MC-LSTM: Mass-Conserving LSTM

Pieter-Jan Hoedt & Frederik Kratzert @ DLCT
ELLIS Unit Linz and LIT AI Lab
Institute for Machine Learning
Johannes Kepler University, Linz, Austria

JOHANNES KEPLER UNIVERSITY LINZ

E-mail: {hoedt,kratzert}@ml.jku.at

Contributors

Johannes Kepler University

Google

Pieter-Jan Hoedt

Frederik Kratzert

Daniel Klotz

Christina Halmich

Markus Holzleitner

Grey Nearing

Sepp Hochreiter

Günter Klambauer

Pre-print: https://arxiv.org/abs/2101.05186

code: https://github.com/ml-jku/mc-lstm

Whereabouts

- @ml_hoedt
- @fkratzert
- @ido87
- @tinahalmich
- @GreyNearing





Linz, AUSTRIA

JKU Institute for Machine Learning

JKU LINZ INSTITUTE OF TECHNOLOGY

ellis unit | LINZ

JKU

Institute blog: https://ml-jku.github.io/

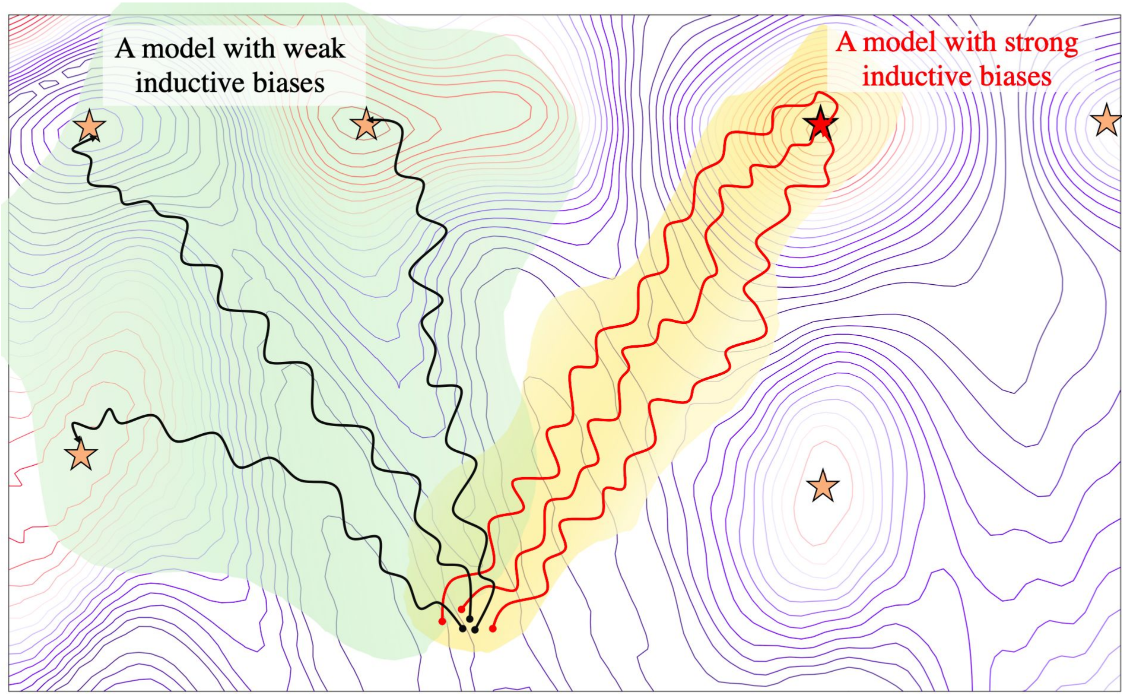
OUTLINE

- Motivation
- Model
- Experiments

OUTLINE

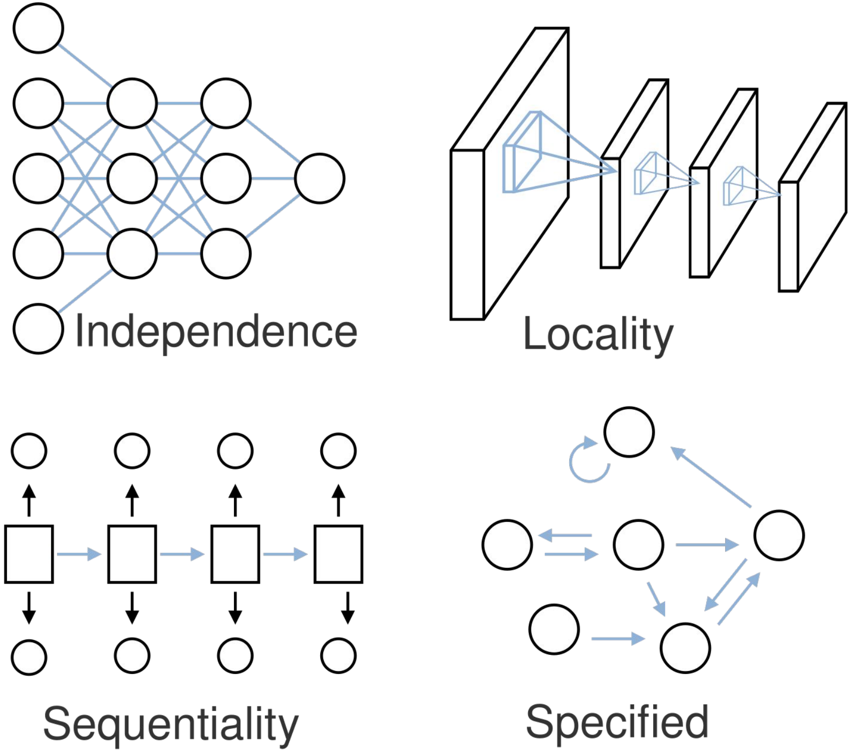
- Motivation
- Model
- Experiments

INDUCTIVE BIAS



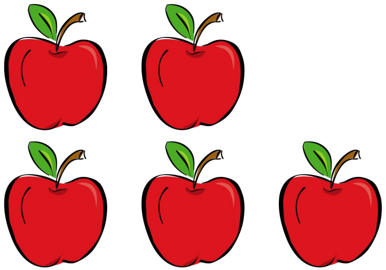
Abnar, Samira (2020), https://samiraabnar.github.io/articles/2020-05/indist

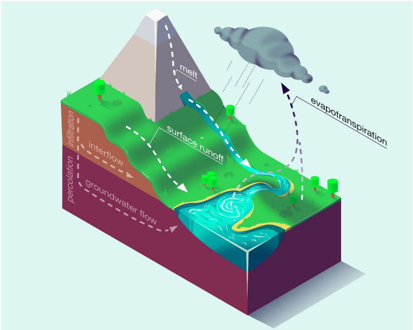
INDUCTIVE BIAS

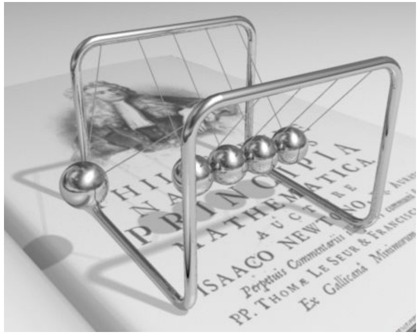


Adapted from Finlayson, Sam (2020). https://sgfin.github.io/2020/06/22/Induction-Intro

CONSERVATION LAWS







JKU

adapted from https://en.wikipedia.org/wiki/Addition
https://en.wikipedia.org/wiki/Newton%27s_cradle
https://www.iarai.ac.at

MASS CONSERVATION

Theorem 1 (Conservation property). *Let* $m_c^\tau = \sum_{k=1}^{K} c_k^\tau$ *be the mass contained in the system and* $m_h^\tau = \sum_{k=1}^{K} h_k^\tau$ *be the mass efflux, or, respectively, the* accumulated mass *in the MC-LSTM storage and the outputs at time* τ. *At any timestep* τ, *we have:*

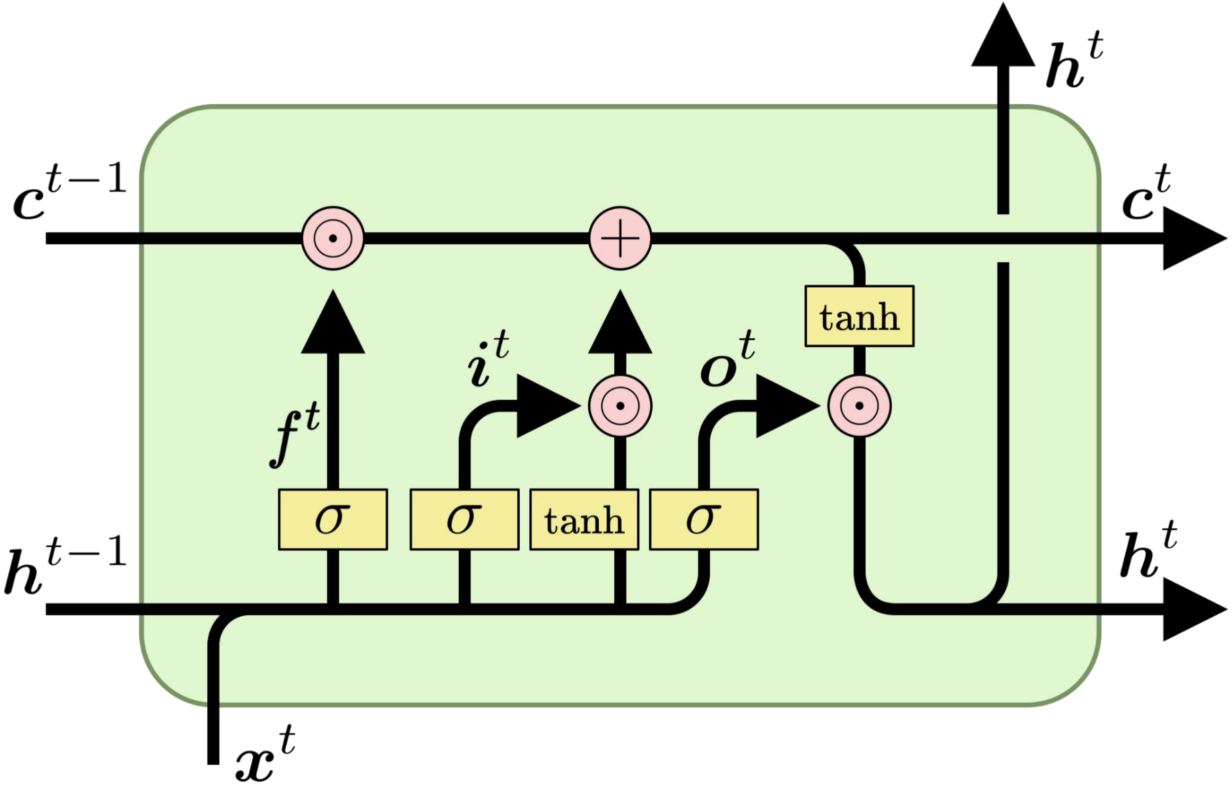
$$m_c^\tau = m_c^0 + \sum_{t=1}^{\tau} x^t - \sum_{t=1}^{\tau} m_h^t. \tag{9}$$

That is, the change of mass in the memory cells is the difference between the input and output mass, accumulated over time.

OUTLINE

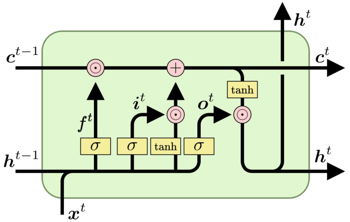
- Motivation
- Model
- Experiments

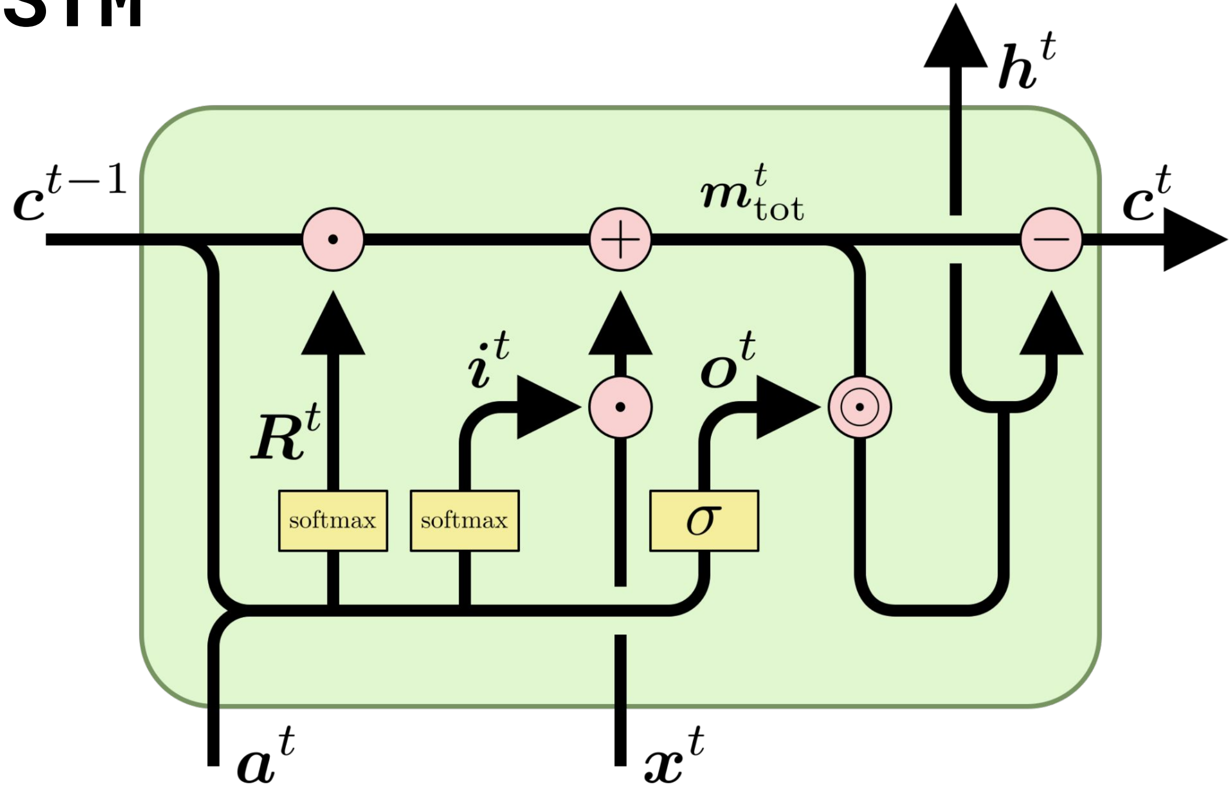
LSTM



Olah, Chris (2015). Understanding LSTM networks.

MC-LSTM





Inspired by Olah, Chris (2015). Understanding LSTM networks.

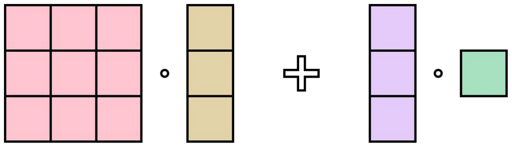
MC-LSTM

Total mass	$\boldsymbol{m}_{\text{tot}}^t = \boldsymbol{R}^t \cdot \boldsymbol{c}^{t-1} + \boldsymbol{i}^t \cdot x^t$
State mass	$\boldsymbol{c}^t = (\mathbf{1} - \boldsymbol{o}^t) \odot \boldsymbol{m}_{\text{tot}}^t$
Output mass	$\boldsymbol{h}^t = \boldsymbol{o}^t \odot \boldsymbol{m}_{\text{tot}}^t.$

Cell State
Mass Input
Auxiliary Input
Parameter

MC-LSTM

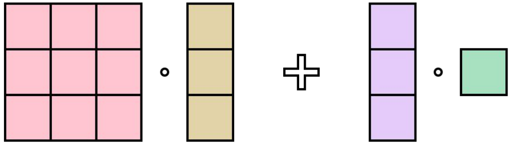
Total mass



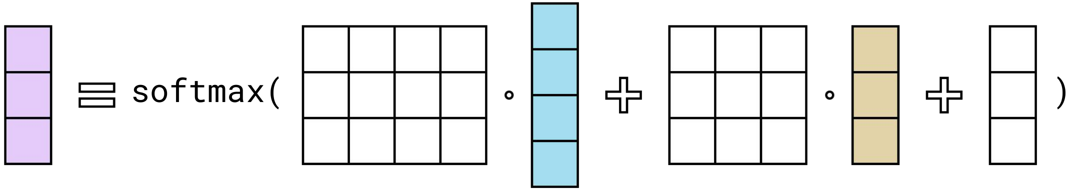
Cell State
Mass Input
Auxiliary Input
Parameter

MC-LSTM

Total mass



Input gate



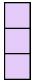
Redistribution
(static)

MC-LSTM

Total mass $\boldsymbol{m}^t_{\text{tot}} = \boldsymbol{R}^t \cdot \boldsymbol{c}^{t-1} + \boldsymbol{i}^t \cdot x^t$

State mass $\boldsymbol{c}^t = (\mathbf{1} - \boldsymbol{o}^t) \odot \boldsymbol{m}^t_{\text{tot}}$

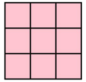
Output mass $\boldsymbol{h}^t = \boldsymbol{o}^t \odot \boldsymbol{m}^t_{\text{tot}}.$

Input gate 

$$\boldsymbol{i}^t = \operatorname{softmax}\left(\boldsymbol{W}_{\text{i}} \cdot \boldsymbol{a}^t + \boldsymbol{U}_{\text{i}} \cdot \frac{\boldsymbol{c}^{t-1}}{\|\boldsymbol{c}^{t-1}\|_1} + \boldsymbol{b}_{\text{i}}\right)$$

Output gate

$$\boldsymbol{o}^t = \sigma\left(\boldsymbol{W}_{\text{o}} \cdot \boldsymbol{a}^t + \boldsymbol{U}_{\text{o}} \cdot \frac{\boldsymbol{c}^{t-1}}{\|\boldsymbol{c}^{t-1}\|_1} + \boldsymbol{b}_{\text{o}}\right)$$

Redistribution (static) 

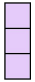
$$\boldsymbol{R}^t = \operatorname{softmax}(\boldsymbol{B}_{\text{r}}),$$

MC-LSTM

Total mass $$\boldsymbol{m}_{\text{tot}}^t = \boldsymbol{R}^t \cdot \boldsymbol{c}^{t-1} + \boldsymbol{i}^t \cdot x^t$$

State mass $$\boldsymbol{c}^t = (\mathbf{1} - \boldsymbol{o}^t) \odot \boldsymbol{m}_{\text{tot}}^t$$

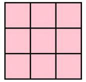
Output mass $$\boldsymbol{h}^t = \boldsymbol{o}^t \odot \boldsymbol{m}_{\text{tot}}^t.$$

Input gate 

$$\boldsymbol{i}^t = \operatorname{softmax}(\boldsymbol{W}_{\text{i}} \cdot \boldsymbol{a}^t + \boldsymbol{U}_{\text{i}} \cdot \frac{\boldsymbol{c}^{t-1}}{\|\boldsymbol{c}^{t-1}\|_1} + \boldsymbol{b}_{\text{i}})$$

Output gate

$$\boldsymbol{o}^t = \sigma(\boldsymbol{W}_{\text{o}} \cdot \boldsymbol{a}^t + \boldsymbol{U}_{\text{o}} \cdot \frac{\boldsymbol{c}^{t-1}}{\|\boldsymbol{c}^{t-1}\|_1} + \boldsymbol{b}_{\text{o}})$$

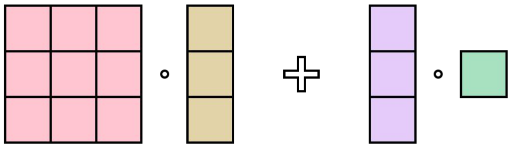
Redistribution (dynamic) 

$$\boldsymbol{R}^t = \operatorname{softmax}\left(\mathbf{W}_{\text{r}} \cdot \boldsymbol{a}^t + \mathbf{U}_{\text{r}} \cdot \frac{\boldsymbol{c}^{t-1}}{\|\boldsymbol{c}^{t-1}\|_1} + \boldsymbol{B}_{\text{r}}\right)$$

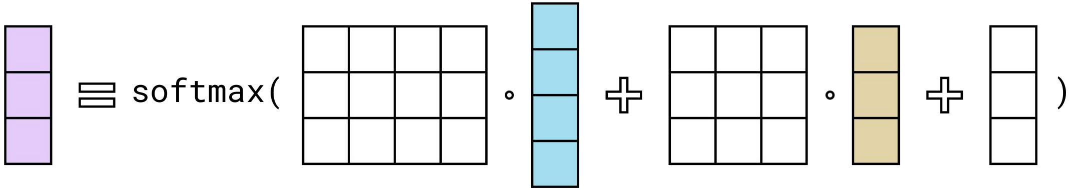
Cell State
Mass Input
Auxiliary Input
Parameter

MC-LSTM

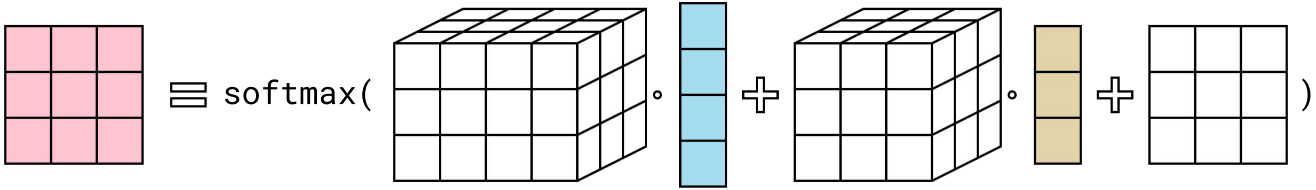
Total mass



Input gate



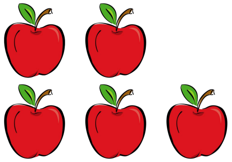
Redistribution
(dynamic)



OUTLINE

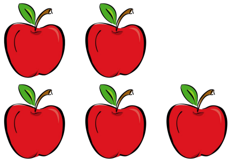
- Motivation
- Model
- Experiments

MC-LSTM ON ARITHMETIC



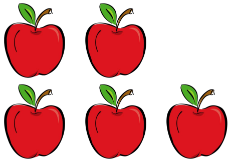
reference: 0.3, 0.4 0.2, 0.0, 0.1, 0.4, 0.3 0.1, 0.5, 0.2

MC-LSTM ON ARITHMETIC



```
reference: 0.3, 0.4 0.2, 0.0, 0.1, 0.4, 0.3 0.1, 0.5, 0.2

length:    0.3, 0.4 0.2, 0.0, 0.1, 0.4, 0.3 0.1, 0.5, 0.2, 0.3, ...

range:     3.4, 4.4 2.2, 0.1, 1.2, 4.4, 3.0 1.5, 5.1, 2.8

count:     0.3, 0.4 0.2, 0.0, 0.1 0.4, 0.3 0.1, 0.5 0.2

combo:     3.4, 4.4 2.2, 0.1, 1.2 4.4, 3.0 1.5, 5.1 2.8, 3.7, ...
```

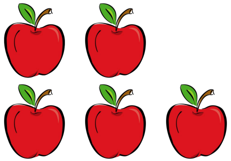
MC-LSTM ON ARITHMETIC



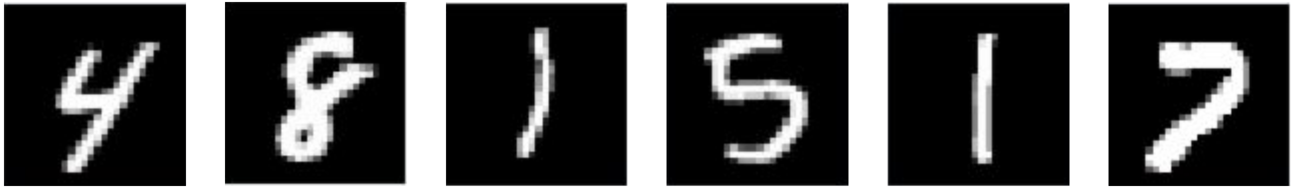
	reference[a]	seq length[b]	input range[c]	count[d]	combo[e]	NaN[f]
MC-LSTM	**0.004** ± 0.003	**0.009** ± 0.004	**0.8** ± 0.5	**0.6** ± 0.4	**4.0** ± 2.5	0
LSTM	0.008 ± 0.003	0.727 ± 0.169	21.4 ± 0.6	9.5 ± 0.6	54.6 ± 1.0	0
NALU	0.060 ± 0.008	0.059 ± 0.009	25.3 ± 0.2	7.4 ± 0.1	63.7 ± 0.6	93
NAU	0.248 ± 0.019	0.252 ± 0.020	28.3 ± 0.5	9.1 ± 0.2	68.5 ± 0.8	24

[a] training regime: summing 2 out of 100 numbers between 0 and 0.5.
[b] longer sequence lengths: summing 2 out of 1 000 numbers between 0 and 0.5.
[c] more *mass* in the input: summing 2 out of 100 numbers between 0 and 5.0.
[d] higher number of summands: summing 20 out of 100 numbers between 0 and 0.5.
[e] combination of previous scenarios: summing 10 out of 500 numbers between 0 and 2.5.
[f] Number of runs that did not converge.

MC-LSTM ON ARITHMETIC

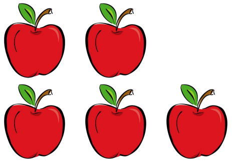


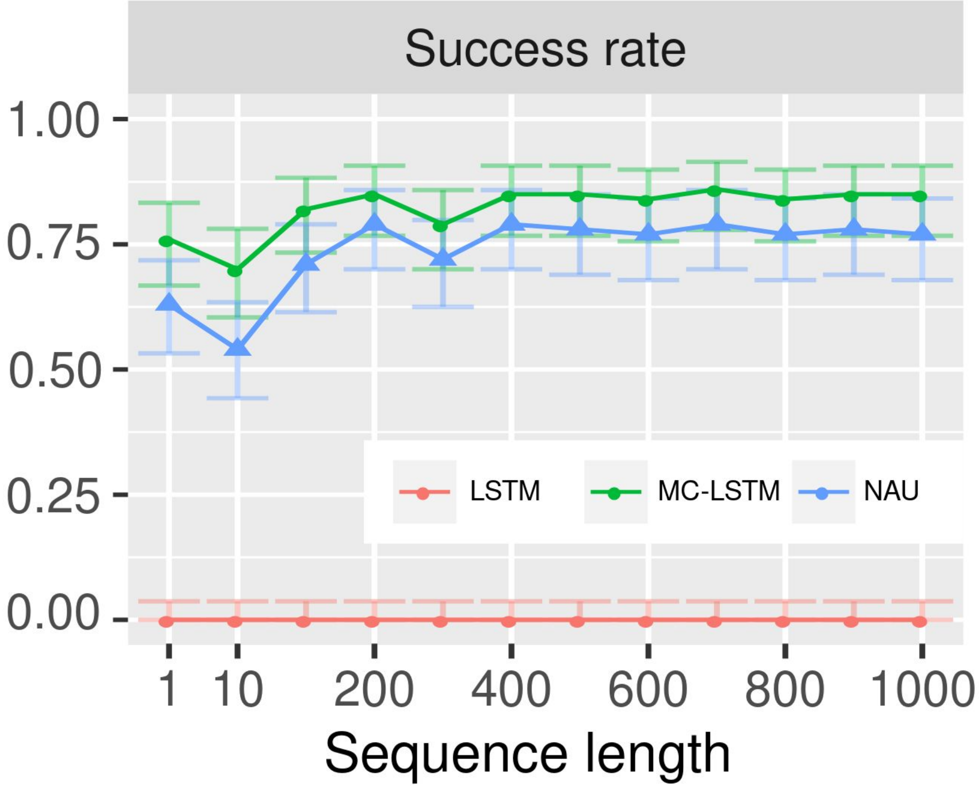
Input:



Output: 26 = 4 + 8 + 1 + 5 + 1 + 7

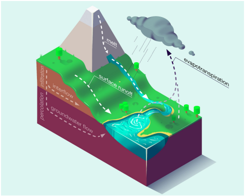
MC-LSTM ON ARITHMETIC

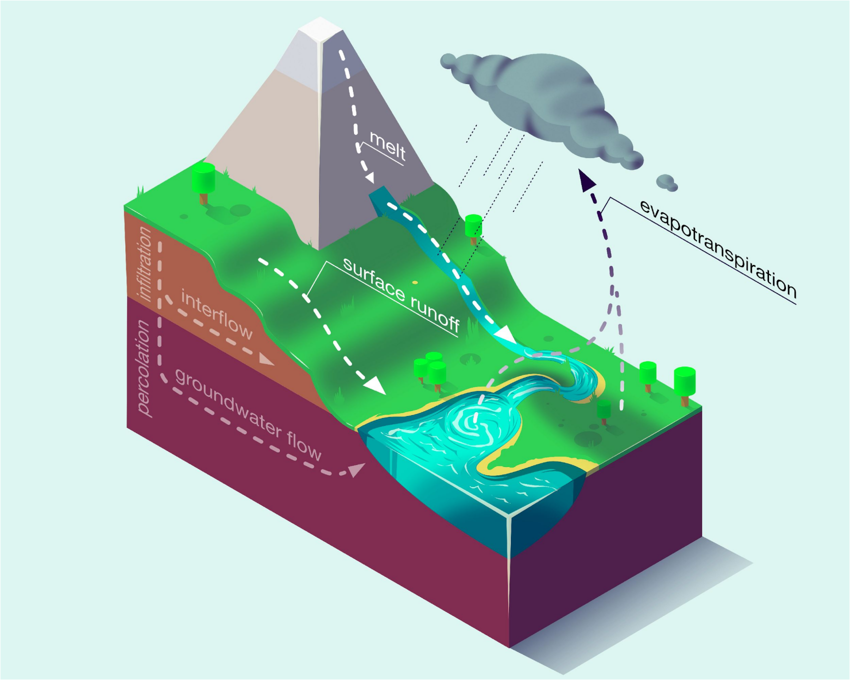




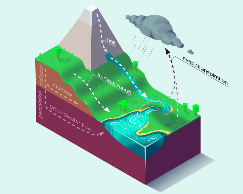
JKU

MC-LSTM ON HYDROLOGY





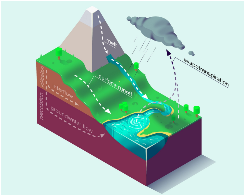
MC-LSTM ON HYDROLOGY

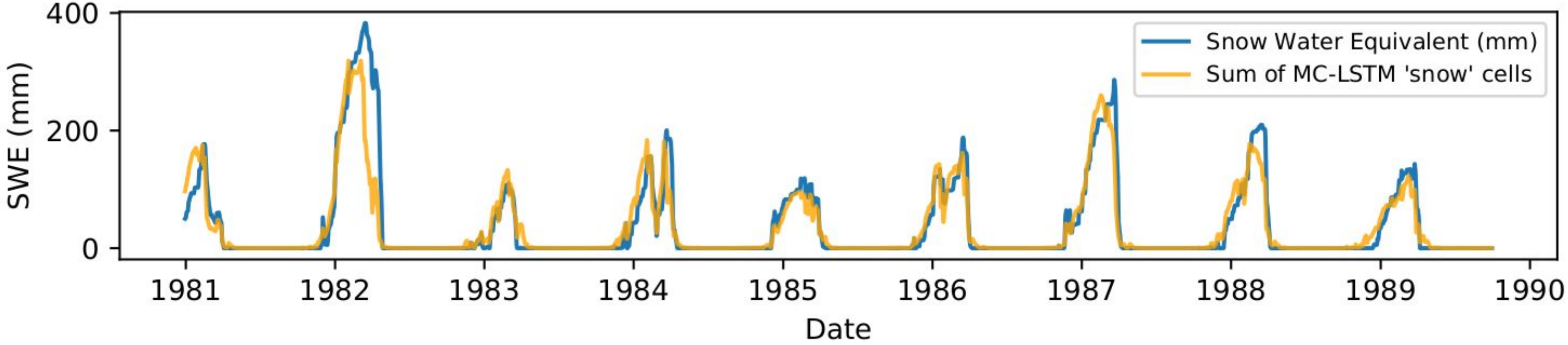


Model	MC[a]	FHV[b]	NSE[c]
MC-LSTM	✓	$\mathbf{-14.7}_{-23.4}^{-7.0}$	$\mathit{0.744}_{0.641}^{0.814}$
LSTM	✗	$\mathit{-15.7}_{-23.8}^{-8.6}$	$\mathbf{0.763}_{0.676}^{0.835}$
mHM	✓	$-18.6_{-27.7}^{-9.5}$	$0.666_{0.588}^{0.730}$
...	...	...	...
HBVub	✓	$-18.5_{-27.8}^{-8.5}$	$0.676_{0.578}^{0.749}$

[a]: *Mass conservation (MC).*
[b]: *Top 2% peak flow bias:* $(-\infty, \infty)$, *values closer to zero are desirable.*
[c]: *Nash-Sutcliffe Efficiency:* $(-\infty, 1]$, *values closer to one are desirable.*

MC-LSTM ON HYDROLOGY





TL;DR: MC-LSTM

- LSTM + inductive bias
- stochastic matrices for conservation
- Better generalisation
- Cell states easier to interpret

Pre-print: https://arxiv.org/abs/2101.05186
code: https://github.com/ml-jku/mc-lstm