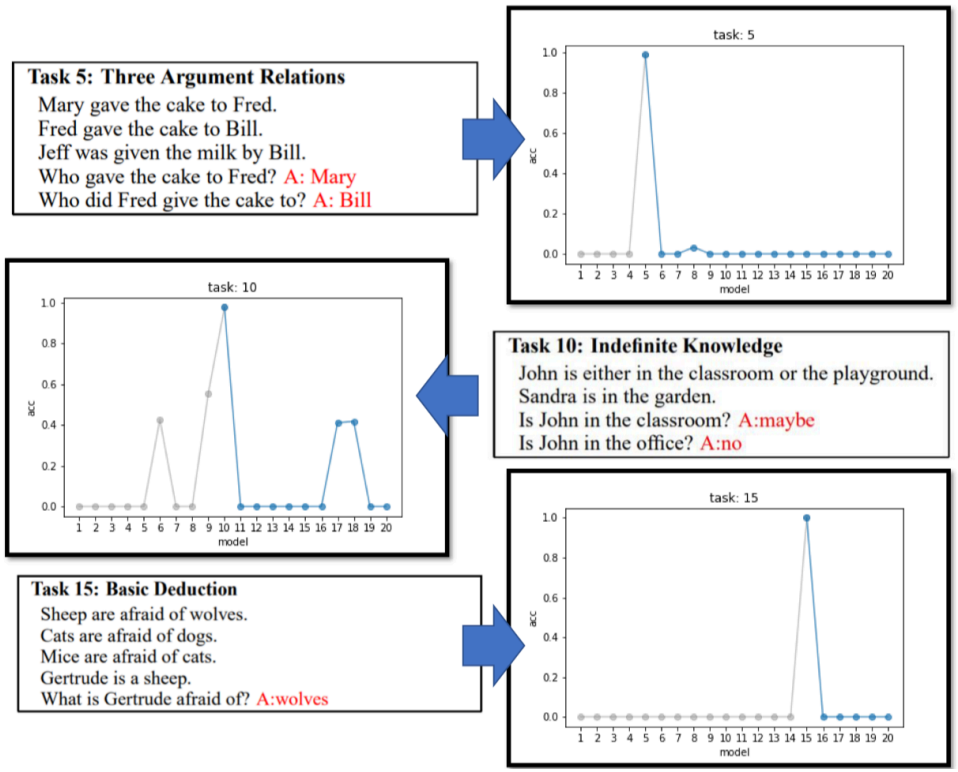
Background

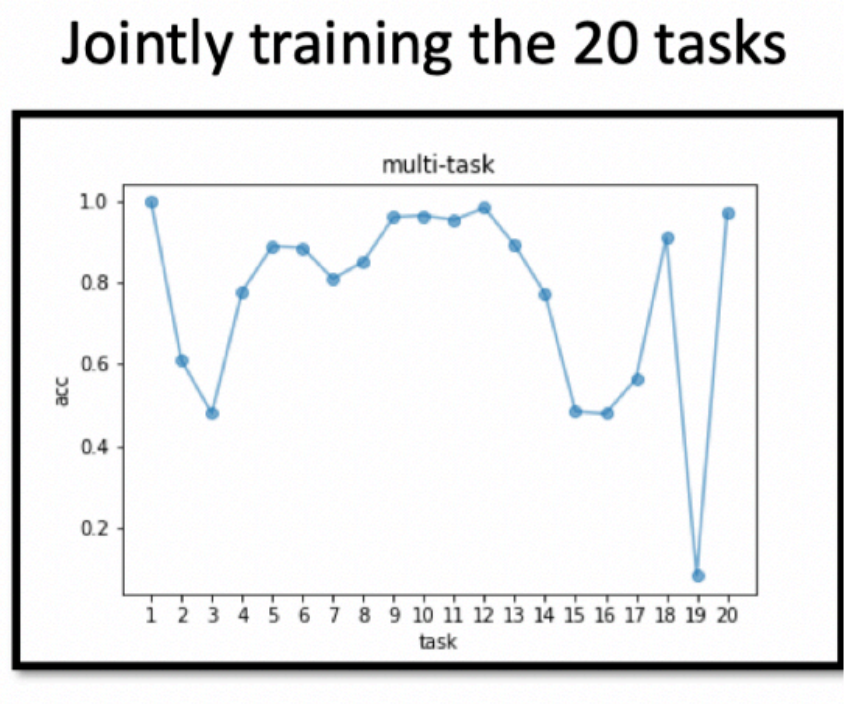
- Life long learning



Background

- Catastrophic Forgetting





Task

- Text classification
- Question answering

(i) Yelp $\rightarrow$ AGNews $\rightarrow$ DBPedia $\rightarrow$ Amazon $\rightarrow$ Yahoo.

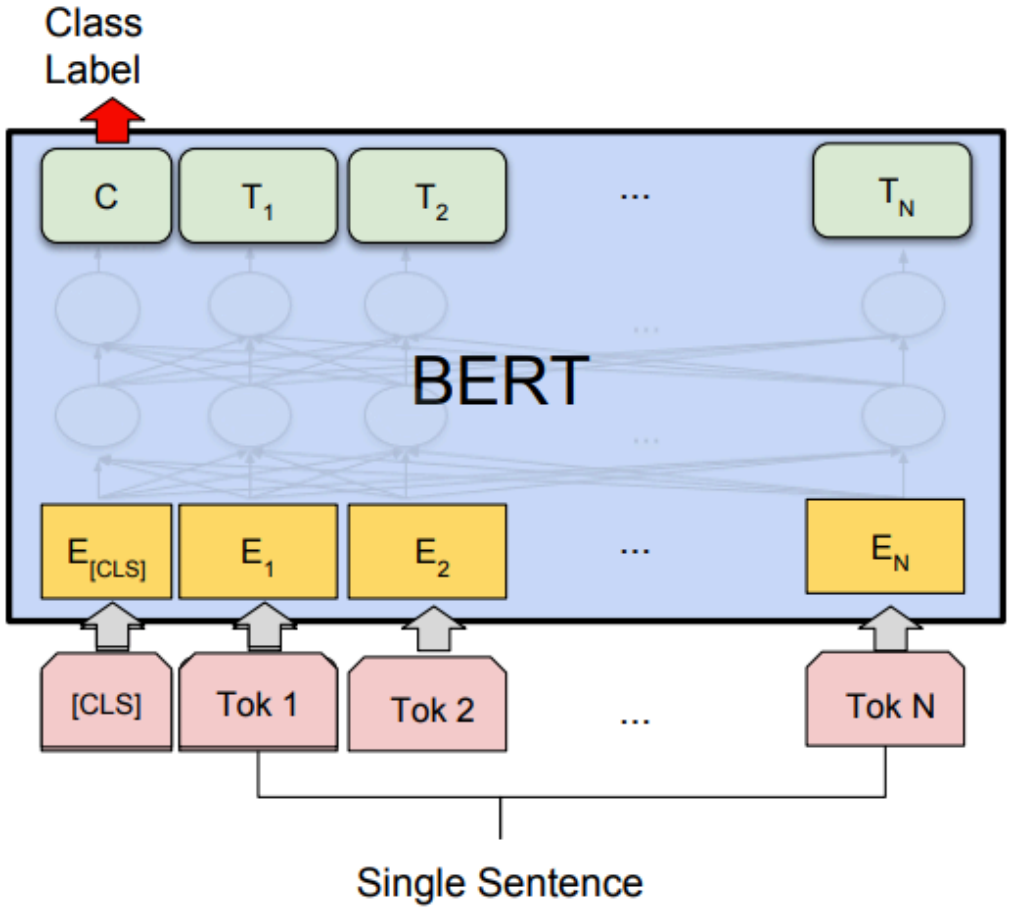
Model

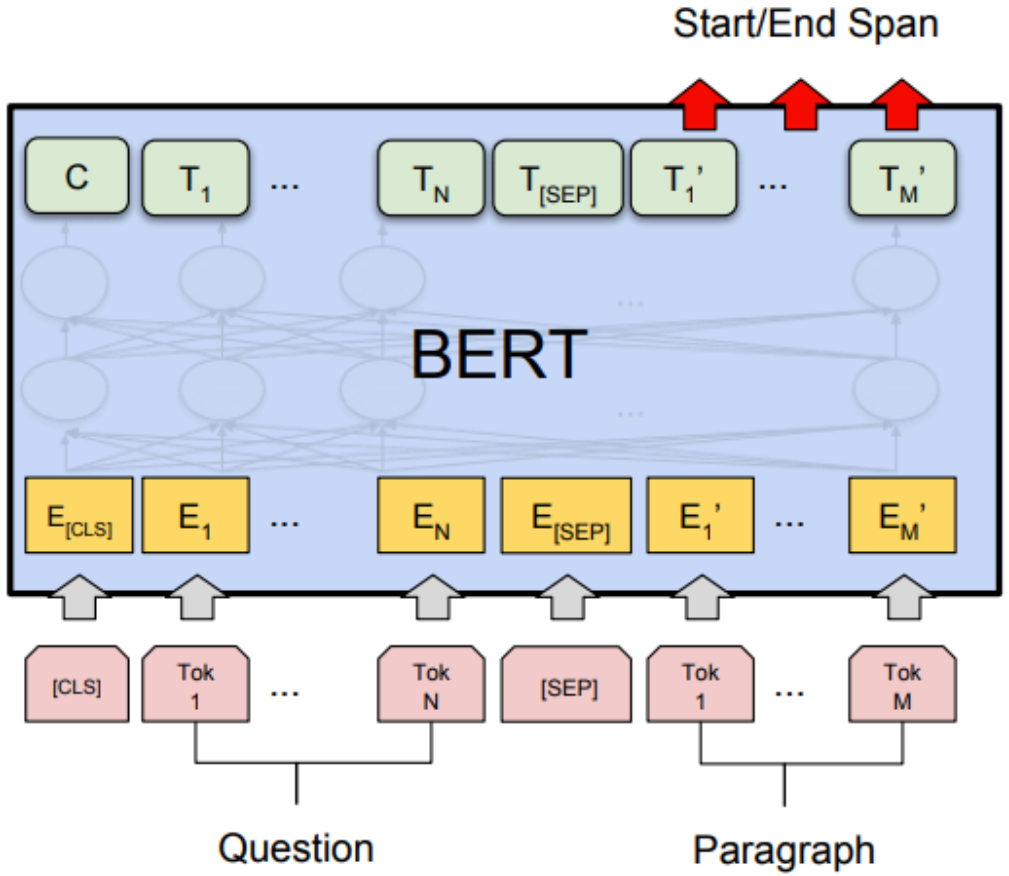
- Example encoder
- Task decoder
- Episodic memory module.

Example encoder && Task decoder

$$p(y_t = c \mid \boldsymbol{x}_t) = \frac{\exp(\mathbf{w}_c^\top \mathbf{x}_{t,0})}{\sum_{y \in \mathcal{Y}} \exp(\mathbf{w}_y^\top \mathbf{x}_{t,0})}$$

$$p(\texttt{start} = x_{t,m}^{\text{context}} \mid \boldsymbol{x}_t) = \frac{\exp(\mathbf{w}_{\text{start}}^\top \mathbf{x}_{t,m}^{\text{context}})}{\sum_{n=0}^{M} \exp(\mathbf{w}_{\text{start}}^\top \mathbf{x}_{t,n}^{\text{context}})}.$$



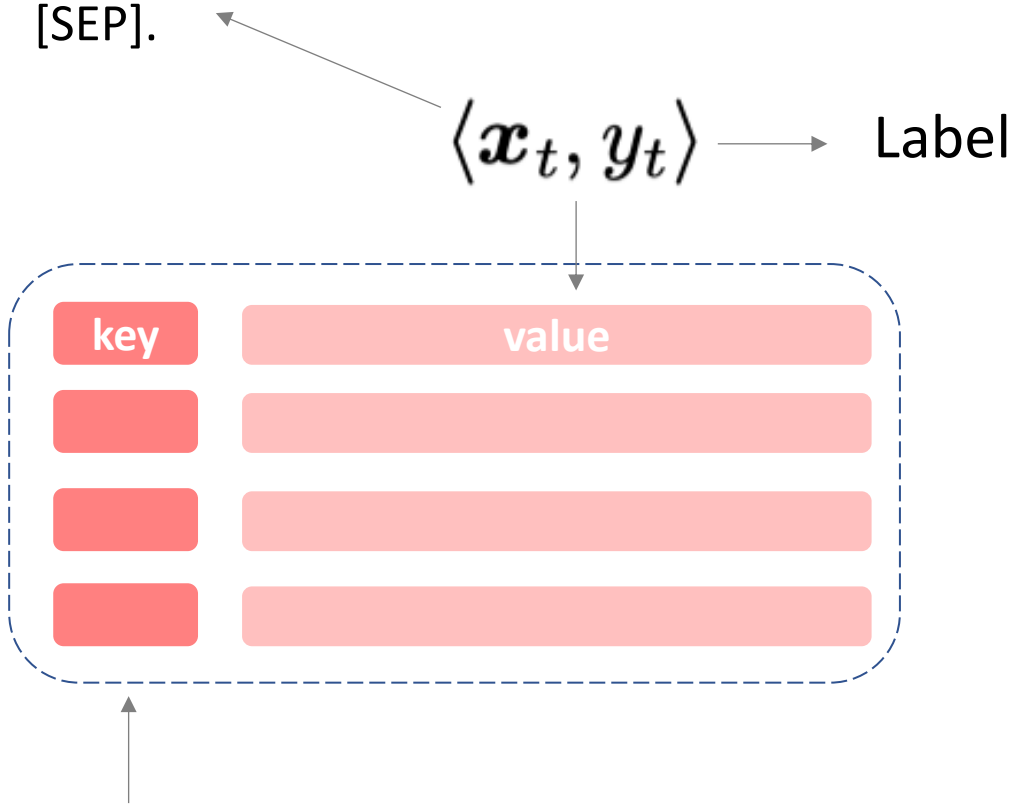


Episodic Memory

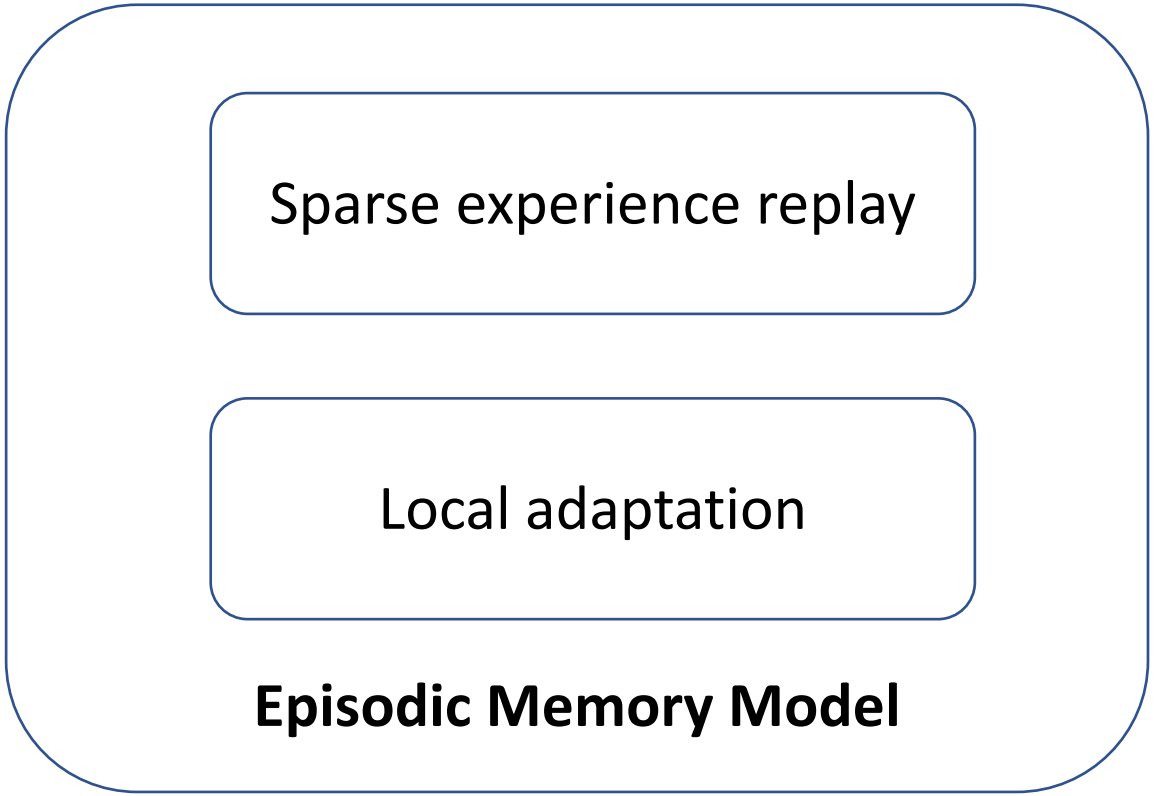
- **key-value memory block**

- Text Classification
  - [CLS]
- Question Answering
  - The first token of question

- Text classification, $\boldsymbol{x_t}$ is a document to be classified
- Question answering, $\boldsymbol{x_t}$ is a concatenation of a context paragraph and a question separated by [SEP].

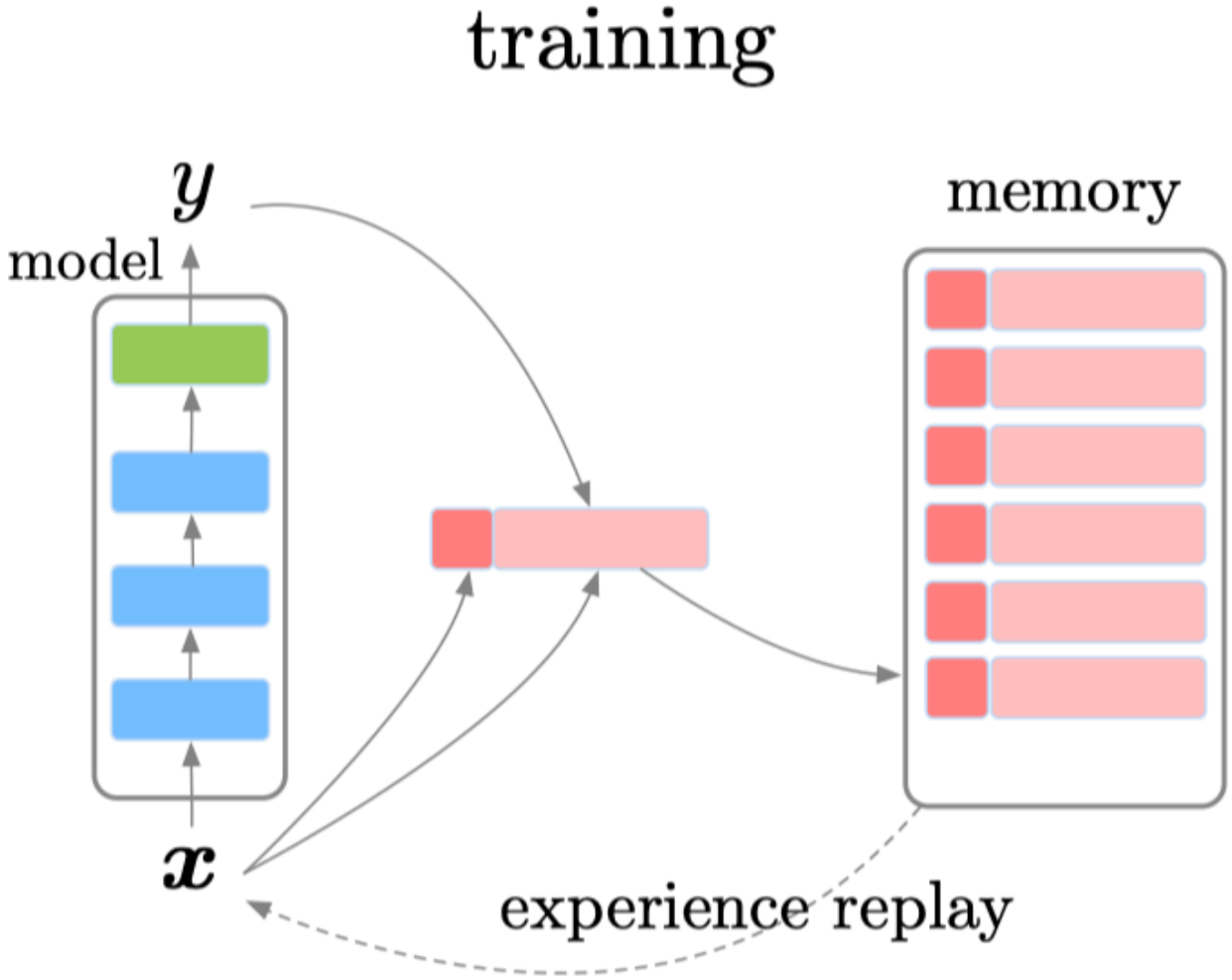


Episodic Memory



Model - Training

- **Write**
  - Based on random write
- **Read** *sparse experience replay*
  - Uniformly random sampling
  - Perform gradient updates based on the retrieved examples
  - **Sparsely :** randomly retrieve 100 examples every 10,000 new examples

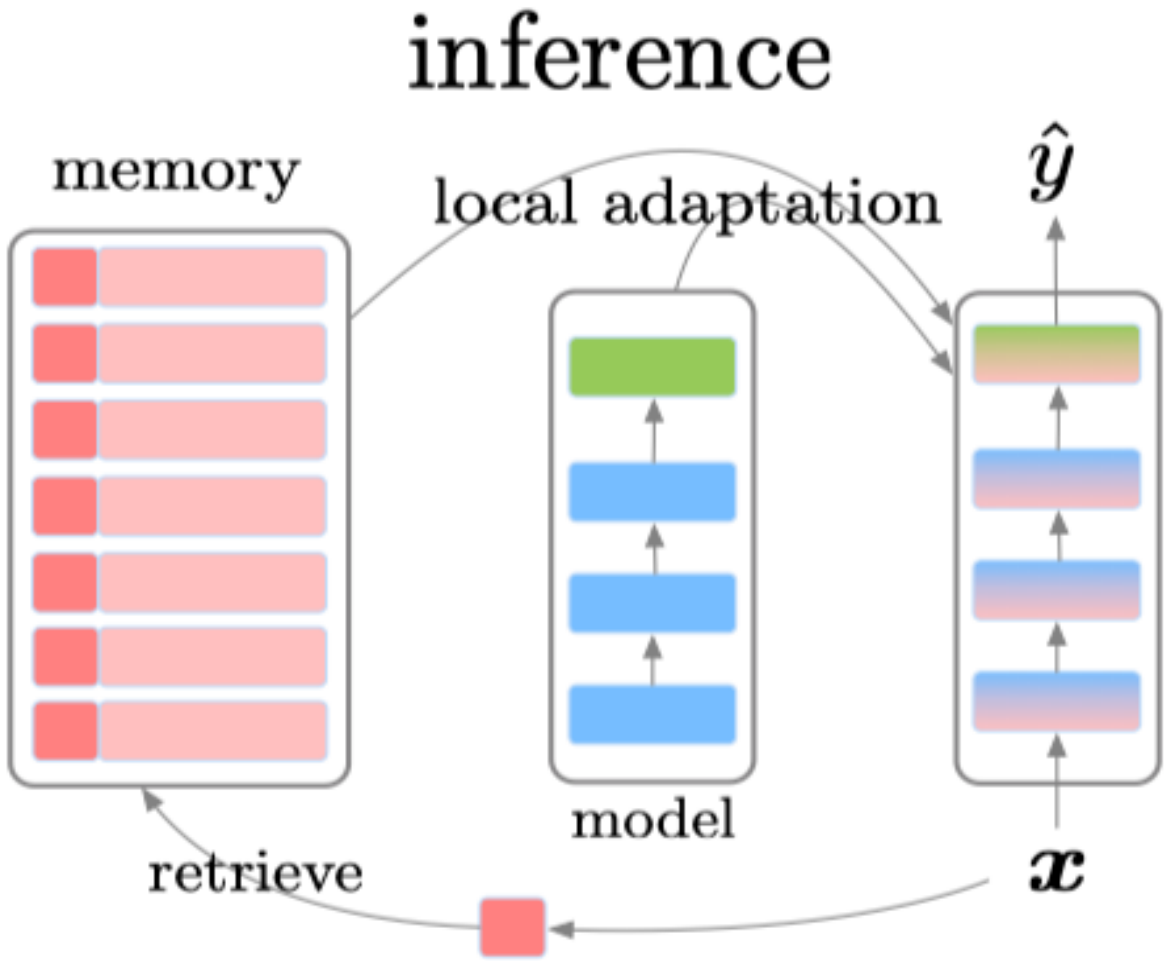


Model - Inference

- Read *local adaptation*
  - Key net→ query vector
  - K-nearest neighbors using the Euclidean distance function

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

$\frac{1}{K}$



Experiments

- **Text classification**
  - News classification (AGNews), sentiment analysis (Yelp, Amazon), Wikipedia article classification (DBPedia), and questions and answers categorization (Yahoo).
  - AGNews (4 classes), Yelp (5 classes), DBPedia (14 classes), Amazon (5 classes), and Yahoo (10 classes) datasets.
  - Yelp and Amazon datasets have similar semantics (product ratings), we merge the classes for these two datasets.
- **Question answering**
  - SQuAD 1.1 ,TriviaQA, QuAC
- Create a **balanced** version all datasets

Results

Text classification

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

QA

(i) QuAC → TrWeb → TrWik → SQuAD.
(ii) SQuAD → TrWik → QuAC → TrWeb.
(iii) TrWeb → TrWik → SQuAD → QuAC.
(iv) TrWik → QuAC → TrWeb → SQuAD.

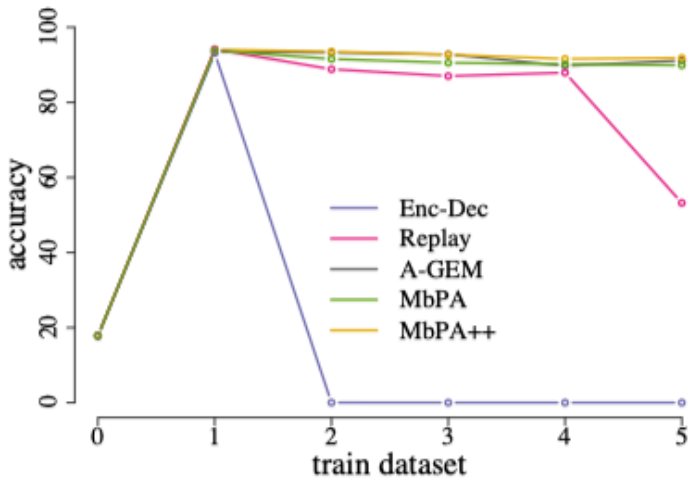
Table 1: Summary of results on text classification (above) and question answering (below) using averaged accuracy and F_1 score respectively (see Appendix A for the dataset orderings).

Order	ENC-DEC	A-GEM	REPLAY	MBPA	MBPA$^{\text{rand}}_{++}$	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
i	57.7	56.1	60.1	60.8	60.0	**62.0**	67.6
ii	55.1	58.4	60.3	60.1	60.0	**62.4**	67.9
iii	41.6	52.4	58.8	58.9	58.8	**61.4**	67.9
iv	58.2	57.9	59.8	61.5	59.8	**62.4**	67.8
QA-avg.	53.1	56.2	57.9	60.3	59.7	**62.4**	67.8

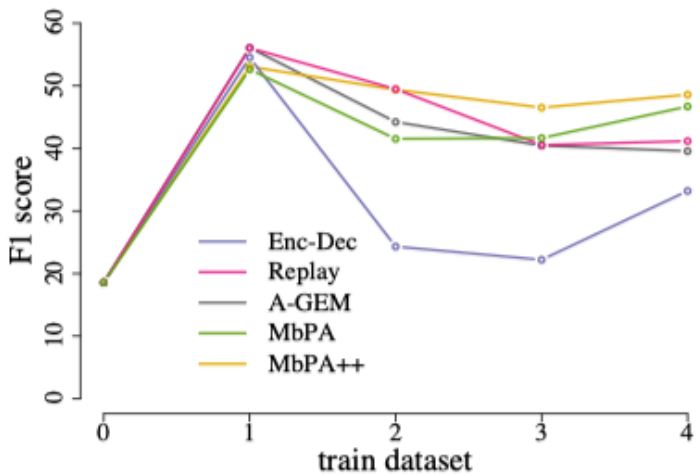
randomly retrieved examples for local adaptation

multitask model

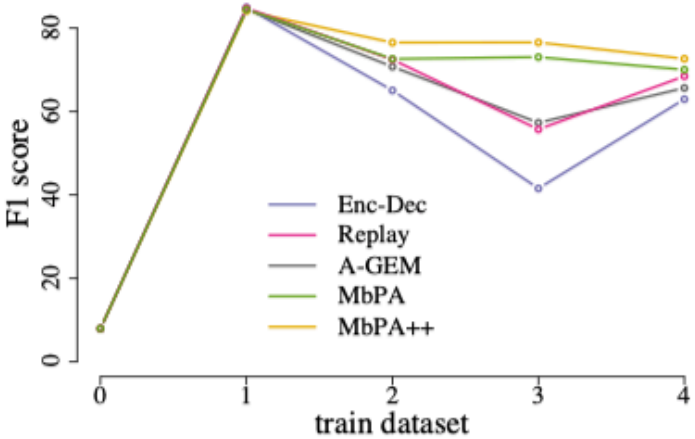
Result



(a) Classification–AGNews



(b) QA–QuAC



(c) QA–SQuAD

Result

Table 2: Results with limited memory capacity.

	10%	50%	100%
class.	67.6	70.3	70.6
QA	61.5	61.6	62.0

store only 50% and 10% of training examples.

Result

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	-	-

Thanks!