Deep Learning for Natural Language Processing

Stephen Clark et al…
DeepMind and University of Cambridge

5. Recurrent Neural Networks

Felix Hill
DeepMind

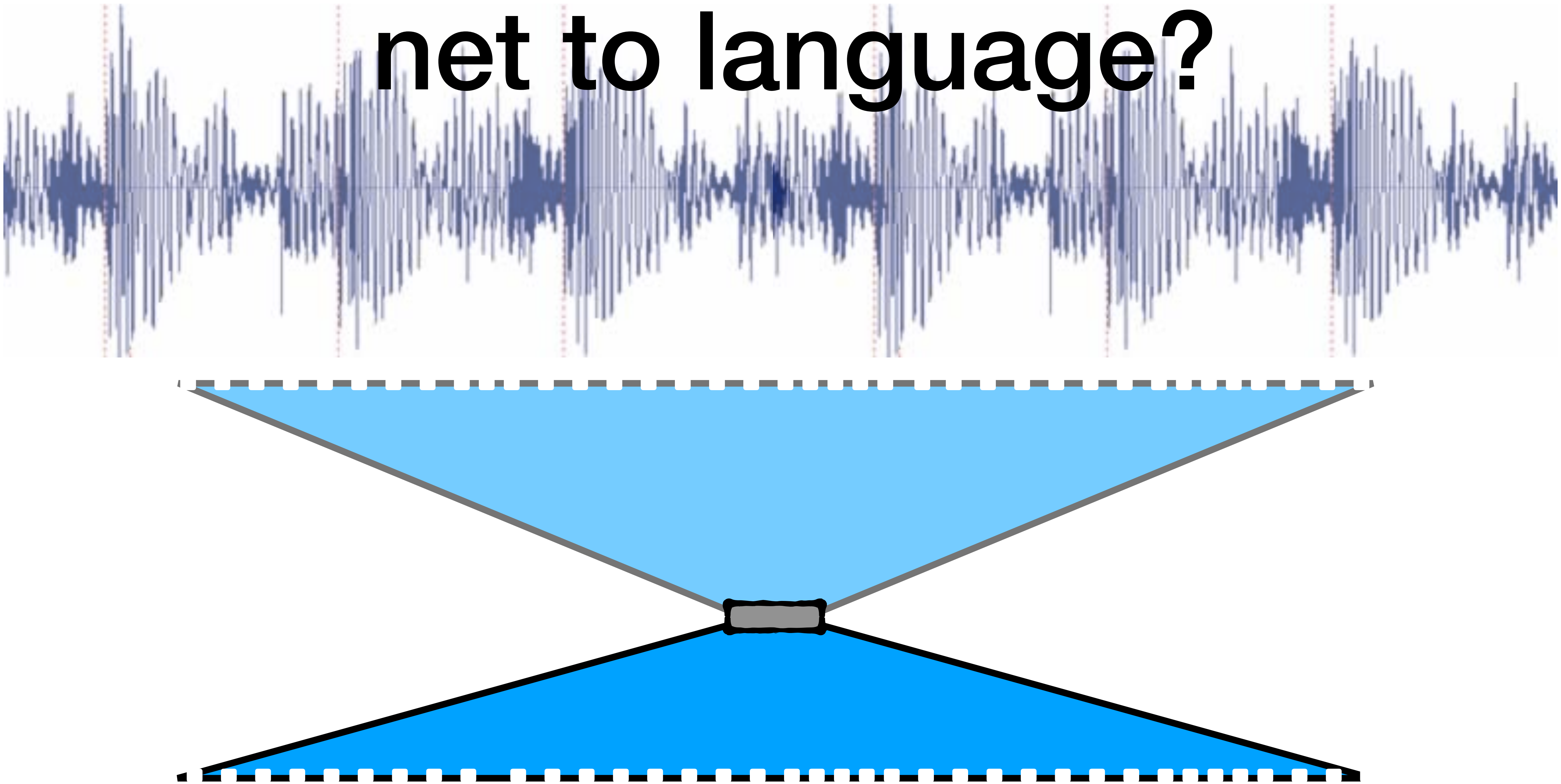
DeepMind

UNIVERSITY OF CAMBRIDGE

What are neural nets for?

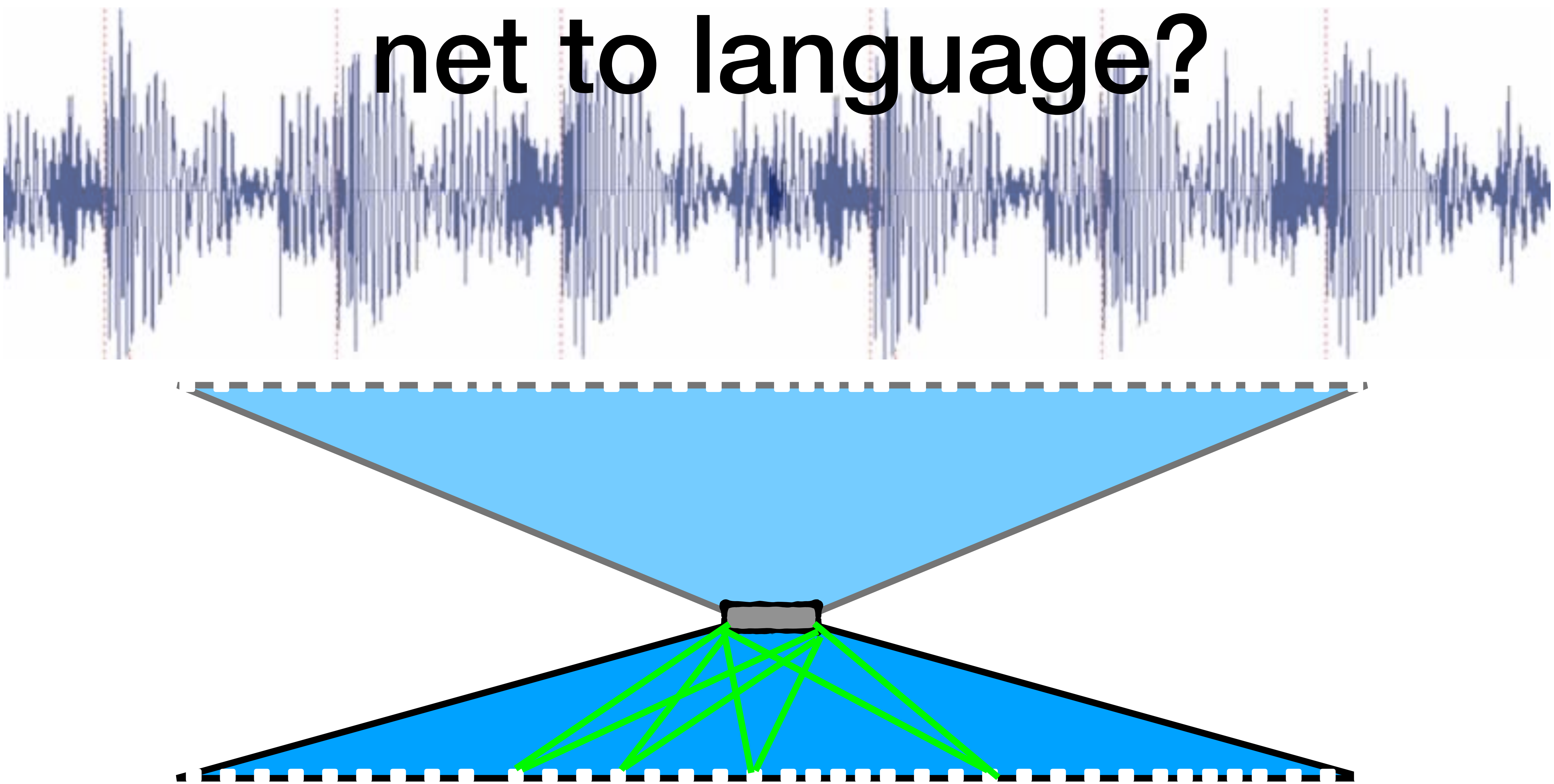
What are neural nets for?

How can you apply a neural net to language?



"language does not naturally go here, ahem, but fortunately….."

How can you apply a neural net to language?

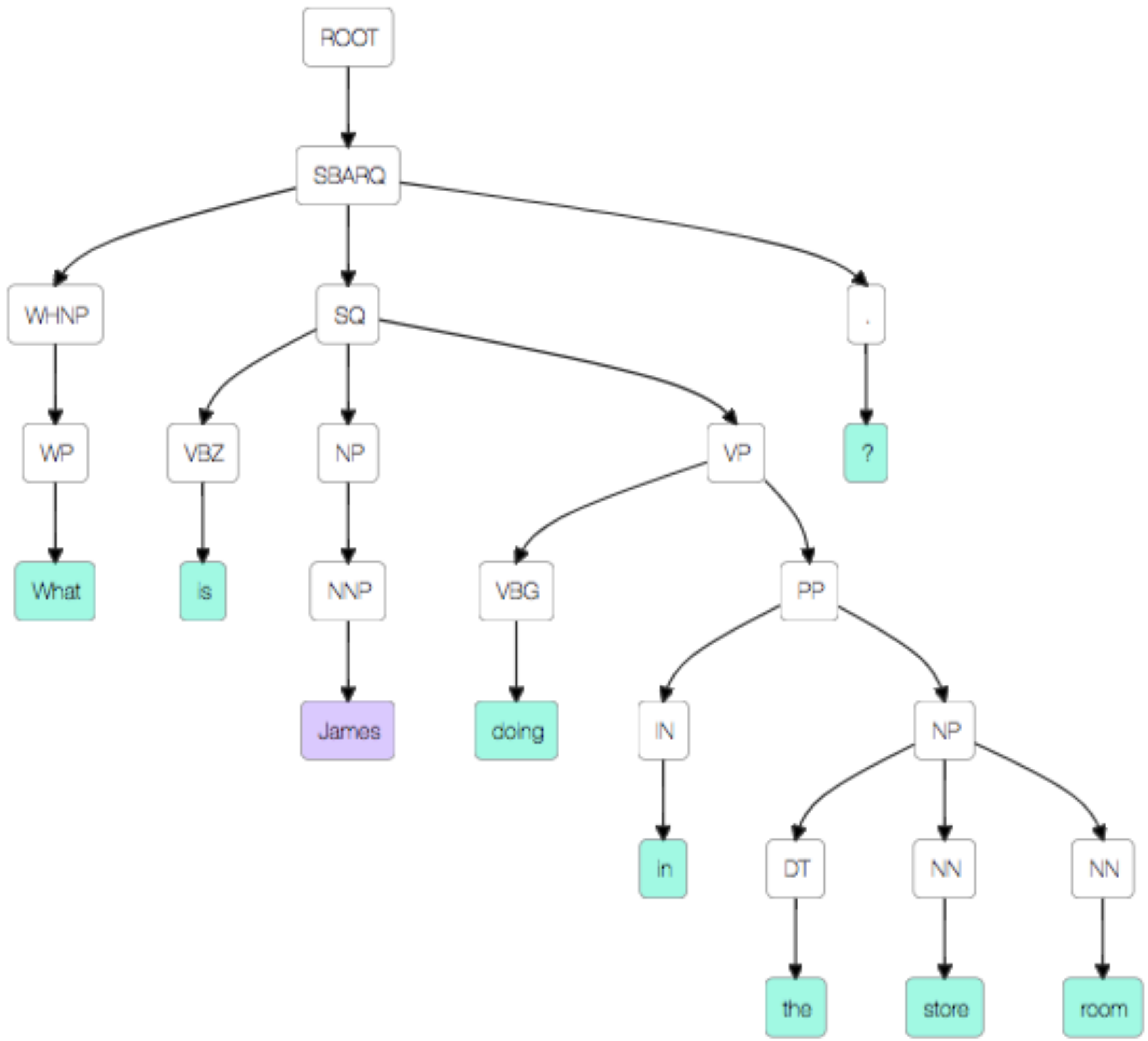


“*language does not naturally go here, ahem, but fortunately…..*”

what’s the issue here????

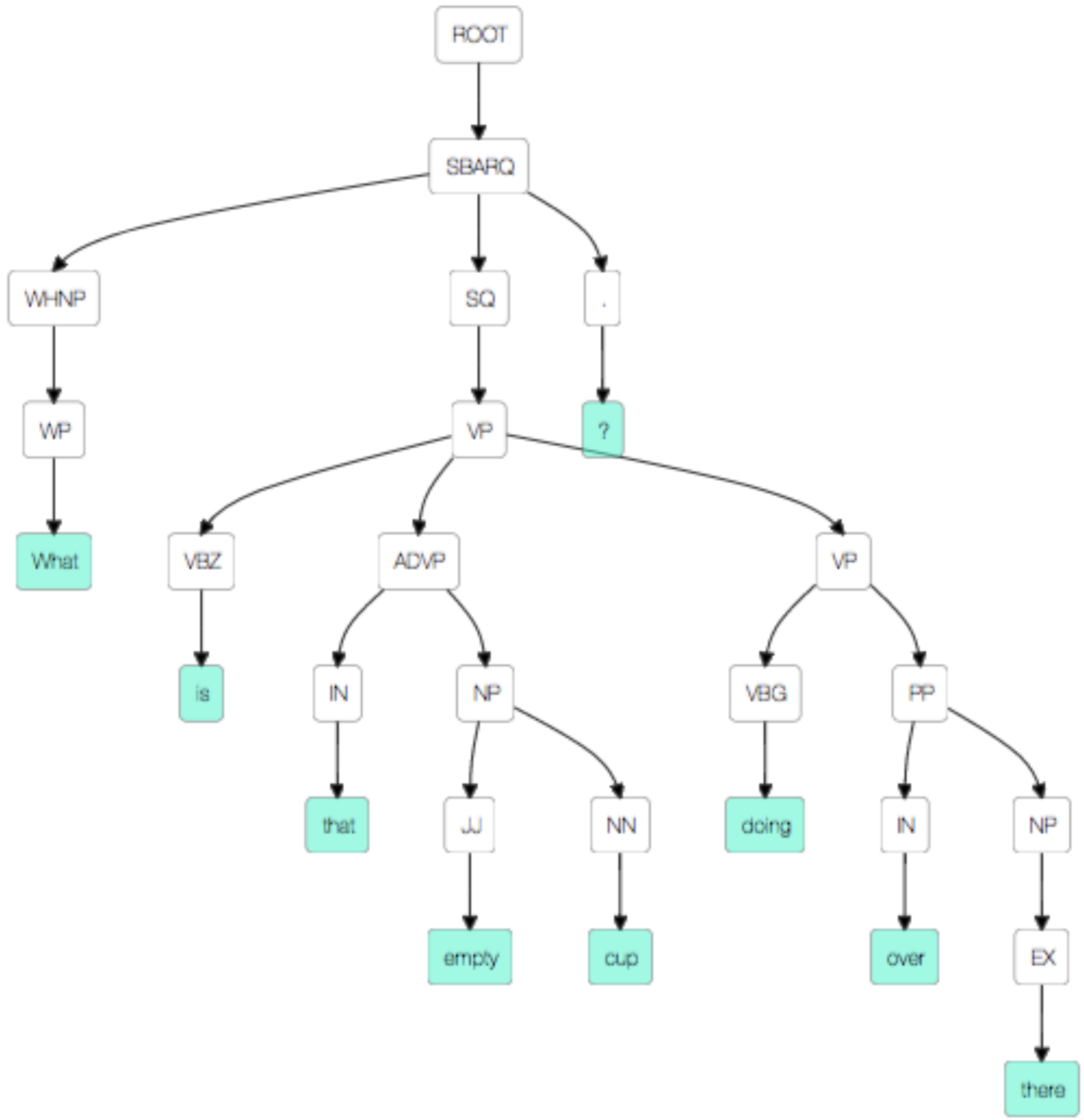
That’s the whole point!!

What is James doing in the store room?



searching for a book…

What is that empty cup doing over there?



err..being a cup?

time flies like an arrow

fruit flies like a banana

The networks that are good at Go and Atari were first developed *for this reason*!

Finding structure in time - Elman, 1990

COGNITIVE SCIENCE 14, 179-211 (1990)

Finding Structure in Time

JEFFREY L. ELMAN

University of California, San Diego

Time underlies many interesting human behaviors. Thus, the question of how to represent time in connectionist models is very important. One approach is to represent time implicitly by its effects on processing rather than explicitly (as in a spatial representation). The current report develops a proposal along these lines first described by Jordan (1986) which involves the use of recurrent links in order to provide networks with a dynamic memory. In this approach, hidden unit patterns are fed back to themselves; the internal representations which develop thus reflect task demands in the context of prior internal states. A set of simulations is reported which range from relatively simple problems (temporal version of XOR) to discovering syntactic/semantic features for words. The networks are able to learn interesting internal representations which incorporate task demands with memory demands; indeed, in this approach the notion of memory is inextricably bound up with task processing. These representations reveal a rich structure, which allows them to be highly context-dependent, while also expressing generalizations across classes of items. These representations suggest a method for representing lexical categories and the type/token distinction.

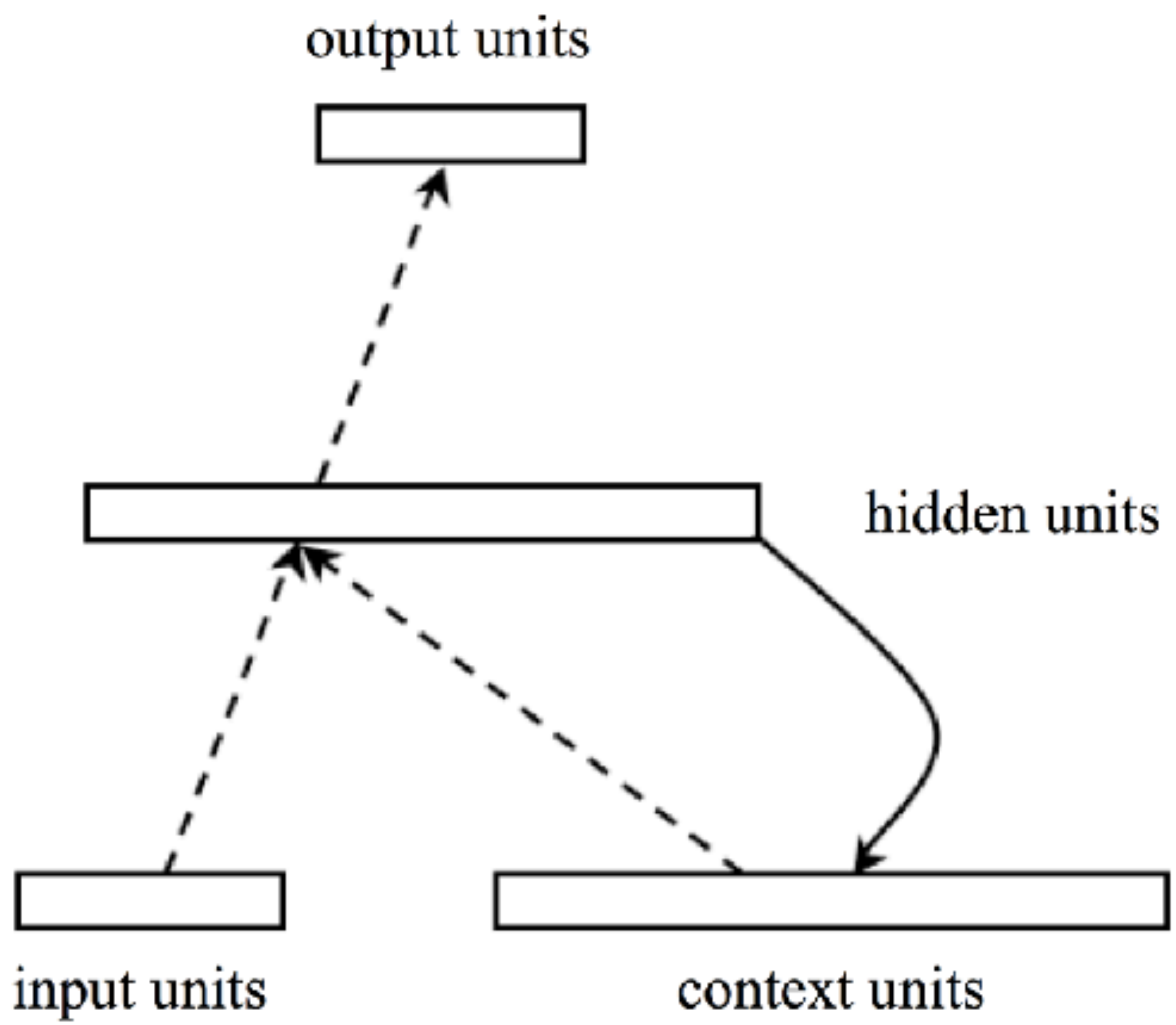
INTRODUCTION

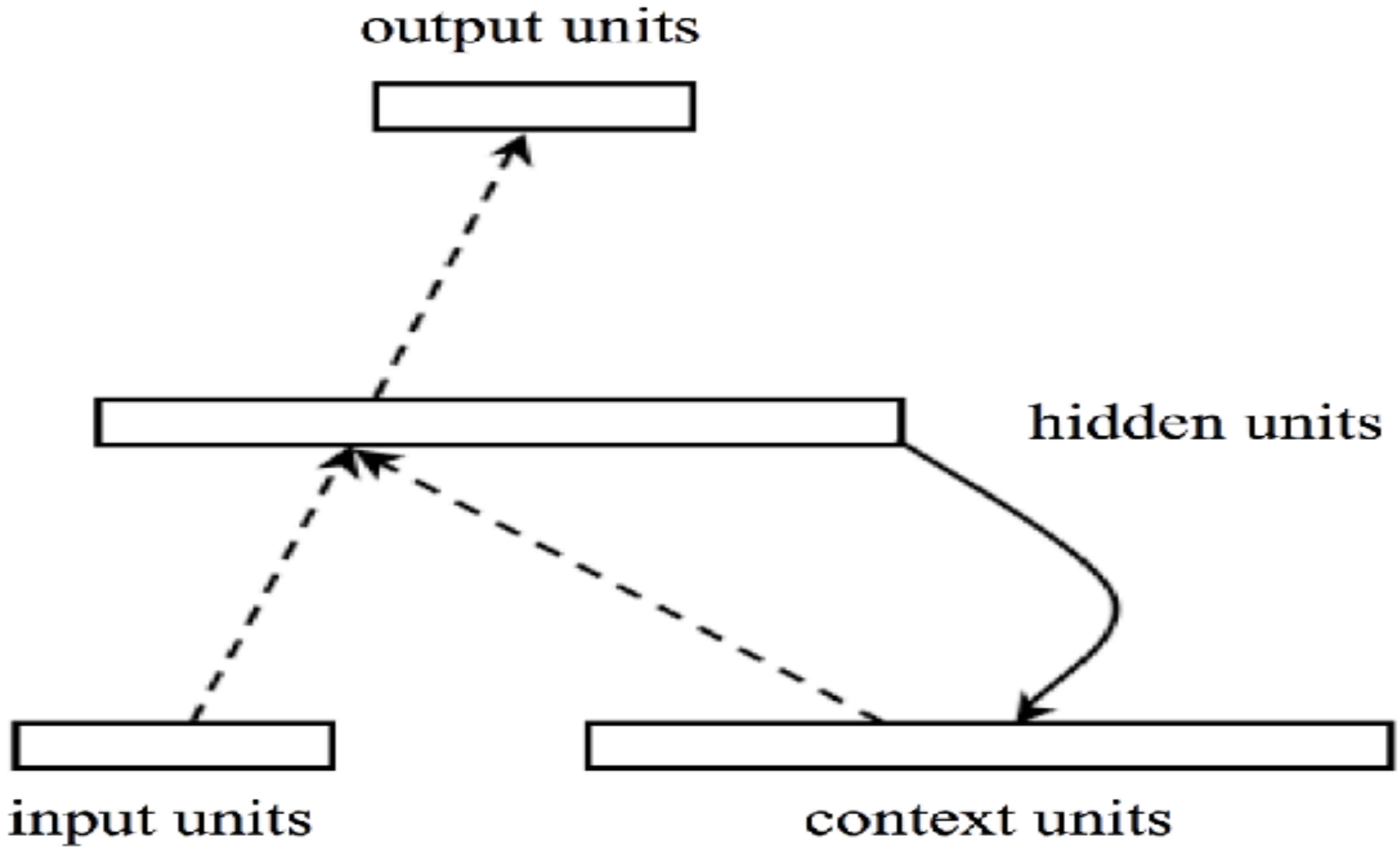
Time is clearly important in cognition. It is inextricably bound up with many behaviors (such as language) which express themselves as temporal sequences. Indeed, it is difficult to know how one might deal with such basic problems as goal-directed behavior, planning, or causation without some way of representing time.

The question of how to represent time might seem to arise as a special problem unique to parallel-processing models, if only because the parallel nature of computation appears to be at odds with the serial nature of tem-

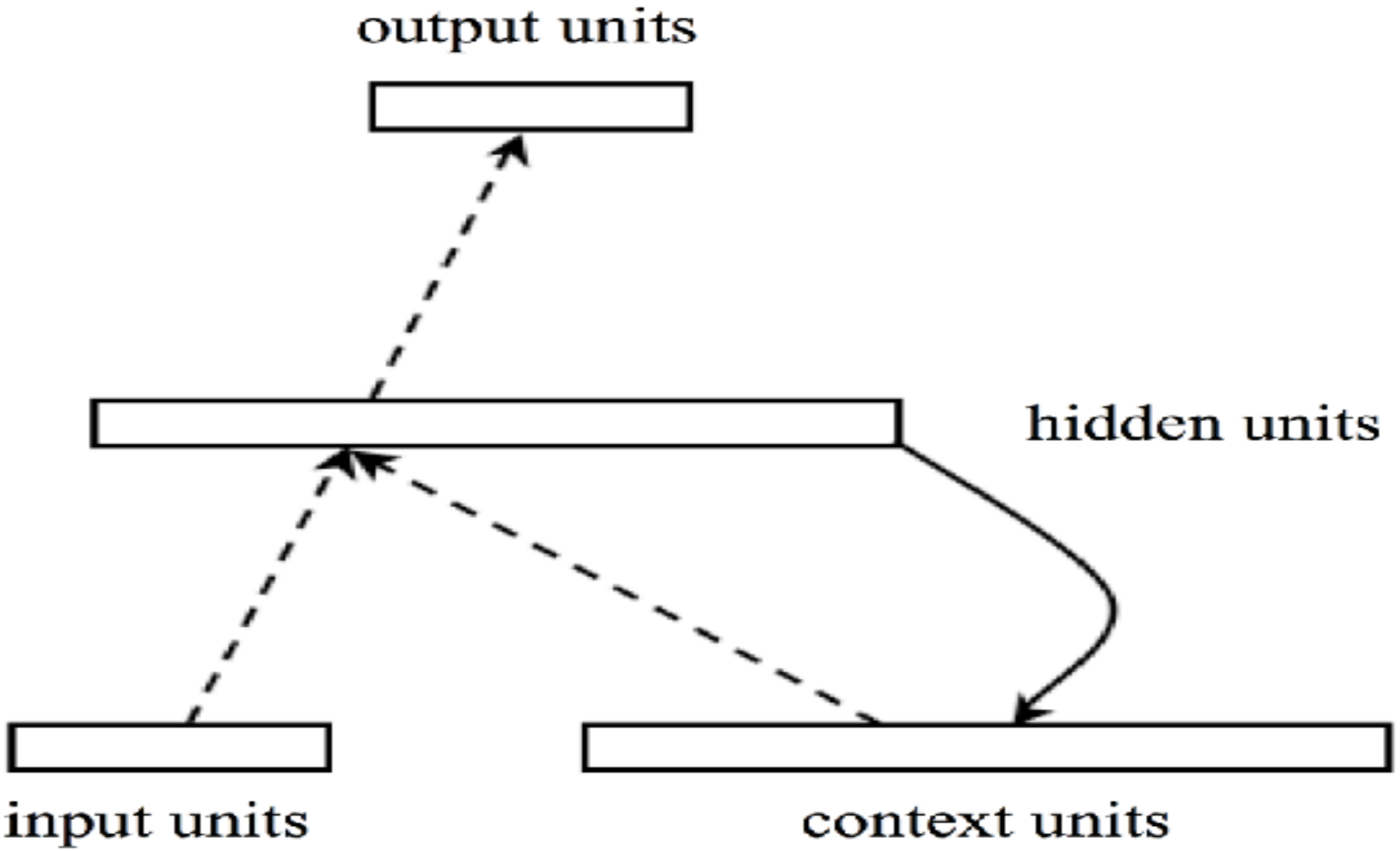
I would like to thank Jay McClelland, Mike Jordan, Mary Hare, Dave Rumelhart, Mike Mozer, Steve Poteet, David Zipser, and Mark Dolson for many stimulating discussions. I thank McClelland, Jordan, and two anonymous reviewers for helpful critical comments on an earlier draft of this article.

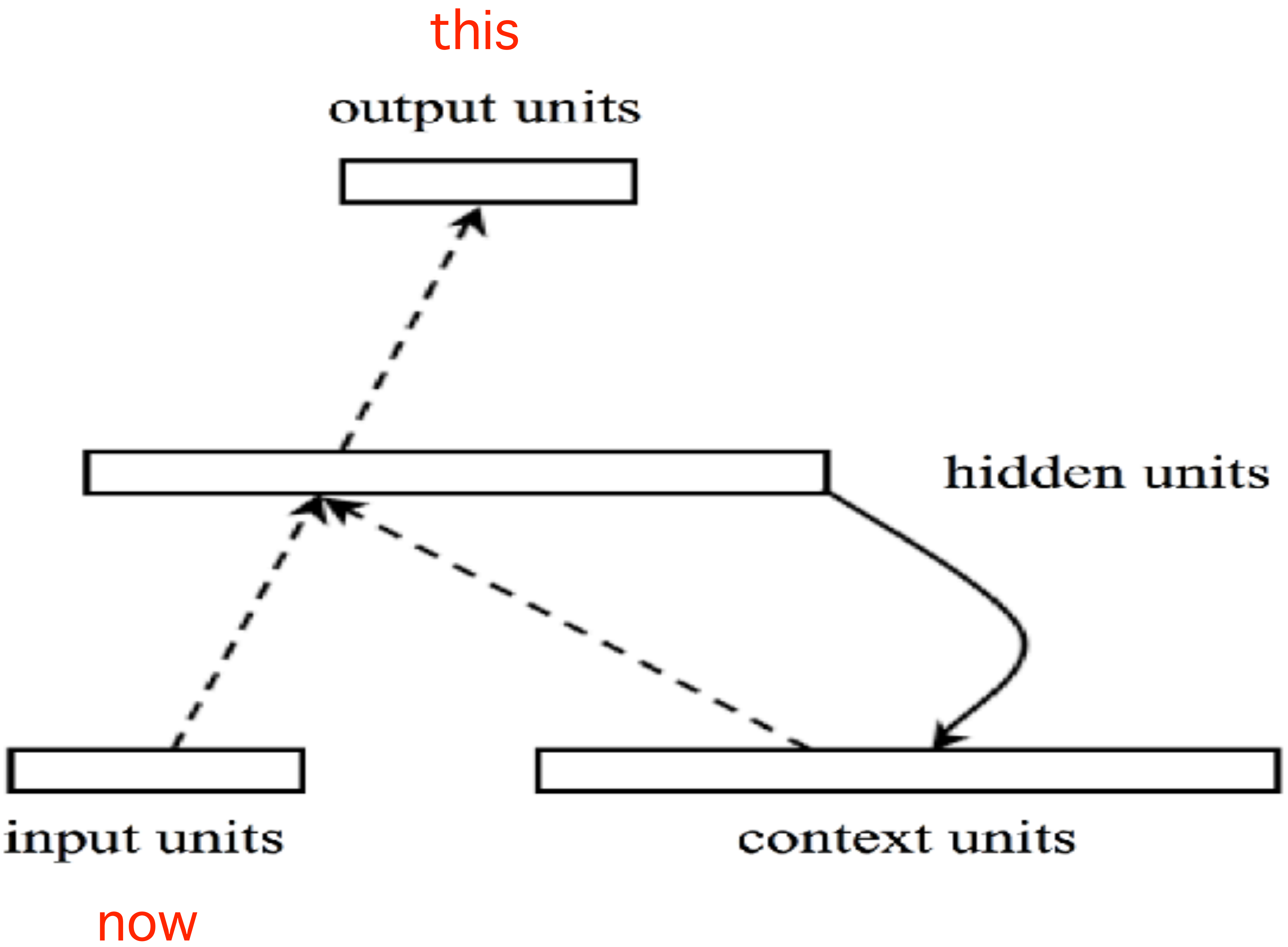
The simple recurrent network (now RNN)

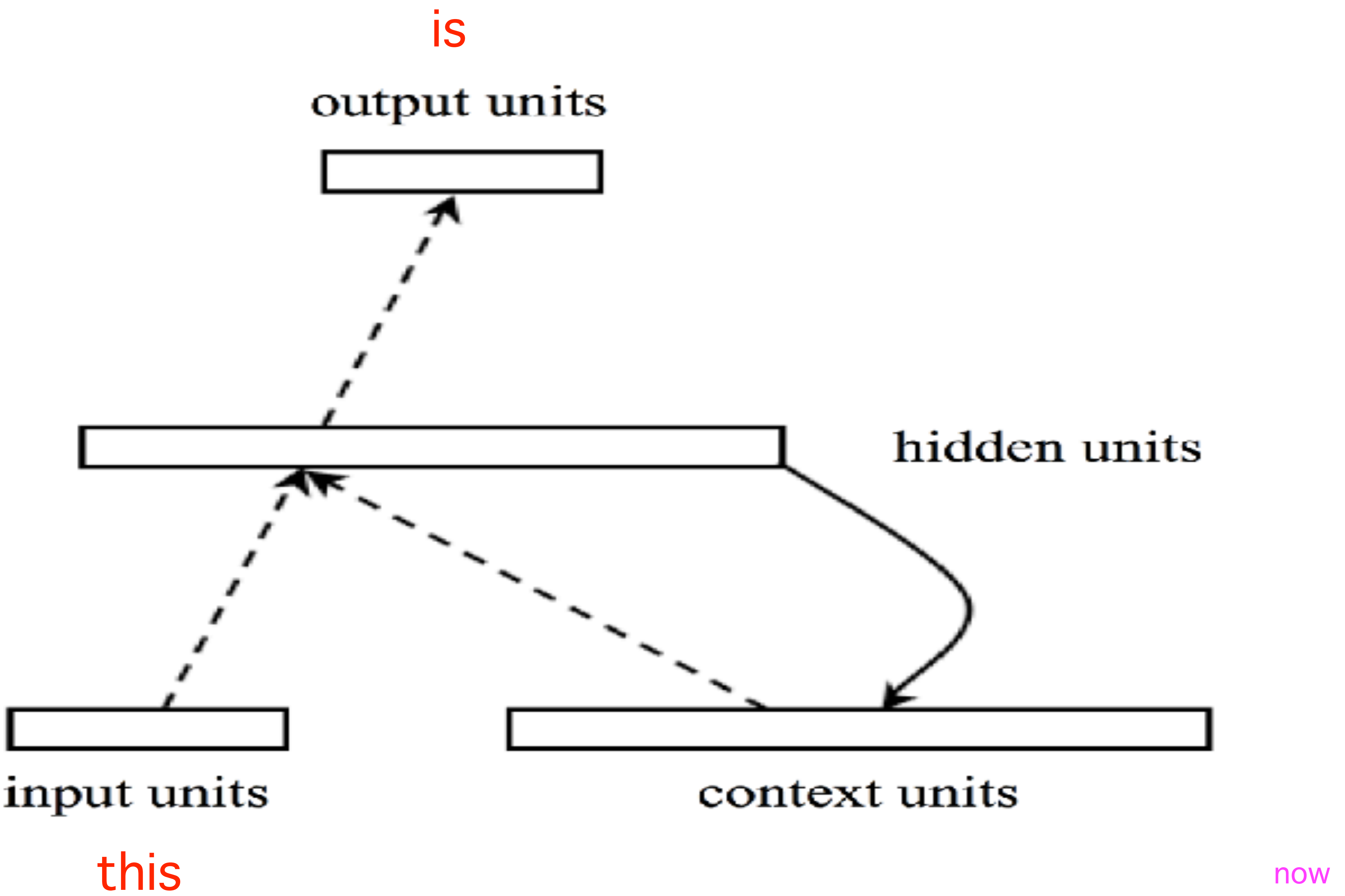


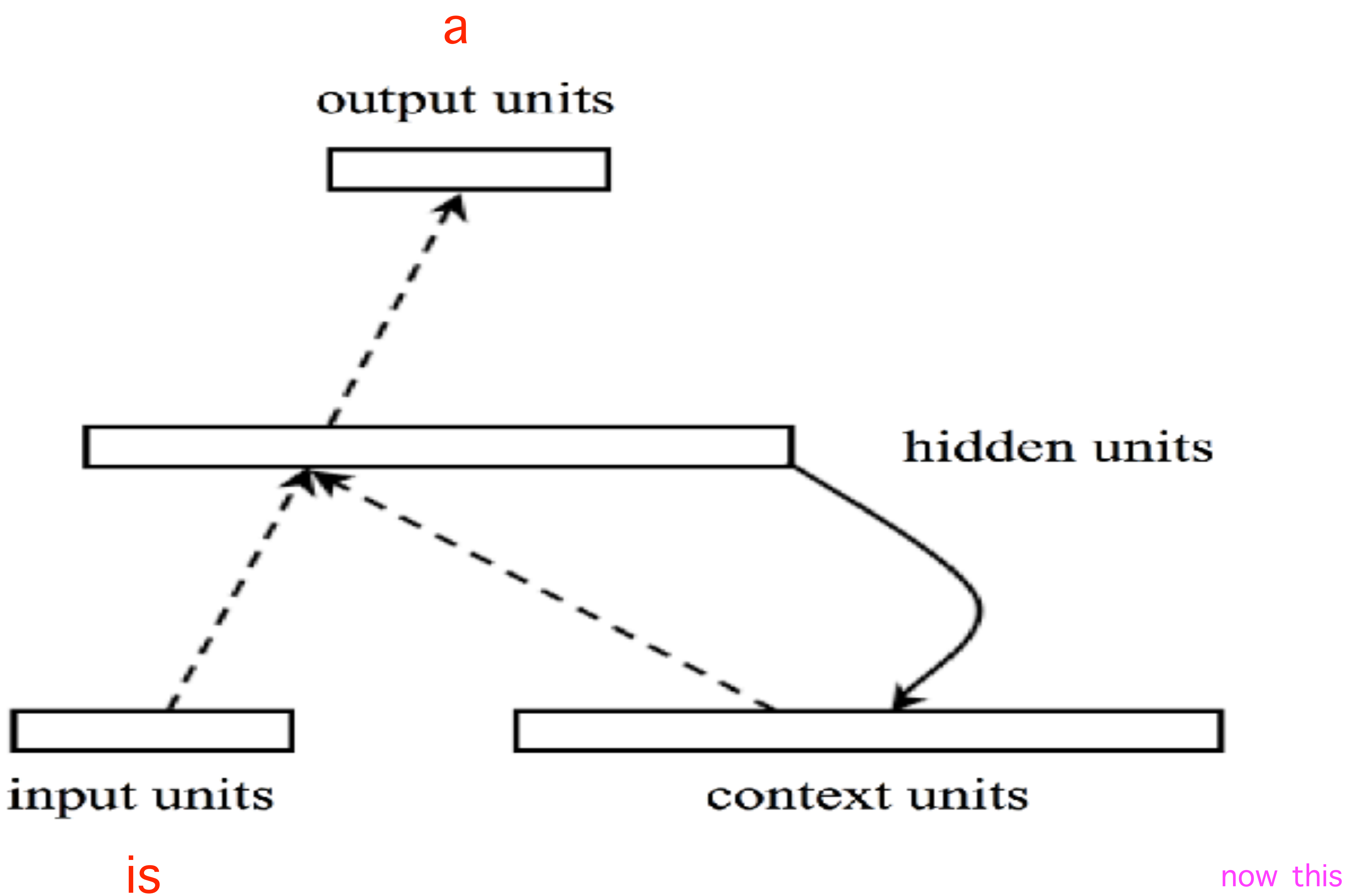

output units
hidden units
input units
context units

now

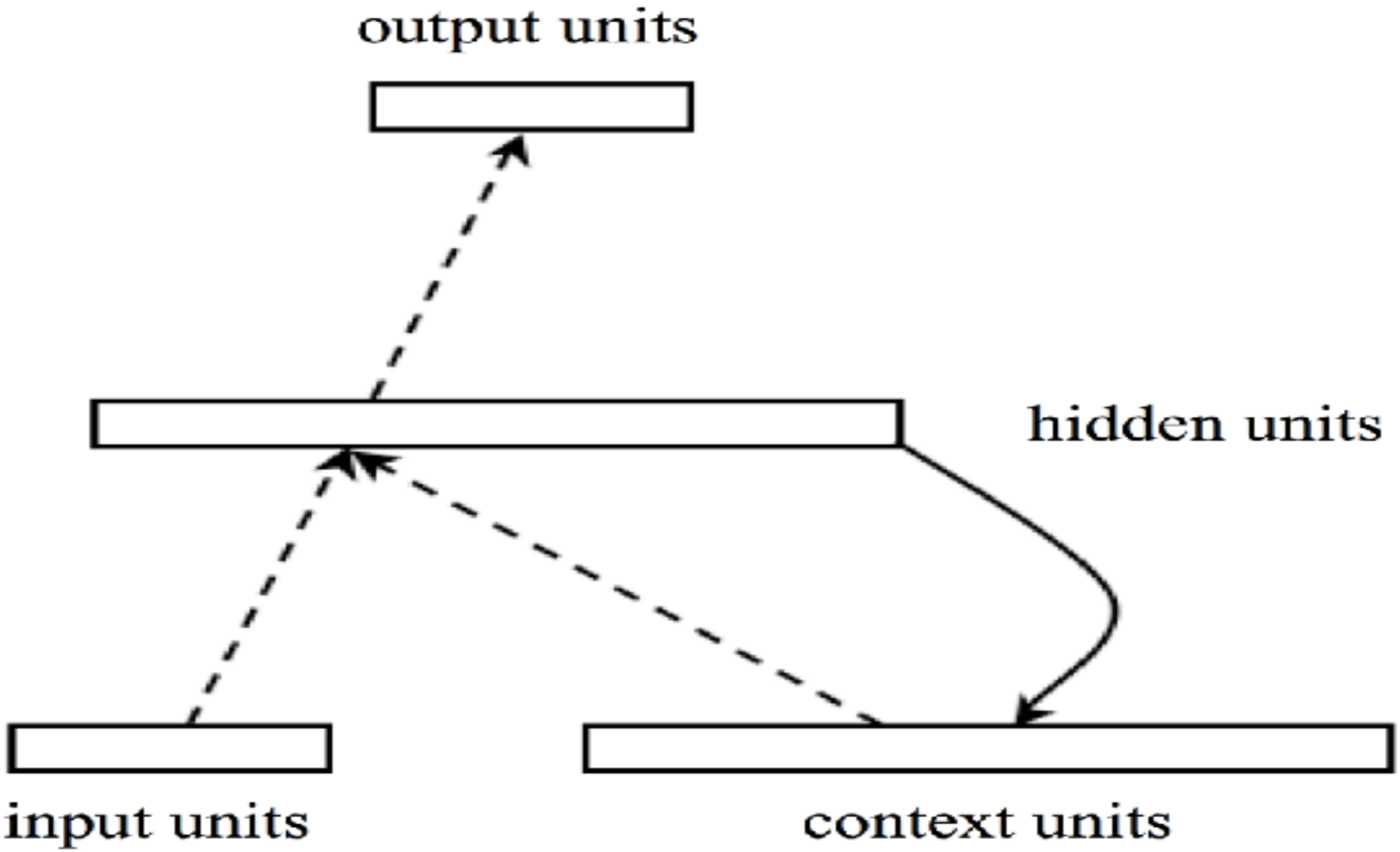



this
output units
hidden units
input units
context units
now


is
output units
hidden units
input units
context units
this
now


a
output units
hidden units
input units
context units
is
now this

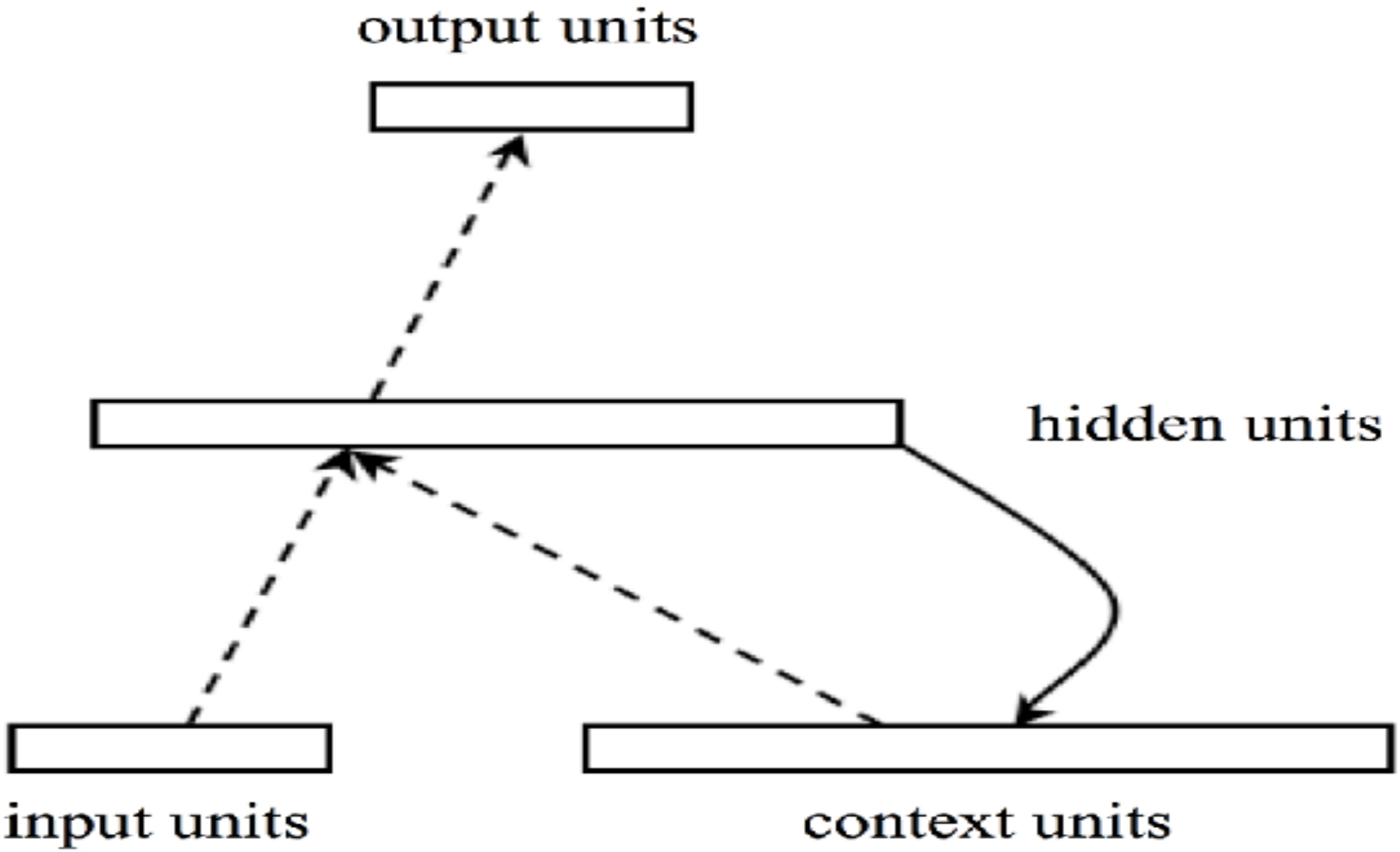
story



a

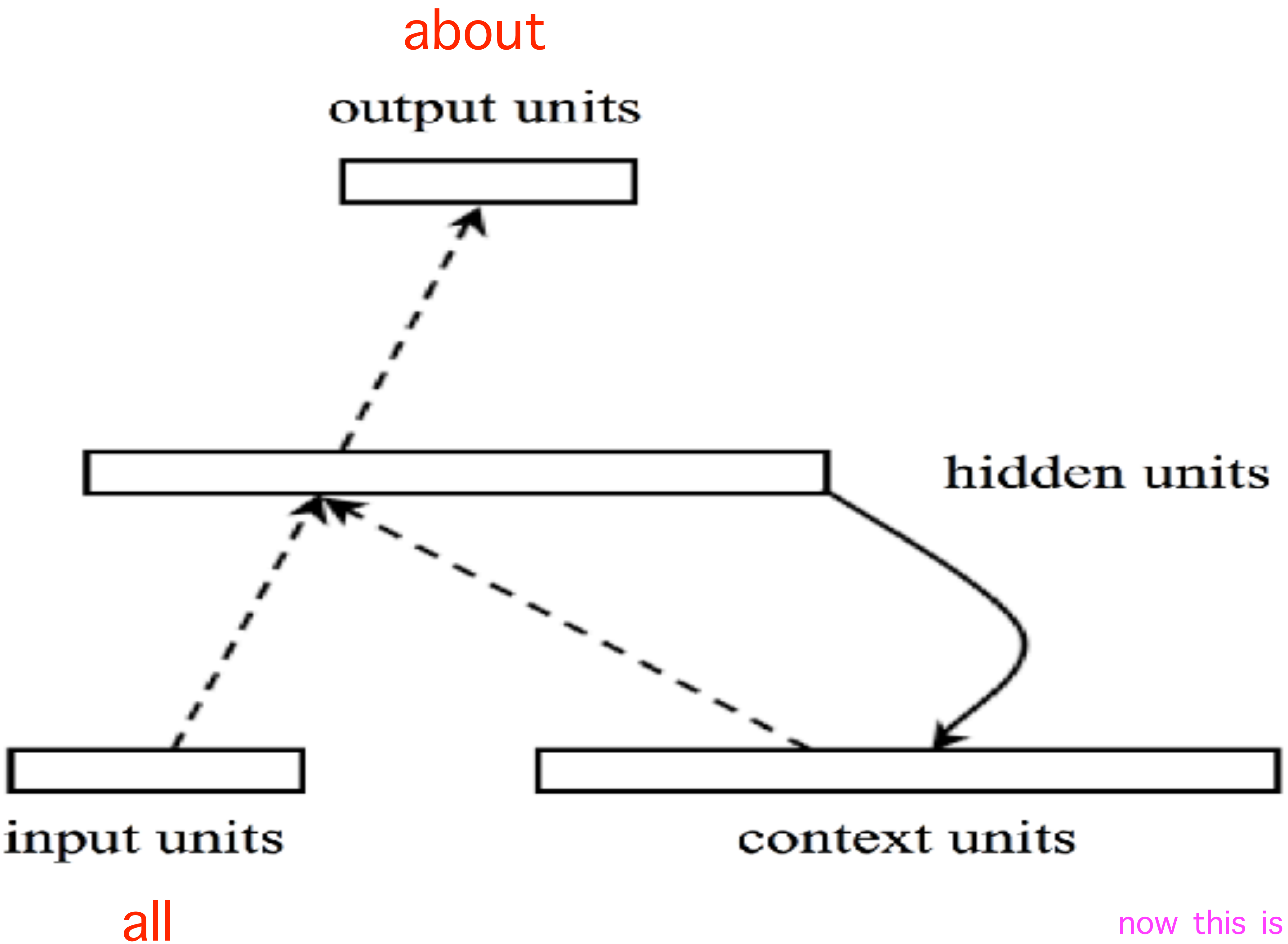
now this is

all



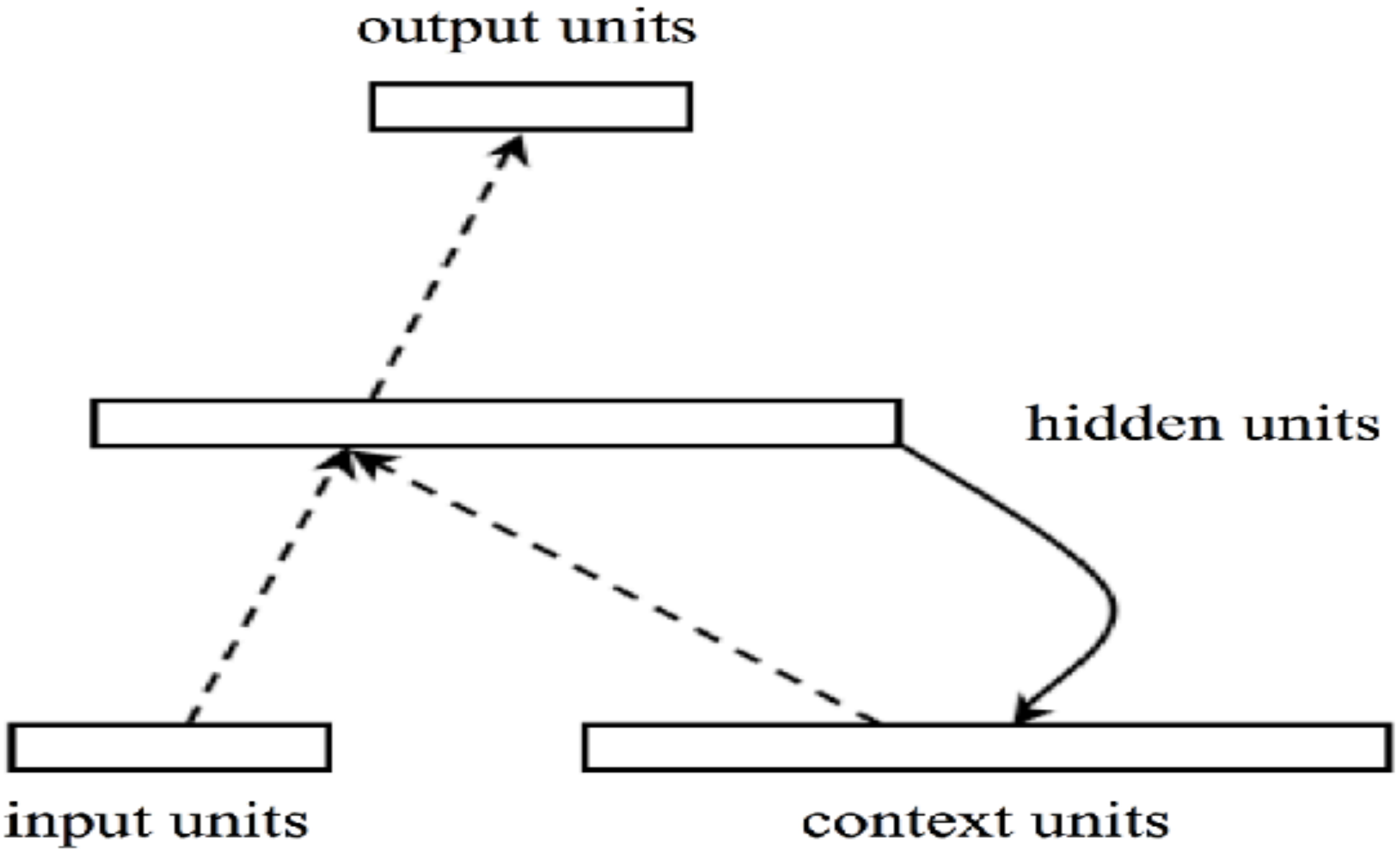
story

now this is a



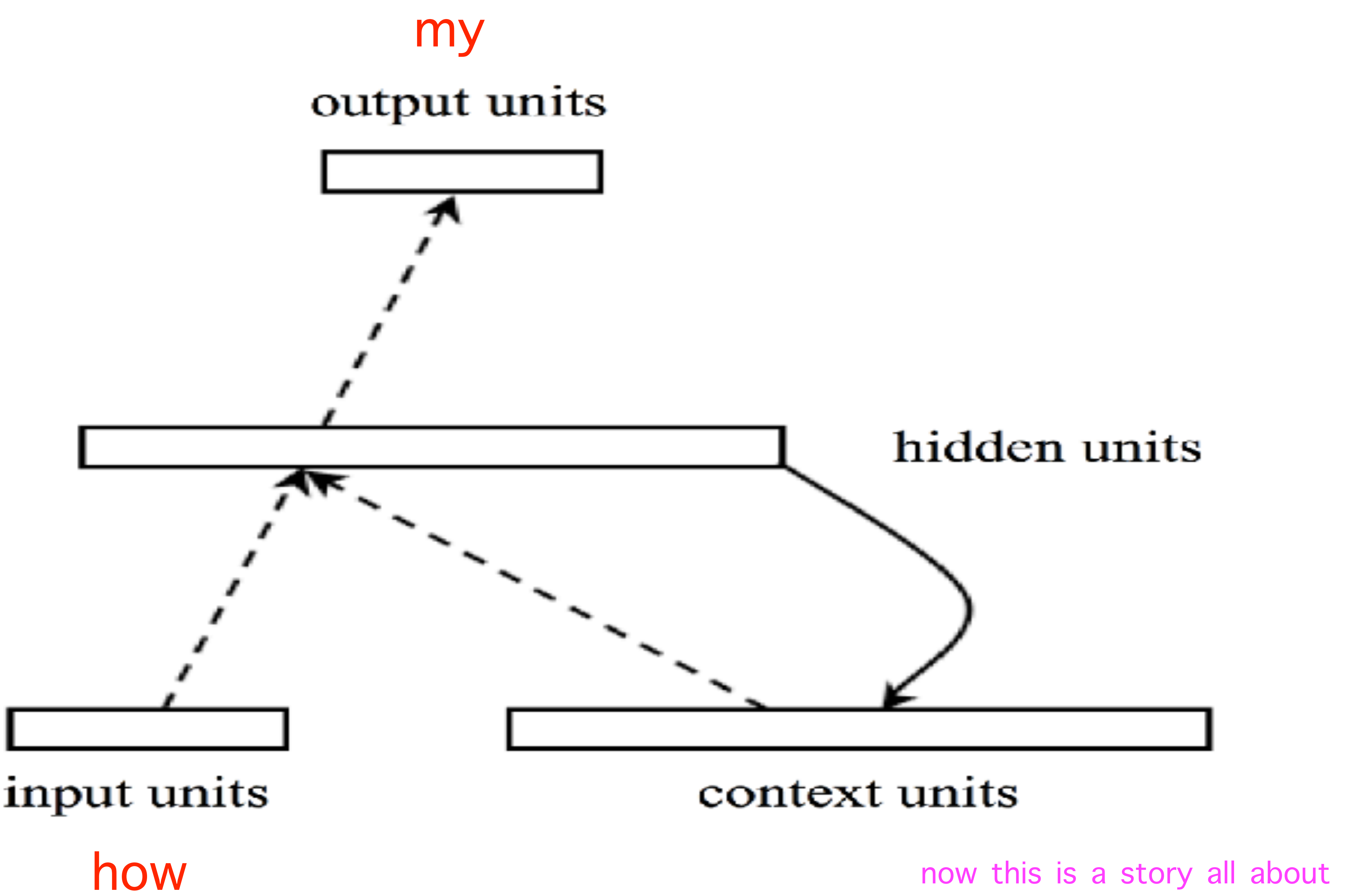
now this is a story

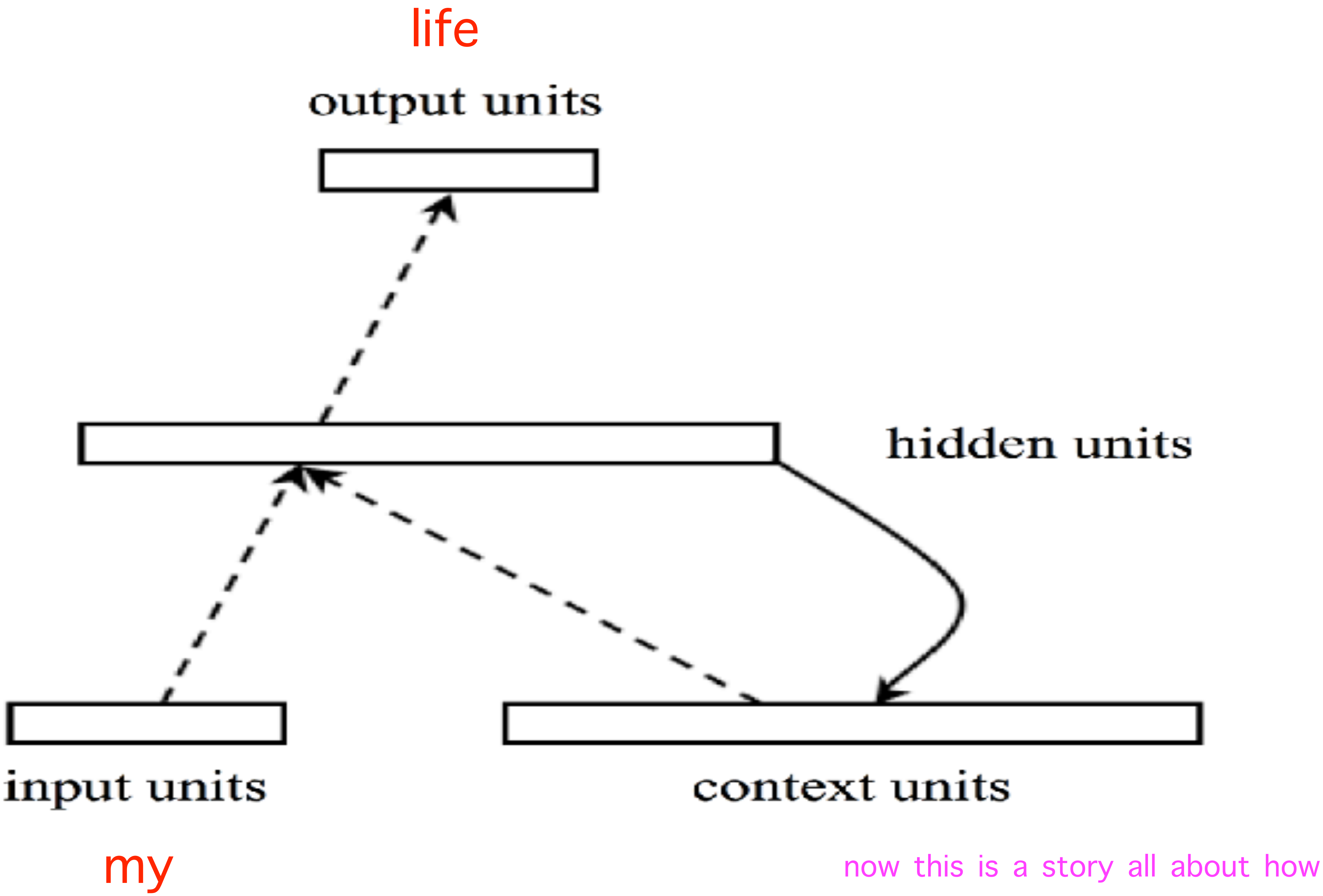
how

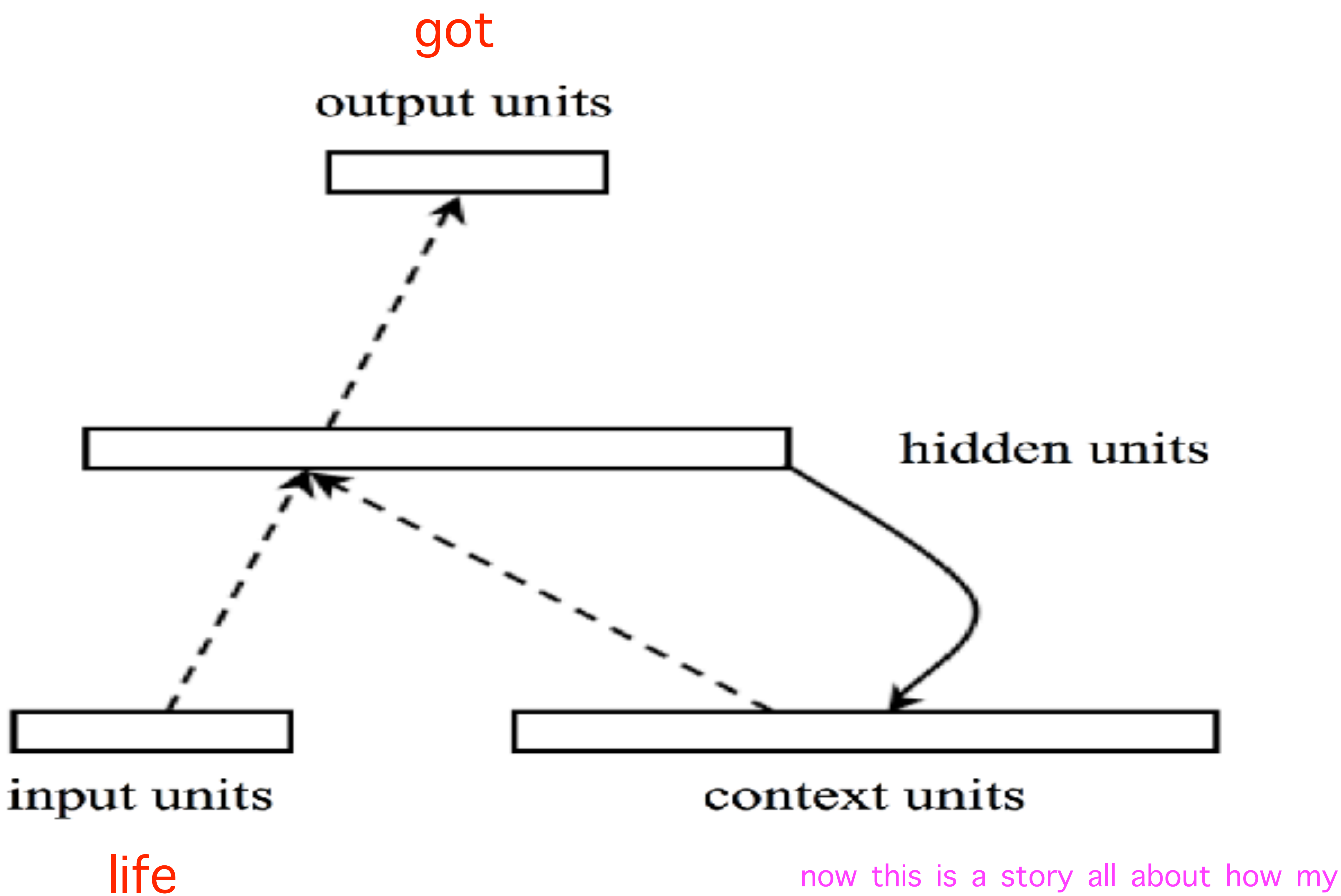


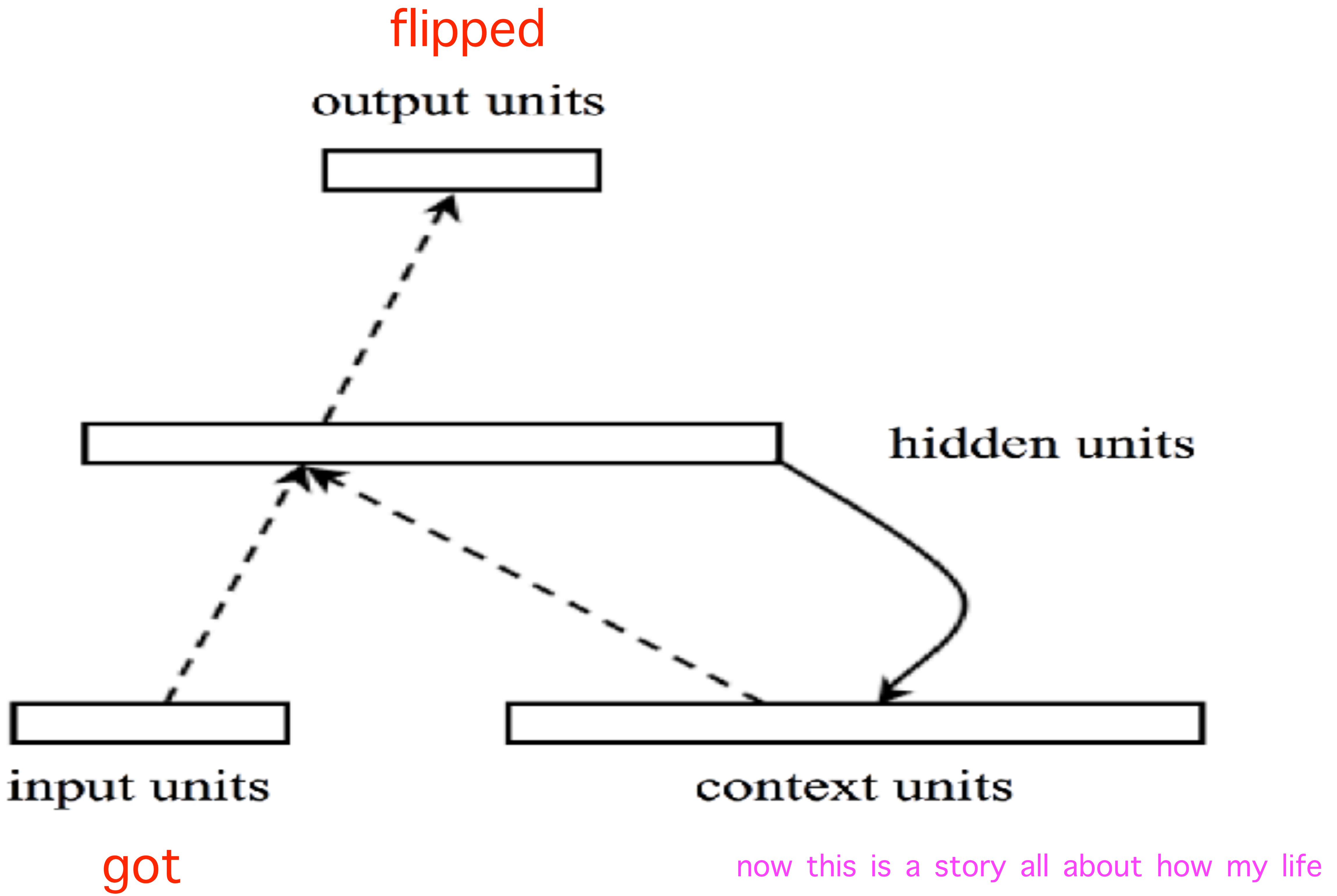
about

now this is a story all

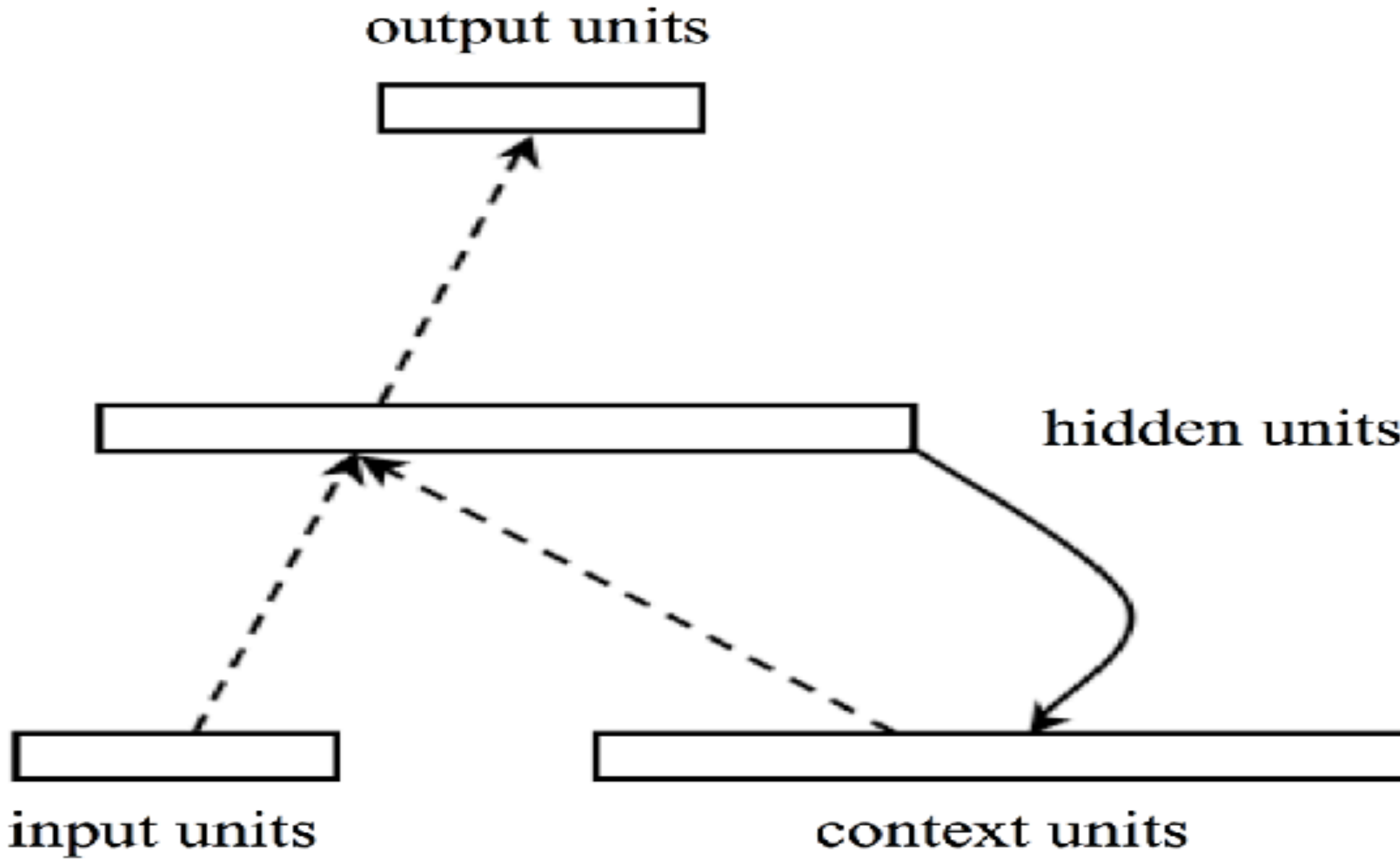

my
output units
hidden units
input units
context units
how
now this is a story all about


life
output units
hidden units
input units
context units
my
now this is a story all about how


got
output units
hidden units
input units
context units
life
now this is a story all about how my

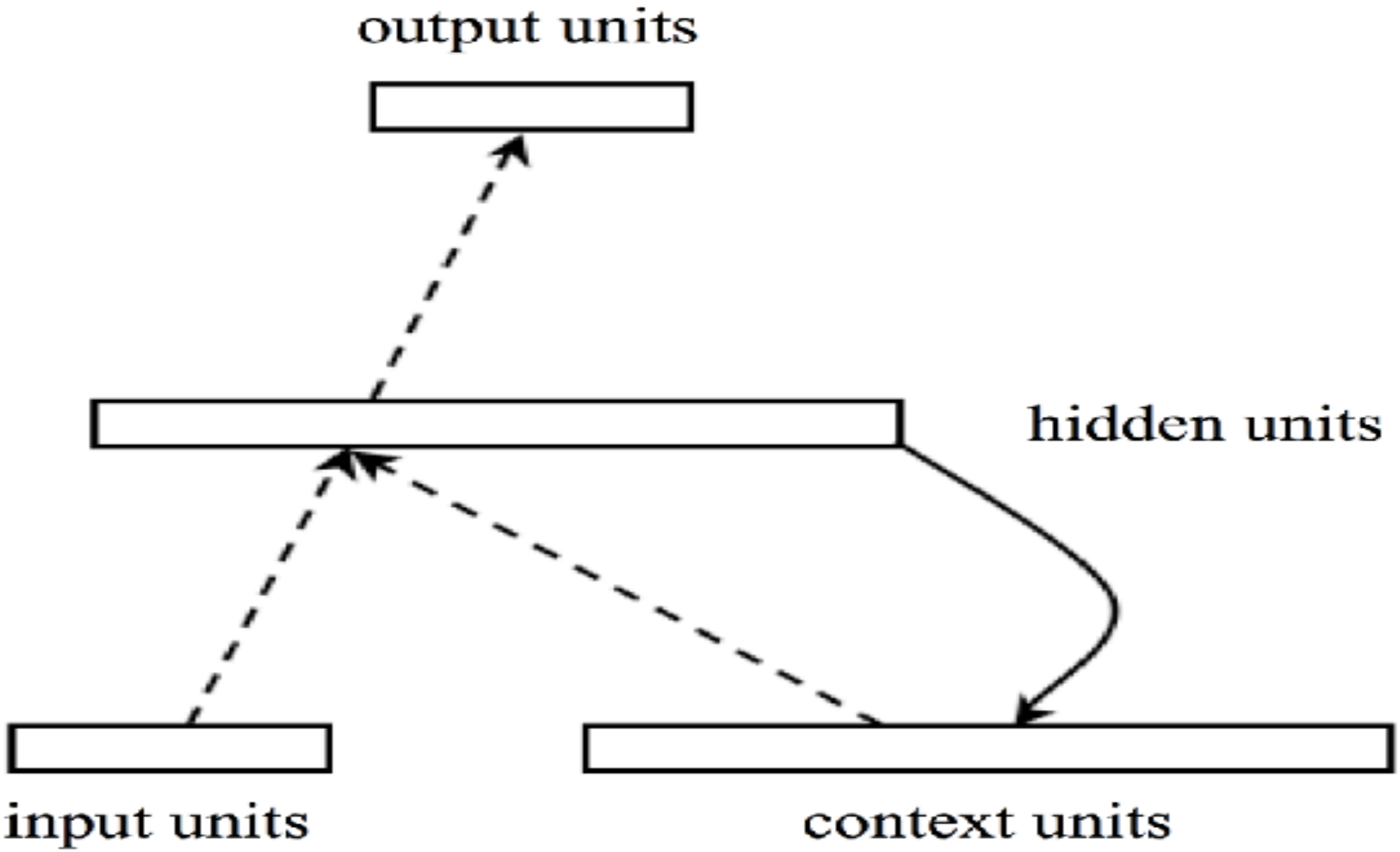

flipped
output units
hidden units
input units
context units
got
now this is a story all about how my life

turned



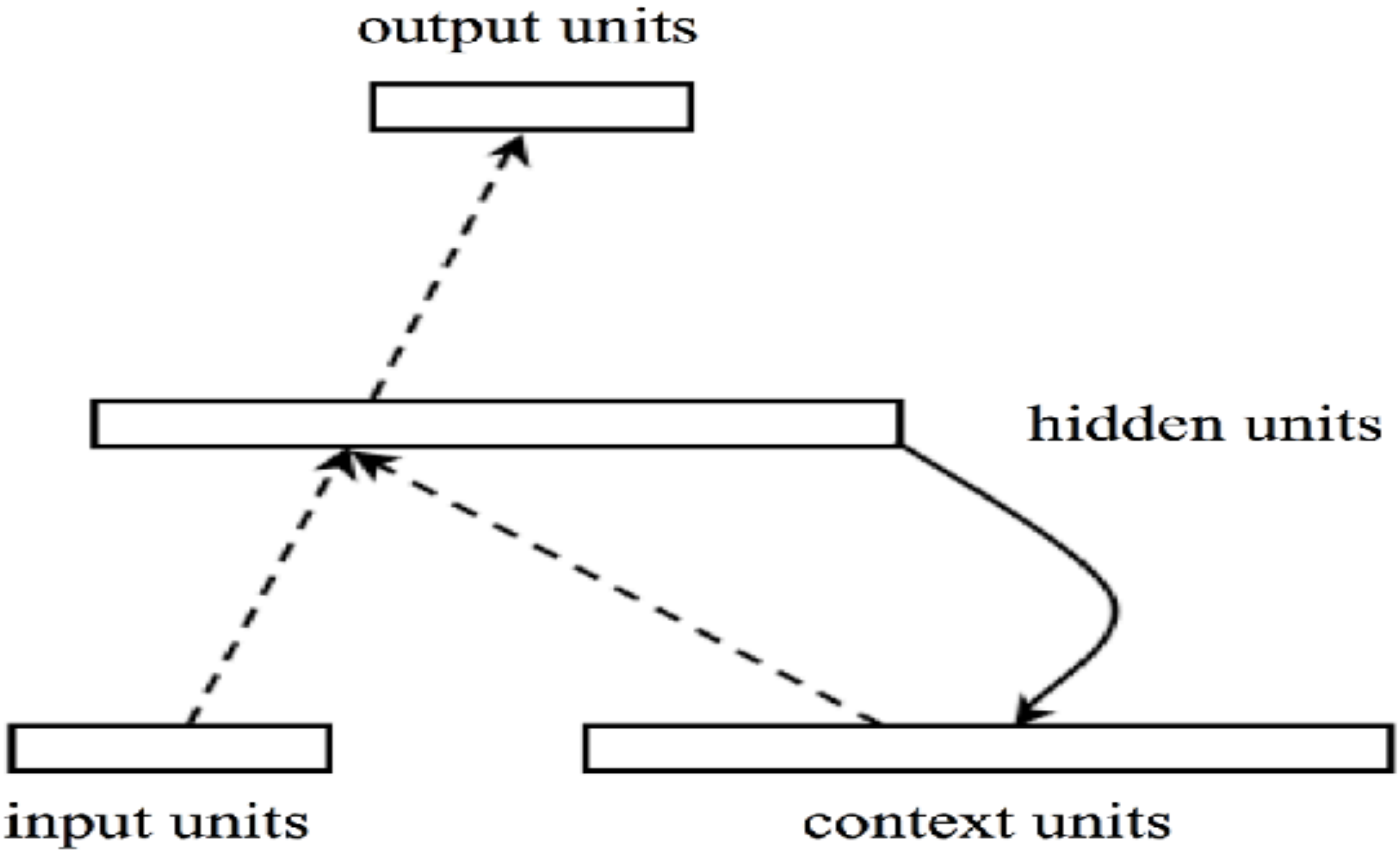
flipped

now this is a story all about how my life got upside



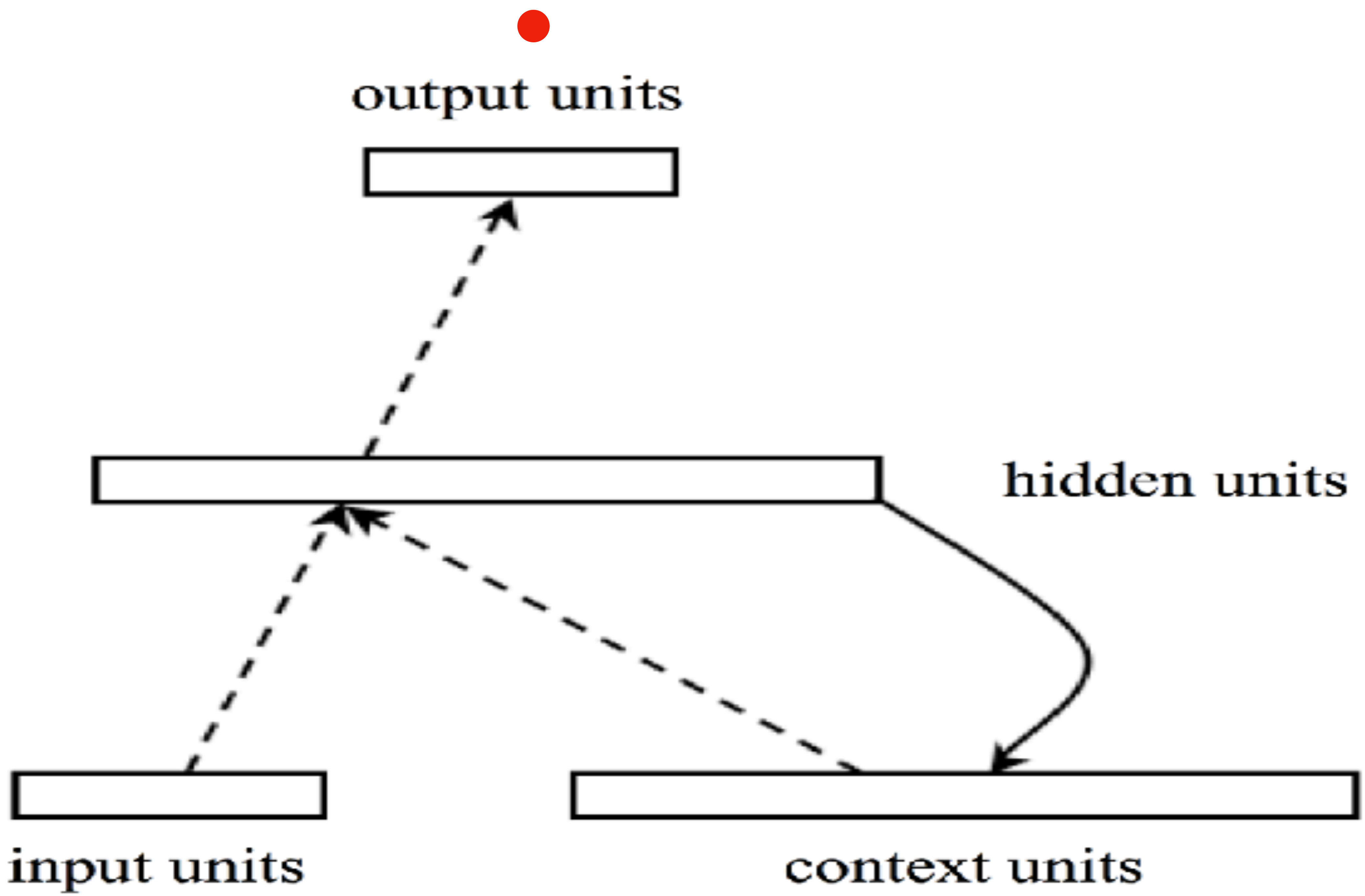
turned

now this is a story all about how my life got flipped down



upside

now this is a story all about how my life got flipped turned



down

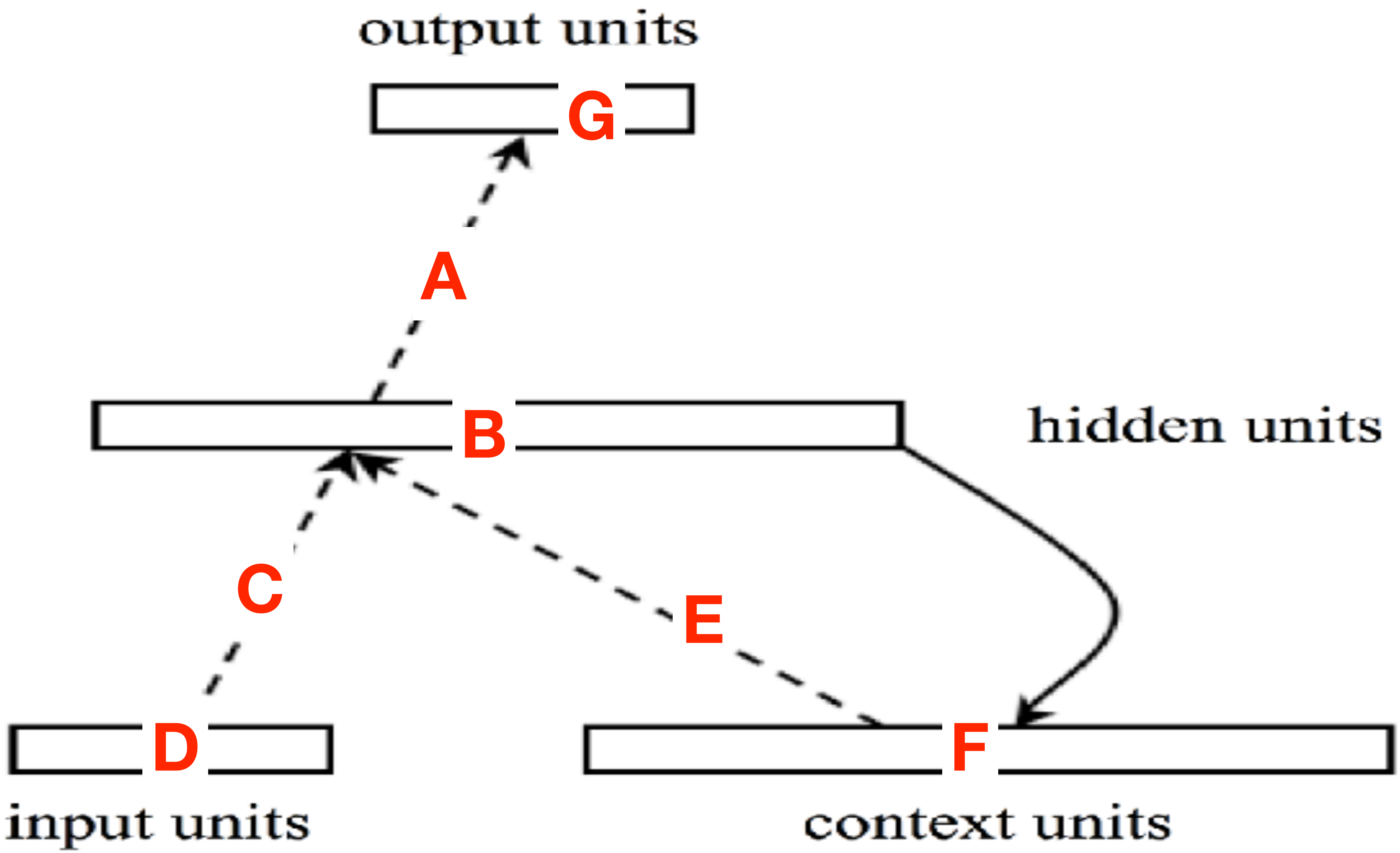
now this is a story all about how my life got flipped turned upside

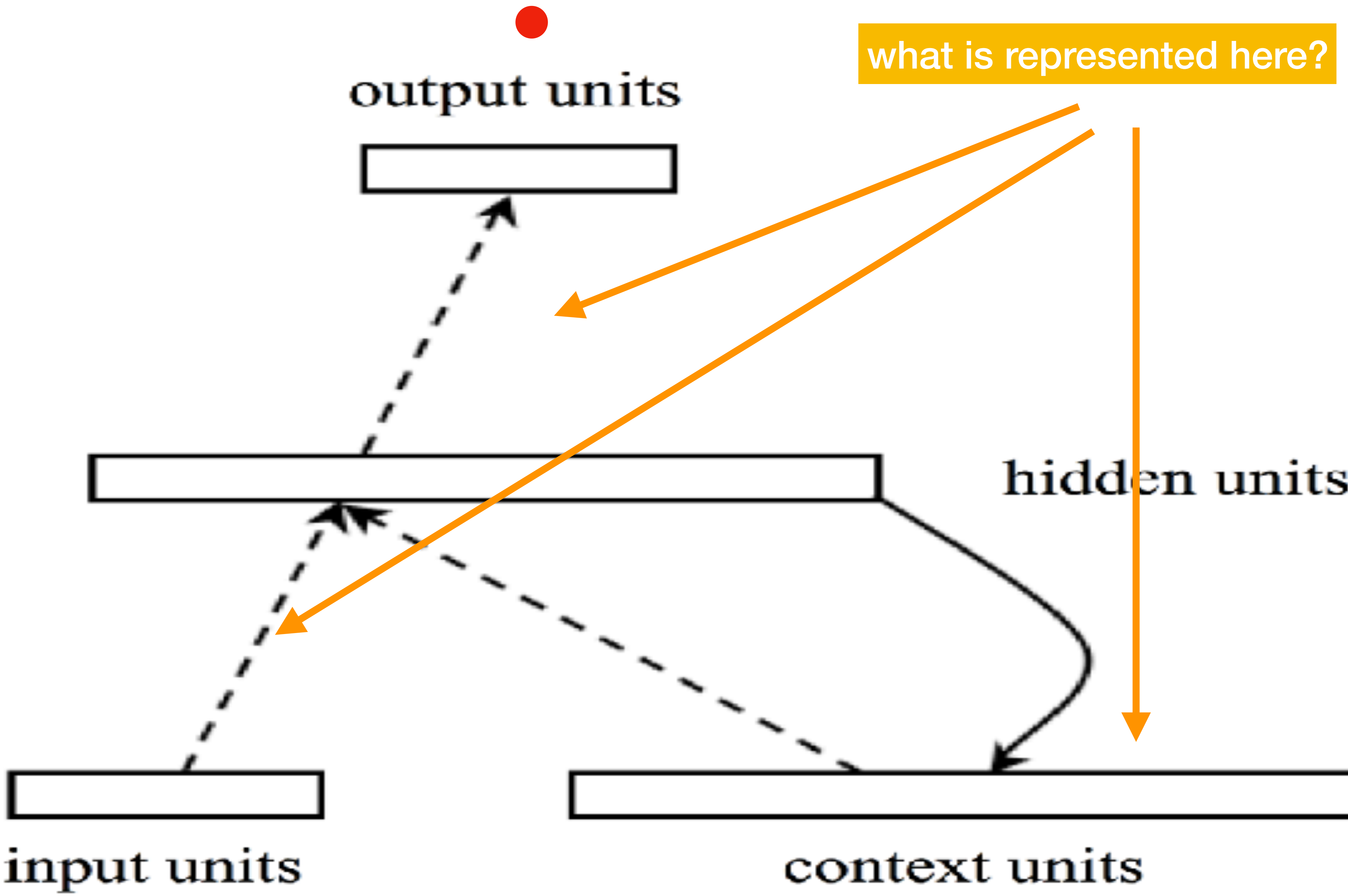
Suppose we have a vocabulary of 100k words.

How many weights are there in Elman’s network?

$$h_t = tanh(Uh_{t-1} + Wx_t)$$

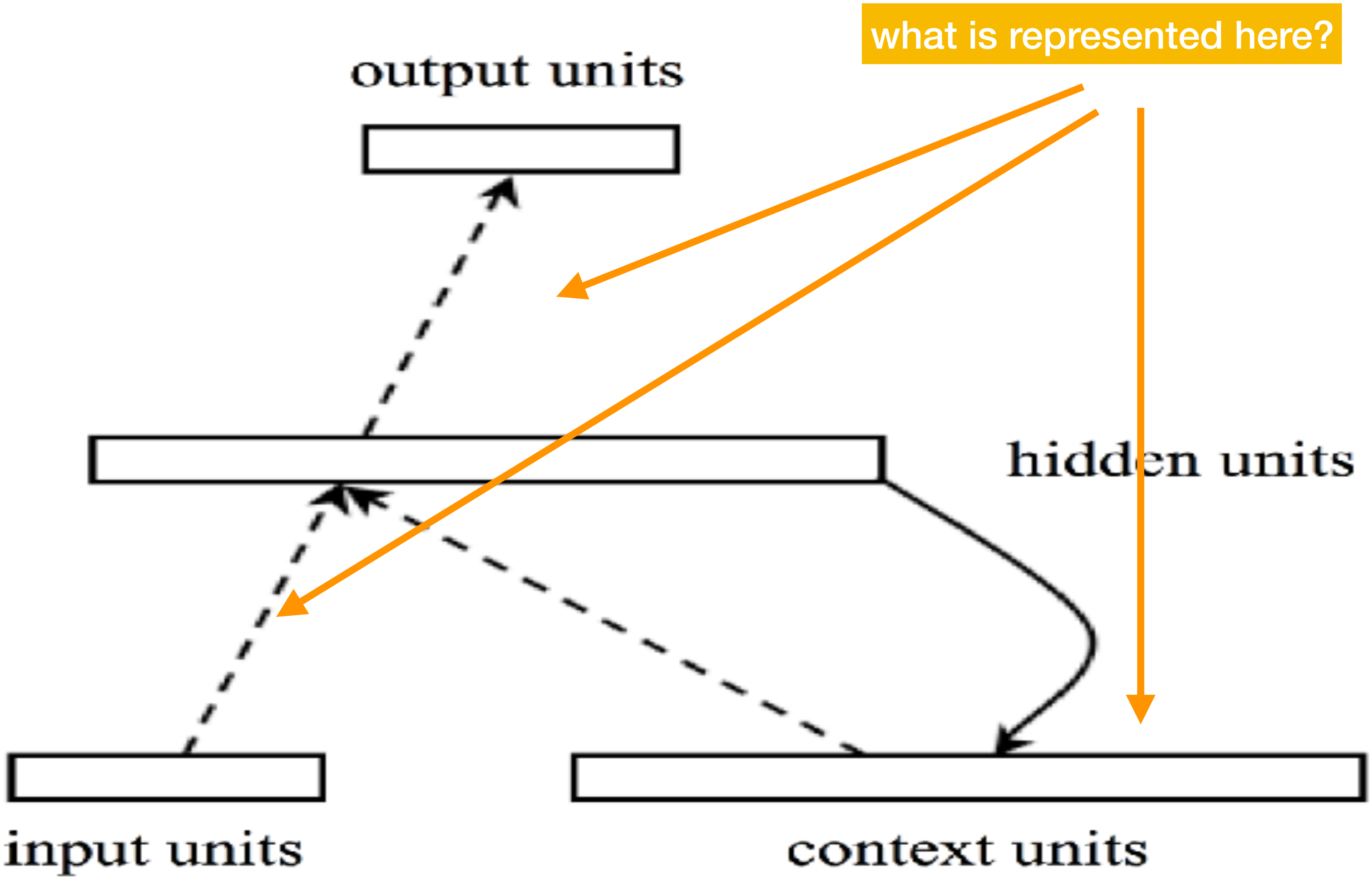
$$y_t = Vh_t$$





down

now this is a story all about how my life got flipped turned upside



now this is a story all about how my life got flipped turned upside dow

Finding structure in time

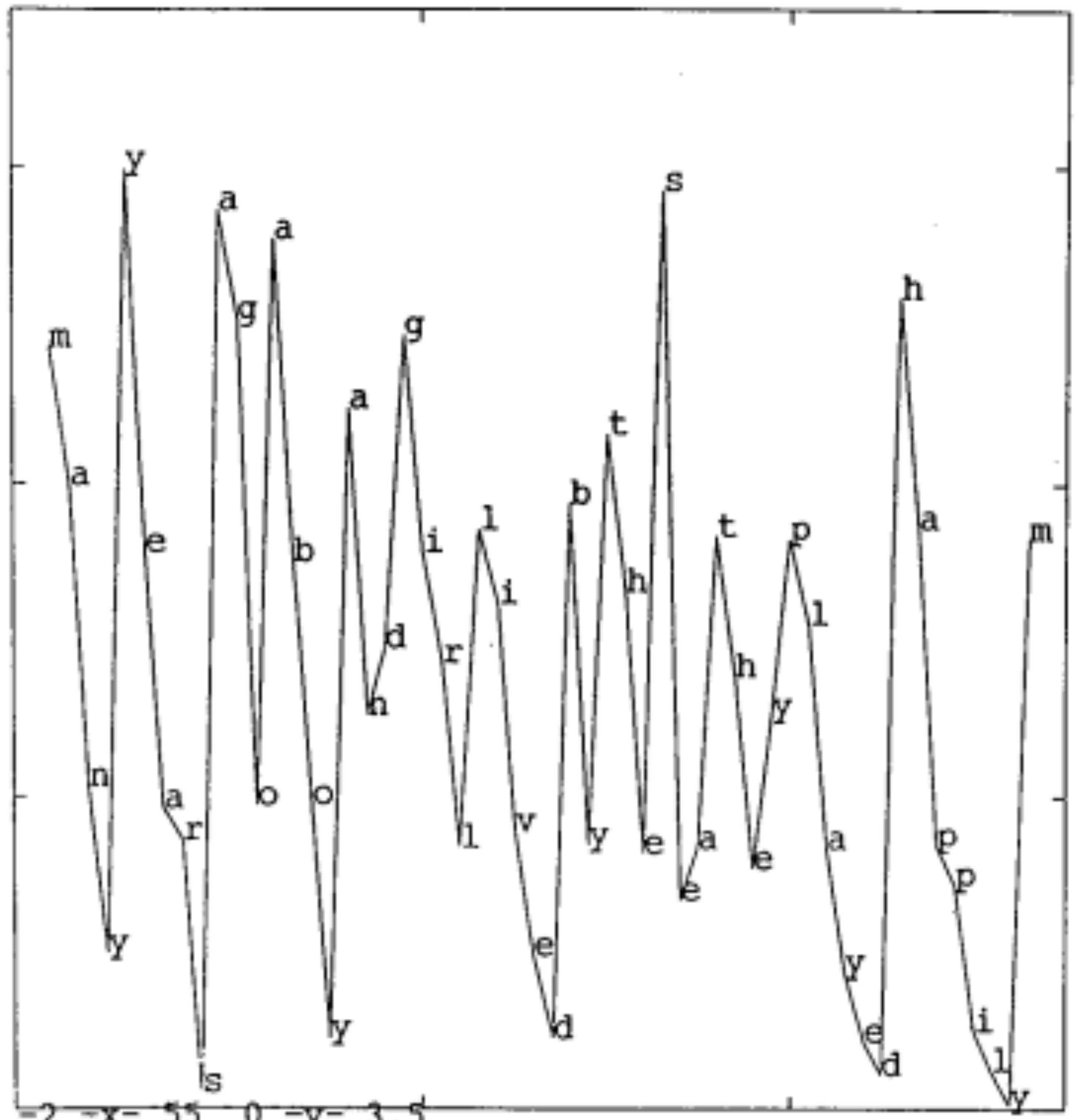


Figure 6. Graph of root mean squared error in letter-in-word precition task.

Finding more structure in time

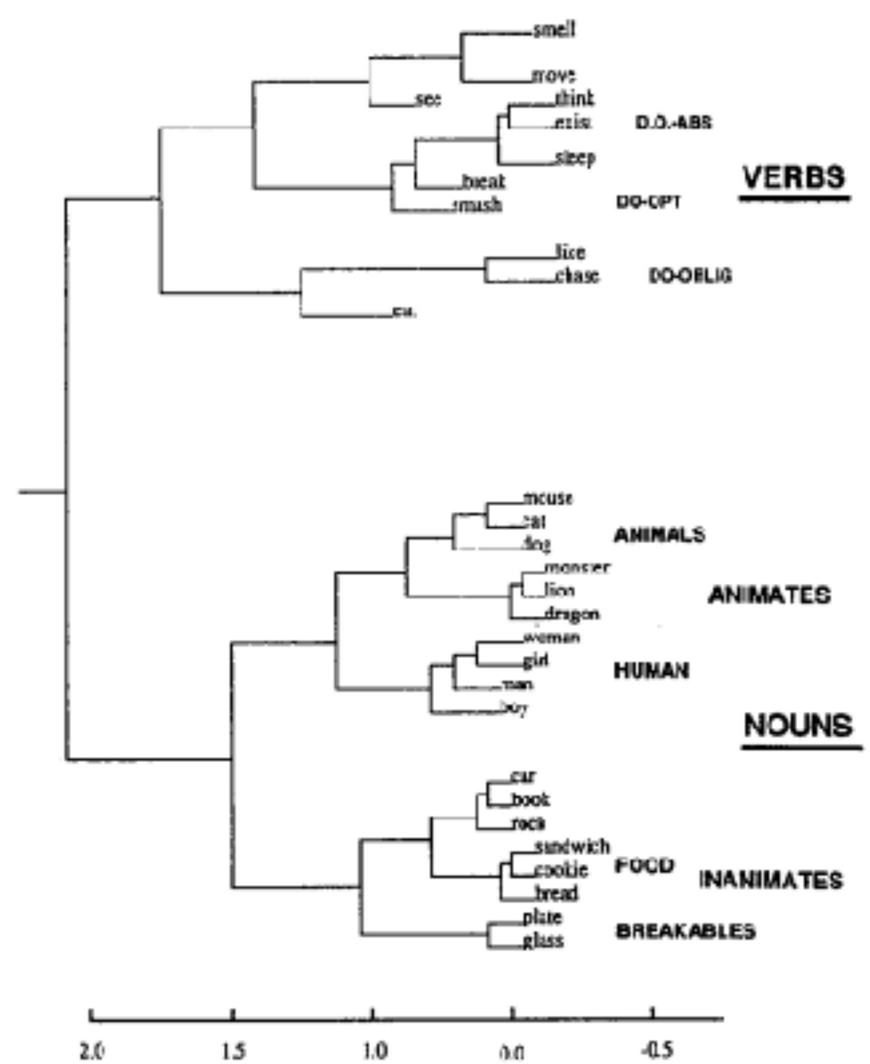
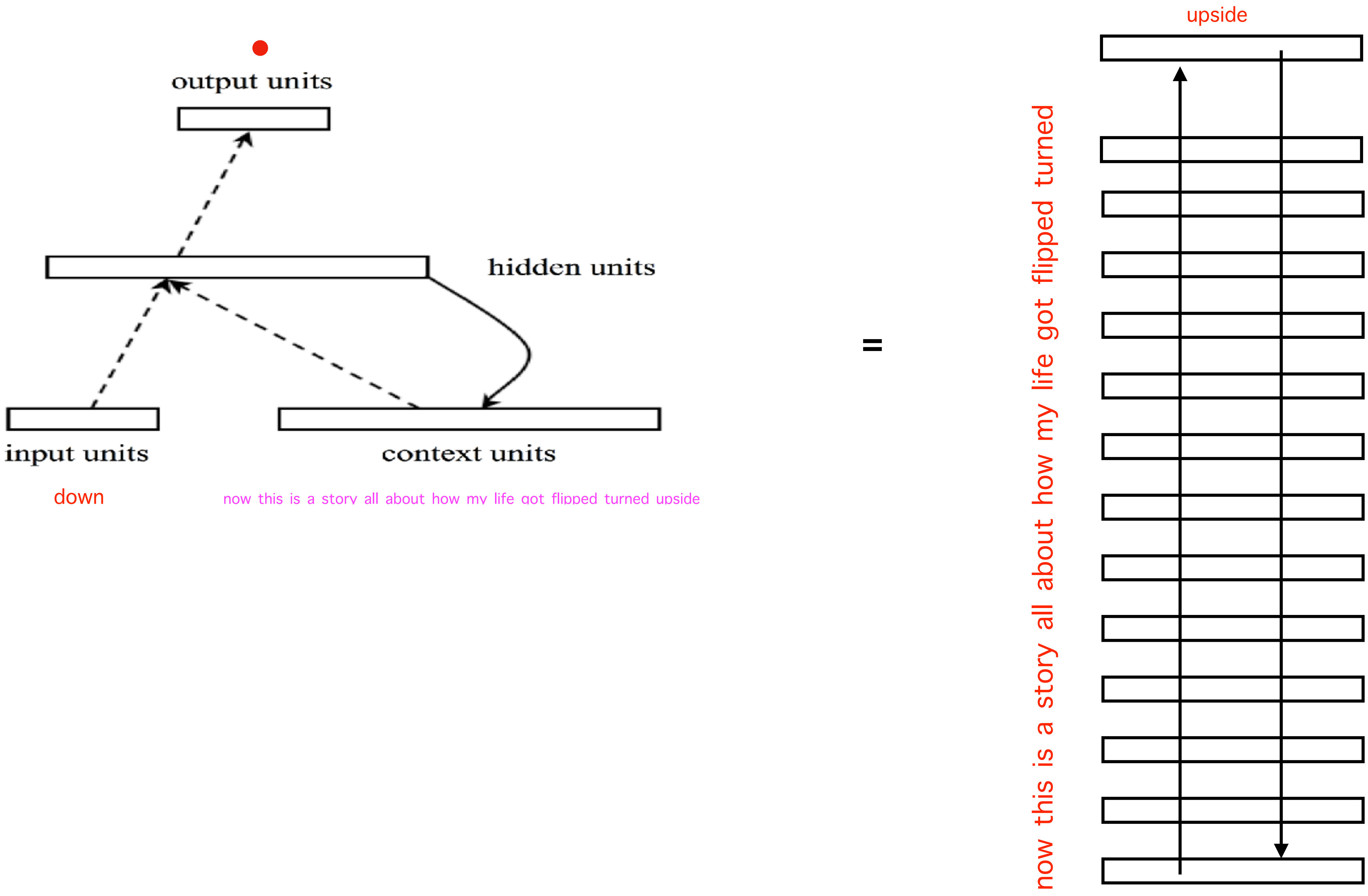


Figure 7. Hierarchical cluster diagram of hidden unit activation vectors in simple sentence prediction task. Labels indicate the inputs which produced the hidden unit vectors; inputs were presented in context, and the hidden unit vectors averaged across multiple contexts.

Any downsides?

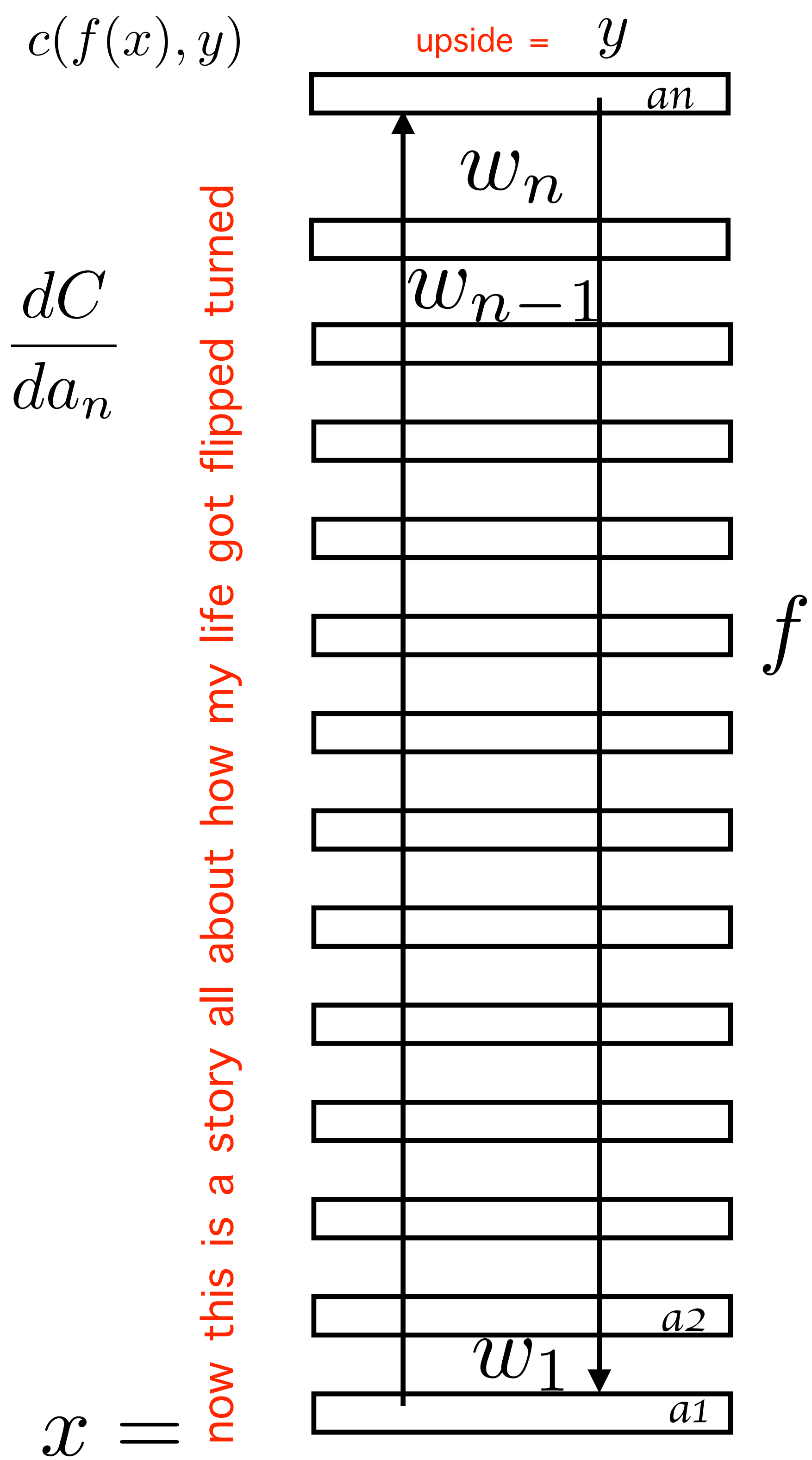


“Vanishing” gradients

$$\frac{dC}{dw_1} \propto \sigma'(z_1) \times w_2 \times \sigma'(z_2) \times w_3 \cdots \times w_n \times \sigma'(z_n) \times \frac{dC}{da_n}$$

where

$$a_i = \sigma(z_i)$$



“Vanishing” gradients

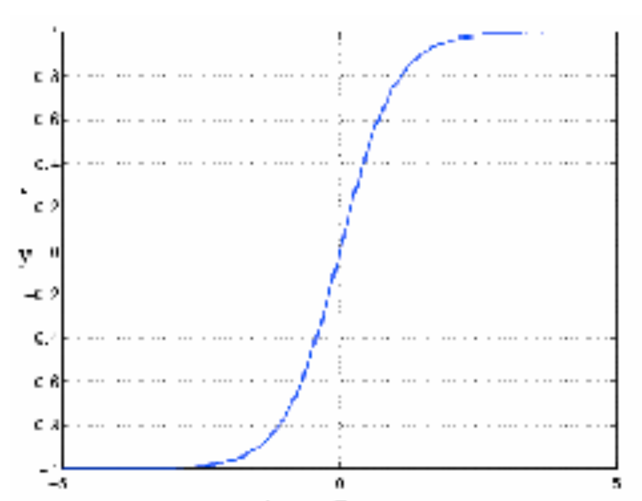
(or exploding)

in an RNN

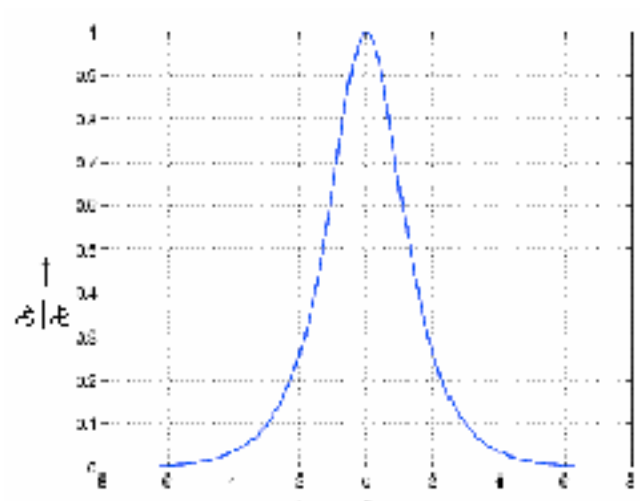
$$\frac{dC}{dw_1} \propto w^n \times \sigma'(z_1) \times \cdots \times \sigma'(z_n) \times \frac{dC}{da_n}$$

small change, big consequences

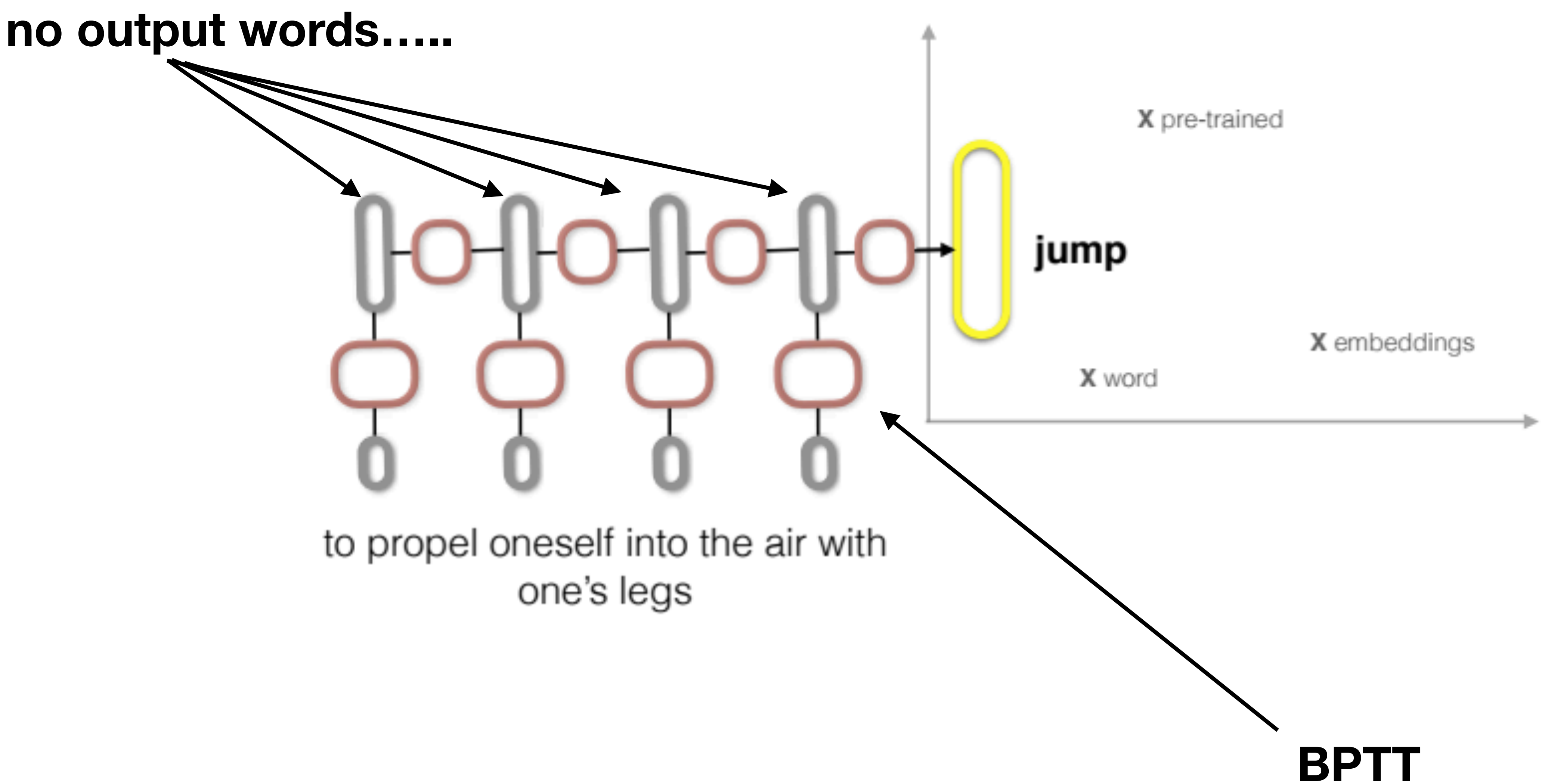
$\sigma(x)$



$\sigma'(x)$



One final thing…



But, more typically…

http://www.cs.toronto.edu/~ilya/rnn.html

References

Finding structure in time (Elman, 1990)

Description and analysis of a recurrent neural network, inference of structure in unsegmented sequences

Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks (Graves et al, 2006)

Scales Elman up to the ML age

Recurrent neural network-based language model (Mikolov et al. 2010)

Scale Graves up to running text

Learning to understand phrases by embedding the dictionary (Hill et al. 2015)

Learns to predict words from dictionary definitions

DeepMind

UNIVERSITY OF CAMBRIDGE