SCC5809 – Redes Neurais

Avaliação Empírica de Unidades Recorrentes na Modelagem de Sequências

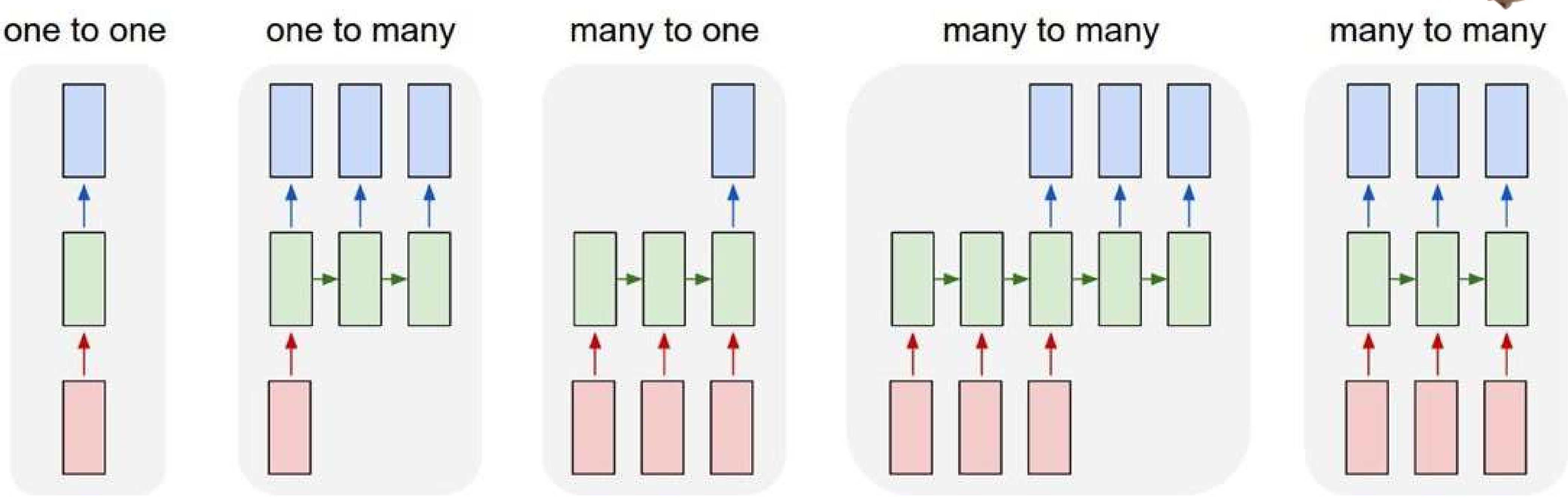
Leonardo Leite de Melo

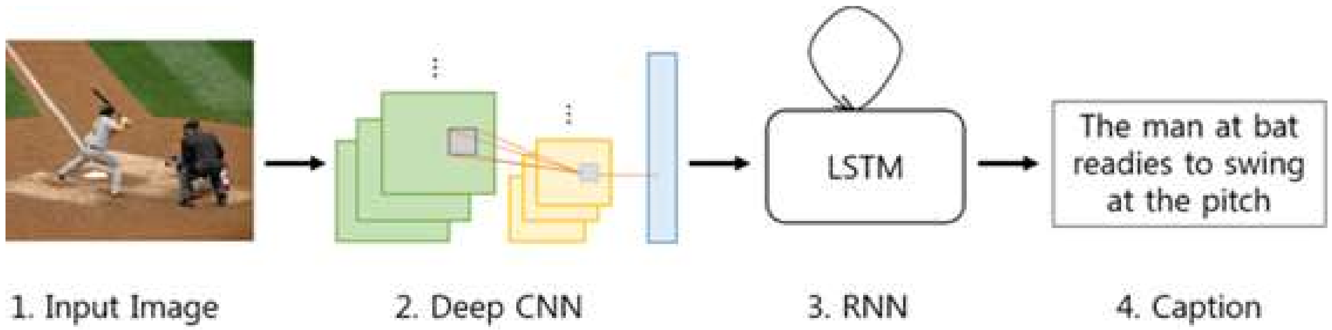
Wagner W. Ávila Bombardelli

04 de outubro
2018

SUMÁRIO

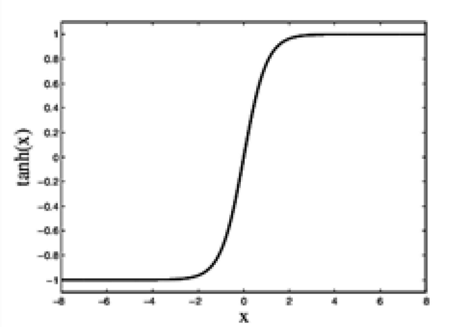
INTRODUÇÃO


one to one
one to many
many to one
many to many
many to many


LSTM
The man at bat readies to swing at the pitch
1. Input Image
2. Deep CNN
3. RNN
4. Caption


SIRI

RESUMO DO ARTIGO



LSTM | GRU

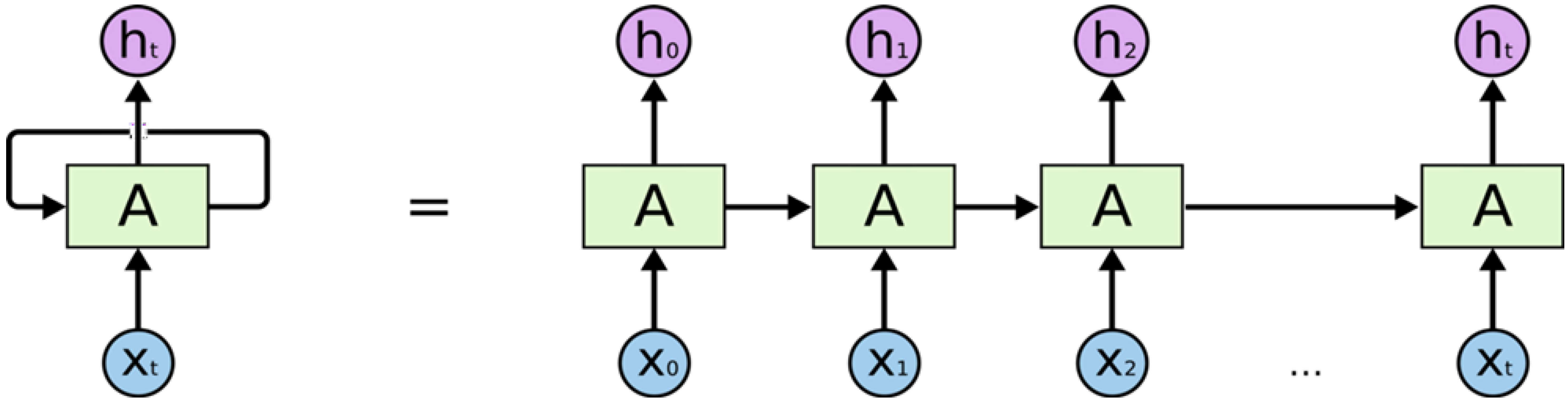
On the Properties of Neural Machine Translation: Encoder–Decoder Approaches

Kyunghyun Cho, Bart van Merriënboer — Université de Montréal
Dzmitry Bahdanau* — Jacobs University, Germany

Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho — Université de Montréal
Yoshua Bengio — Université de Montréal, CIFAR Senior Fellow

CONTEXTUALIZAÇÃO



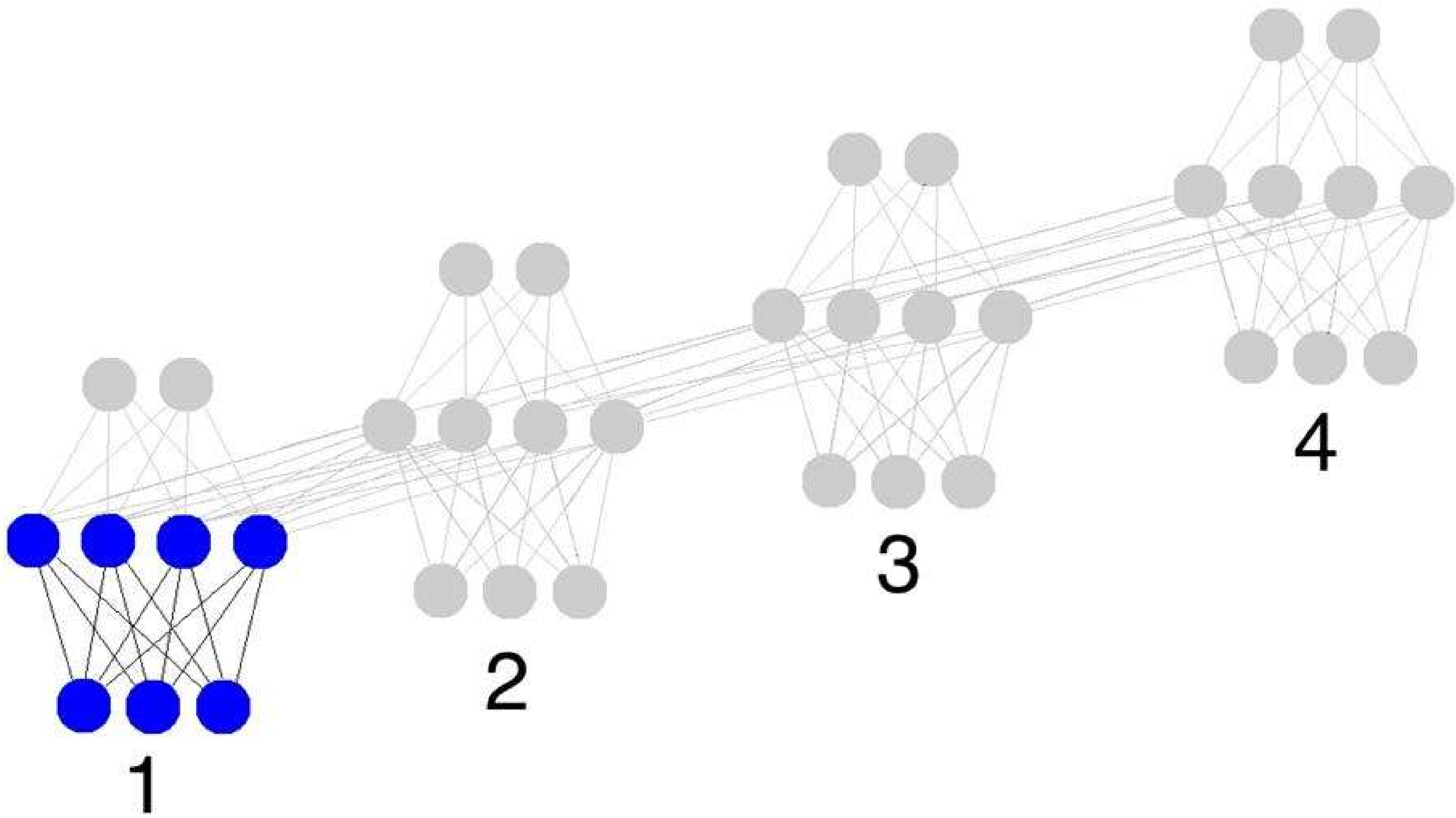
$$\mathbf{h}_t = \begin{cases} 0, & t = 0 \\ \phi\left(\mathbf{h}_{t-1}, \mathbf{x}_t\right), & \text{otherwise} \end{cases}$$

$$\mathbf{h}_t = g\left(W\mathbf{x}_t + U\mathbf{h}_{t-1}\right)$$

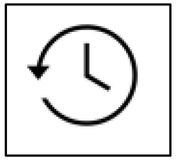
$$p(x_1, \ldots, x_T) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2) \cdots p(x_T \mid x_1, \ldots, x_{T-1})$$

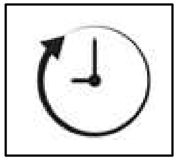
$$p(x_t \mid x_1, \ldots, x_{t-1}) = g(h_t)$$

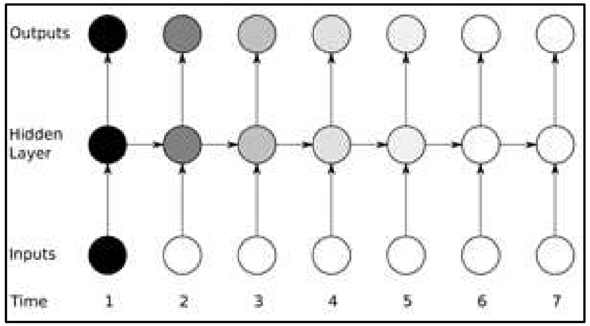
CONTEXTUALIZAÇÃO

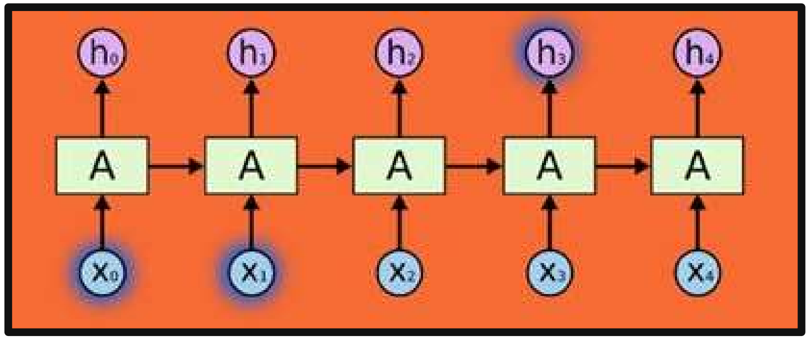


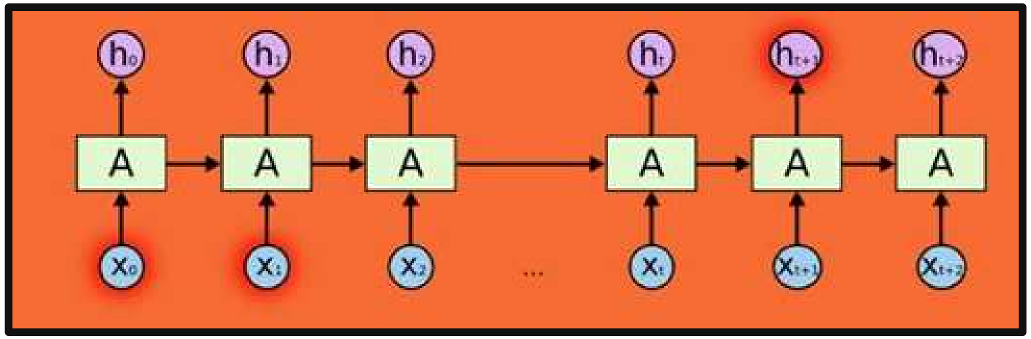
VANISHING GRADIENT PROBLEM

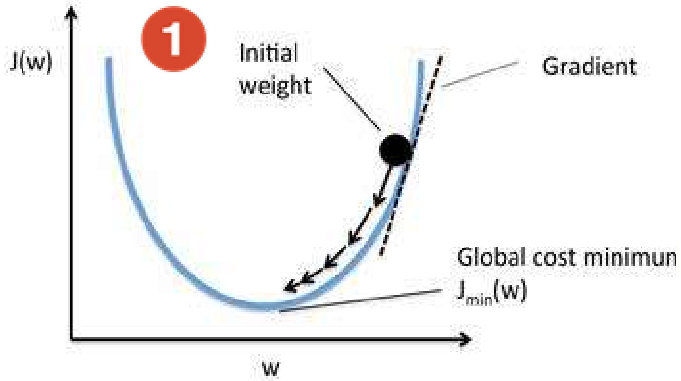


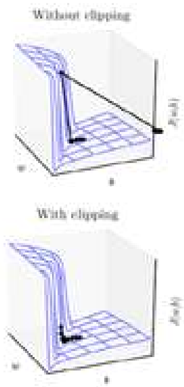


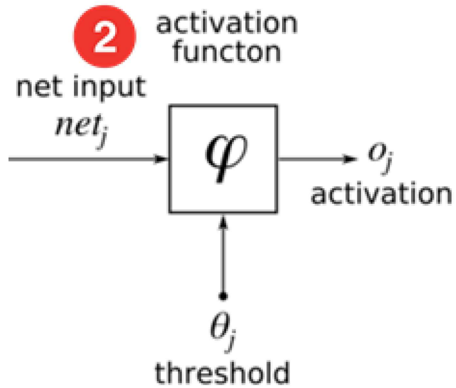




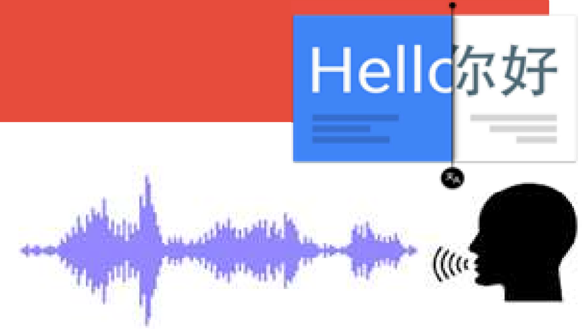


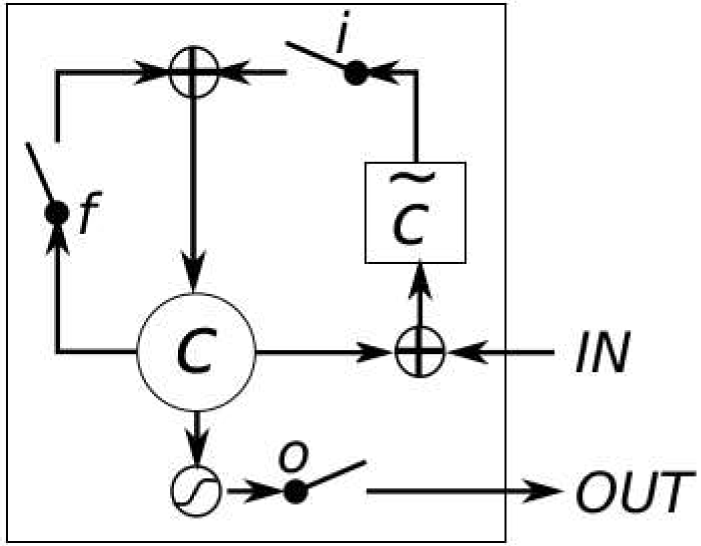






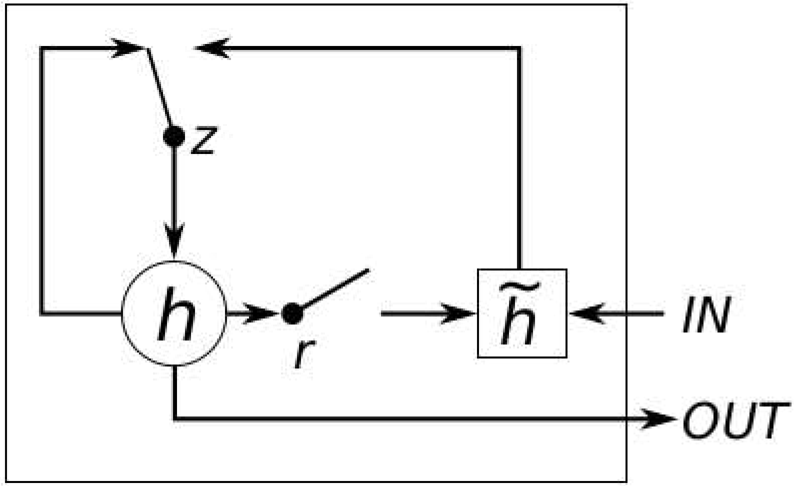
LSTM x GRU





(a) Long Short-Term Memory

[Hochreiter and Schmidhuber, 1997]



(b) Gated Recurrent Unit

[Cho et al., 2014]

LSTM was the winner of the ICDAR (International Conference on Document Analysis and Recognition) 2009 - Handwriting Competition for the best known results in h

LSTM

$$h_t^j = o_t^j \tanh\left(c_t^j\right)$$

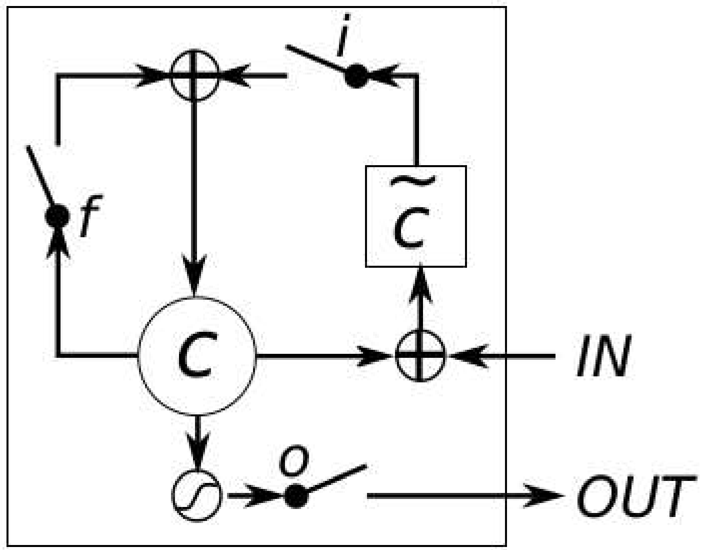
$$o_t^j = \sigma\left(W_o\mathbf{x}_t + U_o\mathbf{h}_{t-1} + V_o\mathbf{c}_t\right)^j$$

$$c_t^j = f_t^j c_{t-1}^j + i_t^j \tilde{c}_t^j$$

$$\tilde{c}_t^j = \tanh\left(W_c\mathbf{x}_t + U_c\mathbf{h}_{t-1}\right)^j$$

$$f_t^j = \sigma\left(W_f\mathbf{x}_t + U_f\mathbf{h}_{t-1} + V_f\mathbf{c}_{t-1}\right)^j$$

$$i_t^j = \sigma\left(W_i\mathbf{x}_t + U_i\mathbf{h}_{t-1} + V_i\mathbf{c}_{t-1}\right)^j$$



(a) Long Short-Term Memory

Introduziu o conceito de célula de memória.
A célula de memória pode manter seu valor por um tempo curto ou longo como uma função de suas entradas, o que permite que a célula se lembre daquilo que é

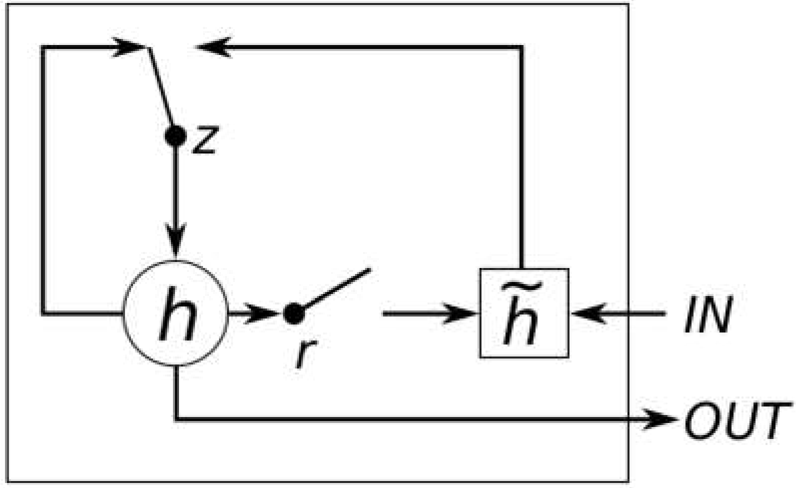
GRU

$$h_t^j = (1 - z_t^j)h_{t-1}^j + z_t^j \tilde{h}_t^j$$

$$z_t^j = \sigma\left(W_z \mathbf{x}_t + U_z \mathbf{h}_{t-1}\right)^j$$

$$\tilde{h}_t^j = \tanh\left(W \mathbf{x}_t + U\left(\mathbf{r}_t \odot \mathbf{h}_{t-1}\right)\right)^j$$

$$r_t^j = \sigma\left(W_r \mathbf{x}_t + U_r \mathbf{h}_{t-1}\right)^j$$



(b) Gated Recurrent Unit

A GRU é mais simples do que a LSTM, pode ser treinada com mais rapidez e pode ser mais eficiente em sua execução. No entanto,a LSTM pode ser mais express

DISCUSSÃO

SEMELHANÇA

(i) É fácil para cada unidade lembrar a existência de um recurso específico no fluxo de entrada

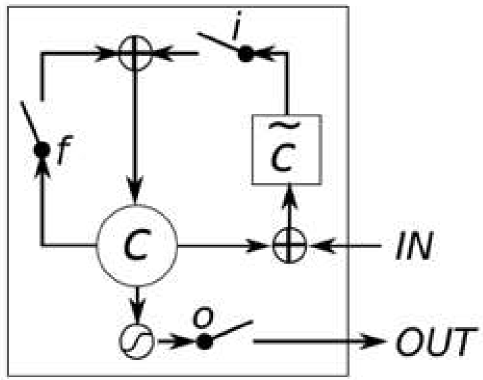
(ii) Essa adição cria efetivamente caminhos de atalho que ignoram várias etapas temporais.

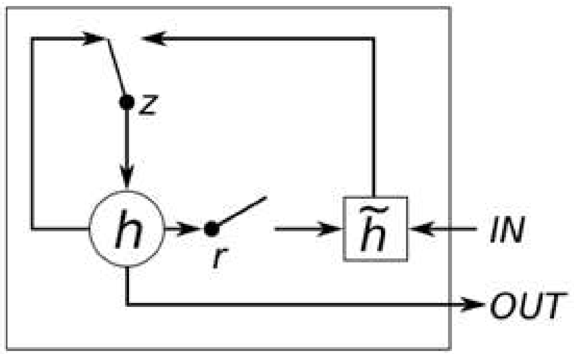
DISCUSSÃO

DIFERENÇAS

(i) Na unidade LSTM, a quantidade de conteúdo da memória acessada por outras unidades n

(ii) Localização do gate de *input* e gate de *reset* correspondente.





INFORMAÇÕES ADICIONAIS

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Recurrent models typically factor computation along the symbol positions of the input and output sequences. Aligning the positions to steps in computation time, they generate a sequence of hidden states h_t, as a function of the previous hidden state h_{t-1} and the input for position t. This inherently sequential nature precludes parallelization within training examples, which becomes critical at longer sequence lengths, as memory constraints limit batching across examples. Recent work has achieved significant improvements in computational efficiency through factorization tricks [21] and conditional computation [32], while also improving model performance in case of the latter.

RNN (LSTM e GRU em particular) foram firmemente estabelecidas como abordagens de última geração em modelagem de sequência, como processamento de lin

Inúmeros esforços, desde então, continuaram a empurrar os limites de modelos de linguagem recorrente e arquiteturas encoder-decoder.

(VASWANI, Ashish et al. Attention is all you need. In: Advances in Neural Information Processing Systems. 2017. p. 5998-6008.)

CONFIGURAÇÃO DO EXPERIMENTO I

- Comparação de LSTM, GRU e tanh na tarefa de modelar uma sequência de dados.

- Aprender a probabilidade de distribuição em uma sequência de dados.

$$\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \log p\left(x_t^n \mid x_1^n, \ldots, x_{t-1}^n; \boldsymbol{\theta}\right),$$

CONFIGURAÇÃO DO EXPERIMENTO II

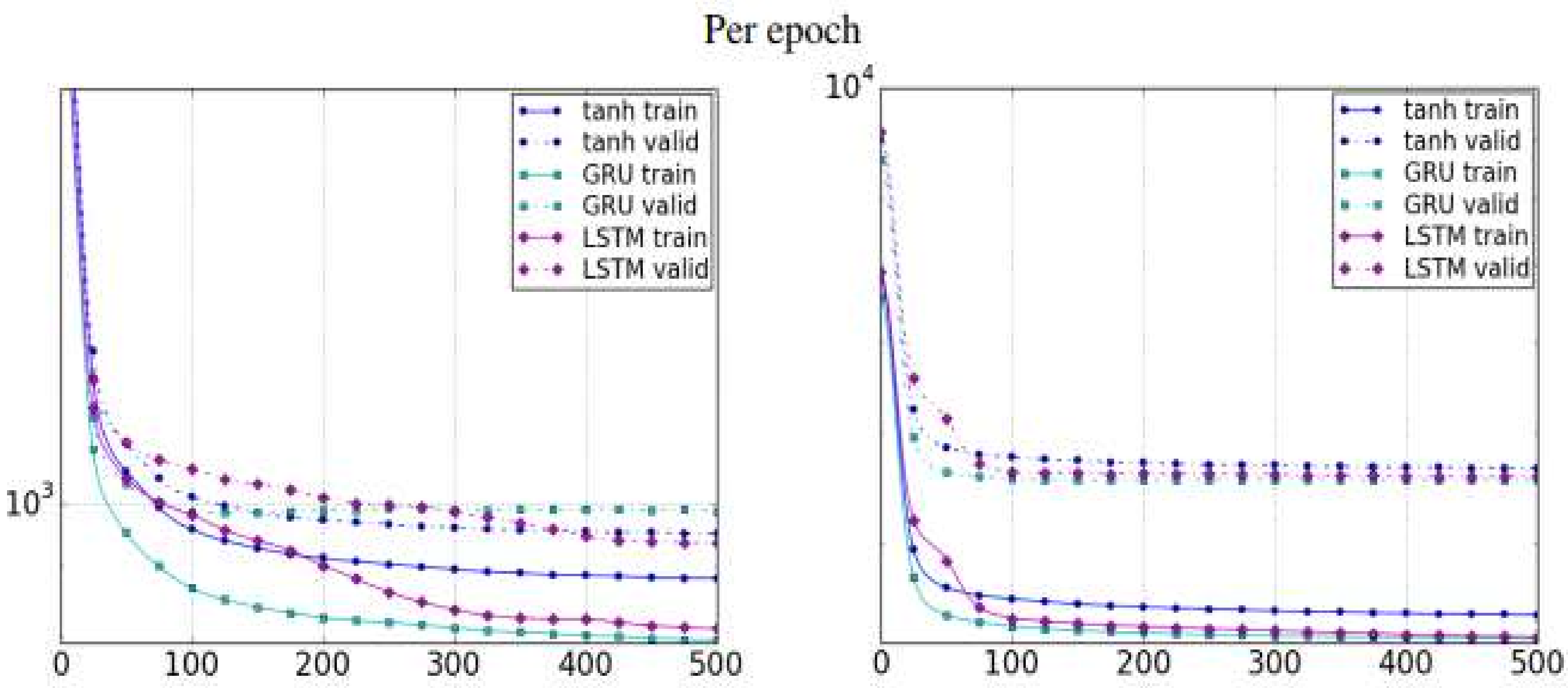
Unit	# of Units	# of Parameters
Polyphonic music modeling		
LSTM	36	$\approx 19.8 \times 10^3$
GRU	46	$\approx 20.2 \times 10^3$
tanh	100	$\approx 20.1 \times 10^3$
Speech signal modeling		
LSTM	195	$\approx 169.1 \times 10^3$
GRU	227	$\approx 168.9 \times 10^3$
tanh	400	$\approx 168.4 \times 10^3$

Table 1: The sizes of the models tested in the experiments.

			tanh	GRU	LSTM
Music Datasets	Nottingham	train	3.22	2.79	3.08
		test	**3.13**	3.23	3.20
	JSB Chorales	train	8.82	6.94	8.15
		test	9.10	**8.54**	8.67
	MuseData	train	5.64	5.06	5.18
		test	6.23	**5.99**	6.23
	Piano-midi	train	5.64	4.93	6.49
		test	9.03	**8.82**	9.03
Ubisoft Datasets	Ubisoft dataset A	train	6.29	2.31	1.44
		test	6.44	3.59	**2.70**
	Ubisoft dataset B	train	7.61	0.38	0.80
		test	7.62	**0.88**	1.26

Table 2: The average negative log-probabilities of the training and test sets.

RESULTADOS E ANÁLISE I

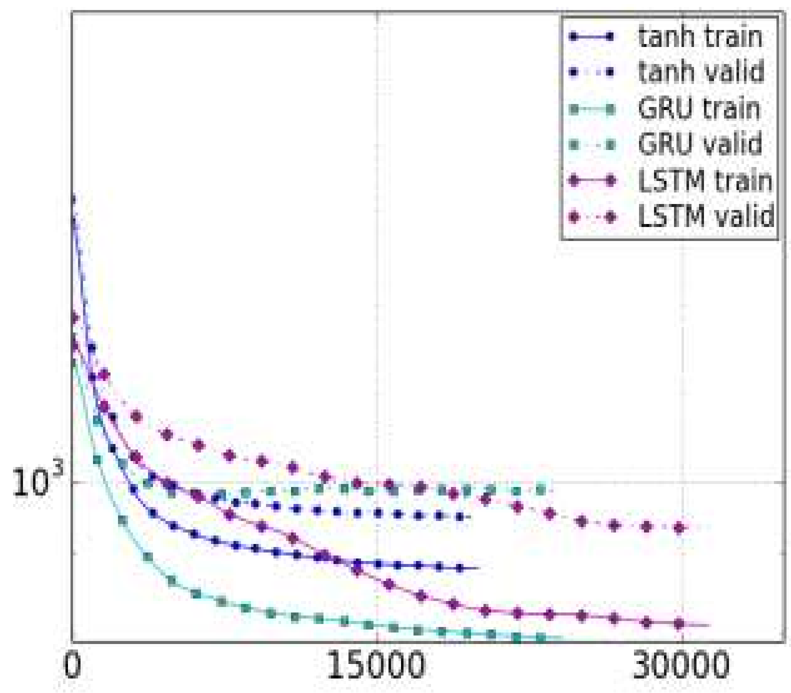


(a) Nottingham Dataset

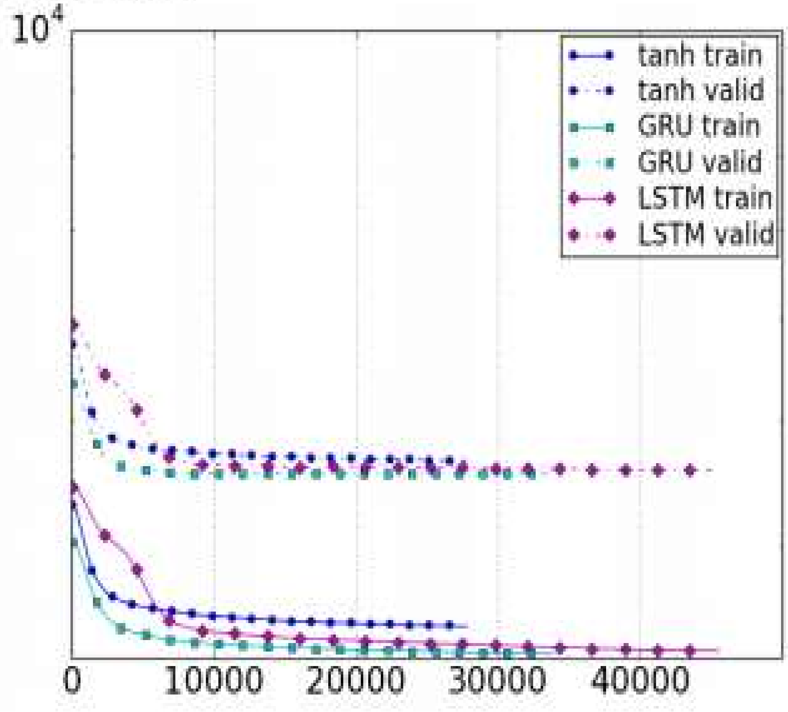
(b) MuseData Dataset

RESULTADOS E ANÁLISE II

Wall Clock Time (seconds)



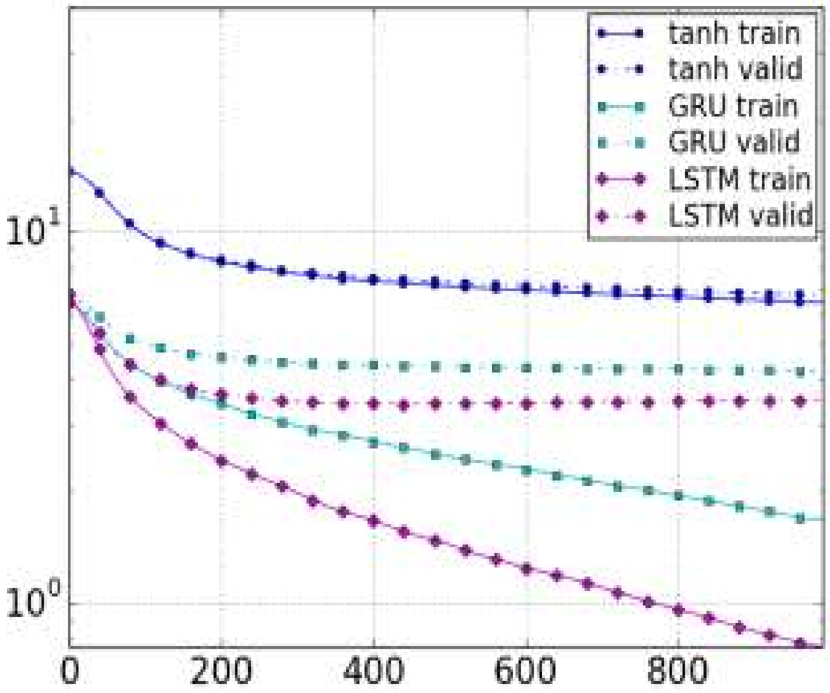
(a) Nottingham Dataset



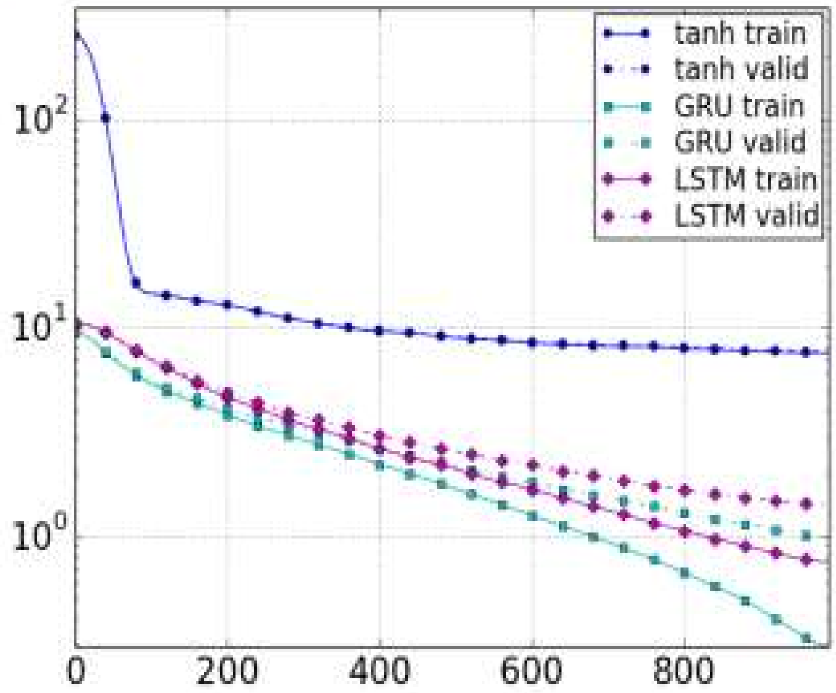
(b) MuseData Dataset

RESULTADOS E ANÁLISE III

Per epoch



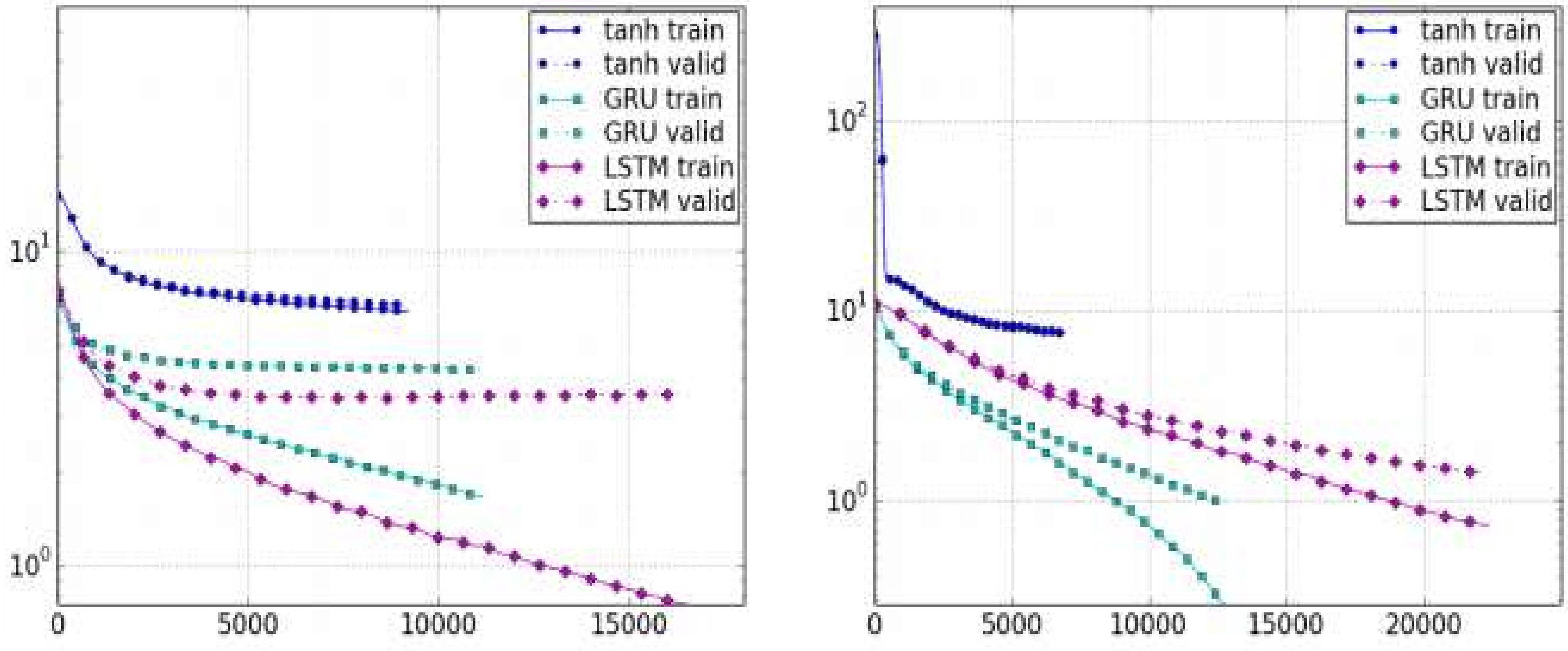
(a) Ubisoft Dataset A



(b) Ubisoft Dataset B

RESULTADOS E ANÁLISE IV

Wall Clock Time (seconds)



(a) Ubisoft Dataset A

(b) Ubisoft Dataset B