Recurrent Neural Networks

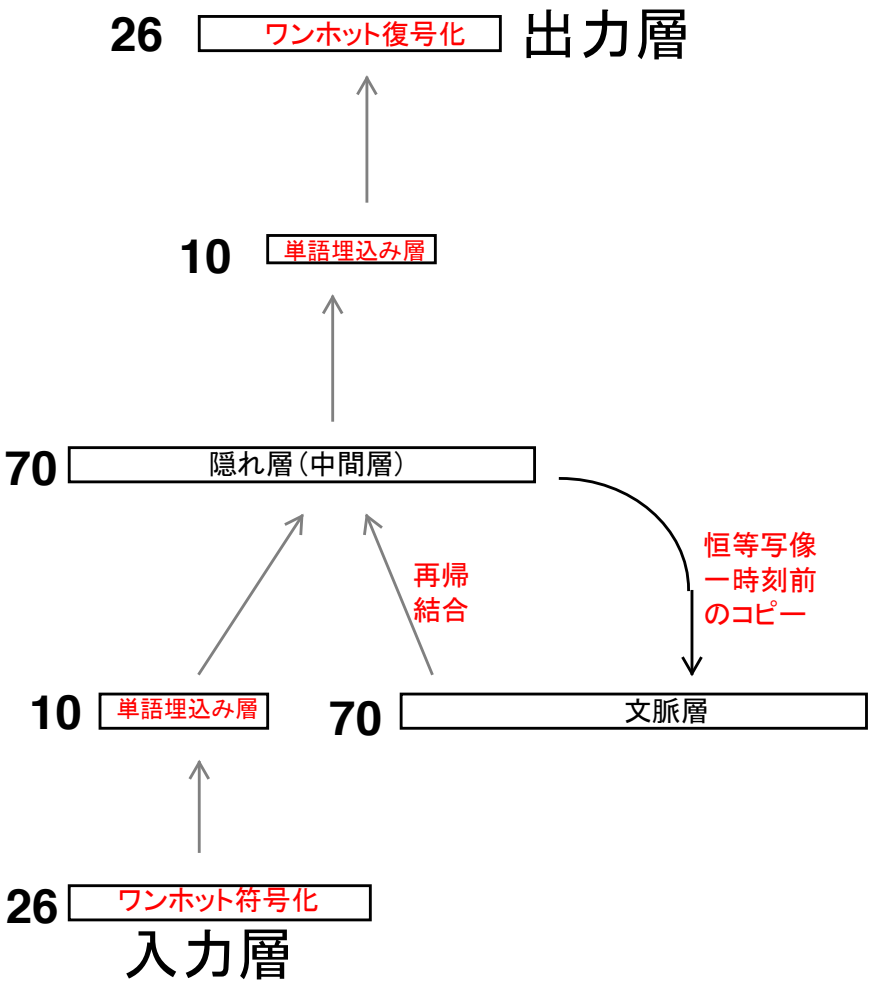


Figure 1: from Elman (1991)

Unfolding of RNN

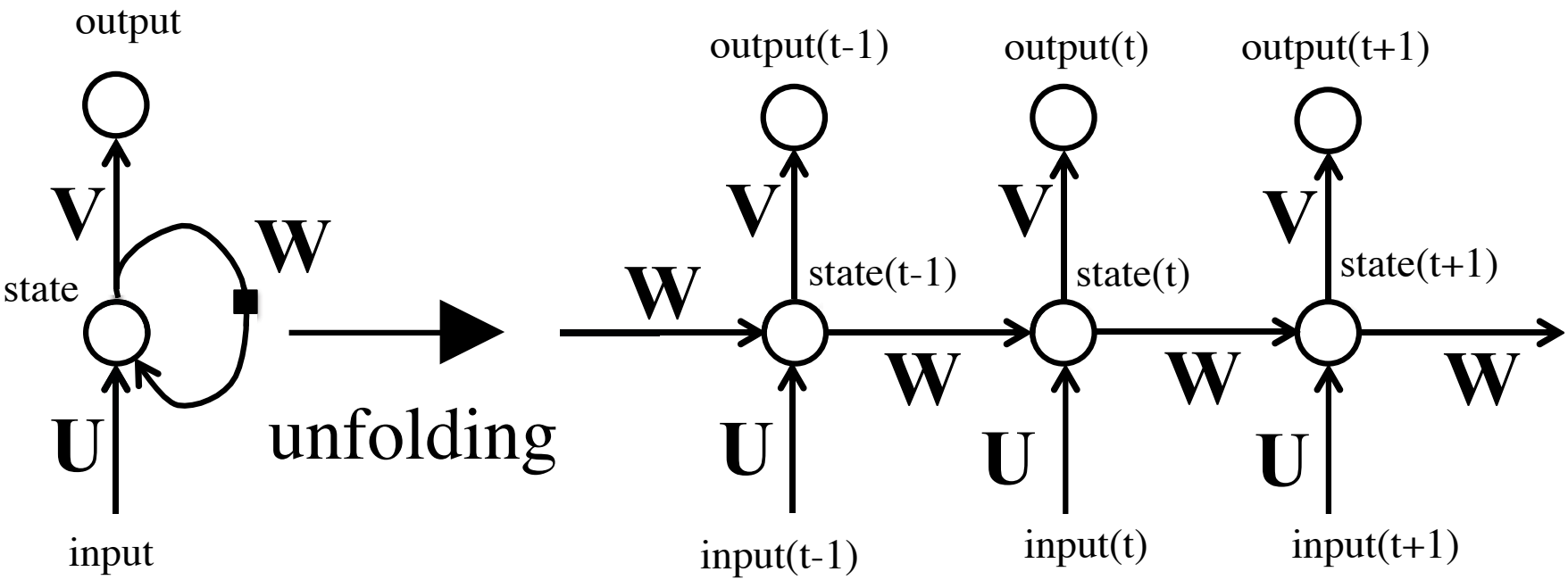


Figure 2: RNN の時間発展

Seq2Seq model

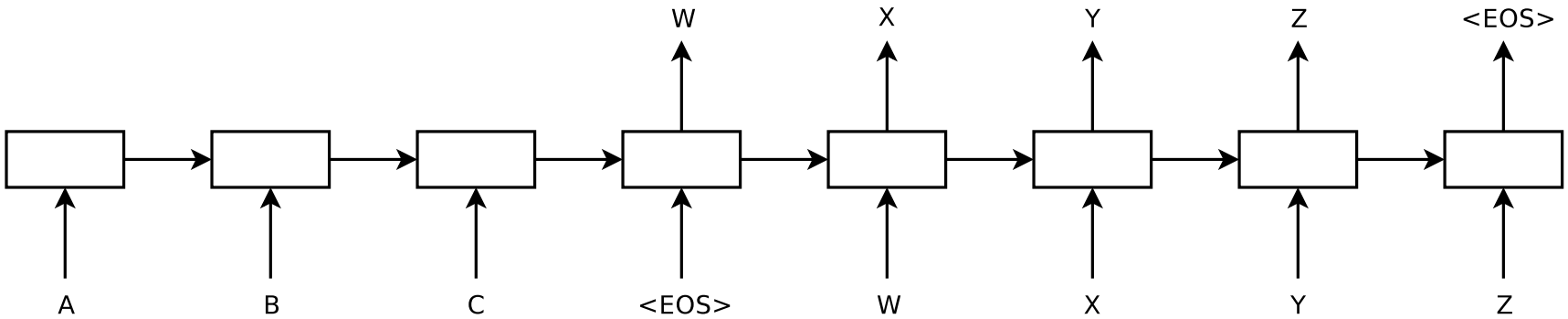


Figure 3: From Sutskever, Vinyals, & Le (2014)

$$\underset{\theta}{\operatorname{argmax}} \left(-\log p\left(w_{t+1}\right)\right) = f\left(w_t \mid \theta\right) \tag{1}$$

Turing completeness

(Siegelmann & Sontag, 1991) said Turing completeness of RNN.

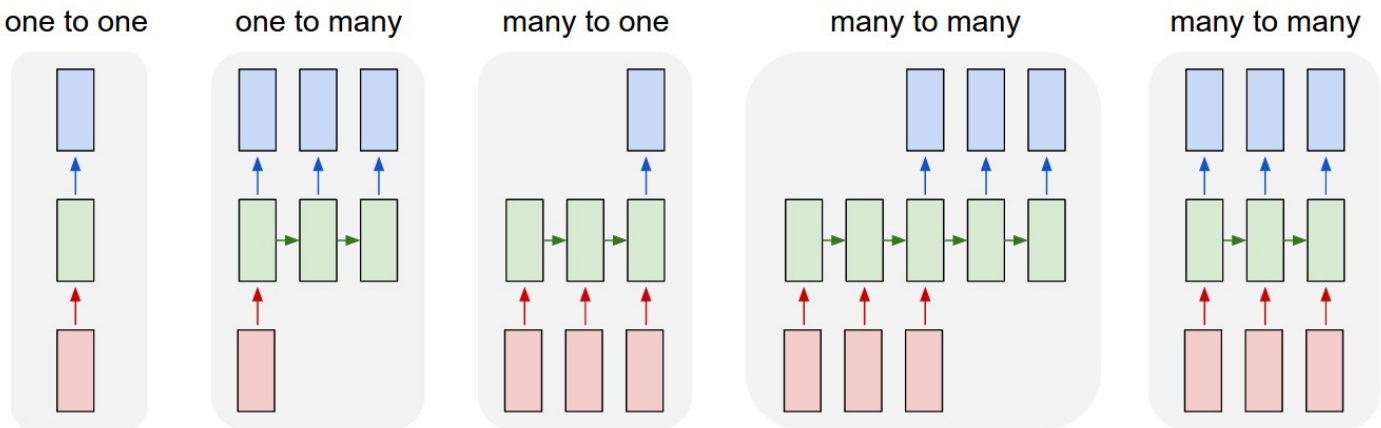


Figure 4: RNN variations from
http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Universal computer

Siegelmann & Sontag (1991) mentioned about the Turing completeness of RNN.

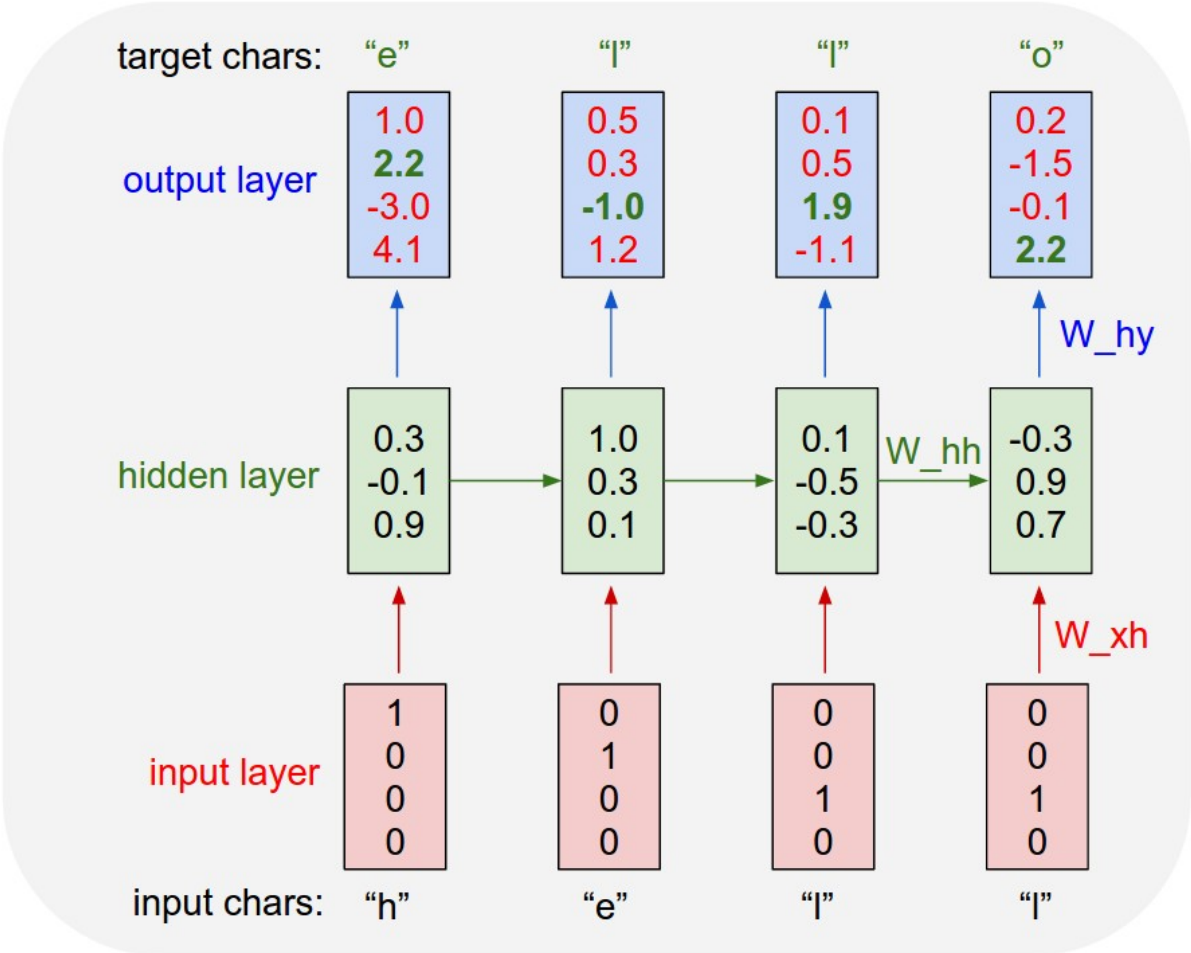


Figure 5: RNN variations from
http://karpathy.github.io/2015/05/21/rnn-effectiveness/

LSTM

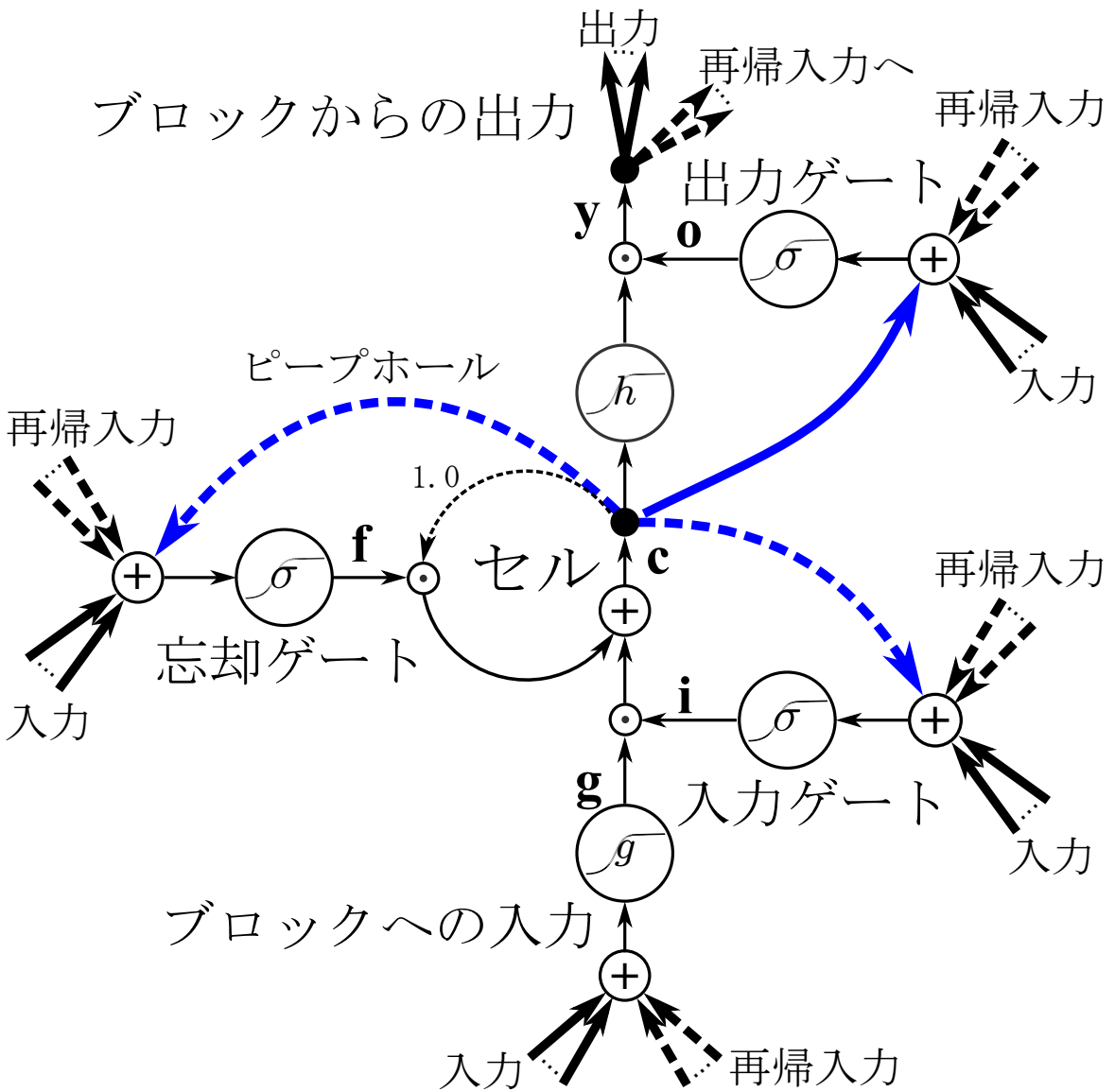


Figure 6: LSTM from 浅川 (2016) より

LSTM in detail

The LSTM (left figure) can be described as the input signals $\boldsymbol{x}_t$ at time t, the output signals $\boldsymbol{o}_t$, the forget gate $\boldsymbol{f}_t$, and the output signal $\boldsymbol{y}_t$, the memory cell $\boldsymbol{c}_t$, then we can get the following:

$$
\begin{aligned}
i_t &= \sigma\left(W_{xi}x_t + W_{hi}y_{t-1} + b_i\right), && (2)\\
f_t &= \sigma\left(W_{xf}x_t + W_{hf}y_{t-1} + b_f\right), && (3)\\
o_t &= \sigma\left(W_{xo}x_t + W_{ho}y_{t-1} + b_o\right), && (4)\\
g_t &= \phi\left(W_{xc}x_t + W_{hc}y_{t-1} + b_c\right), && (5)\\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t, && (6)\\
h_t &= o_t \odot \phi\left(c_t\right) && (7)
\end{aligned}
$$

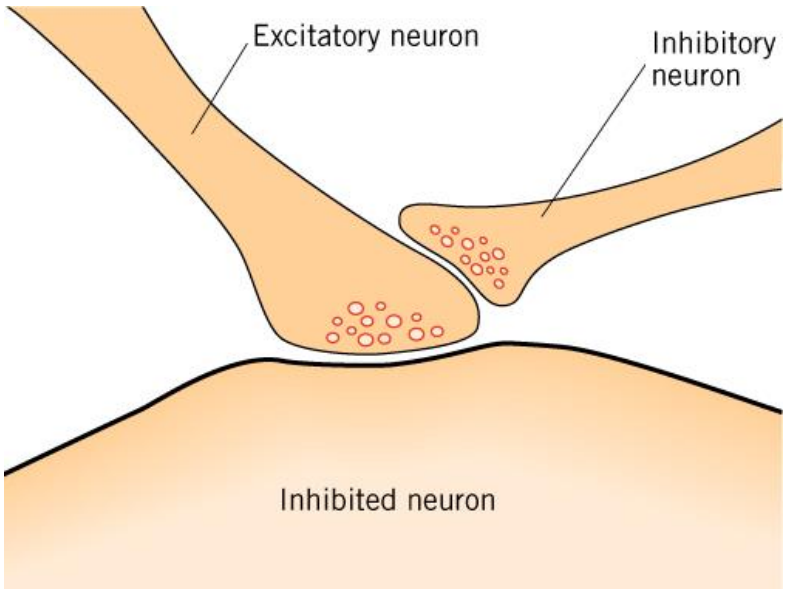
where

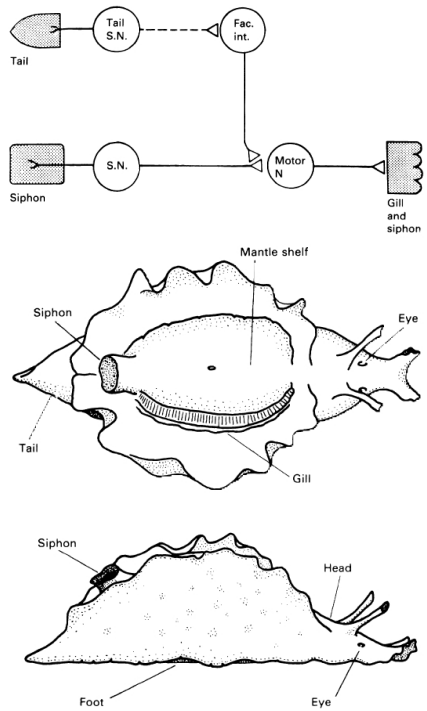
$\sigma(x) = \dfrac{1}{1+\exp(-x)}$ (logistic function)

$\phi(x) = \dfrac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)}$ (hyper tangent)

and $\odot$ menas Hadamard (element–wise) product.

Physiological correlates of gates in LSTM





from
http://kybele.psych.cornell.edu/~edelman/Psych-2140/week-2-2.html

Neural Image Captioning

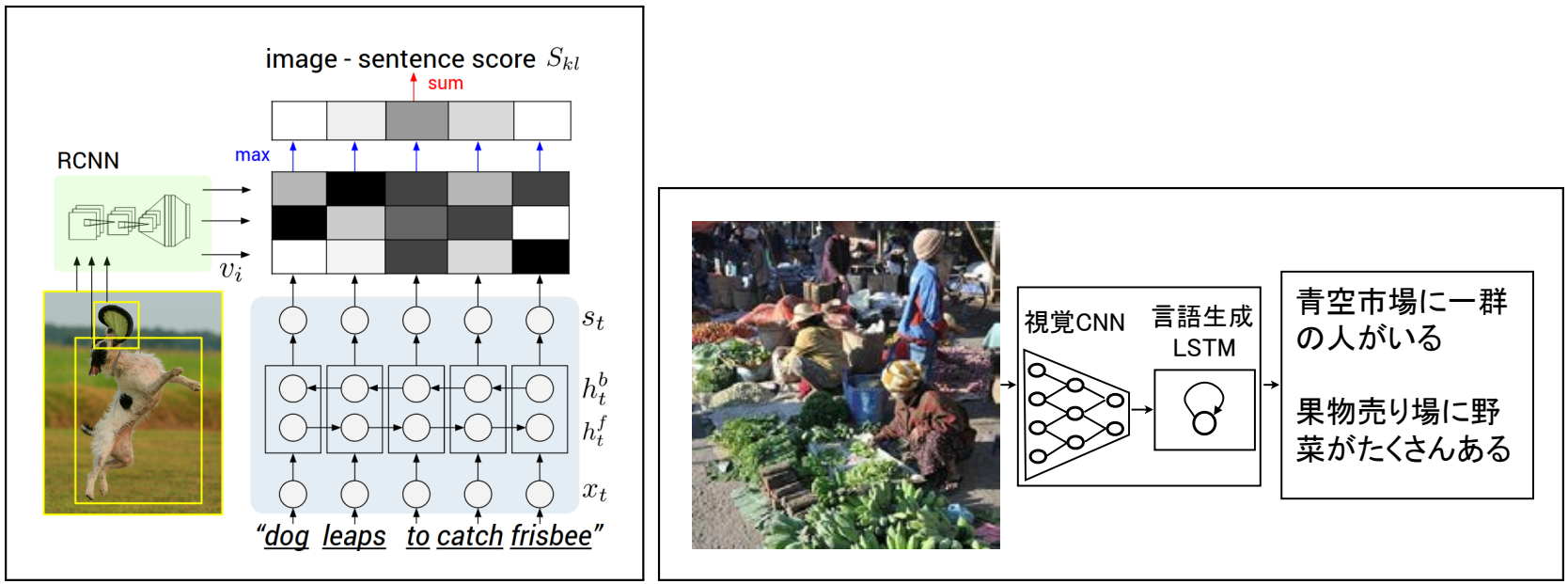
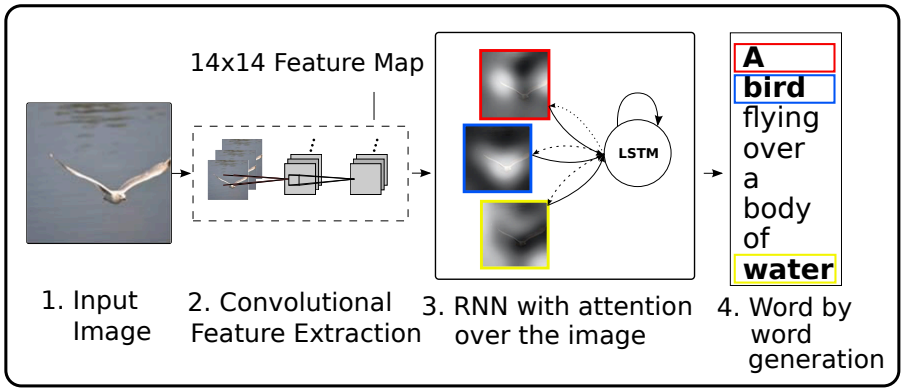


Figure 7: right:(Karpathy & Fei-Fei, 2015), left:(Vinyals, Toshev, Bengio, & Erhan, 2015)

RNN with Attention



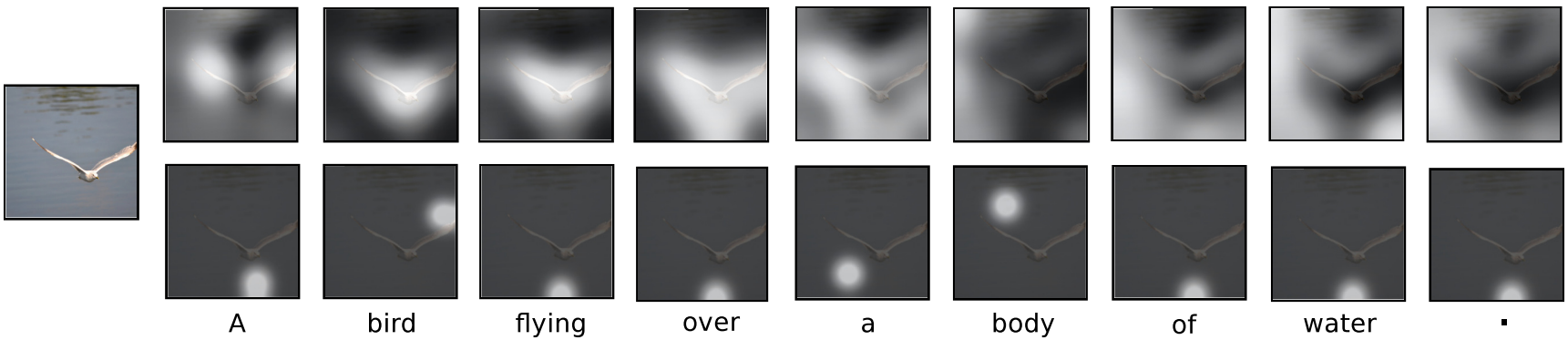


Figure 8: Xu et al. (2015)

Examples of NIC with attention

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

Figure 9: Xu et al. (2015)

Glimpse model

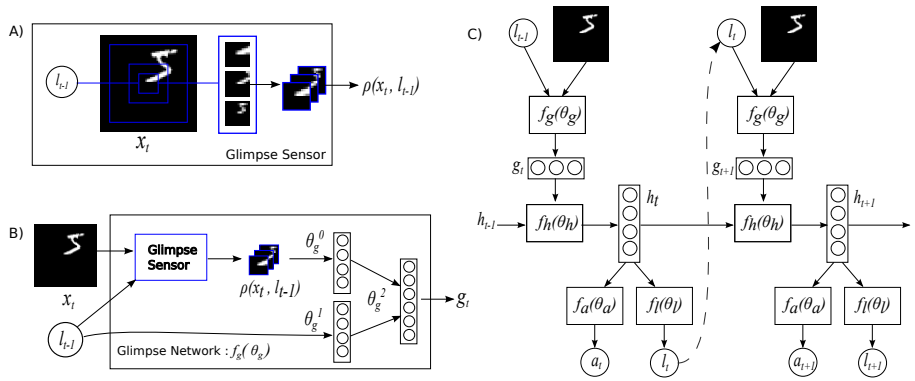


Figure 10: A)**Glimpse Sensor**: Given the coordinates of the glimpse and an input image, the sensor extracts a *retina-like* representation $\rho\left(x_t, l_{t-1}\right)$) centered at l_{t-1} that contains multiple resolution patches. B) **Glimpse Network**: Given the location $\left(l_{t-1}\right)$ and input image (x_t), uses the glimpse sensor to extract retina representation $\rho\left(x_t, l_{t-1}\right)$. The retina representation and glimpse location is then mapped into a hidden space using independent linear layers parameterized by θ_g^0 and θ_g^1 respectively using rectified units followed by another linear layer θ_2^2 to combine the information from both components. The glimpse network $f_g\left(;\left\{\theta_g^0, \theta_g^1, \theta_g^2\right\}\right)$ defines a trainable bandwidth limited sensor for the attention network producing the glimpse representation gt. C)**Model Architecture**: Overall, the model is an RNN. The core network of the model $f_h\left(;\theta_h\right)$ takes the glimpse representation g_t as input and combining with the internal representation at previous time step $h_t - 1$, produces the new internal state of the model h_t. The location network $f_l\left(;\theta_a\right)$ and the action network $f_a\left(;\theta_a\right)$ use the internal state h_t of the model to produce the next location to attend to l_t and the action/classification at respectively. This basic RNN iteration is repeated for a variable number of steps. Mnih et al. (2014)

World model by RNN

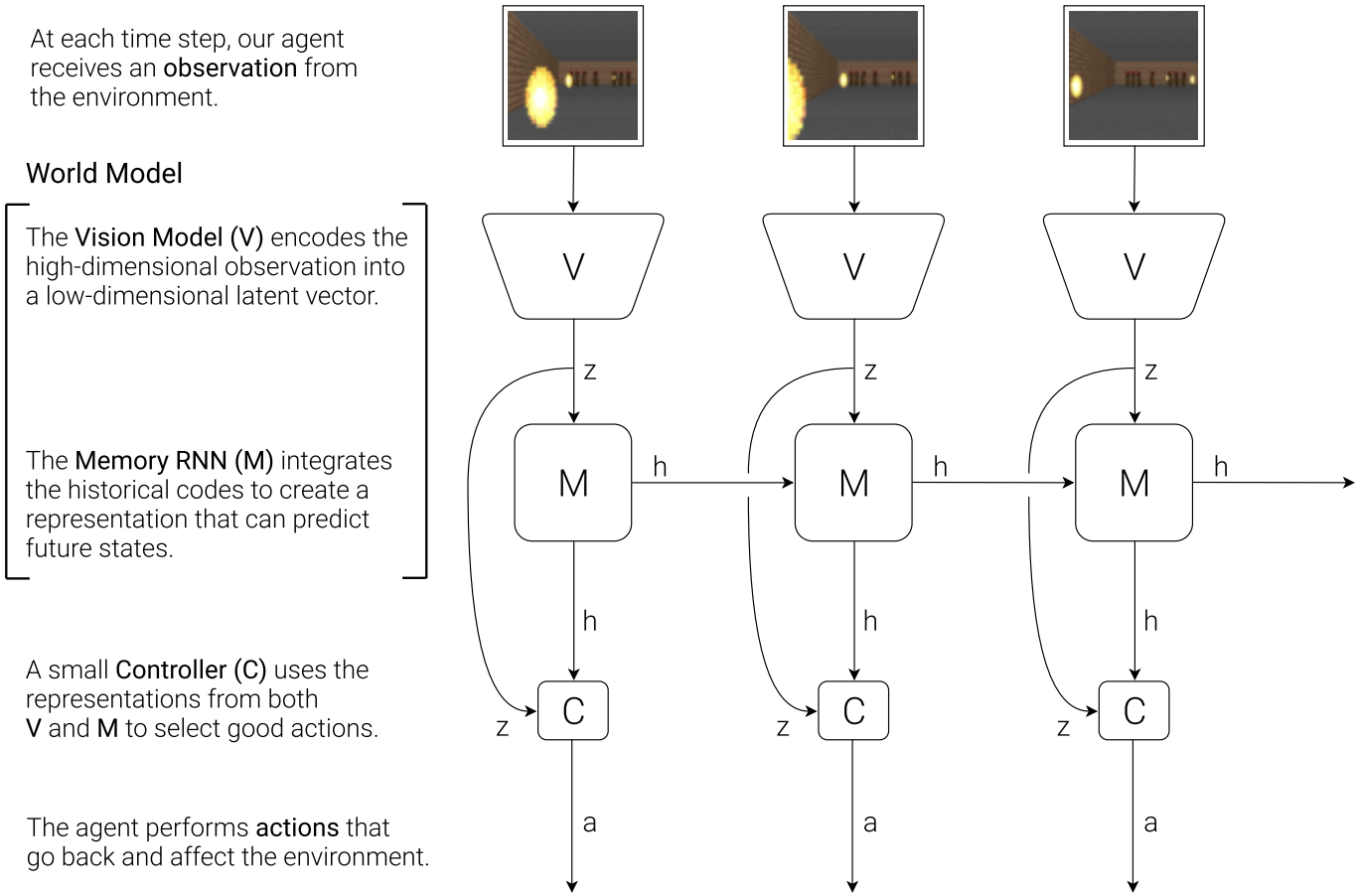


Figure 11: Ha & Schmithuber (2018) Fig.1

World Models

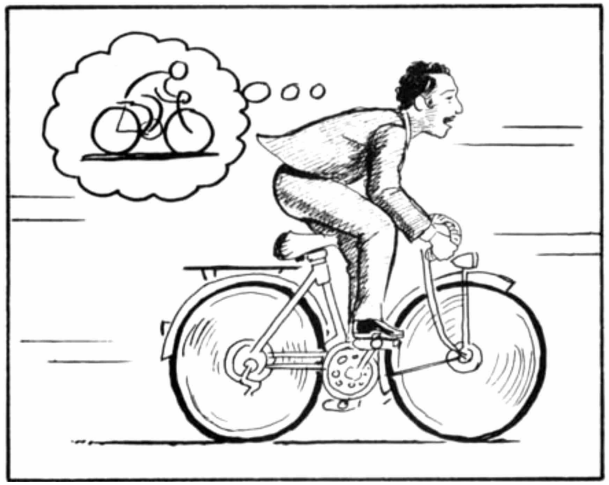


Figure 12: A World Model, from Scott McCloud's Understanding Comics. (McCloud, 1993; E, 2012)

Jay Wright Forrester, the father of system dynamics, described a mental model as:

> *The image of the world around us, which we carry in our head, is just a model. Nobody in his head imagines all the world, government or country. He has only selected concepts, and relationships between them, and uses those to represent the real system. (Forrester, 1971)*

文献

浅川 伸一. (2016). **リカレントニューラルネットワーク．『人工知能学事典新版』．東京**: **共立出版**.

Elman, J. L. (1991). *Incremental learing, or the importance of starting small* (Vol. 9101; Tech. Rep.). San Diego, CA: University of California, San Diego.

Ha, D., & Schmithuber, J. (2018). World models. *arXiv preprint*.

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Boston, MA, USA.

Mnih, V., Heess, N., Graves, A., & Kavukcuoglu, K. (2014). Recurrent models of visual attention. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger (Eds.), *Advances in neural information processing systems 27* (pp. 2204–2212). Curran Associates, Inc.

Siegelmann, H. T., & Sontag, E. D. (1991). Turing computability with neural nets. *Applied Mathematics Letter b*, *4*, 77–80.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger (Eds.), *Advances in Neural Information Processing Systems (NIPS)* (Vol. 27, pp. 3104–3112). Montreal, BC, Canada.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR).* Boston, MA, USA.

Xu, K., Ba, J. L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. S., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *arXiv:1502.03044*.