SCUOLA NORMALE SUPERIORE

UNIVERSITÀ DI PISA

Continual Learning for Recurrent Neural Networks

Review + Empirical Evaluation

Andrea Cossu, Antonio Carta, Vincenzo Lomonaco, Davide Bacciu

Is sequential data processing important for CL?

Human activity recognition (sensors, videos)

Robot control (temporally-correlated raw data)

Finance (next stock value prediction)

Natural Language Processing (domain shift, translation)

…

Why do you focus on RNNs?

e.g CNNs? Transformers? → different questions

- Variable number of layers → unrolling
- Weight sharing over time steps
- Backpropagation through time (for deep RNNs)

Organized review of RNN in CL

Let’s look at what is **already** here (*shallow* taxonomy)

- **Seminal works** simple studies on synthetic benchmarks
- **NLP** application-specific
- **Bio-inspired / alternative recurrent paradigms** Custom learning algorithms, ad-hoc architectures
- **Deep networks** LSTM, GRU…

The vast majority of works focus on **new** models / strategies

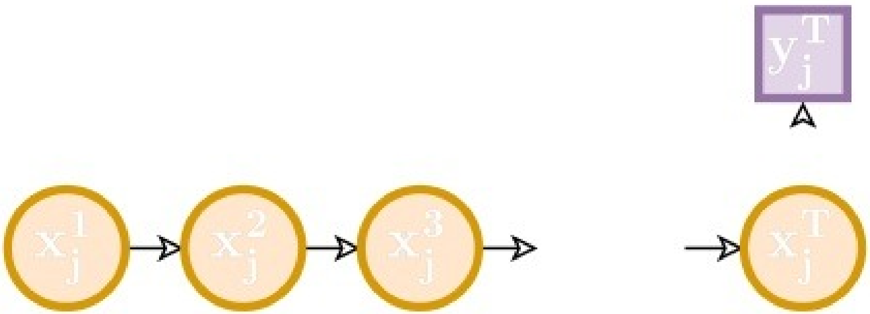
Benchmarks description

Dataset	Application	Scenario
Copy Task [103, 34]	synthetic	MT+NI
Delay/Memory Pro/Anti [32]	synthetic, neuroscience	MT+NI
Seq. Stroke MNIST [103, 34]	stroke classification	SIT+(NI/NC)
Quick, Draw! †	stroke classification	SIT+NC
MNIST-like [27] [26] †	object classification	SIT+(NI/NC)
CORe50 [92]	object recognition	SIT+(NI/NC)
MNLI [10]	domain adaptation	SIT+NI
MDSD [81]	sentiment analysis	SIT+NI
WMT17 [14]	NMT	MT+NC
OpenSubtitles18 [76]	NMT	MT+NC
WIPO COPPA-V2 [63] [107]	NMT	MT+NC
CALM [66]	language modeling	Online
WikiText-2 [118]	language modeling	SIT+NI/NC
Audioset [27, 34]	sound classification	SIT+NC
LibriSpeech, Switchboard [119]	speech recognition	(SIT/MT)+NC
Synthetic Speech Commands †	sound classification	SIT+NC
Acrobot [65]	reinforcement learning	MT+NI

SCUOLA NORMALE SUPERIORE
UNIVERSITÀ DI PISA

Our objective

- Study the behavior of RNNs
  - with popular CL strategies
    not designed for sequential data processing
  - on application-agnostic benchmarks



Experimental evaluation

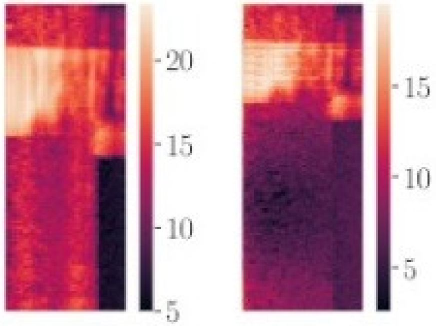
- Class-incremental (no task labels), single-head
- 6 strategies + Naive + Joint Training
  - EWC, MAS, LwF, GEM, A-GEM, Replay (random sampling)
- No architectural strategies
- Grid search protocol with held-out experiences

Benchmarks

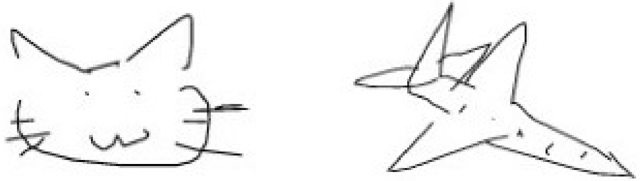
- Split / Permuted MNIST (really?)

  Application-agnostic, “closer” to real-world

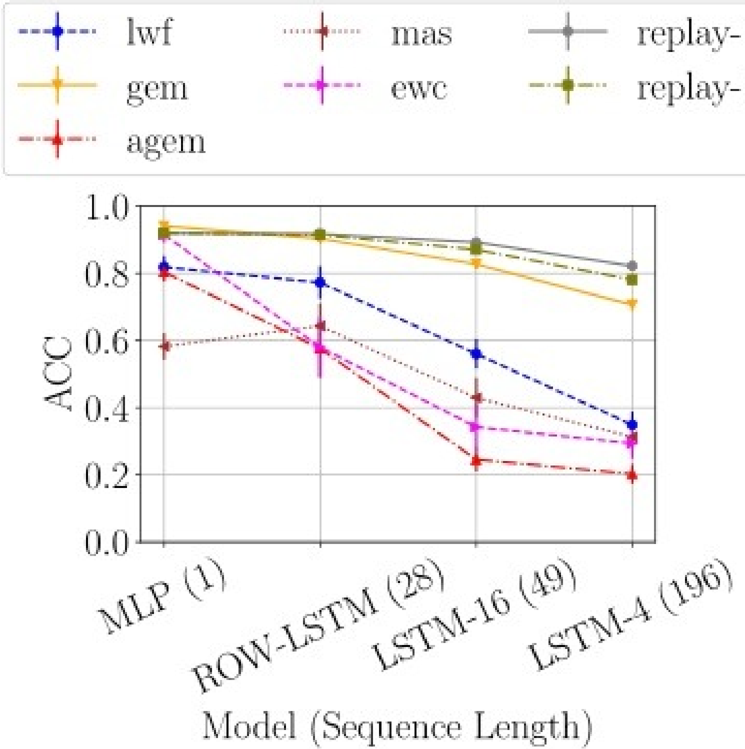
- Synthetic Speech Commands – audio with 101 time steps

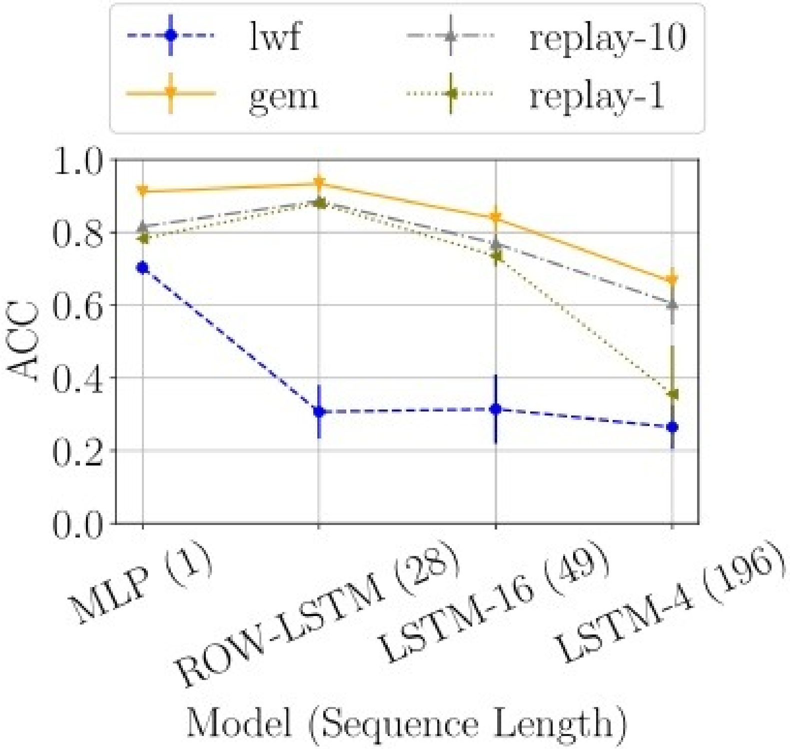


- Quick, Draw! - variable sequence length



Sequence length affects forgetting





SSC	MLP	LSTM
EWC	$0.10_{\pm 0.00}$	$0.10_{\pm 0.00}$
LWF	$0.05_{\pm 0.00}$	$0.12_{\pm 0.01}$
MAS	$0.10_{\pm 0.00}$	$0.10_{\pm 0.00}$
GEM	$0.55_{\pm 0.00}$	$0.53_{\pm 0.01}$
A-GEM	$0.05_{\pm 0.00}$	$0.09_{\pm 0.01}$
REPLAY	$\mathbf{0.81_{\pm 0.03}}$	$\mathbf{0.73_{\pm 0.04}}$
NAIVE	$0.10_{\pm 0.00}$	$0.10_{\pm 0.00}$
Joint Training	$0.93_{\pm 0.00}$	$0.89_{\pm 0.02}$

QD	LSTM
EWC	$0.12_{\pm 0.02}$
LWF	$0.12_{\pm 0.01}$
MAS	$0.10_{\pm 0.00}$
GEM	$0.47_{\pm 0.03}$
A-GEM	$0.10_{\pm 0.00}$
REPLAY	$\mathbf{0.49_{\pm 0.02}}$
NAIVE	$0.10_{\pm 0.00}$
Joint Training	$0.96_{\pm 0.00}$

SCUOLA NORMALE SUPERIORE
Università di Pisa

Impact on more realistic benchmarks

What for the future?

Just scratched the surface

- Adapt existing CL strategies
  - Orthogonal projections seem promising – find a better tradeoff
- Improve recurrent models and learning algorithms
  - BPTT alternatives – local algorithms
- Applications: place something somewhere... and leave it there!

Do you believe RNNs are worth studying in CL?

https://arxiv.org/abs/2103.07492