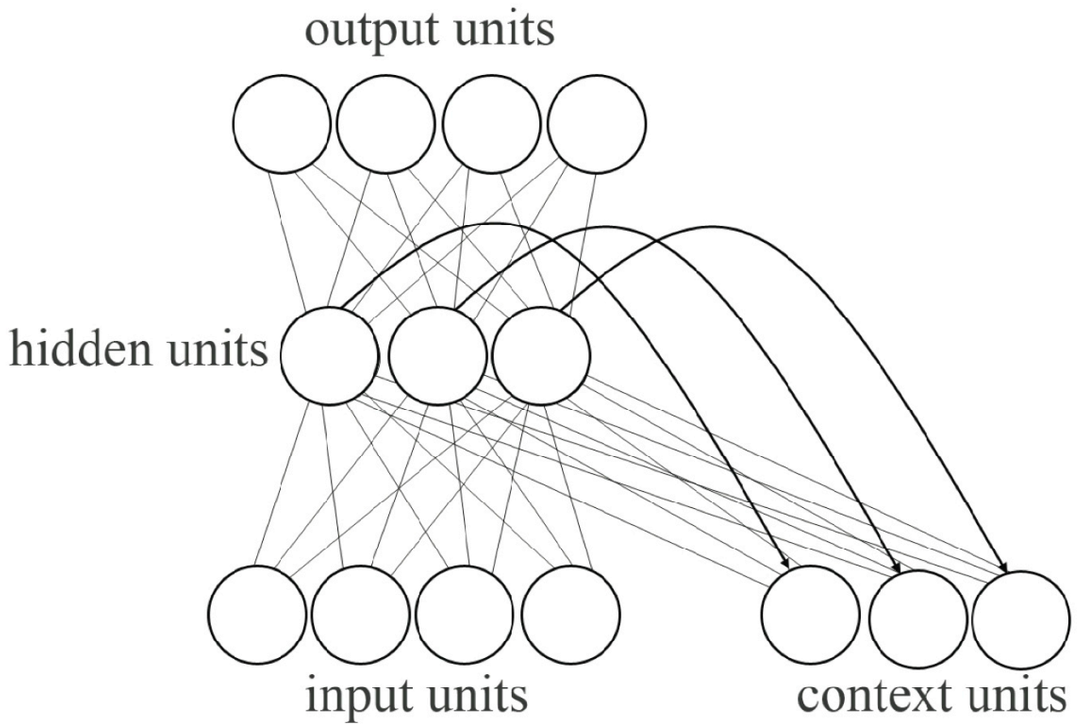
Computational Cognitive Modeling

Neural networks and deep learning (part 2)

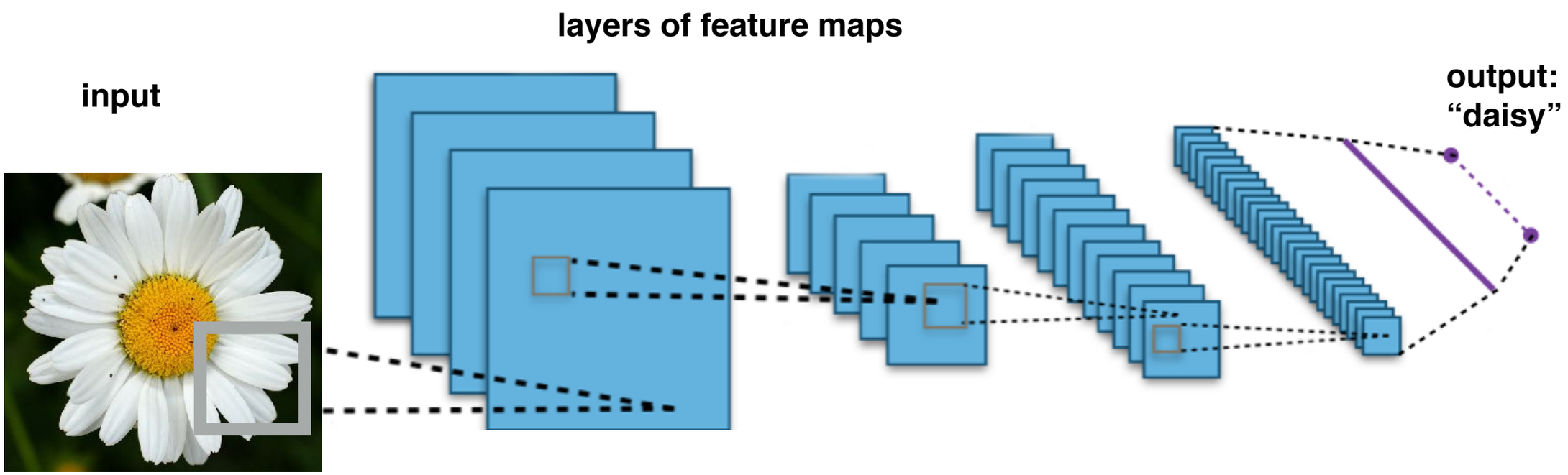
Brenden Lake & Todd Gureckis

Two very important types of neural networks

Recurrent neural networks (RNNs)



Convolutional neural networks (ConvNets)

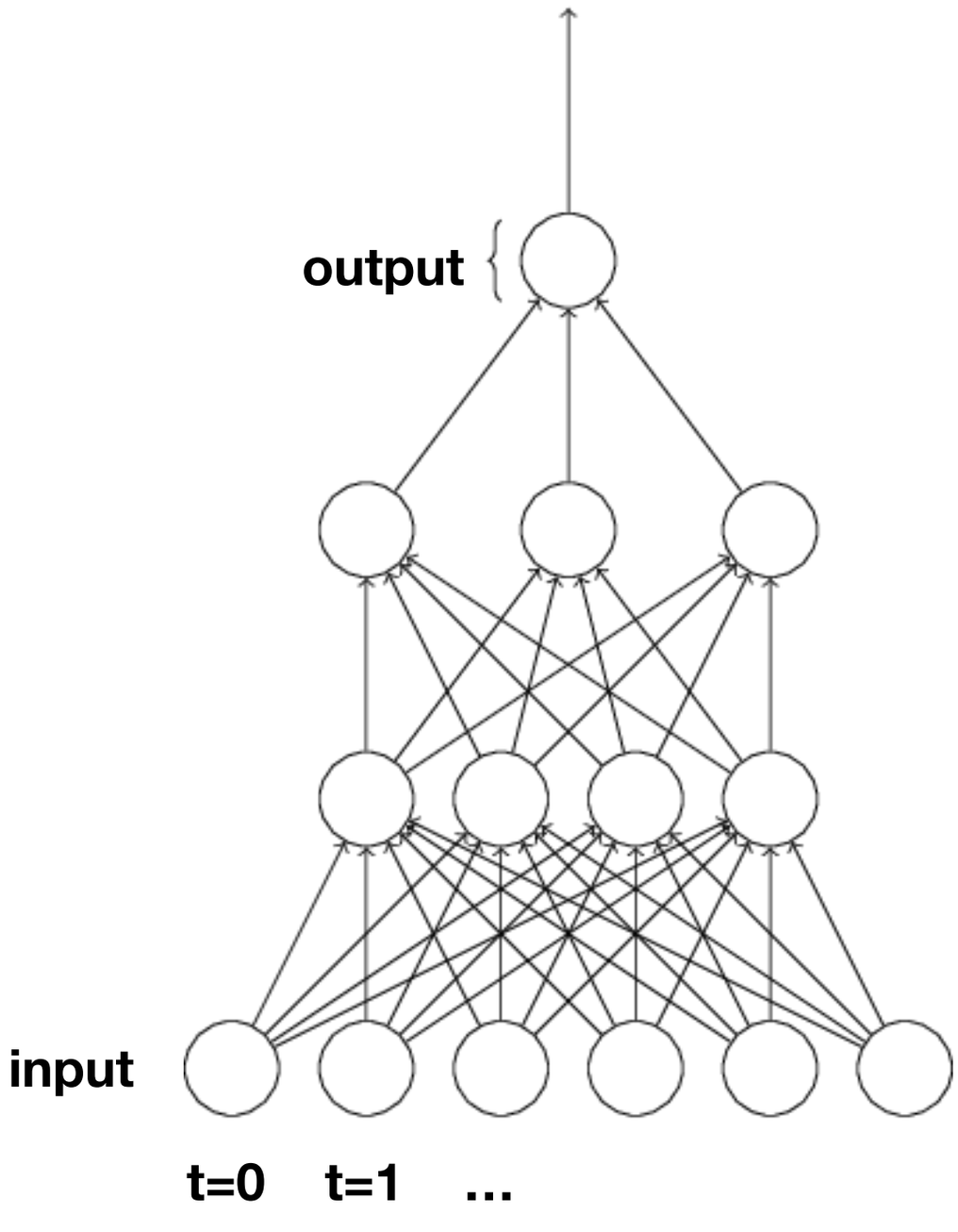


Data often has important temporal structure

- Language/text
- Speech
- Video
- Financial
- Weather
- All of human behavior unfolds in time

How do we represent time in a neural network?

Naive approach: represent time spatially in a standard network



Problems with this approach:

- The network has a fixed buffer. How large do you make the buffer?
- Two identical patterns translated in time, such as **[0 1 1 0 0 0], [0 0 0 1 1 0]** have no natural overlap in the (untrained) architecture

COGNITIVE SCIENCE **14,** 179–211 (1990)

Finding Structure in Time

JEFFREY L. ELMAN
University of California, San Diego

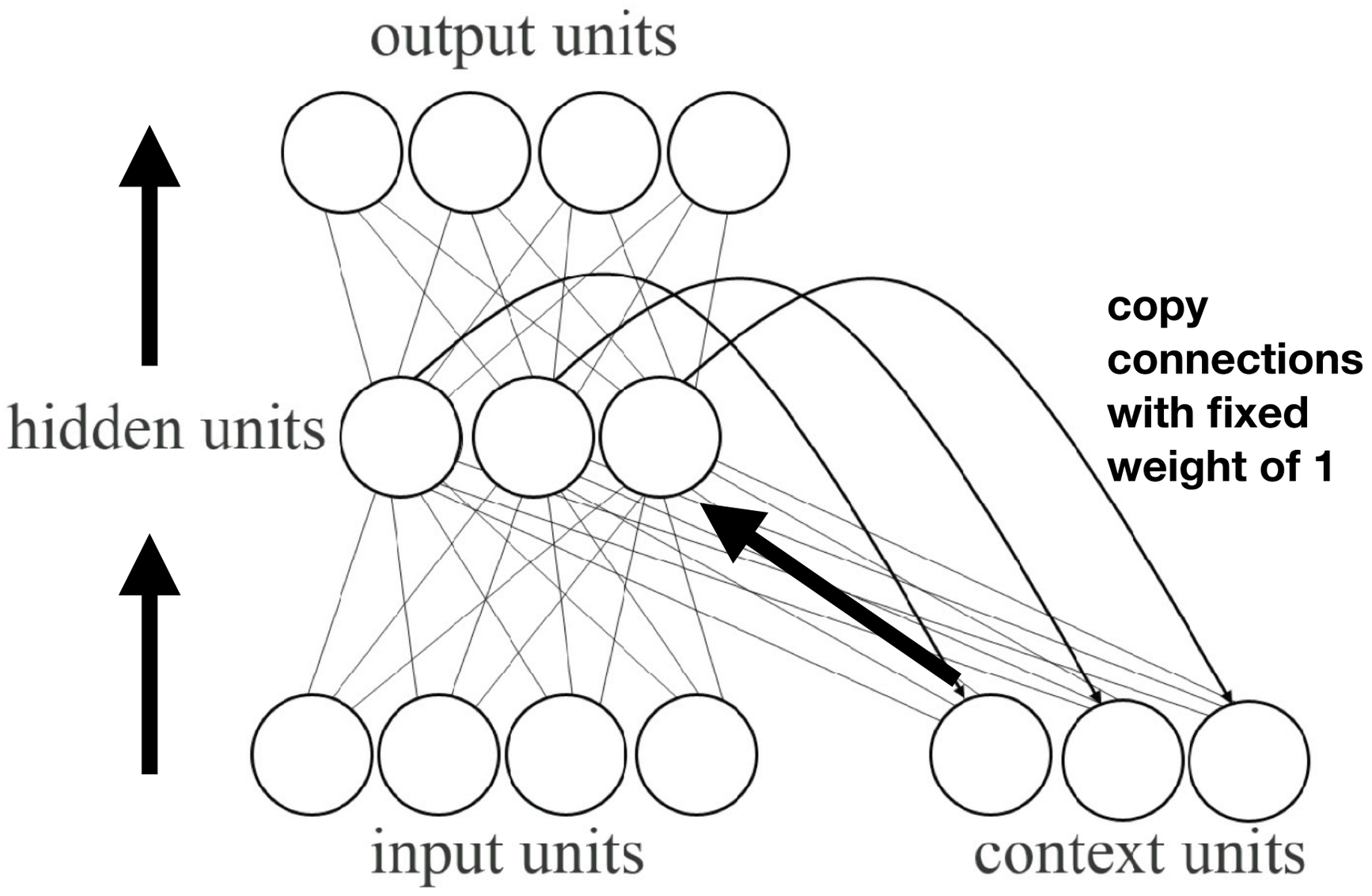
Time underlies many interesting human behaviors. Thus, the question of how to represent time in connectionist models is very important. One approach is to represent time implicitly by its effects on processing rather than explicitly (as in a spatial representation). The current report develops a proposal 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

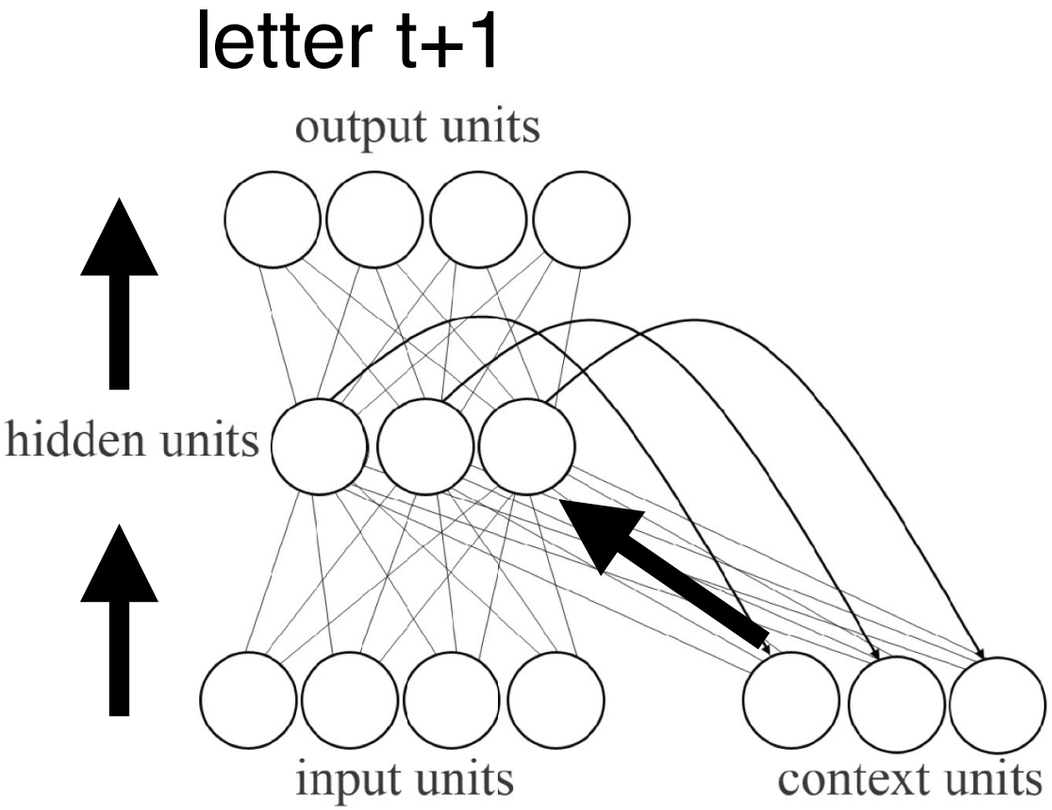
Time is clearly important in cognition. It is inextricably bound up with many behaviors (such as language) which express themselves as temporal sequences. Indeed, it is difficult to know how one might deal with such basic problems as goal-directed behavior, planning, or causation without some way of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

Simple recurrent network (SRN; or Elman network)



A simple example task



input data:

diibaguuubadiidiiguuuguuudiidiibadii….
(random mix of “ba”, “dii”, and “guuu”)

task:

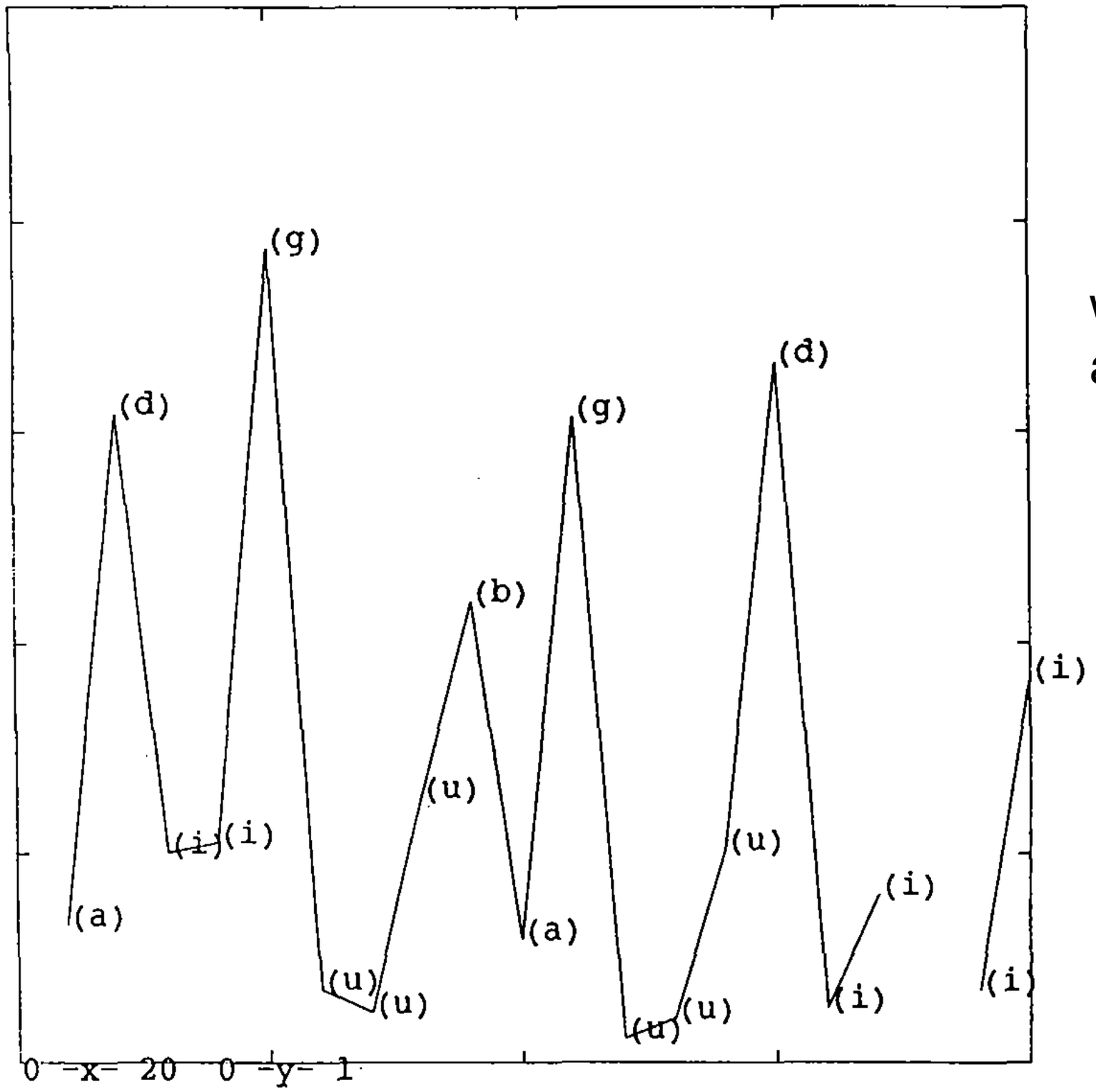
predict the next letter

representation of input

	Consonant	Vowel	Interrupted	High	Back	Voiced
b	1	0	1	0	0	1
d	1	0	1	1	0	1
g	1	0	1	0	1	1
a	0	1	0	0	1	1
i	0	1	0	1	0	1
u	0	1	0	1	1	1

Performance of the trained network

error (root mean squared error)



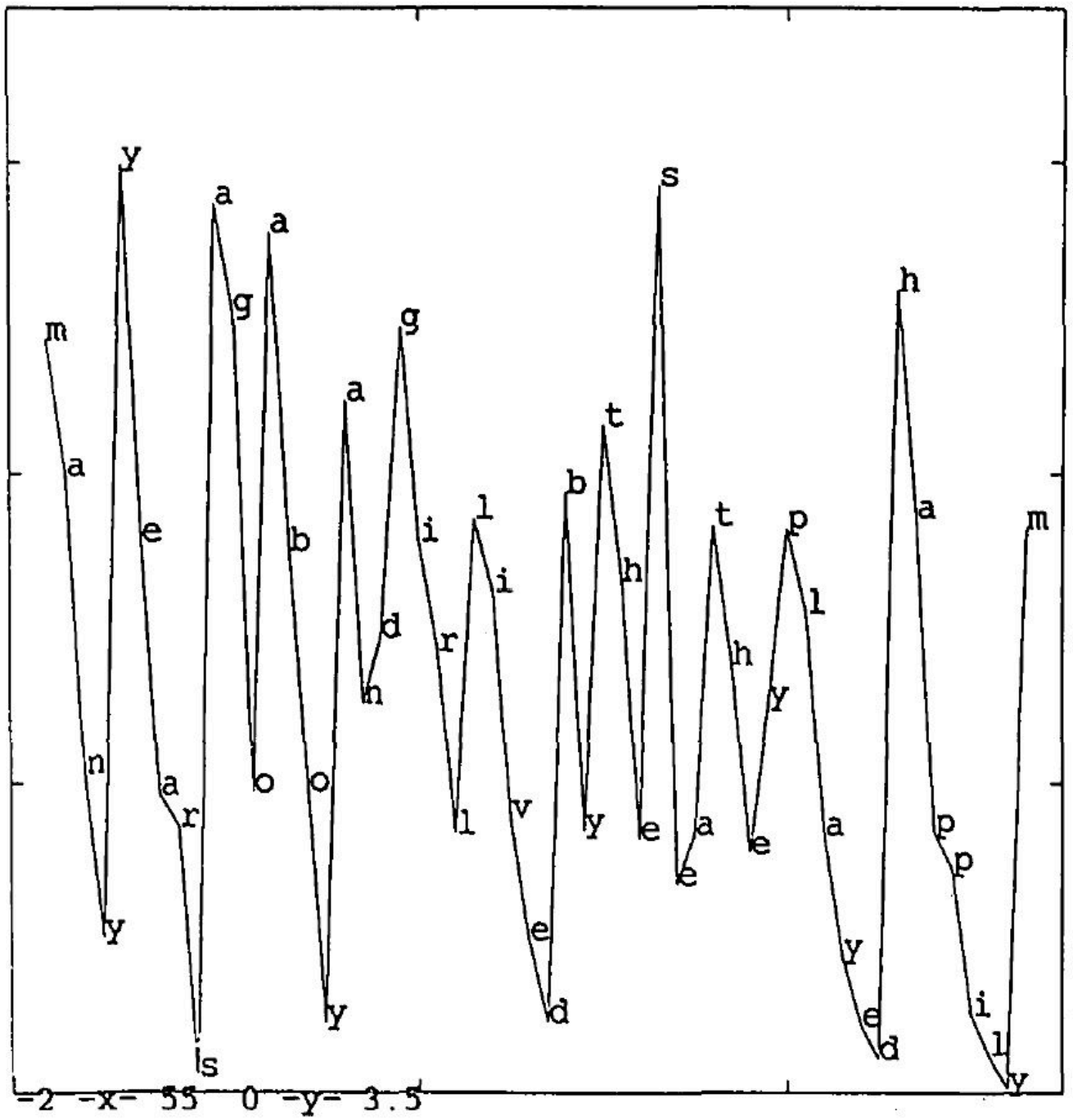
time

words “ba”, “dii”, and “guuu”

Network shows higher error when predicting unpredictable characters (b, d, g)

What is a “word”?

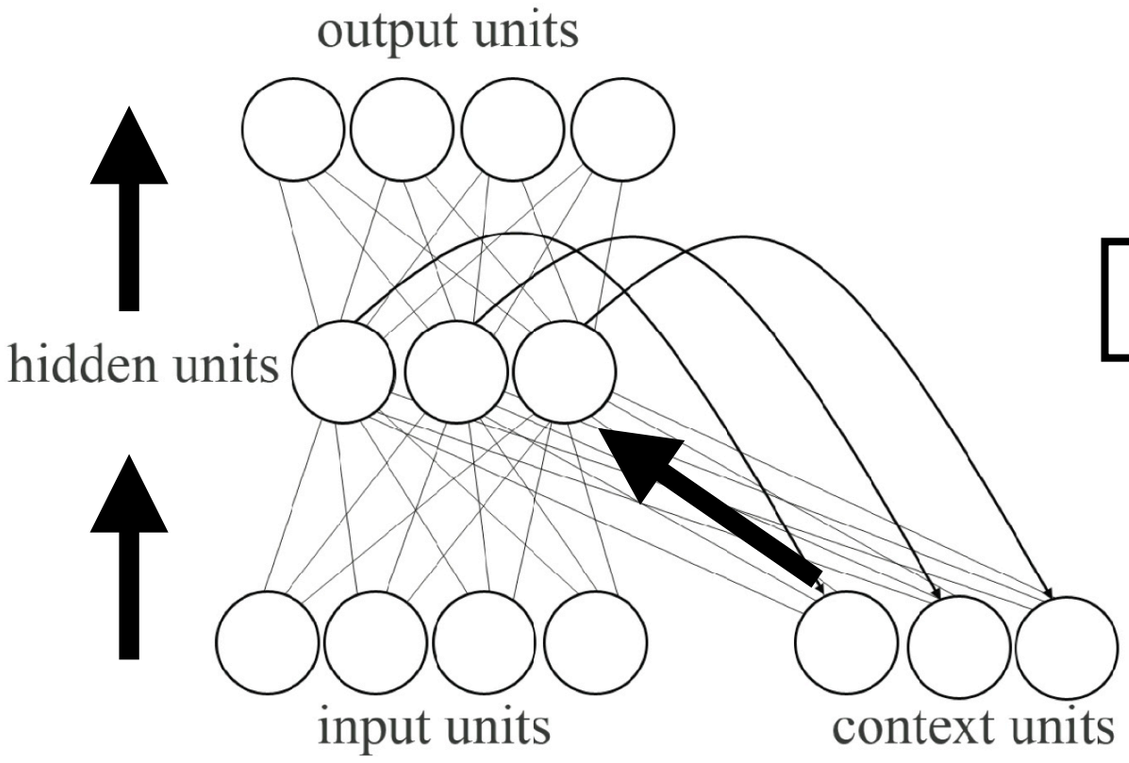
A similar simulation with a sequence of artificial sentences such as “manyyearsagoaboyandgirllivedbythesea..”



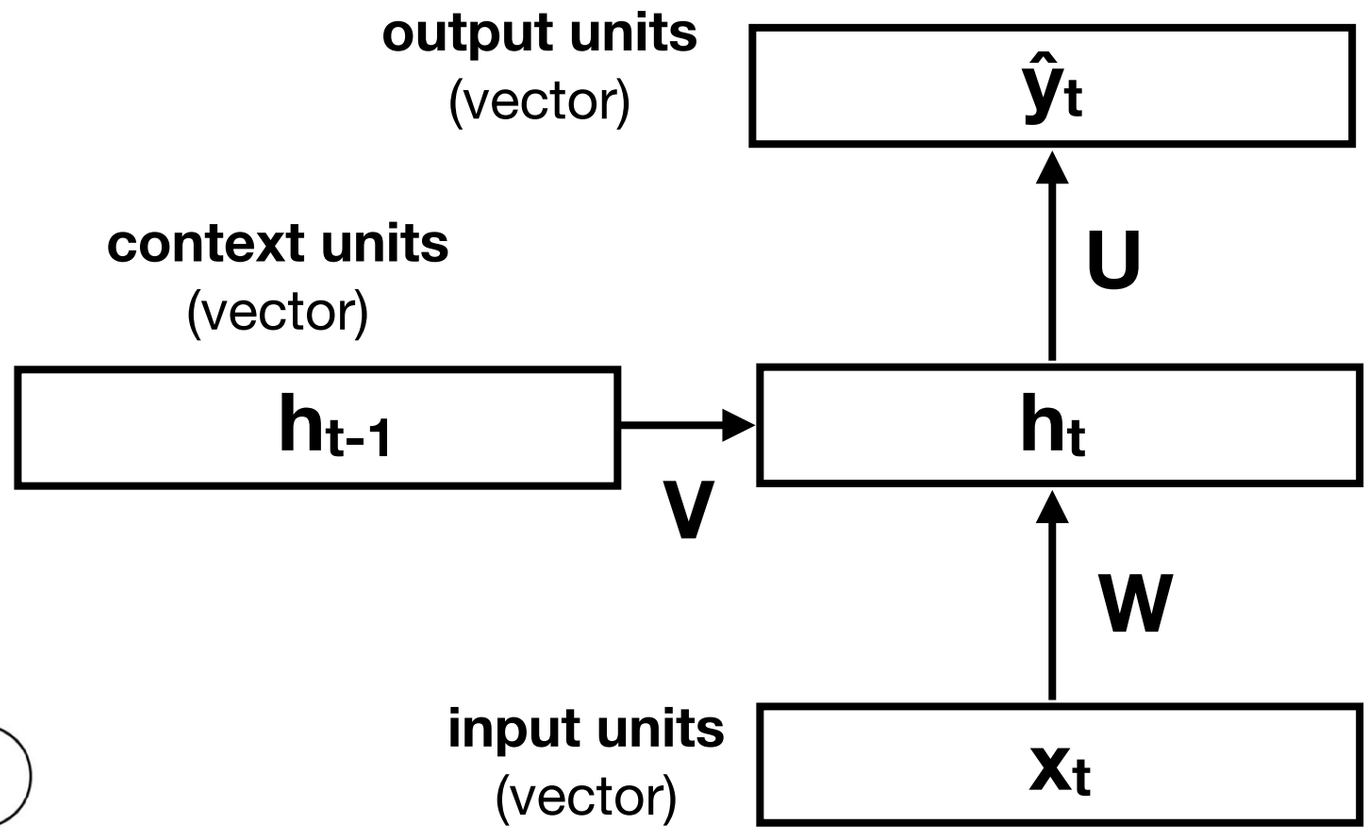
- Much previous work in cognitive science and linguistics assumed basic linguistic units such as phonemes, morphemes, words, etc.
- But the definition of a “word” isn’t that clear
- How many words is each of these phrases? (examples from PDP Handbook)
  - ‘line drive’, ‘flagpole’, ‘carport’, ‘gonna’, ‘wanna’, ‘hafta’, ‘isn’t’ and ‘didn’t’. Here’s a thought: maybe “words” are just peaks in the error prediction signal?
- Elman’s paper and the SRN was among the first to break away from commitments to basic linguistic units (phonemes, morphemes, words, etc.)

Training a recurrent neural network

old notation



new notation

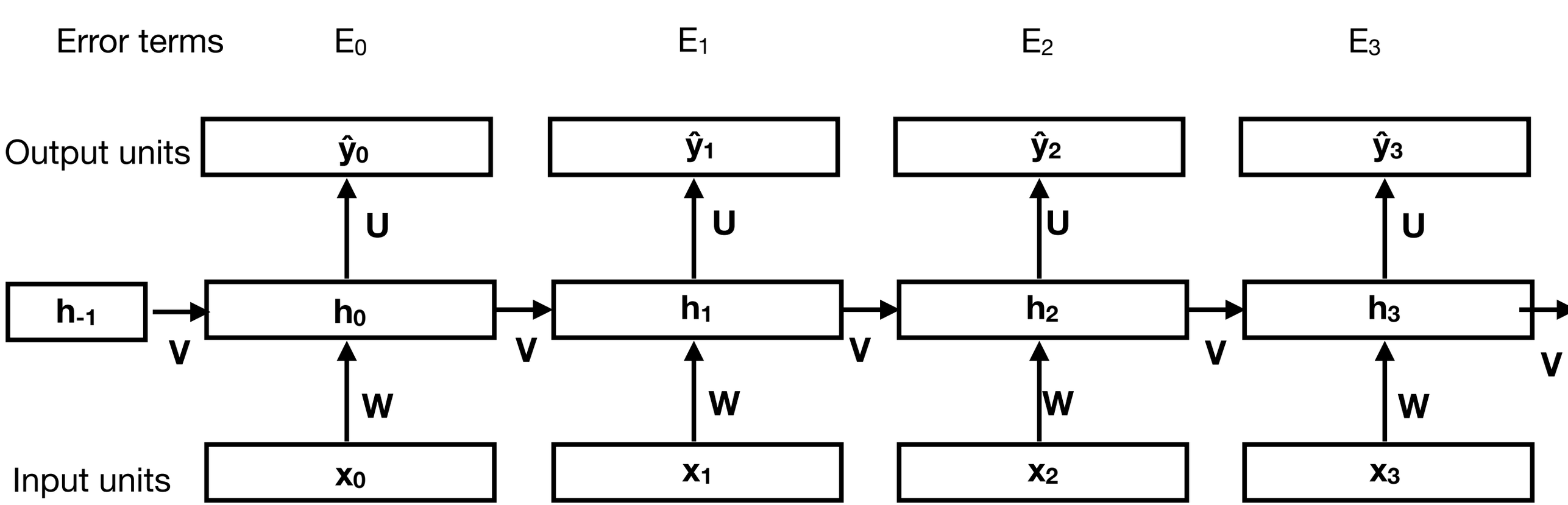


weight matrices **V, W, U**

index **t** is represents the step in time

Unrolling a recurrent network in time

Global error $E = \sum_t E_t$



weight matrices V, W, U

Unrolling a recurrent network in time

Global error $E = \sum_t E_t$

Error terms E_0 E_1 E_2 E_3

i i b a

Output units $\hat{y}_0$ $\hat{y}_1$ $\hat{y}_2$ $\hat{y}_3$

U U U U

h_{-1} h_0 h_1 h_2 h_3

V V V V V

W W W W

Input units x_0 x_1 x_2 x_3

d i i b

Input sequence:

diibaguuu

weight matrices V, W, U

Reminder: Backpropagation algorithm for computing gradient

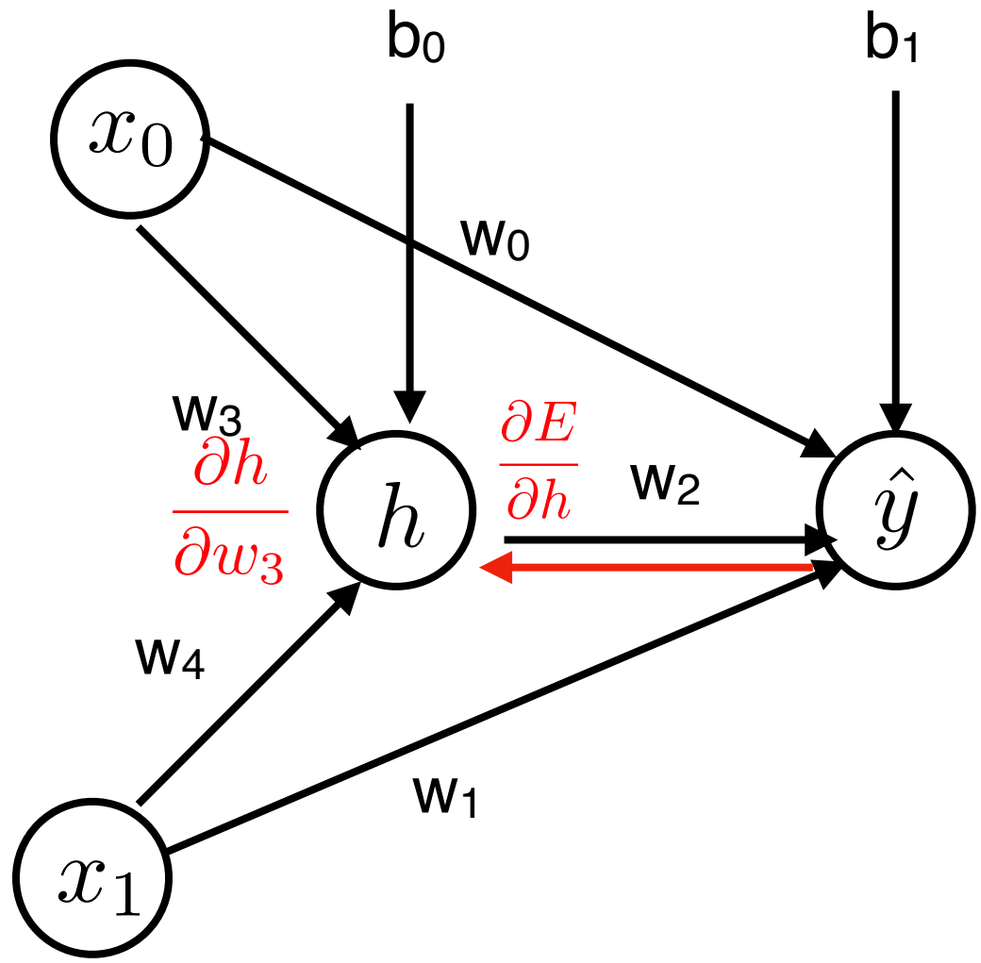
$$E(w,b) = (\hat{y} - y)^2$$
$$= (g(\text{net}) - y)^2$$

Multi-step strategy:

$$\frac{\partial E}{\partial w_3} = \frac{\partial E}{\partial h}\frac{\partial h}{\partial w_3}$$

Step 1) Compute how error changes as a function of hidden unit activation

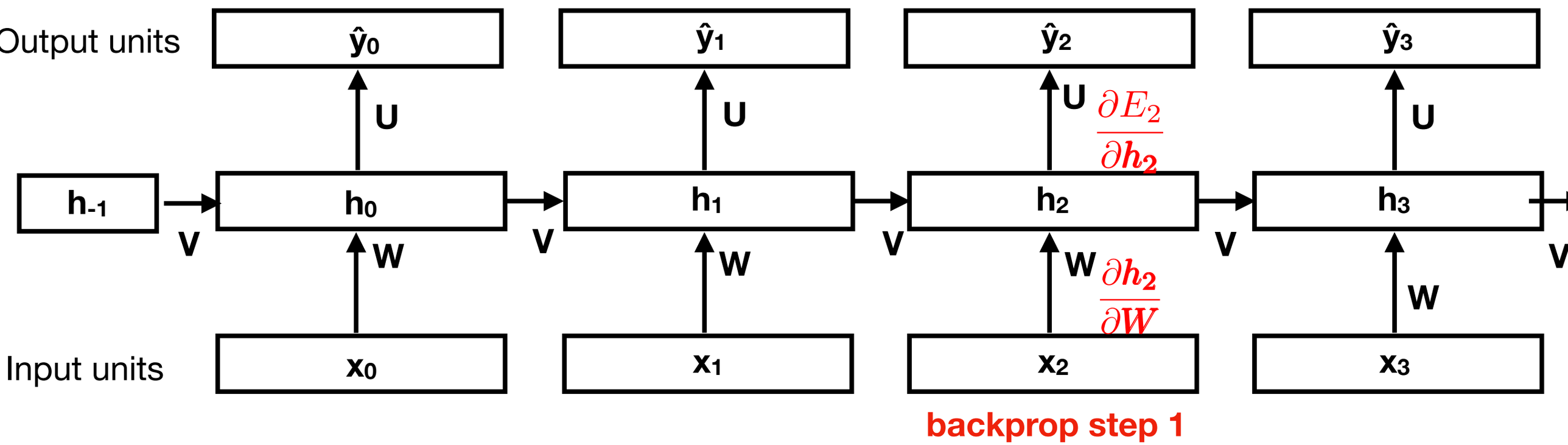
Step 2) Compute how hidden unit activation changes as a function of weight



Backpropagation through time

Global error $E = \sum_t E_t$ $\quad \frac{\partial E}{\partial W} = \sum_t \frac{\partial E_t}{\partial W}$

Illustration for computing $\frac{\partial E_2}{\partial \boldsymbol{W}}$



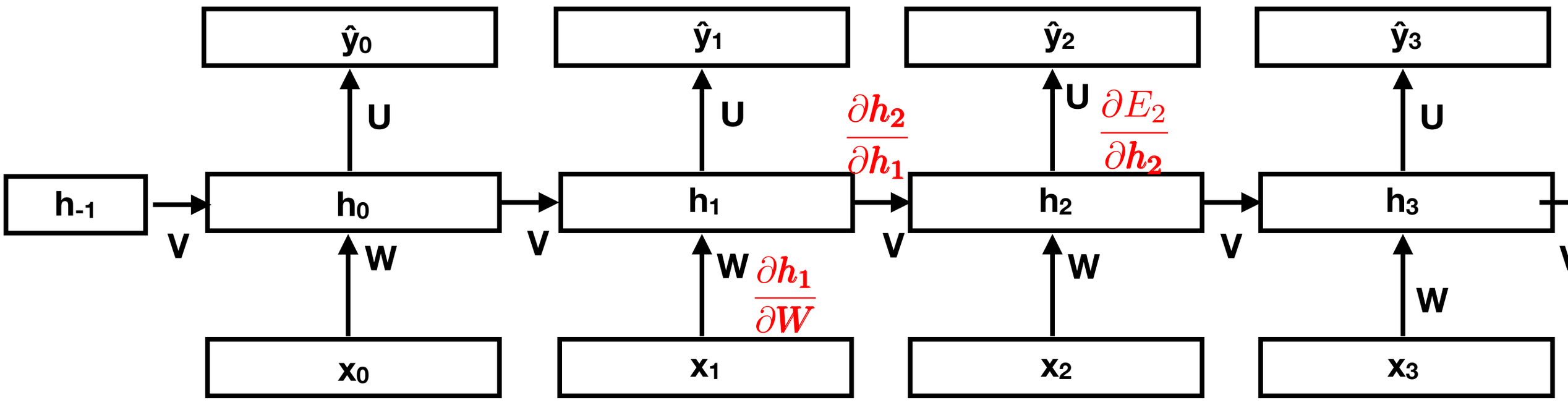
backprop step 1

weight matrices V, W, U

Backpropagation through time

Global error $E = \sum_t E_t \qquad \frac{\partial E}{\partial W} = \sum_t \frac{\partial E_t}{\partial W}$

$$\text{Illustration for computing } \frac{\partial E_2}{\partial \boldsymbol{W}}$$



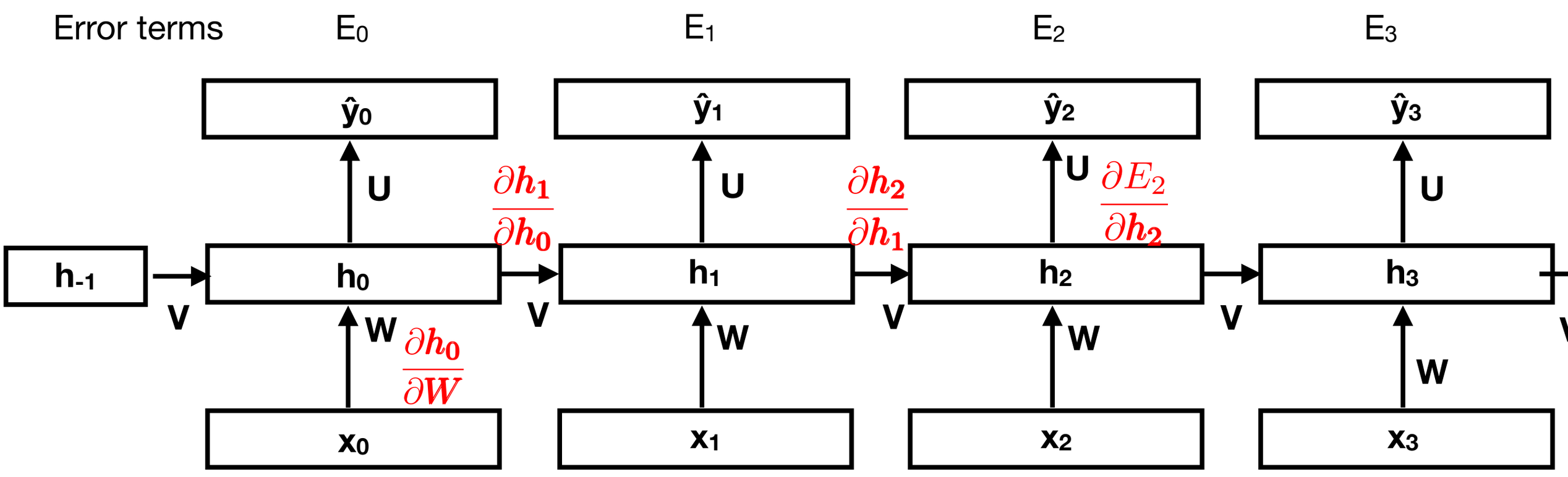
backprop step 2

weight matrices V, W, U

Backpropagation through time

Global error $E = \sum_t E_t$ $\quad \frac{\partial E}{\partial W} = \sum_t \frac{\partial E_t}{\partial W}$

Illustration for computing $\frac{\partial E_2}{\partial W}$



backprop step 3

weight matrices V, W, U

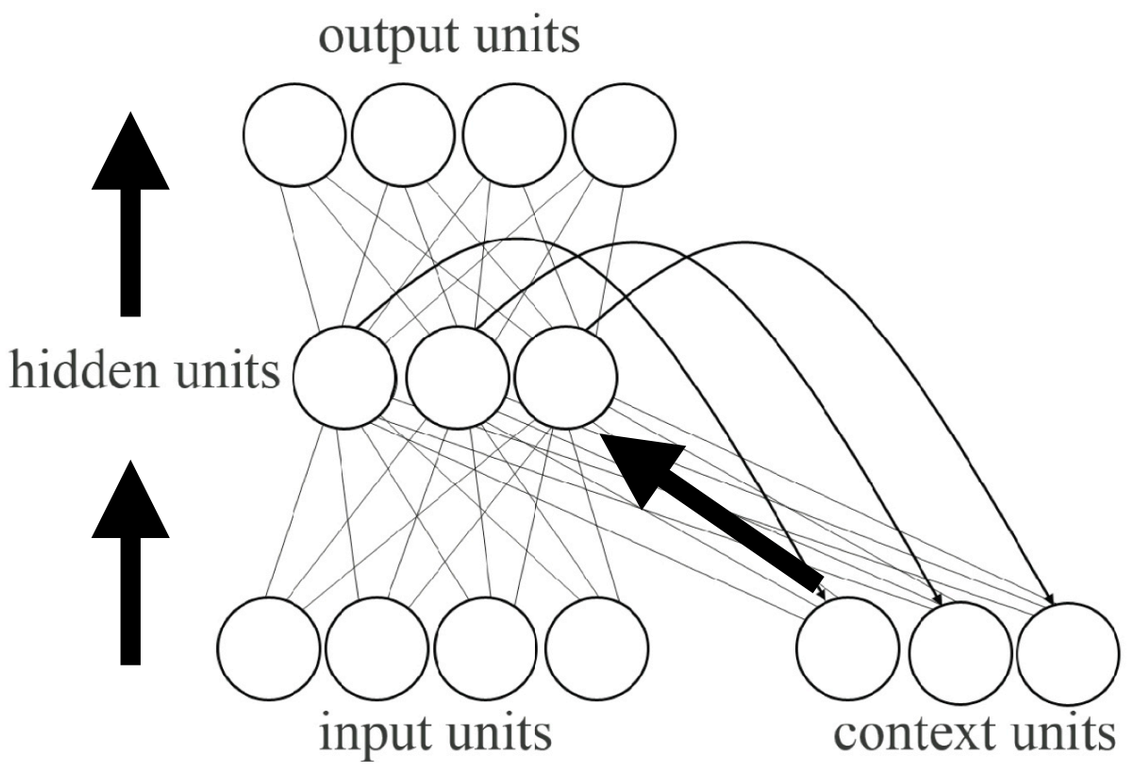
Now, we sum the gradients from the last three slides.

And compute gradients for other time steps, E_0 , E_1 again summing over shared weights…

Conceptually, still no different than wiggling W and seeing how error changes!

Discovering lexical classes from simple sentences

(also Elman, 1990)



“man eat cookie”
“woman see book”
“dragon eat human”
…

Can the SRN discover lexical classes like nouns and verbs?

Discovering lexical classes from simple sentences

Template for generating simple sentences

Categories of Lexical Items Used in Sentence Simulation

Category	Examples
NOUN-HUM	man, woman
NOUN-ANIM	cat, mouse
NOUN-INANIM	book, rock
NOUN-AGRESS	dragon, monster
NOUN-FRAG	glass, plate
NOUN-FOOD	cookie, break
VERB-INTRAN	think, sleep
VERB-TRAN	see, chase
VERB-AGPAT	move, break
VERB-PERCEPT	smell, see
VERB-DESTROY	break, smash
VERB-EAT	eat

Templates for Sentence Generator

WORD 1	WORD 2	WORD 3
NOUN-HUM	VERB-EAT	NOUN-FOOD
NOUN-HUM	VERB-PERCEPT	NOUN-INANIM
NOUN-HUM	VERB-DESTROY	NOUN-FRAG
NOUN-HUM	VERB-INTRAN	
NOUN-HUM	VERB-TRAN	NOUN-HUM
NOUN-HUM	VERB-AGPAT	NOUN-INANIM
NOUN-HUM	VERB-AGPAT	
NOUN-ANIM	VERB-EAT	NOUN-FOOD
NOUN-ANIM	VERB-TRAN	NOUN-ANIM
NOUN-ANIM	VERB-AGPAT	NOUN-INANIM
NOUN-ANIM	VERB-AGPAT	
NOUN-INANIM	VERB-AGPAT	
NOUN-AGRESS	VERB-DESTROY	NOUN-FRAG
NOUN-AGRESS	VERB-EAT	NOUN-HUM
NOUN-AGRESS	VERB-EAT	NOUN-ANIM
NOUN-AGRESS	VERB-EAT	NOUN-FOOD

“man eat cookie”
“woman see book”
“dragon eat human”
...

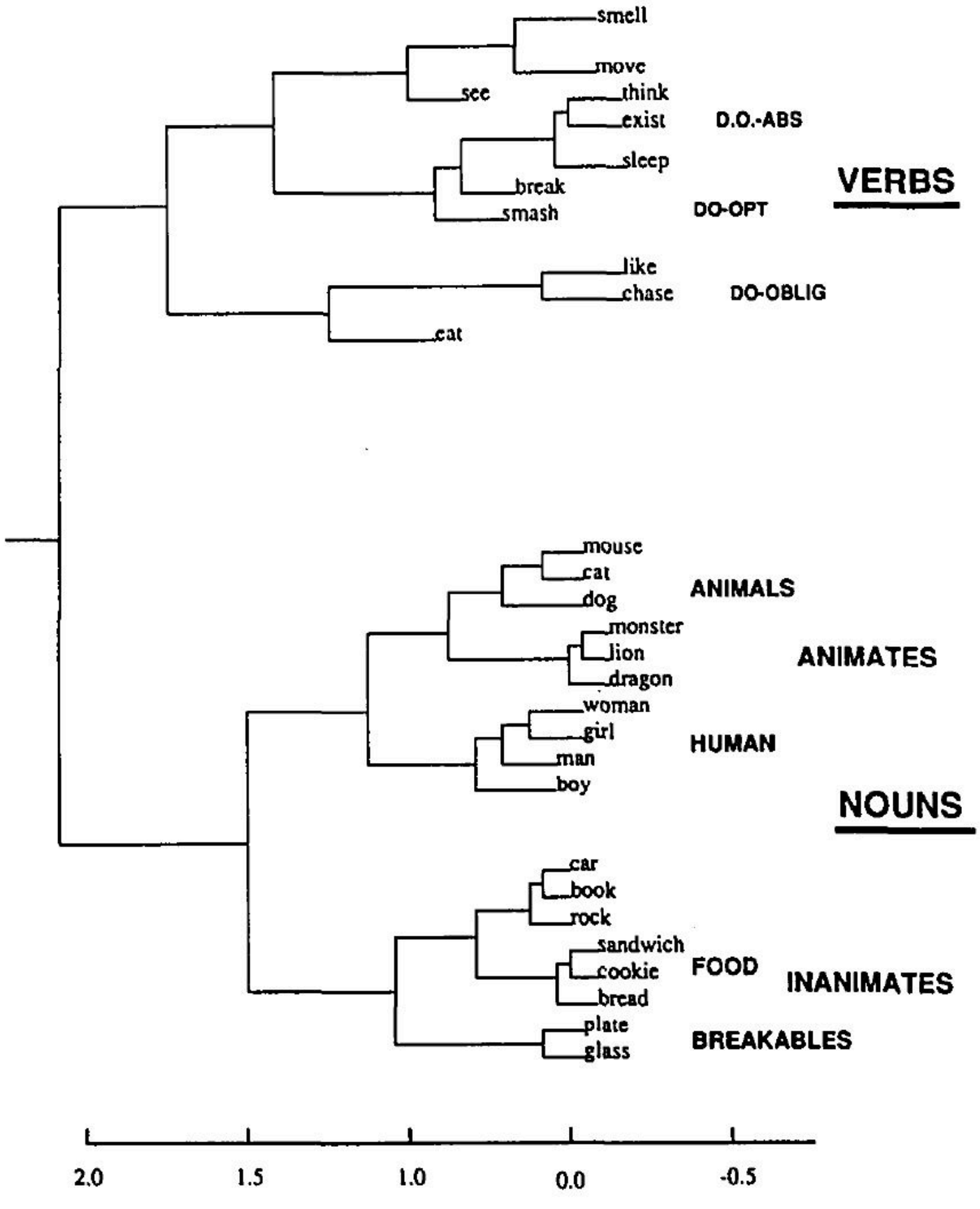
Discovering lexical classes from simple sentences

"one-hot" encoding for words

TABLE 5
Fragment of Training Sequences for Sentence Simulation

Input	Output
00000000000000000000000000000010 (*woman*)	00000000000000000000000000010000 (*smash*)
00000000000000000000000000010000 (*smash*)	00000000000000000000001000000000 (*plate*)
00000000000000000000001000000000 (*plate*)	00000100000000000000000000000000 (*cat*)
00000100000000000000000000000000 (*cat*)	00000000000000000000100000000000 (*move*)
00000000000000000000100000000000 (*move*)	00000000000000000100000000000000 (*man*)
00000000000000000100000000000000 (*man*)	00010000000000000000000000000000 (*break*)
00010000000000000000000000000000 (*break*)	00001000000000000000000000000000 (*car*)
00001000000000000000000000000000 (*car*)	01000000000000000000000000000000 (*boy*)
01000000000000000000000000000000 (*boy*)	00000000000000000000100000000000 (*move*)
00000000000000000000100000000000 (*move*)	00000000000010000000000000000000 (*girl*)
00000000000010000000000000000000 (*girl*)	00000000001000000000000000000000 (*eat*)
00000000001000000000000000000000 (*eat*)	00100000000000000000000000000000 (*bread*)
00100000000000000000000000000000 (*bread*)	00000000100000000000000000000000 (*dog*)
00000000100000000000000000000000 (*dog*)	00000000000000000000100000000000 (*move*)
00000000000000000000100000000000 (*move*)	00000000000000000001000000000000 (*mouse*)
00000000000000000001000000000000 (*mouse*)	00000000000000000001000000000000 (*mouse*)
00000000000000000001000000000000 (*mouse*)	00000000000000000000100000000000 (*move*)
00000000000000000000100000000000 (*move*)	10000000000000000000000000000000 (*book*)
10000000000000000000000000000000 (*book*)	00000000000000001000000000000000 (*lion*)

Discovering lexical classes from simple sentences



- Rich structure from just learning to **predict the next word** from the **previous words**
- This diagram shows results of clustering the average pattern over the hidden units for each word (across many presentations)
- From just the prediction task, the network “discovers” nouns vs. verbs, and also animate vs inanimate, etc. without building in these lexical classes in any way

Finding Structure in One Child's Linguistic Experience

Wentao Wang, **Wai Keen Vong**, **Najoung Kim**, and **Brenden M. Lake**
Center for Data Science, New York University

Abstract

Neural network models have recently made striking progress in natural language processing, but they are typically trained on orders of magnitude more language input than children receive. What can these neural networks, which are primarily distributional learners, learn from a naturalistic subset of a single child's experience? We examine this question using a recent longitudinal dataset collected from a single child, consisting of egocentric visual data paired with text transcripts. We train both language-only and vision-and-language neural networks and analyze the linguistic knowledge they acquire. In parallel with findings from Elman's (1990) seminal work, the neural networks form emergent clusters of words corresponding to syntactic (nouns, transitive and intransitive verbs) and semantic categories (e.g., animals and clothing), based solely on one child's linguistic input. The networks also acquire sensitivity to acceptability contrasts from linguistic phenomena such as determiner-noun agreement and argument structure. We find that incorporating visual information produces an incremental gain in predicting words in context, especially for syntactic categories that are comparatively more easily grounded such as nouns and verbs, but the underlying linguistic representations are not fundamentally altered. Our findings demonstrate which kinds of linguistic knowledge are learnable from a snapshot of a single child's real developmental experience, and which kinds may benefit from stronger inductive biases or richer sources of data.

1 Introduction

In the first three years of life, children's linguistic development progresses rapidly. Young children begin understanding words at around 6 months (Tincoff and Jusczyk, 1999, 2012; Bergelson and Swingley, 2012, 2015). The vocabulary that they can comprehend and produce increases gradually until around 12–14 months, at which a non-linear comprehension boost

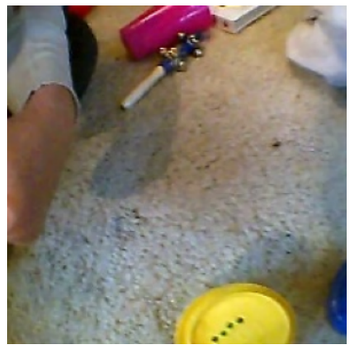
Finding structure in one child's linguistic experience

ok, see the ball?

there's the ball.

where's the ball?

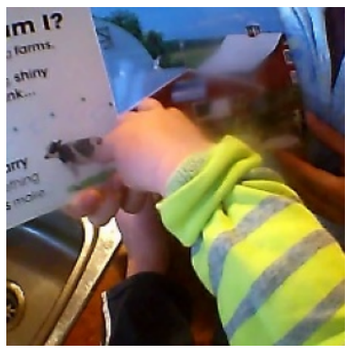
oooh.



where's the ball?

is that the ball?

and here is a farm with a cow on it.

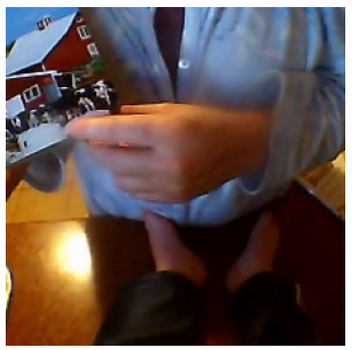


and the cow has an udder, and then milk comes out of the udder.

with out hands yeah.

do you want to go back to the farm sometime?

yeah we might go this weekend sometime to the farm again.

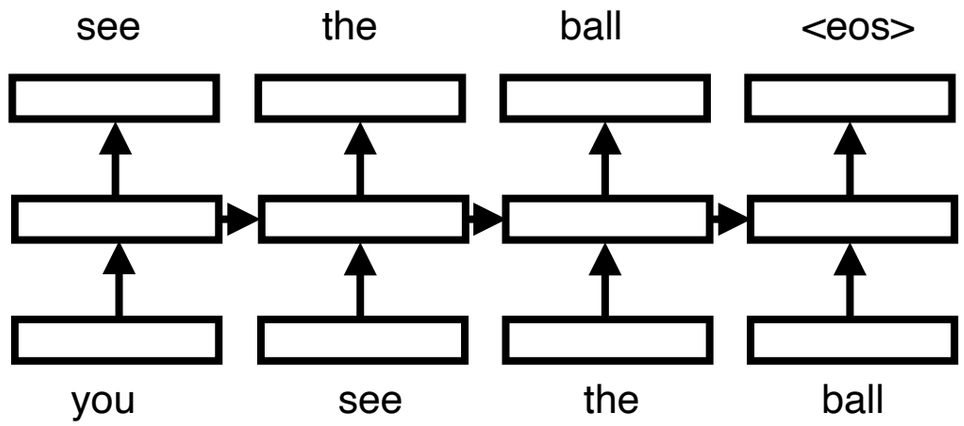


and then we have to pour the milk, you pour the milk that is in buckets into a big milk truck.

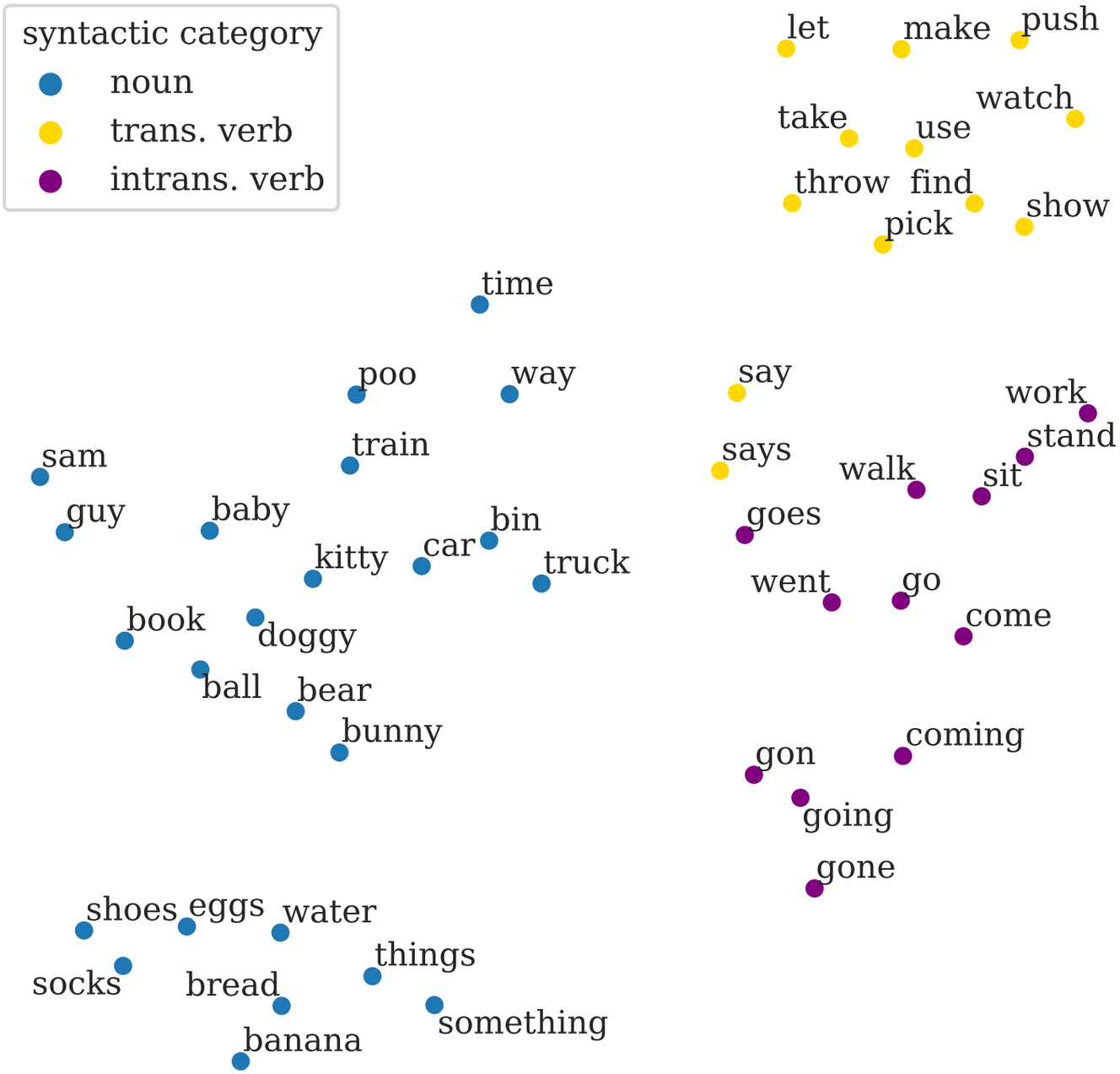
Dataset (SAYCam; Sullivan et al., 2021)

- Egocentric video from one child between 0.5 and 2.5 years old
- 37,000 child-directed utterances transcribed

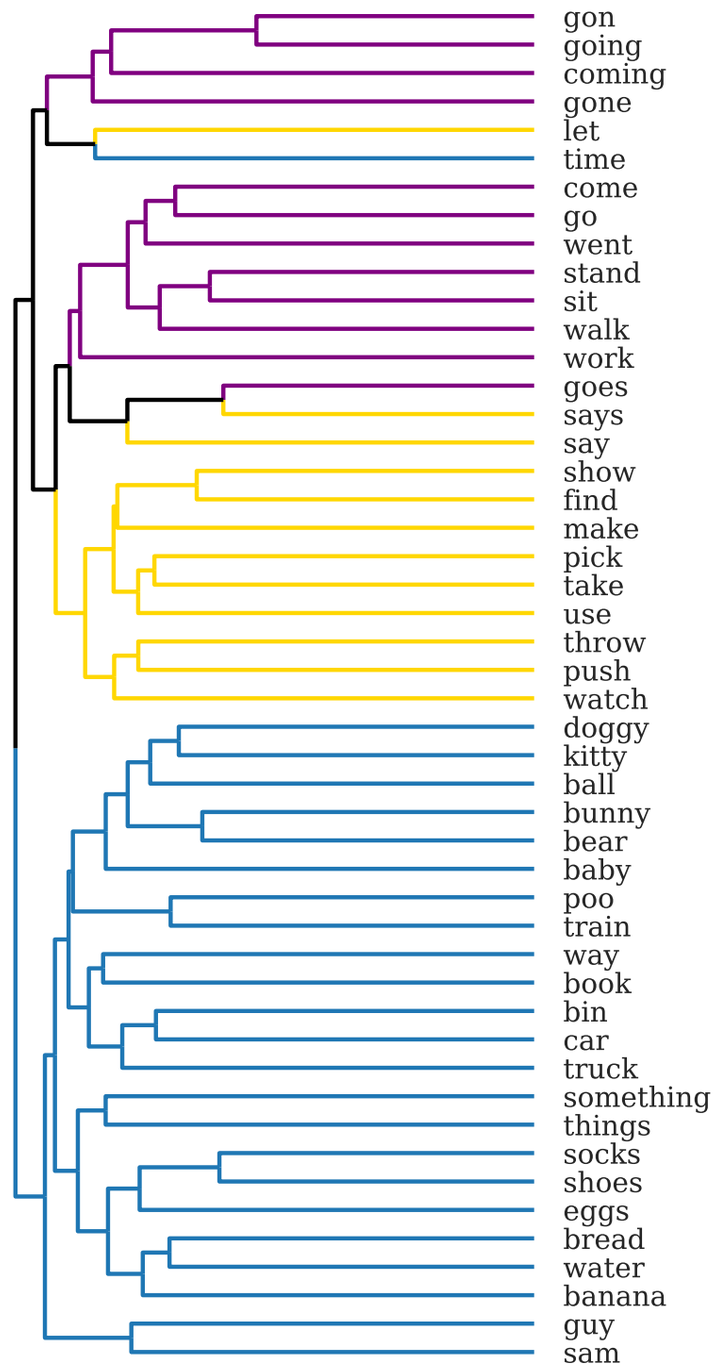
Recurrent neural network



Finding syntactic structure in one child's linguistic experience



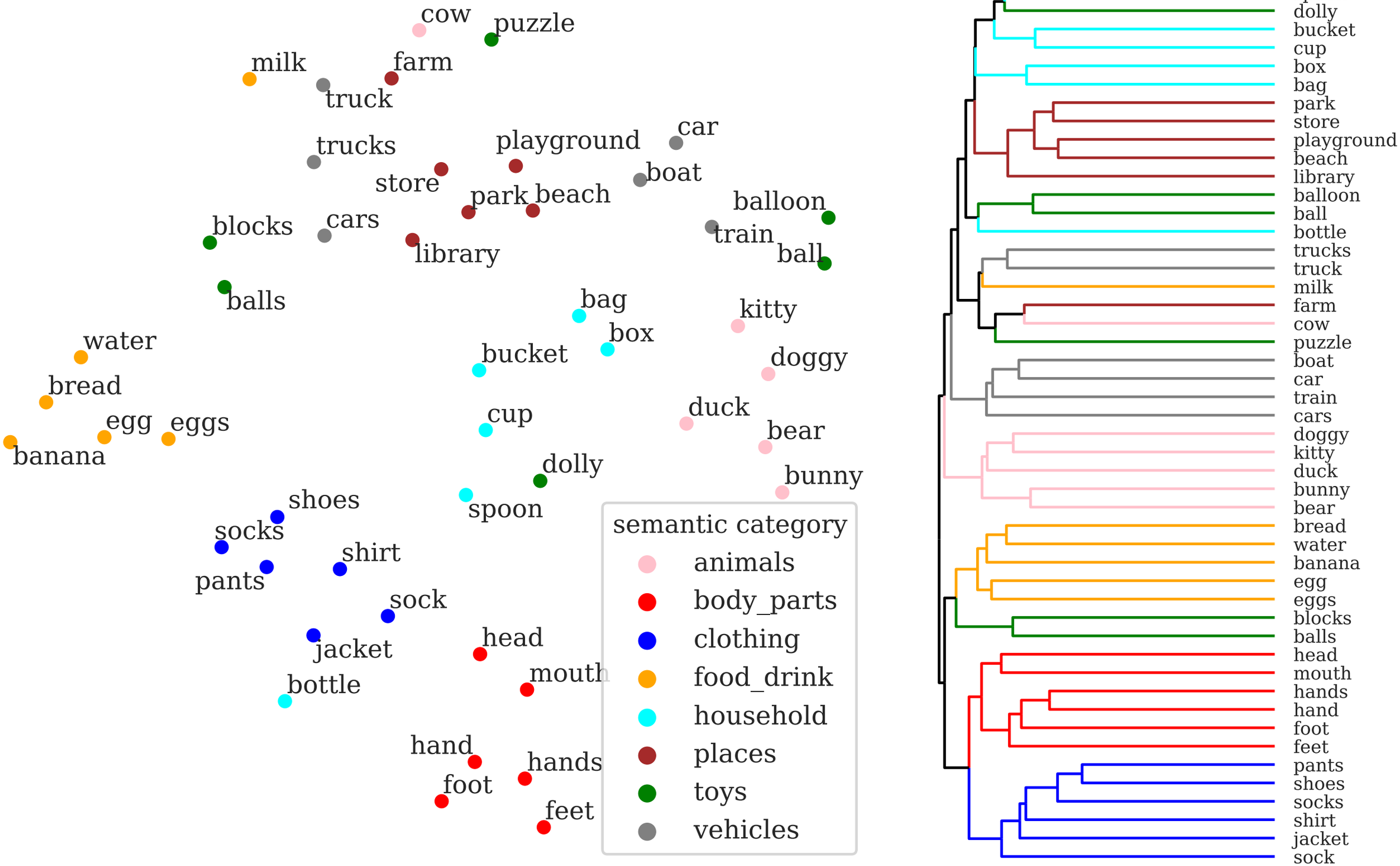
(a) t-SNE



(b) Dendrogram clustering

visualization based on input embedding vectors and cosine similarity

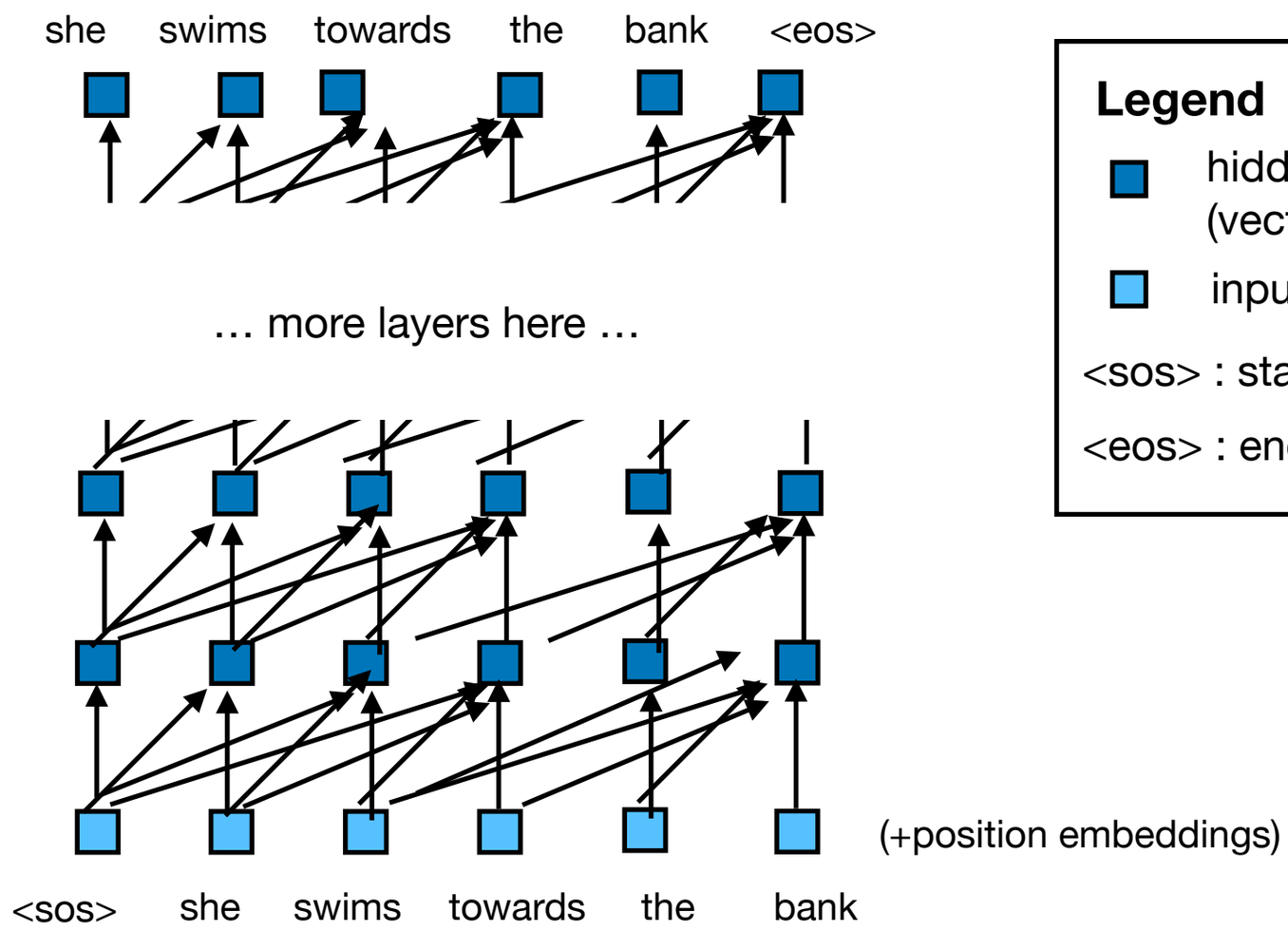
Finding semantic structure in one child's linguistic experience



(a) t-SNE (b) Dendrogram clustering

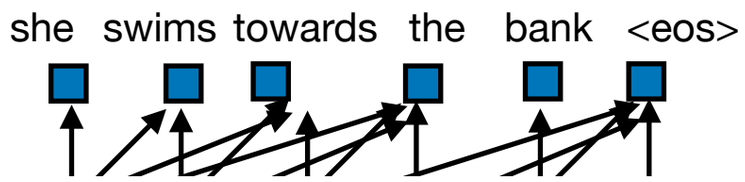
visualization based on input embedding vectors and cosine similarity

State-of-the-art text processing with Transformers



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Transformer layer, step 1: self-attention



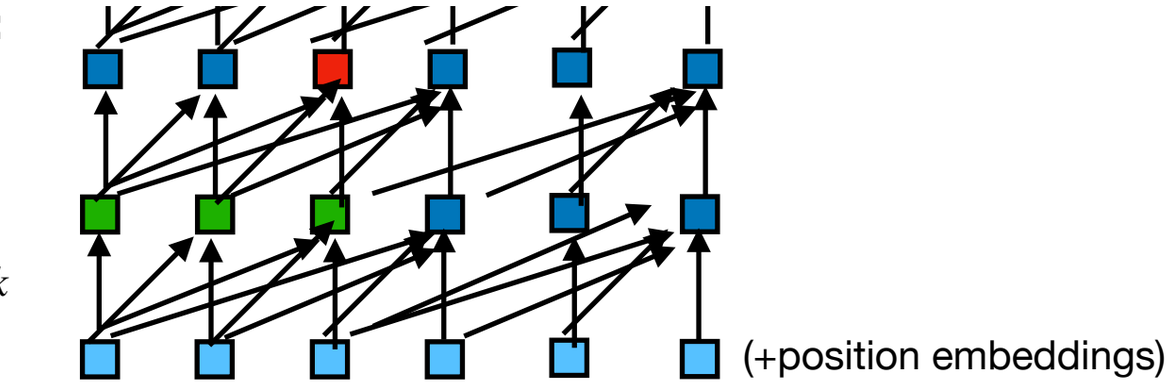
… more layers here …

Let's compute this vector:

z_k **for token** k

Given these vectors

x_j **for tokens** $j \leq k$



Legend

- hidden (vector) embedding
- input embedding

<sos> : start of sequence

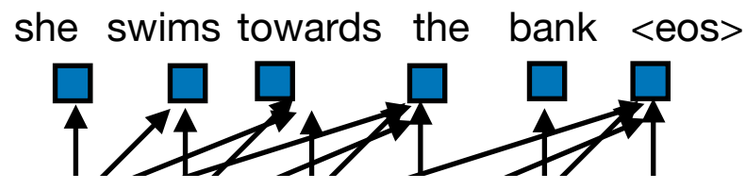
<eos> : end of sequence

High-level intuition: embedding (index k) at next layer is soft-combination of all previous layer's embeddings (index <= k)

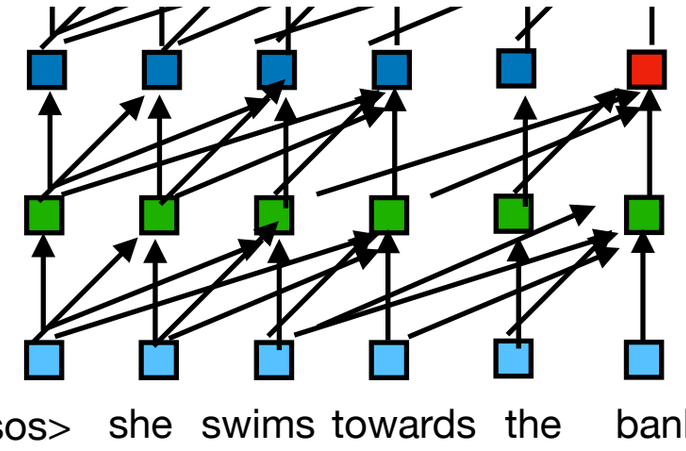
$$\underset{\mathbf{k}}{\blacksquare} \leftarrow \underset{\mathbf{k}}{\blacksquare} + \sum_{j \leq k} a_{kj} \underset{\mathbf{j}}{\blacksquare}$$

a : attention weight

Transformer layer, step 1: self-attention



… more layers here …



Let's compute this vector:

z_k **for token** k

Given these vectors

x_j **for tokens** $j \leq k$

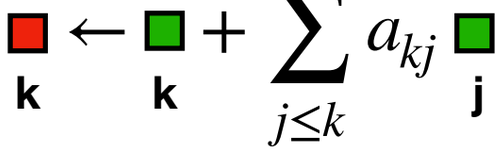
Legend

- hidden (vector) embedding
- input embedding

<sos> : start of sequence

<eos> : end of sequence

High-level intuition



Details

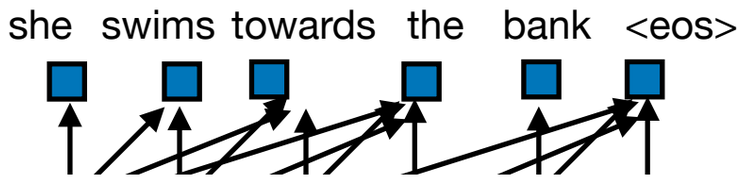
Computing the attention weights (assuming just one attention head)

$$\textbf{sim}_{kj} = (W_Q x_k)^T (W_K x_j) \qquad a_{kj} = \frac{e^{\textbf{sim}_{kj}}}{\sum_{i \leq k} e^{\textbf{sim}_{ki}}}$$

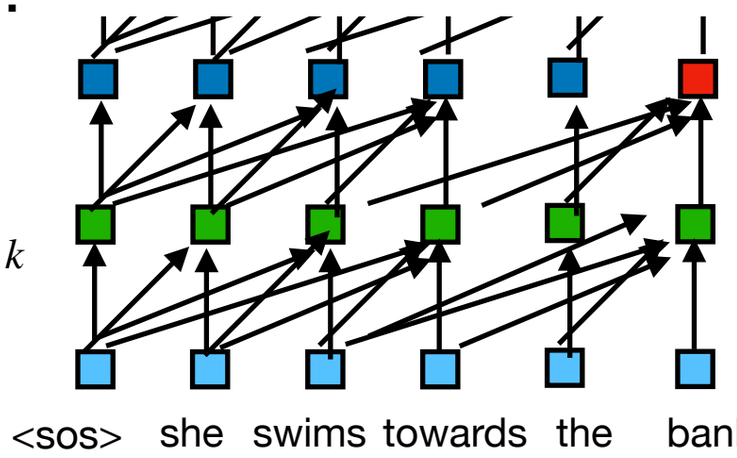
Self-attention output

$$z_k \leftarrow \textbf{LayerNorm}(x_k + \sum_{j \leq k} a_{kj}(W_v x_j))$$

Transformer layer, step 2: feedfoward sub-layer



… more layers here …



Let's compute this vector:

z_k **for token** k

Given these vectors

x_j **for tokens** $j \leq k$

Step 1: Mixing between tokens with self-attention

$$z_k \leftarrow \textbf{LayerNorm}(x_k + \sum_{j \leq k} a_{kj}(W_v x_j))$$

Step 2: Tokenwise processing with feedforward net

$$z_k \leftarrow \textbf{LayerNorm}(x_k + \textbf{FeedForward}(z_k))$$

Legend

hidden (vector) embedding

input embedding

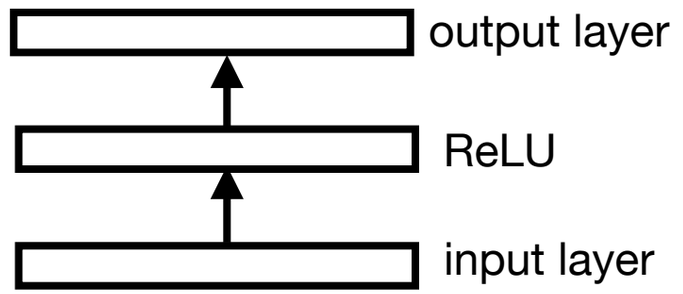
<sos> : start of sequence

<eos> : end of sequence

LayerNorm(·)

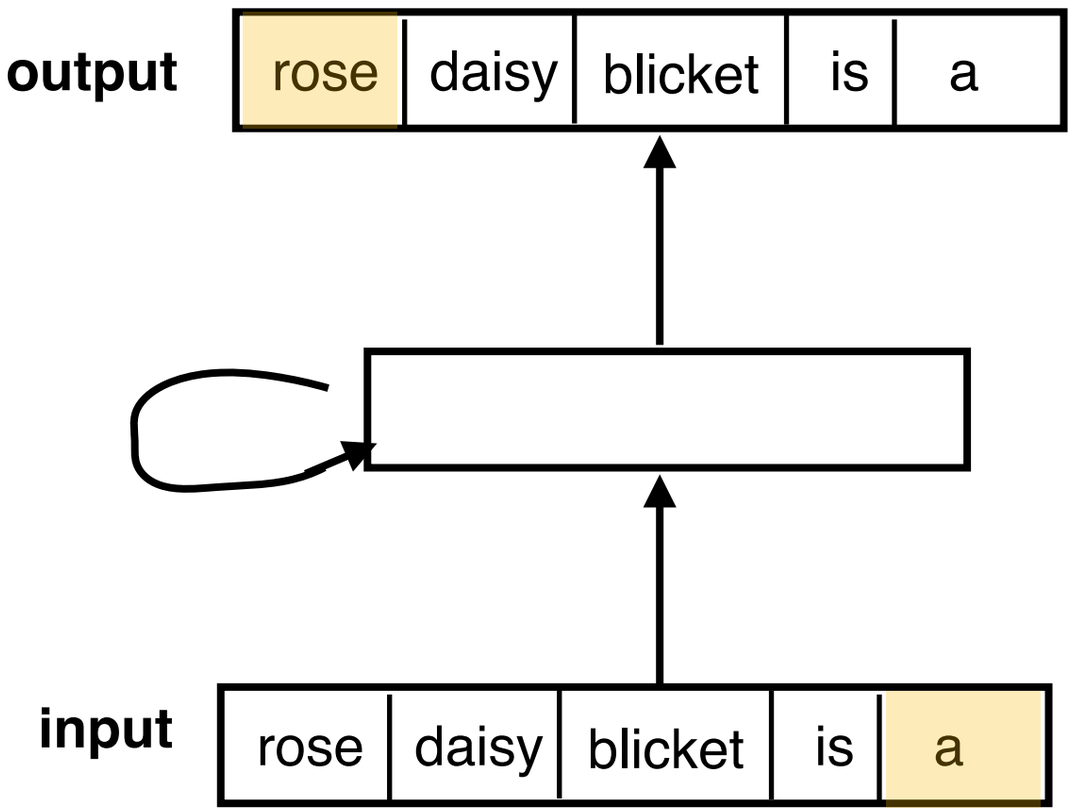
z-score each embedding and transform with learned mean and SD parameter

FeedForward(·)



Critiques of recurrent / transformer networks

(Marcus, 1998, Rethinking eliminative connectionism)

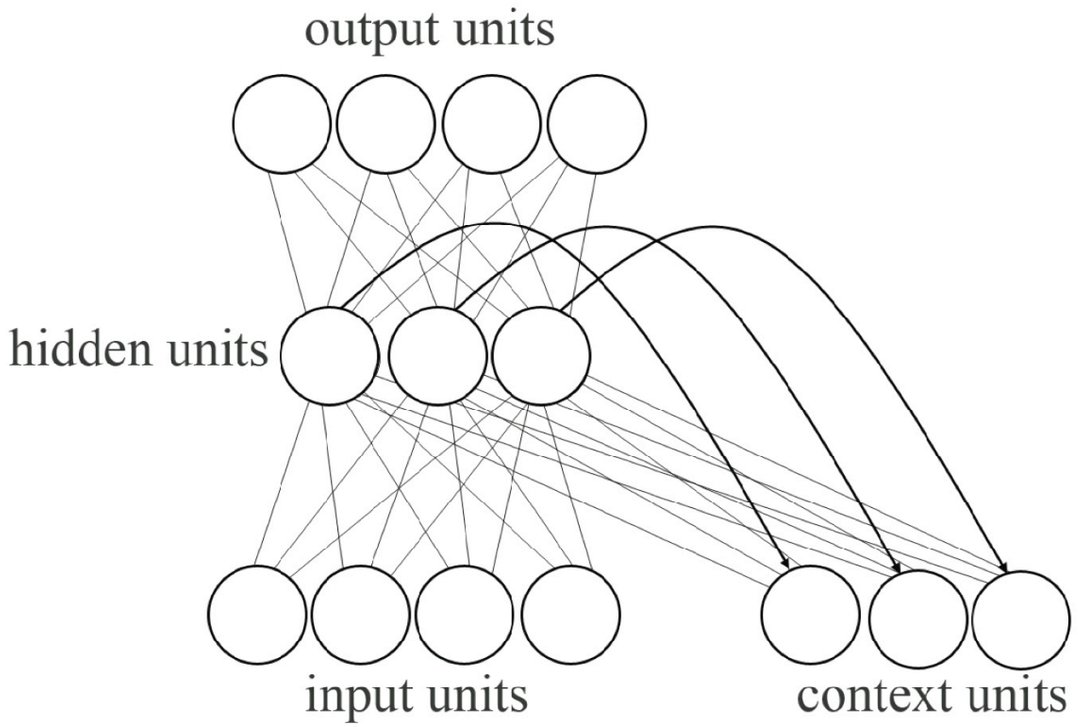


Sentence input… “A rose is a”

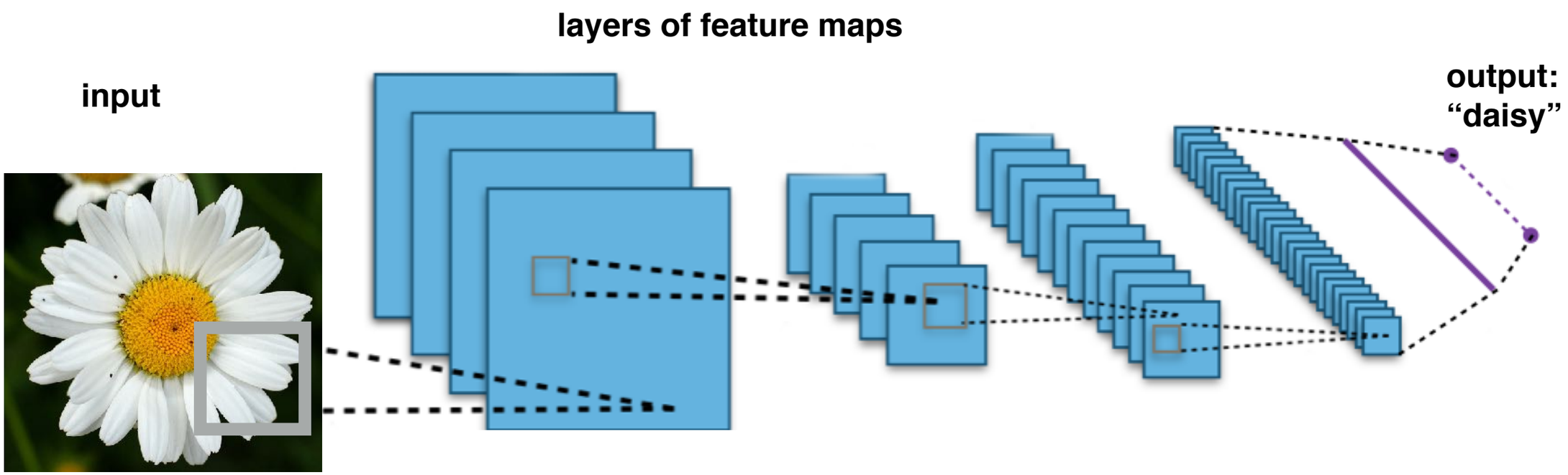
- Sequence-based neural networks generalize very well WITHIN a specific set of words / symbols
- However, they generalize very poorly to new words OUTSIDE the training set
- Say the network is trained on sentences such as, “a rose is a rose”, “a daisy is a daisy”, and “a violet is a violet”
- It will NOT generalize to a new word, “a blicket is a [blicket]”
- People can easily learn rules that apply to arbitrary new variables

Two very important types of neural networks

Recurrent neural networks (RNNs)

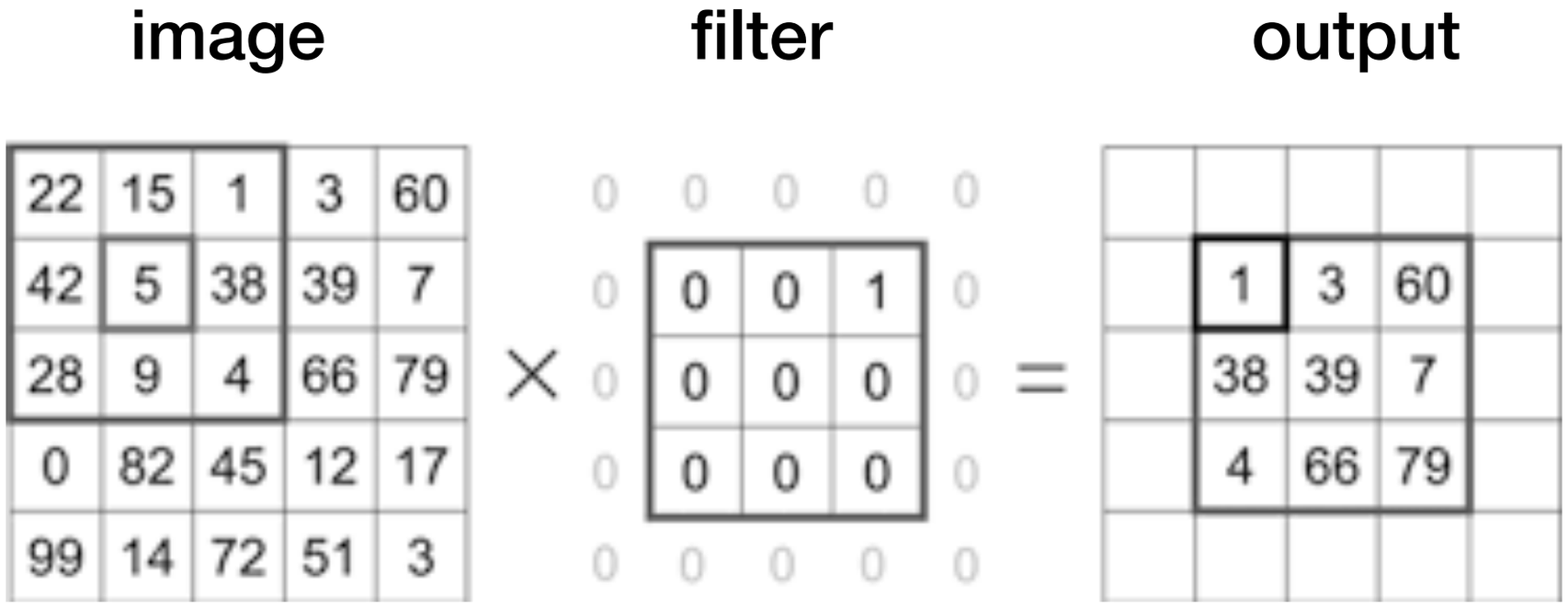


Convolutional neural networks (ConvNets)



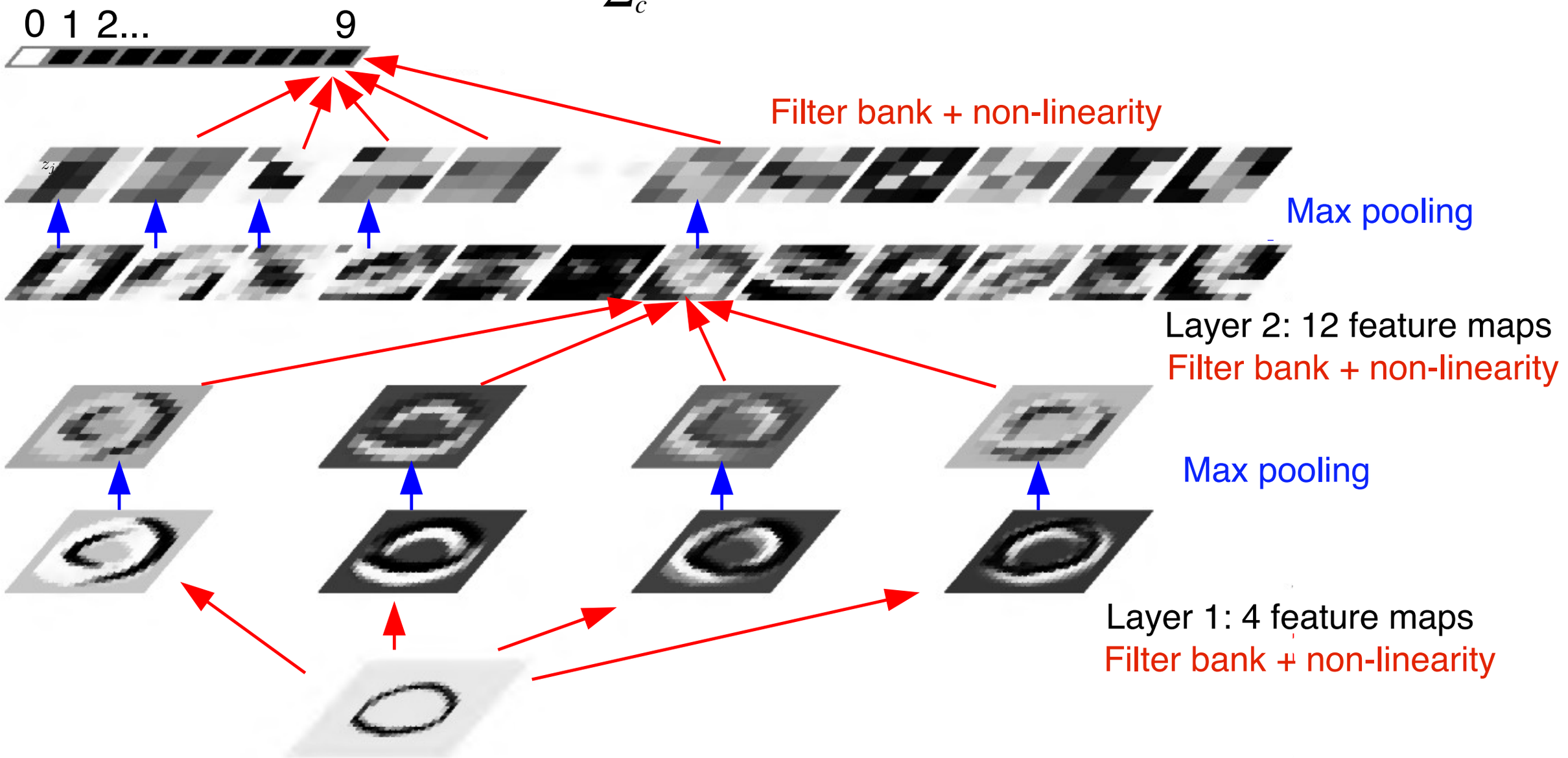
convolving an image with a filter

We slide the filter across the image to produce an output image

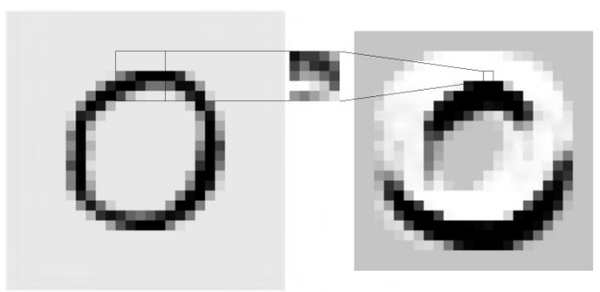


Deep convolutional neural network (convnet) for vision

$$g(\mathbf{net}_i) = \frac{e^{\mathbf{net}_i}}{\sum_c e^{\mathbf{net}_c}}$$

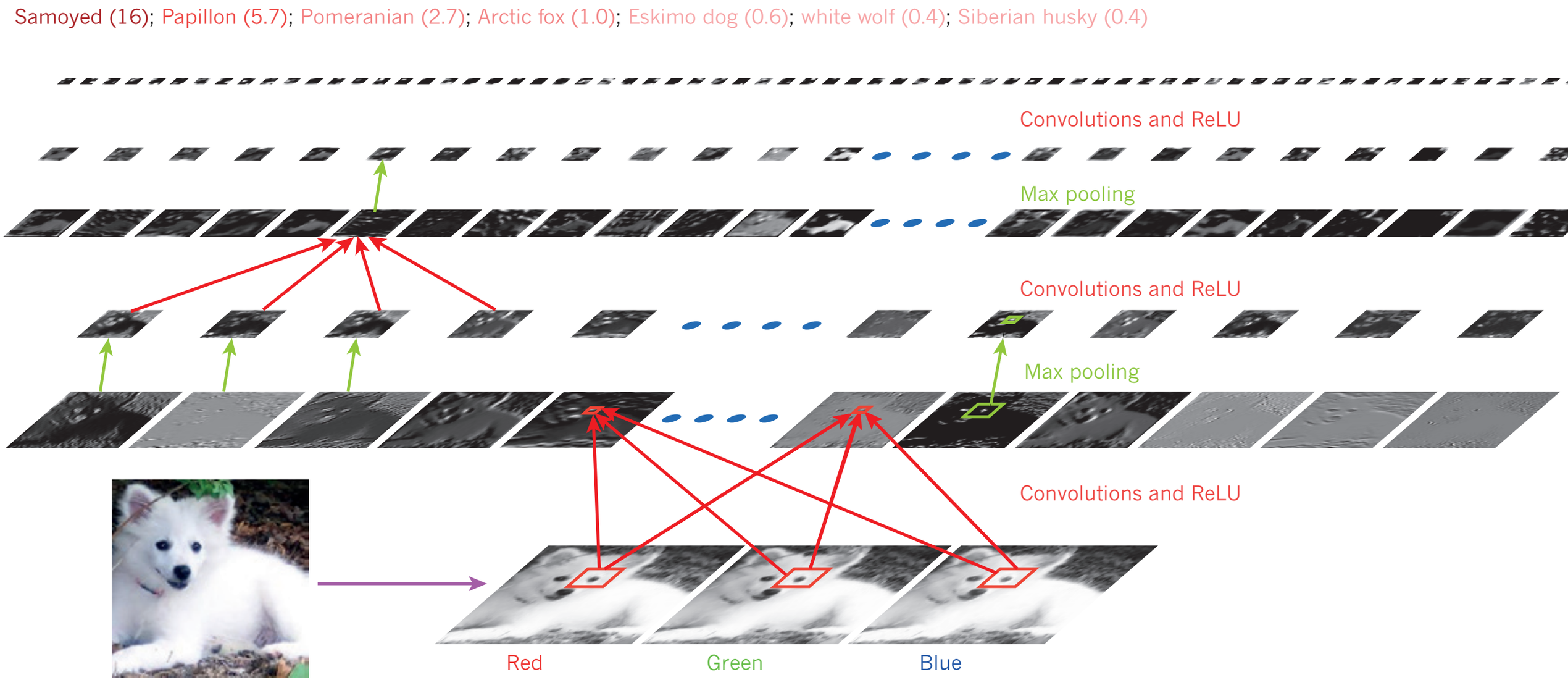


example convolution



Slide courtesy of Yann LeCun

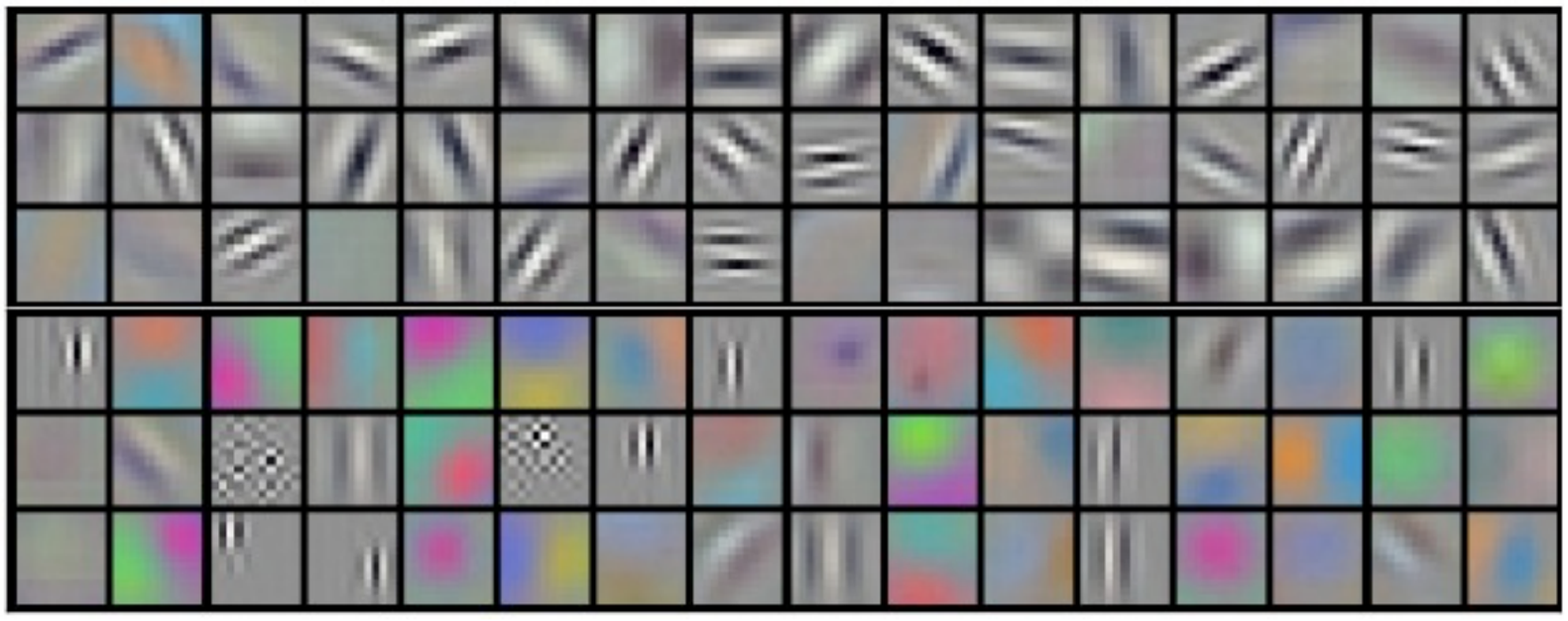
Deep convolutional neural network



From LeCun, Bengio, & Hinton (2015).

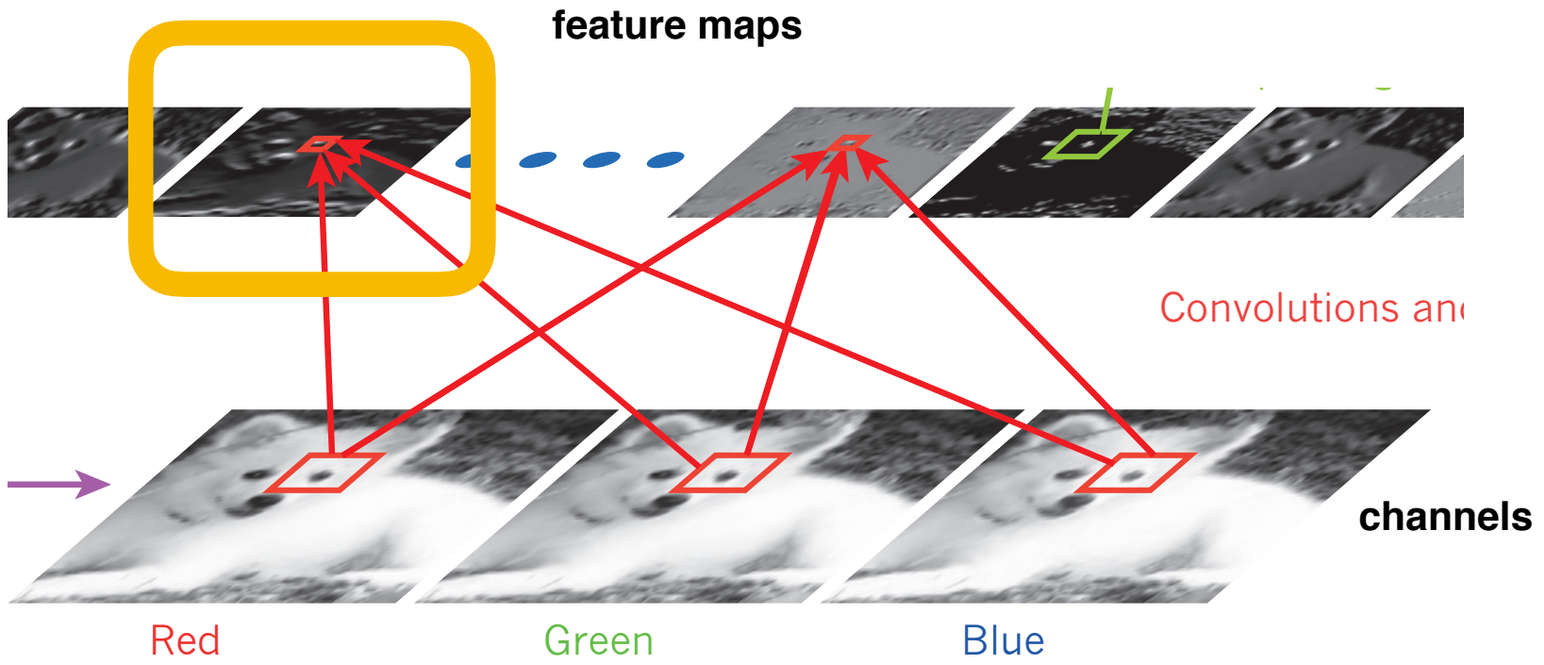
Learned filters in a deep convnet

- Key assumption of “translation invariance”: If a filter (e.g., a horizontal edge detector) is useful in one part of the image, it is probably useful anywhere in the image



(some filters learned from Krizhevsky et al., 2012)

Let’s go through computing with convolutions to build a feature map…



Filter 0 **Filter 1** **Next layer of network**

Channel 0
(e.g., red)

Channel 1
(e.g., green)

Channel 2
(e.g., blue)

Input Volume (+pad 1) (7x7x3)

x[:,:,0]

0	0	0	0	0	0	0
0	2	0	0	1	0	0
0	2	2	0	0	2	0
0	1	0	2	0	1	0
0	2	0	0	0	2	0
0	0	0	2	0	2	0
0	0	0	0	2	1	0

x[:,:,1]

0	0	0	0	0	0	0
0	1	0	2	0	1	0
0	0	0	0	0	2	0
0	0	0	2	1	0	0
0	2	1	1	1	2	0
0	1	1	0	0	0	0
0	0	0	0	0	0	0

x[:,:,2]

0	0	0	0	0	0	0
0	1	0	0	0	1	0
0	0	1	1	0	0	0
0	1	1	2	2	0	0
0	2	2	1	0	0	0
0	0	0	1	1	2	0
0	0	0	0	0	0	0

dot product (element-wise multiply then sum)

dot product

dot product

Filter W0 (3x3x3)

w0[:,:,0]

0	1	0
-1	1	0
0	1	1

w0[:,:,1]

-1	-1	1
0	1	0
0	1	0

w0[:,:,2]

1	-1	-1
0	0	0
0	0	0

Bias b0 (1x1x1)

b0[:,:,0]

1

Filter W1 (3x3x3)

w1[:,:,0]

0	-1	-1
0	0	1
1	1	-1

w1[:,:,1]

1	1	-1

w1[:,:,2]

0	-1	0
-1	1	0
-1	-1	0

Bias b1 (1x1x1)

b1[:,:,0]

0

sum

Output Volume (3x3x2)

o[:,:,0]

8	6	5
7	9	2
-1	7	3

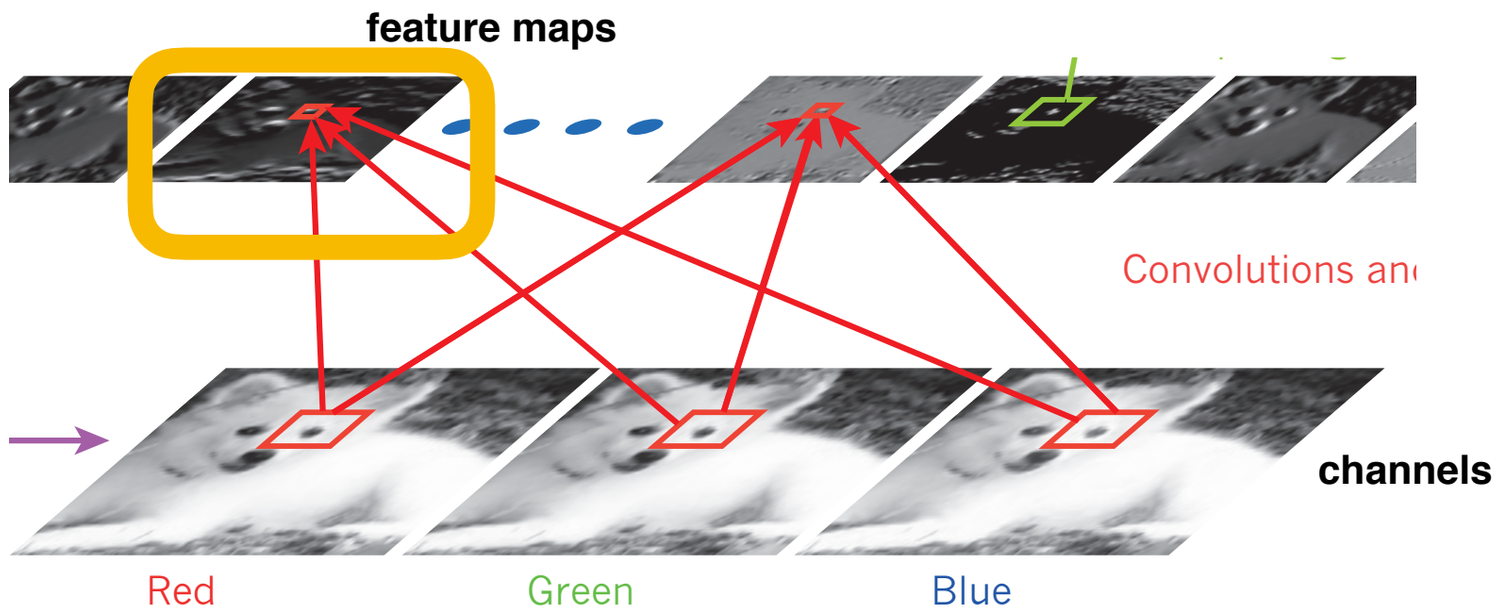
o[:,:,1]

2	3	7
-2	1	-2
-5	-1	-2

Feature map 0

Feature map 1

Credit for demo: http://cs231n.github.io/convolutional-networks/



Filter 0 **Filter 1** **Next layer of network**

Channel 0 (e.g., red)

Channel 1 (e.g., green)

Channel 2 (e.g., blue)

Input Volume (+pad 1) (7x7x3)

x[:,:,0]

0	0	0	0	0	0	0
0	2	0	0	2	2	0
0	2	2	1	2	1	0
0	1	0	2	1	0	0
0	2	0	2	0	1	0
0	0	0	2	0	2	0
0	0	0	0	0	0	0

x[:,:,1]

0	0	0	0	0	0	0
0	1	0	2	0	1	0
0	0	0	0	0	2	0
0	0	0	2	1	0	0
0	2	1	1	1	2	0
0	1	1	2	2	2	0
0	0	0	0	0	0	0

x[:,:,2]

0	0	0	0	0	0	0
0	1	0	0	0	1	0
0	0	1	1	0	0	0
0	1	1	2	2	0	0
0	2	2	1	0	0	0
0	0	0	1	1	2	0
0	0	0	0	0	0	0

Filter W0 (3x3x3)

w0[:,:,0]

0	1	0
-1	1	0
0	1	1

w0[:,:,1]

-1	-1	1
0	1	0
0	1	0

w0[:,:,2]

1	-1	-1
0	0	0
0	0	0

Bias b0 (1x1x1)

b0[:,:,0]

1

Filter W1 (3x3x3)

w1[:,:,0]

0	-1	-1
0	0	1
1	1	-1

w1[:,:,1]

0	-1	-1
-1	1	1
1	1	-1

w1[:,:,2]

0	-1	0
-1	1	0
-1	-1	0

Bias b1 (1x1x1)

b1[:,:,0]

0

Output Volume (3x3x2)

o[:,:,0]

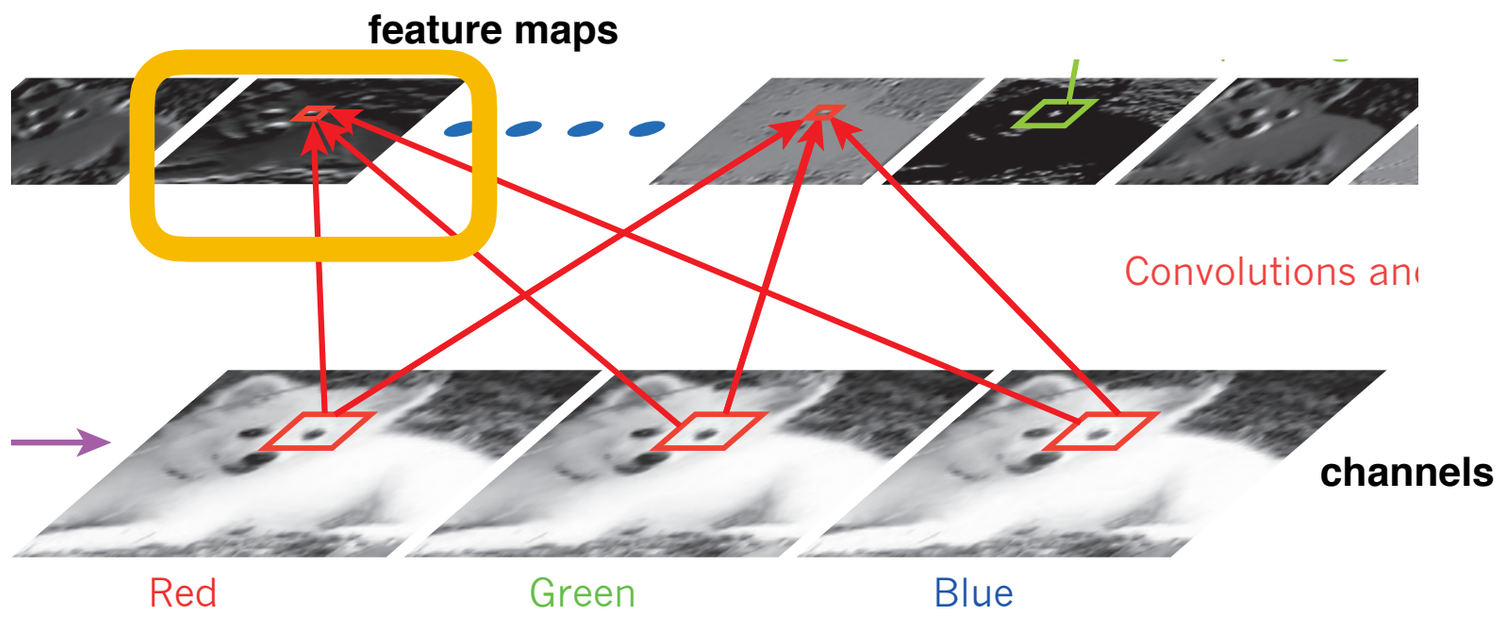
8	6	5
7	9	2
-1	7	3

Feature map 0

o[:,:,1]

2	3	7
-2	1	-2
-5	-1	-2

Feature map 1



Filter 0 **Filter 1** **Next layer of network**

Channel 0 (e.g., red)

Channel 1 (e.g., green)

Channel 2 (e.g., blue)

Input Volume (+pad 1) (7x7x3)

x[:,:,0]

0	0	0	0	0	0	0
0	2	0	0	2	2	0
0	2	2	1	2	1	0
0	1	0	2	1	0	0
0	2	0	2	0	1	0
0	0	0	2	0	2	0
0	0	0	0	0	0	0

x[:,:,1]

0	0	0	0	0	0	0
0	1	0	2	0	1	0
0	0	0	0	0	2	0
0	0	0	2	1	0	0
0	2	1	1	1	2	0
0	1	1	2	2	2	0
0	0	0	0	0	0	0

x[:,:,2]

0	0	0	0	0	0	0
0	1	0	0	0	1	0
0	0	1	1	0	0	0
0	1	1	2	2	0	0
0	2	2	1	0	0	0
0	0	0	1	1	2	0
0	0	0	0	0	0	0

Filter W0 (3x3x3)

w0[:,:,0]

0	1	0
-1	1	0
0	1	1

w0[:,:,1]

-1	-1	1
0	1	0
0	1	0

w0[:,:,2]

1	-1	-1
0	0	0
0	0	0

Bias b0 (1x1x1)

b0[:,:,0]

1

Filter W1 (3x3x3)

w1[:,:,0]

0	-1	-1
0	0	1
1	1	-1

w1[:,:,1]

0	-1	-1
-1	1	1
1	1	-1

w1[:,:,2]

0	-1	0
-1	1	0
-1	-1	0

Bias b1 (1x1x1)

b1[:,:,0]

0

Output Volume (3x3x2)

o[:,:,0]

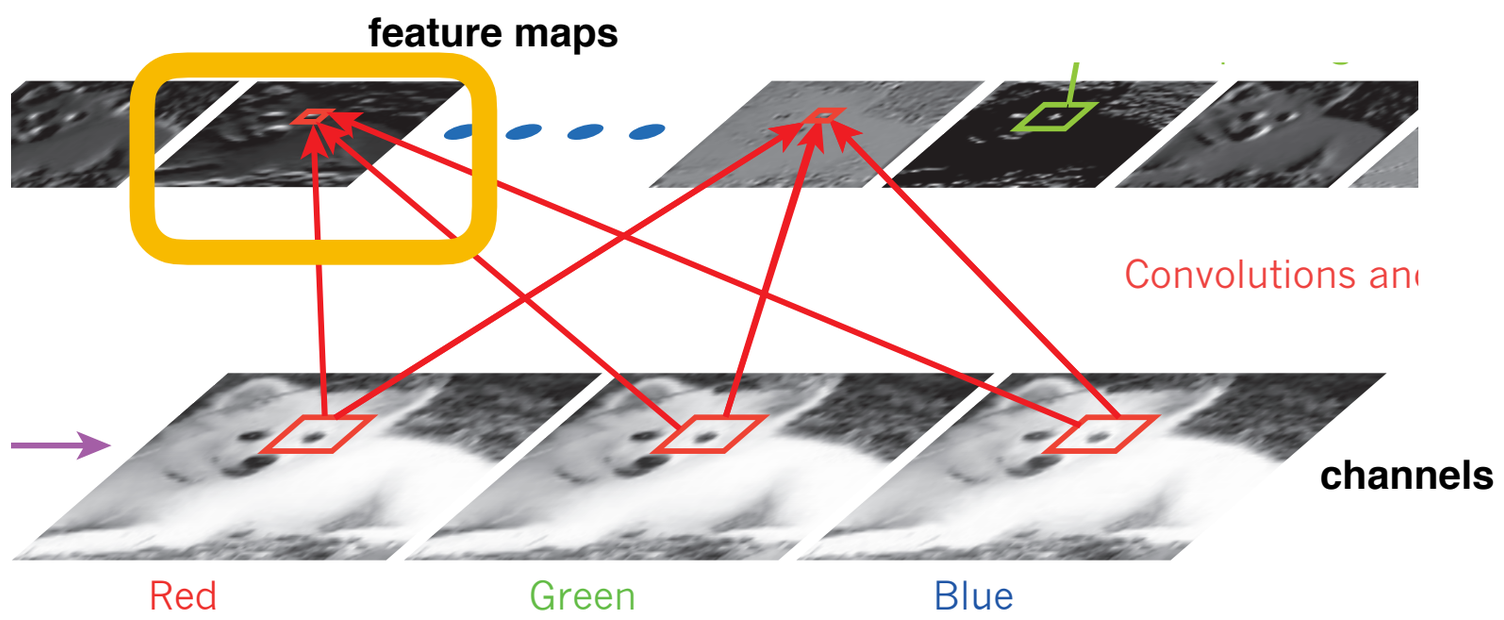
8	6	5
7	9	2
-1	7	3

Feature map 0

o[:,:,1]

2	3	7
-2	1	-2
-5	-1	-2

Feature map 1



Filter 0 **Filter 1** **Next layer of network**

Channel 0 (e.g., red)

Channel 1 (e.g., green)

Channel 2 (e.g., blue)

Input Volume (+pad 1) (7x7x3)

x[:,:,0]

0	0	0	0	0	0	0
0	2	0	0	2	2	0
0	2	2	1	2	1	0
0	1	0	2	1	0	0
0	2	0	2	0	1	0
0	0	0	2	0	2	0
0	0	0	0	0	0	0

x[:,:,1]

0	0	0	0	0	0	0
0	1	0	2	0	1	0
0	0	0	0	0	2	0
0	0	0	2	1	0	0
0	2	1	1	1	2	0
0	1	1	2	2	2	0
0	0	0	0	0	0	0

x[:,:,2]

0	0	0	0	0	0	0
0	1	0	0	0	1	0
0	0	1	1	0	0	0
0	1	1	2	2	0	0
0	2	2	1	0	0	0
0	0	0	1	1	2	0
0	0	0	0	0	0	0

Filter W0 (3x3x3)

w0[:,:,0]

0	1	0
-1	1	0
0	1	1

w0[:,:,1]

-1	-1	1
0	1	0
0	1	0

w0[:,:,2]

1	-1	-1
0	0	0
0	0	0

Filter W1 (3x3x3)

w1[:,:,0]

0	-1	-1
0	0	1
1	1	-1

w1[:,:,1]

0	-1	-1
-1	1	1
1	1	-1

w1[:,:,2]

0	-1	0
-1	1	0
-1	-1	0

Output Volume (3x3x2)

o[:,:,0]

8	6	5
7	9	2
-1	7	3

Feature map 0

o[:,:,1]

2	3	7
-2	1	-2
-5	-1	-2

Feature map 1

Bias b0 (1x1x1)

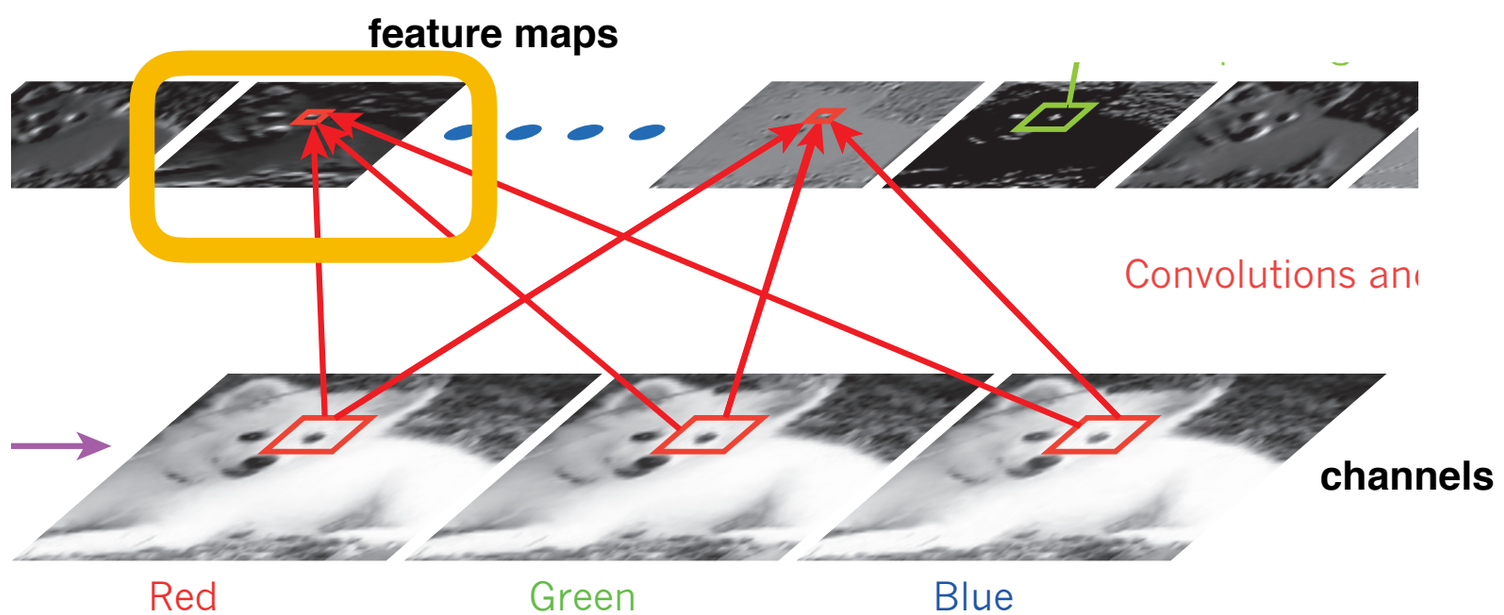
b0[:,:,0]

1

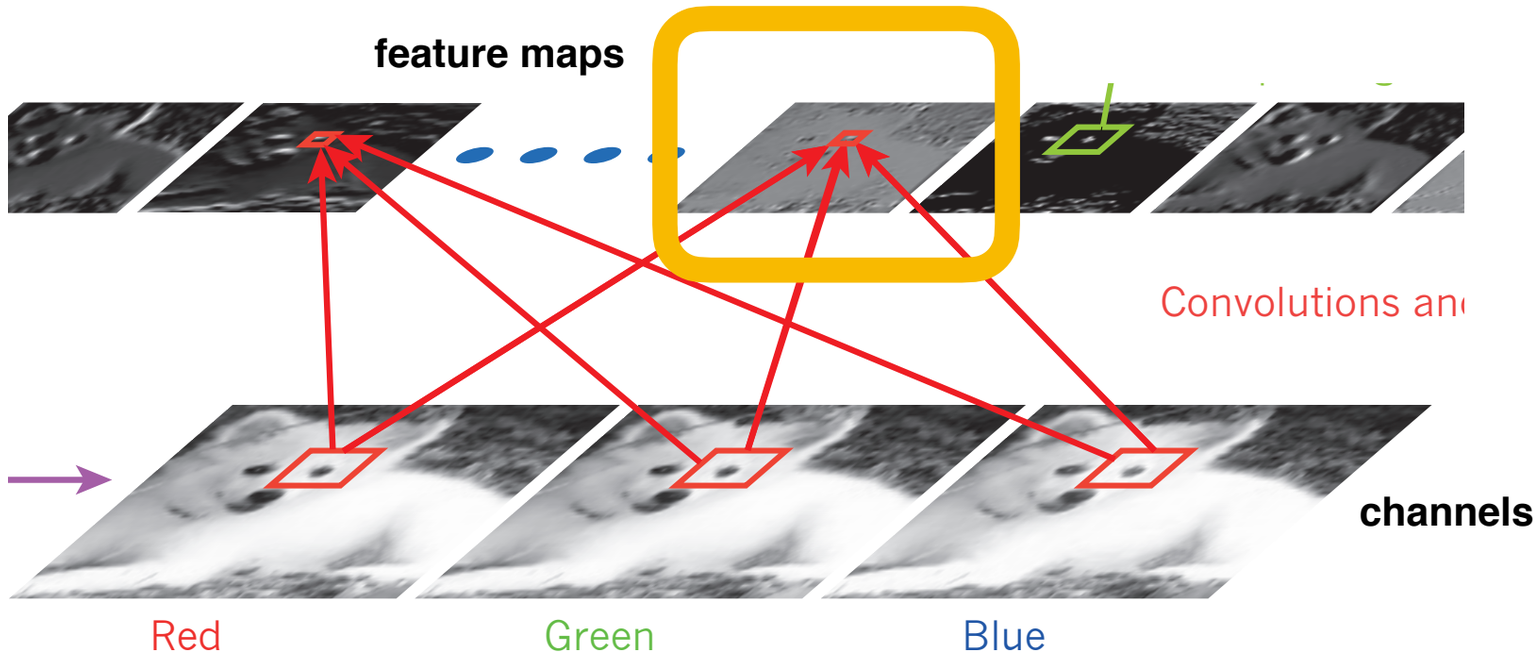
Bias b1 (1x1x1)

b1[:,:,0]

0



Let’s skip to compute the next feature map…



Channel 0

(e.g., red)

Channel 1

(e.g., green)

Channel 2

(e.g., blue)

Filter 0

Filter 1

Next layer of network

Input Volume (+pad 1) (7x7x3)

`x[:,:,0]`

0	0	0	0	0	0	0
0	2	0	0	2	2	0
0	2	2	1	2	1	0
0	1	0	2	1	0	0
0	2	0	2	0	1	0
0	0	0	2	0	2	0
0	0	0	0	0	0	0

`x[:,:,1]`

0	0	0	0	0	0	0
0	1	0	2	0	1	0
0	0	0	0	0	2	0
0	0	0	2	1	0	0
0	2	1	1	1	2	0
0	1	1	2	2	2	0
0	0	0	0	0	0	0

`x[:,:,2]`

0	0	0	0	0	0	0
0	1	0	0	0	1	0
0	0	1	1	0	0	0
0	1	1	2	2	0	0
0	2	2	1	0	0	0
0	0	0	1	1	2	0
0	0	0	0	0	0	0

Filter W0 (3x3x3)

`w0[:,:,0]`

0	1	0
-1	1	0
0	1	1

`w0[:,:,1]`

-1	-1	1
0	1	0
0	1	0

`w0[:,:,2]`

1	-1	-1
0	0	0
0	0	0

Bias b0 (1x1x1)

`b0[:,:,0]`

1

Filter W1 (3x3x3)

`w1[:,:,0]`

0	-1	-1
0	0	1
1	1	-1

`w1[:,:,1]`

0	-1	-1
-1	1	1
1	1	-1

`w1[:,:,2]`

0	-1	0
-1	1	0
-1	-1	0

Bias b1 (1x1x1)

`b1[:,:,0]`

0

Output Volume (3x3x2)

`o[:,:,0]`

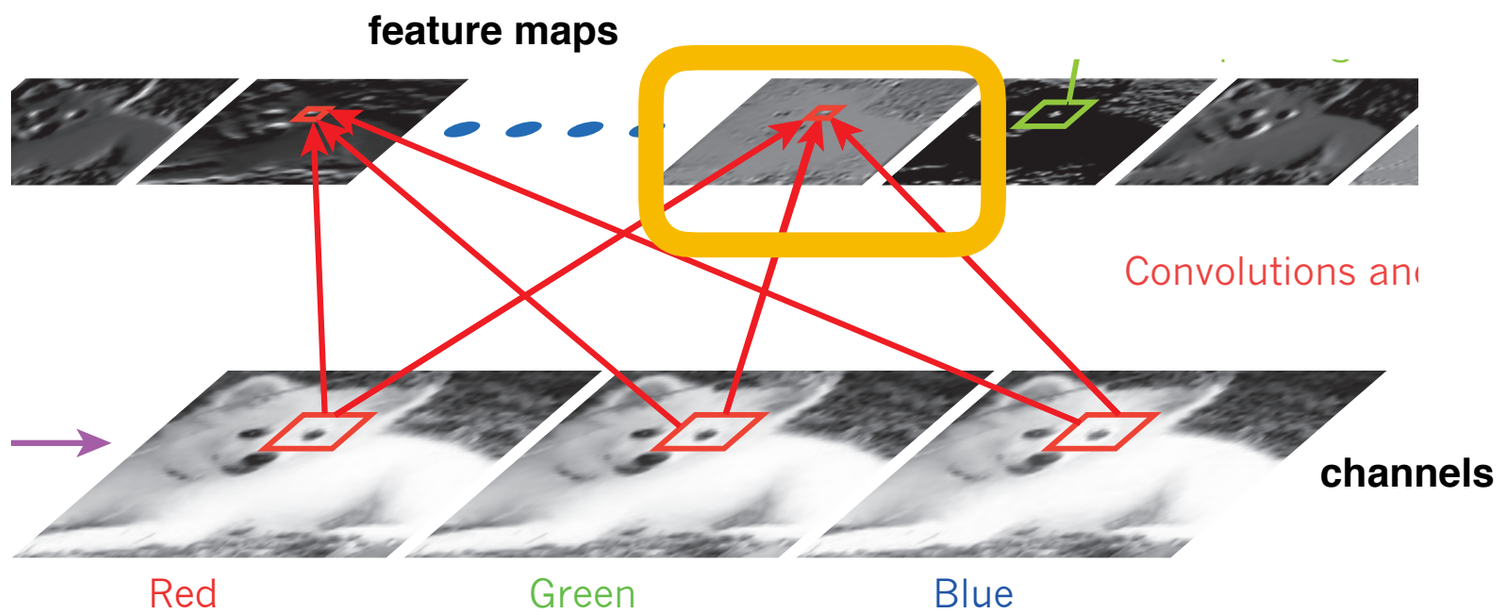
8	6	5
7	9	2
-1	7	3

Feature map 0

`o[:,:,1]`

2	3	7
-2	1	-2
-5	-1	-2

Feature map 1



Filter 0 **Filter 1** **Next layer of network**

Input Volume (+pad 1) (7x7x3)

`x[:,:,0]`

0	0	0	0	0	0	0
0	2	0	0	2	2	0
0	2	2	1	2	1	0
0	1	0	2	1	0	0
0	2	0	2	0	1	0
0	0	0	2	0	2	0
0	0	0	0	0	0	0

Channel 0

(e.g., red)

`x[:,:,1]`

0	0	0	0	0	0	0
0	1	0	2	0	1	0
0	0	0	0	0	2	0
0	0	0	2	1	0	0
0	2	1	1	1	2	0
0	1	1	2	2	2	0
0	0	0	0	0	0	0

Channel 1

(e.g., green)

`x[:,:,2]`

0	0	0	0	0	0	0
0	1	0	0	0	1	0
0	0	1	1	0	0	0
0	1	1	2	2	0	0
0	2	2	1	0	0	0
0	0	0	1	1	2	0
0	0	0	0	0	0	0

Channel 2

(e.g., blue)

Filter W0 (3x3x3)

`w0[:,:,0]`

0	1	0
-1	1	0
0	1	1

`w0[:,:,1]`

-1	-1	1
0	1	0
0	1	0

`w0[:,:,2]`

1	-1	-1
0	0	0
0	0	0

Bias b0 (1x1x1)

`b0[:,:,0]`

1

Filter W1 (3x3x3)

`w1[:,:,0]`

0	-1	-1
0	0	1
1	1	-1

`w1[:,:,1]`

0	-1	-1
-1	1	1
1	1	-1

`w1[:,:,2]`

0	-1	0
-1	1	0
-1	-1	0

Bias b1 (1x1x1)

`b1[:,:,0]`

0

Output Volume (3x3x2)

`o[:,:,0]`

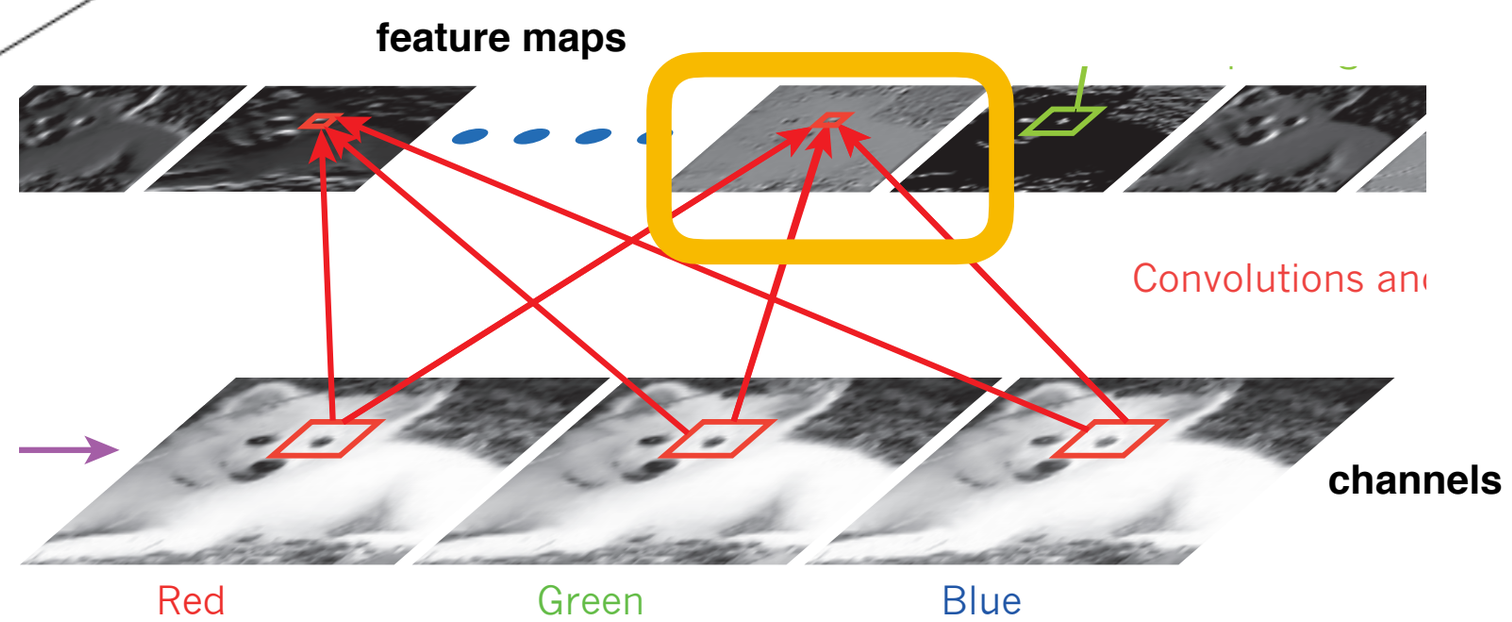
8	6	5
7	9	2
-1	7	3

Feature map 0

`o[:,:,1]`

2	3	7
-2	1	-2
-5	-1	-2

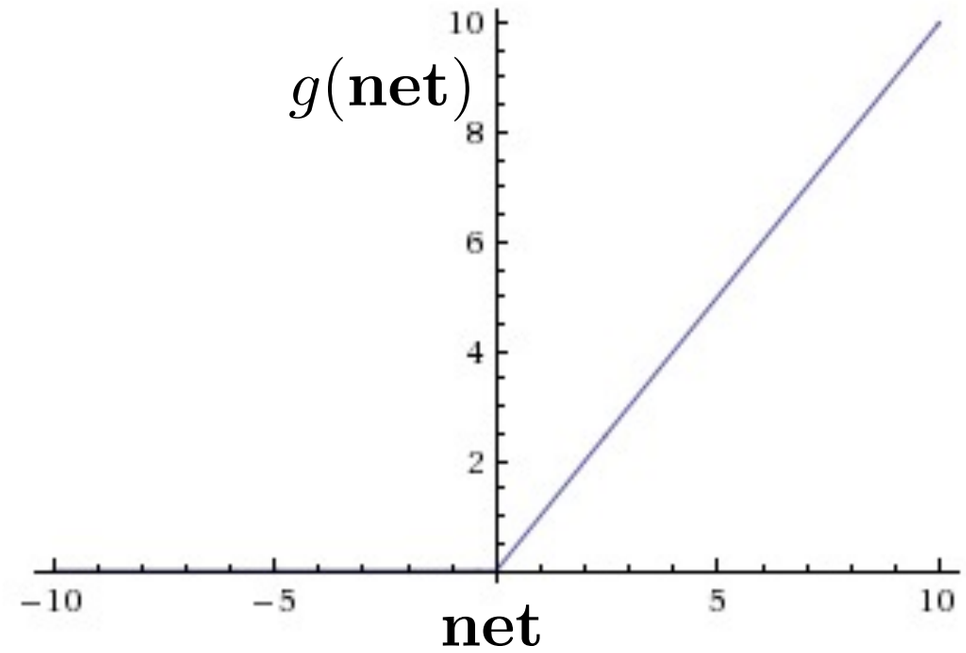
Feature map 1



After convolutions, apply a non-linear activation function (e.g., ReLUs)

$$g(\mathbf{net}) = \begin{cases} \mathbf{net} & \mathbf{net} \geq 0 \\ 0 & \mathbf{net} < 0 \end{cases}$$

Rectified linear unit (ReLU)



Feature map 0

8	6	5
7	9	2
-1	7	3

Feature map 1

2	3	7
-2	1	-2
-5	-1	-2

$g(\mathbf{net})$ →

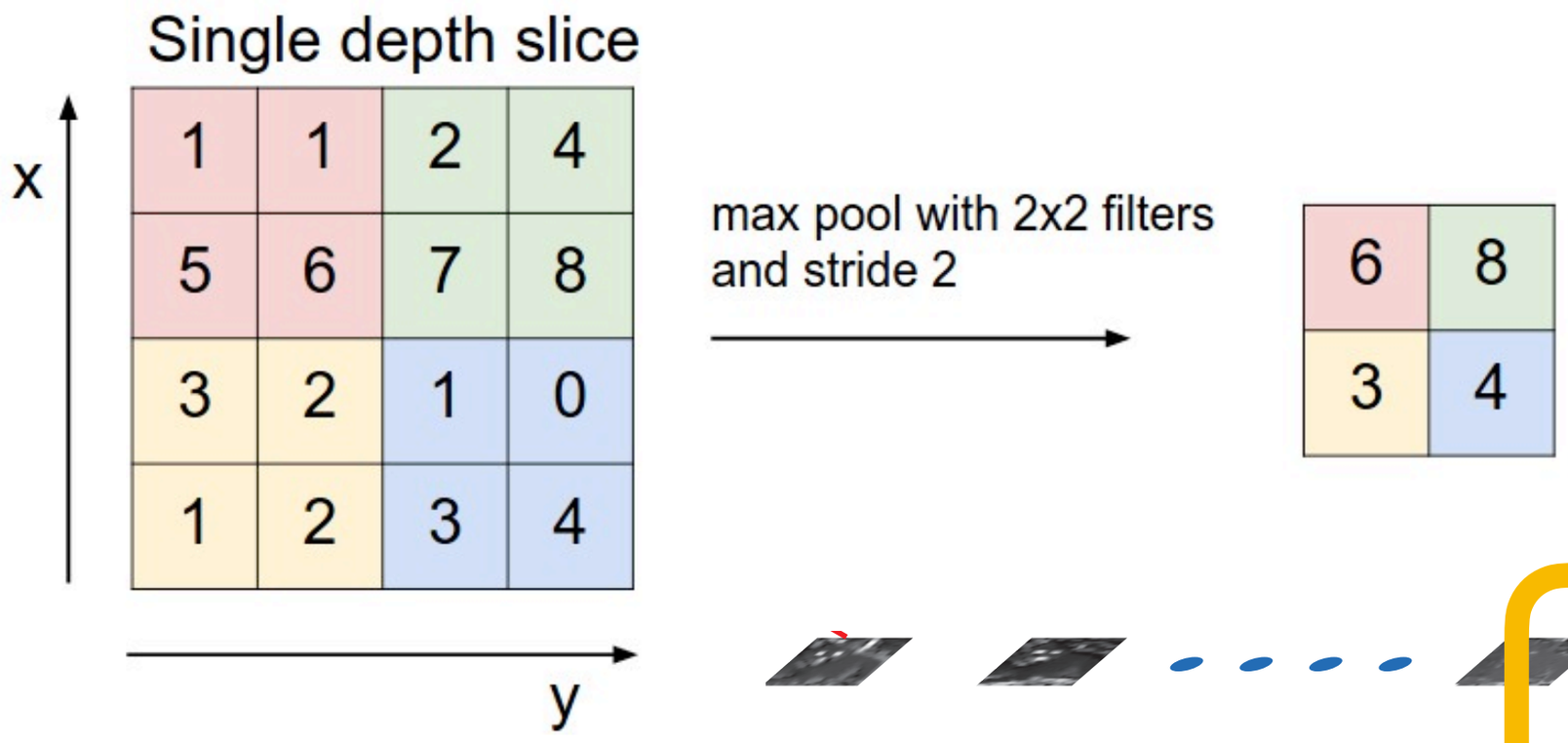
8	6	5
7	9	2
0	7	3

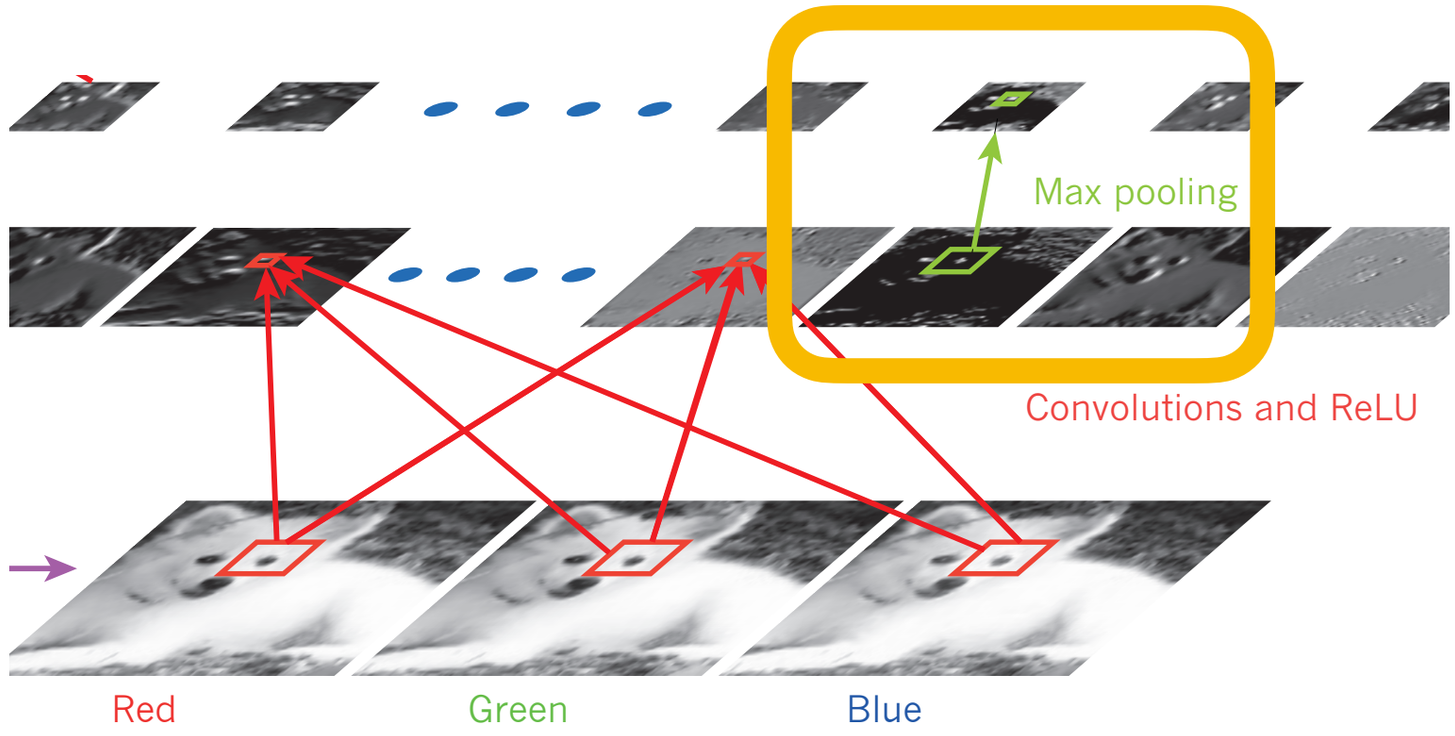
2	3	7
0	1	0
0	0	0

Next step: Pooling

- Downsamples a feature map to a coarser resolution
- Provides additional translation invariance

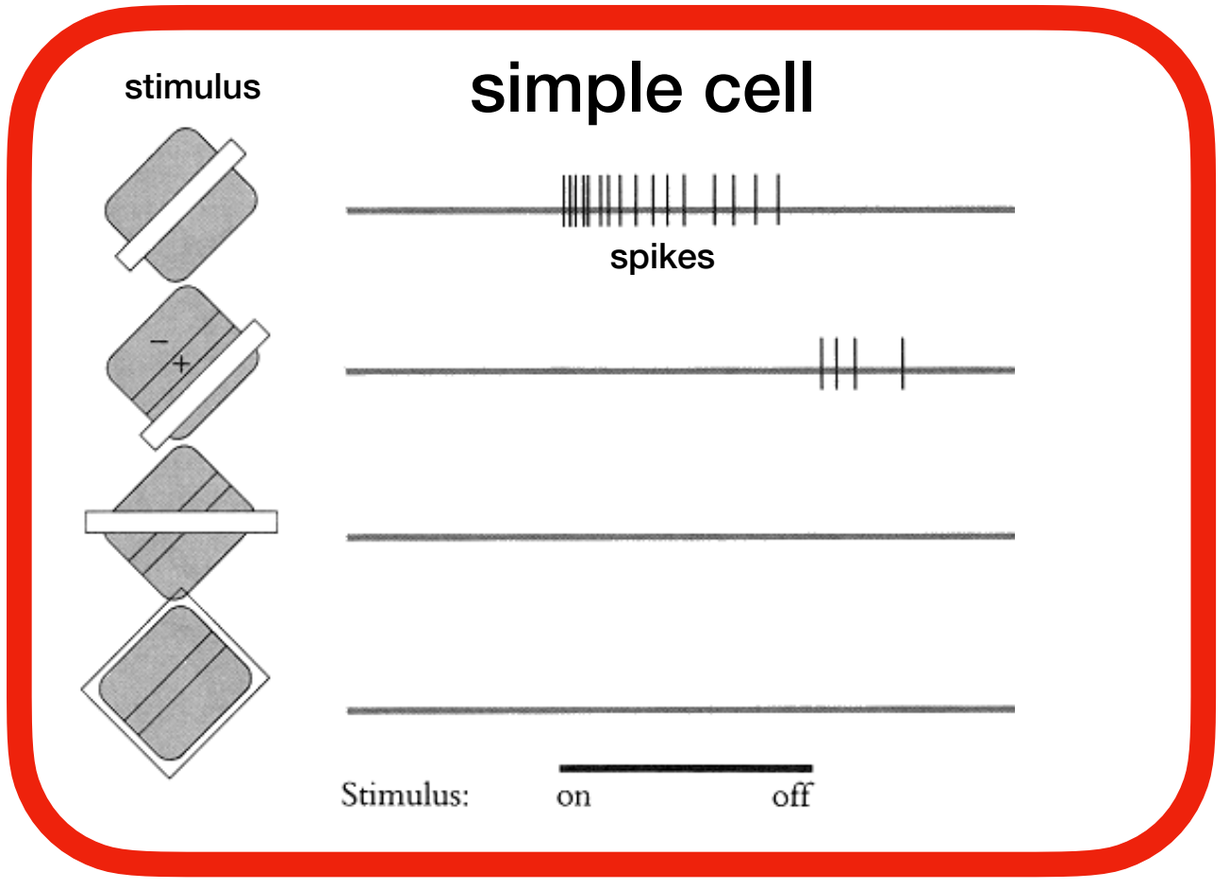
Max pooling



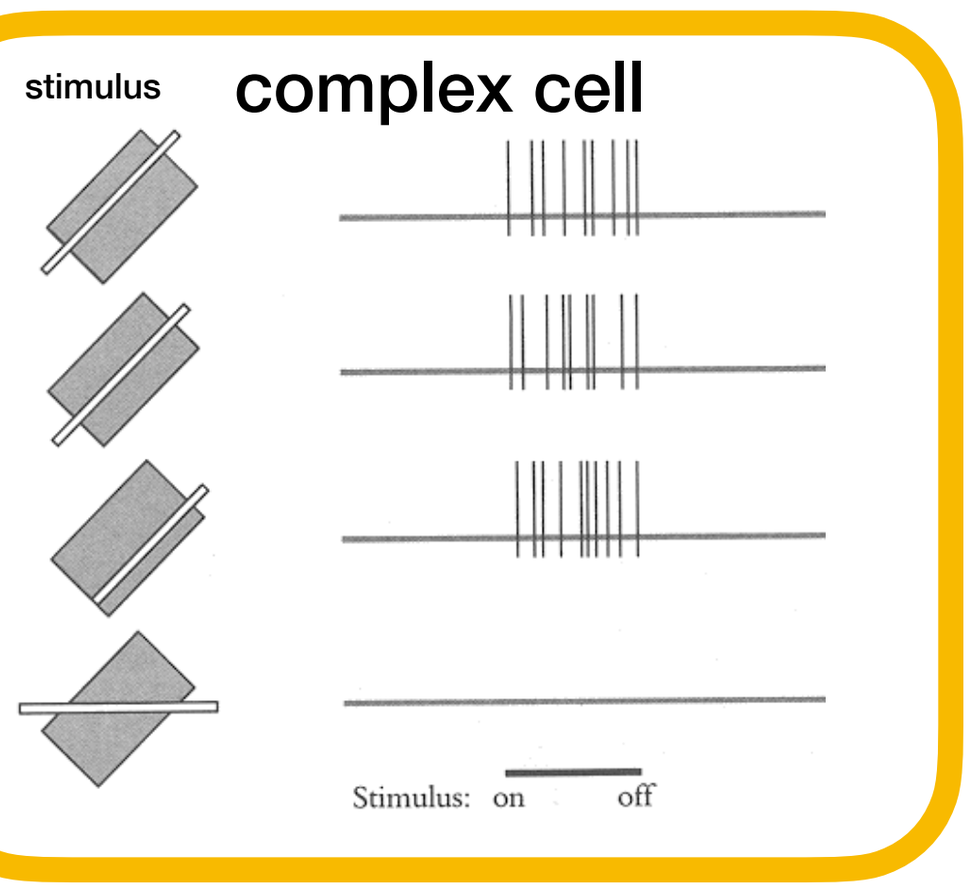


Connection to biological architecture of primary visual cortex

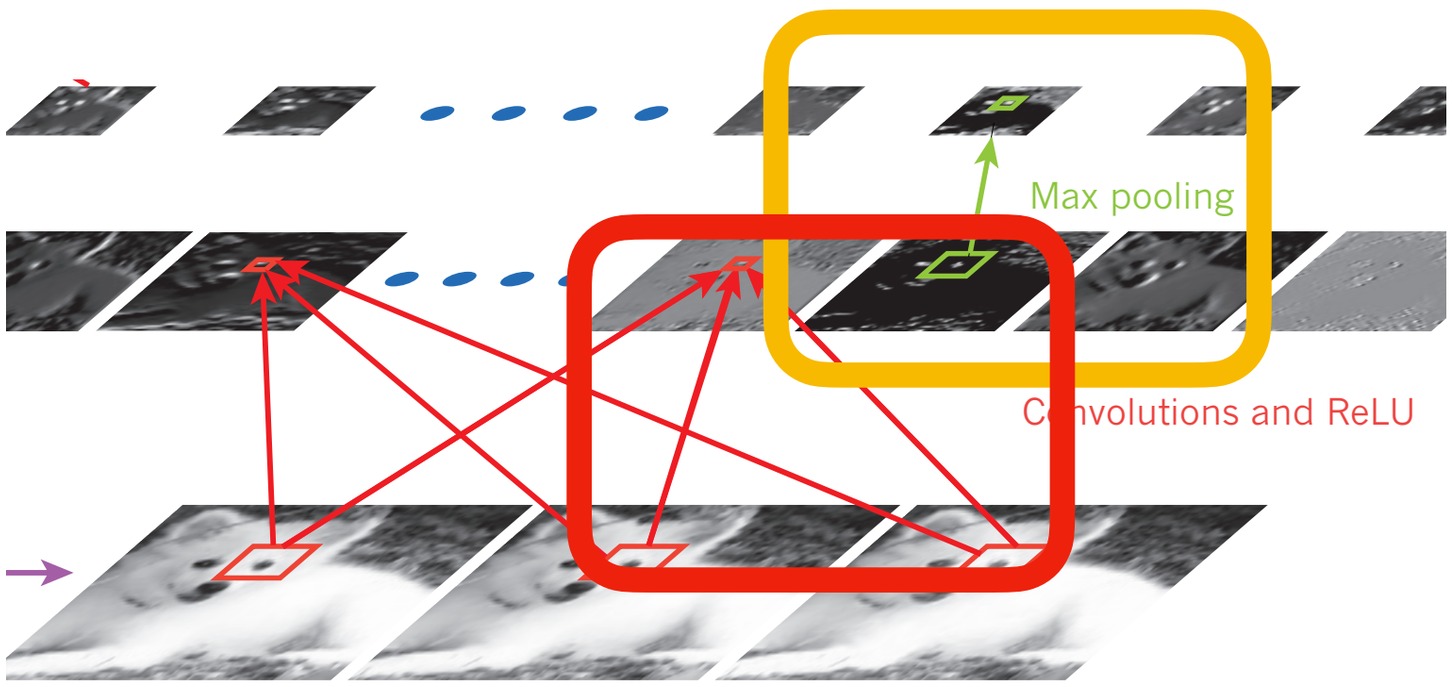
(Hubel & Wiesel)



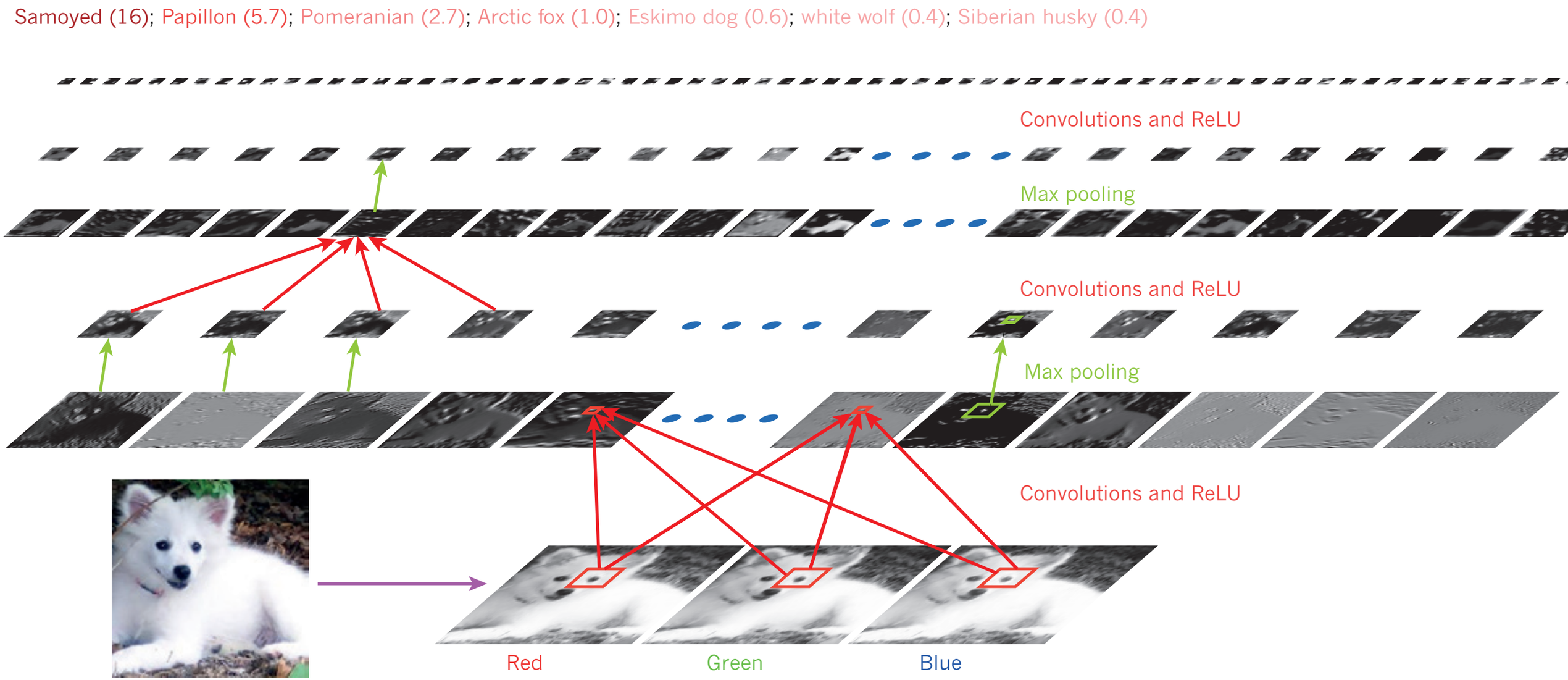
localized edge detector



invariance through pooling

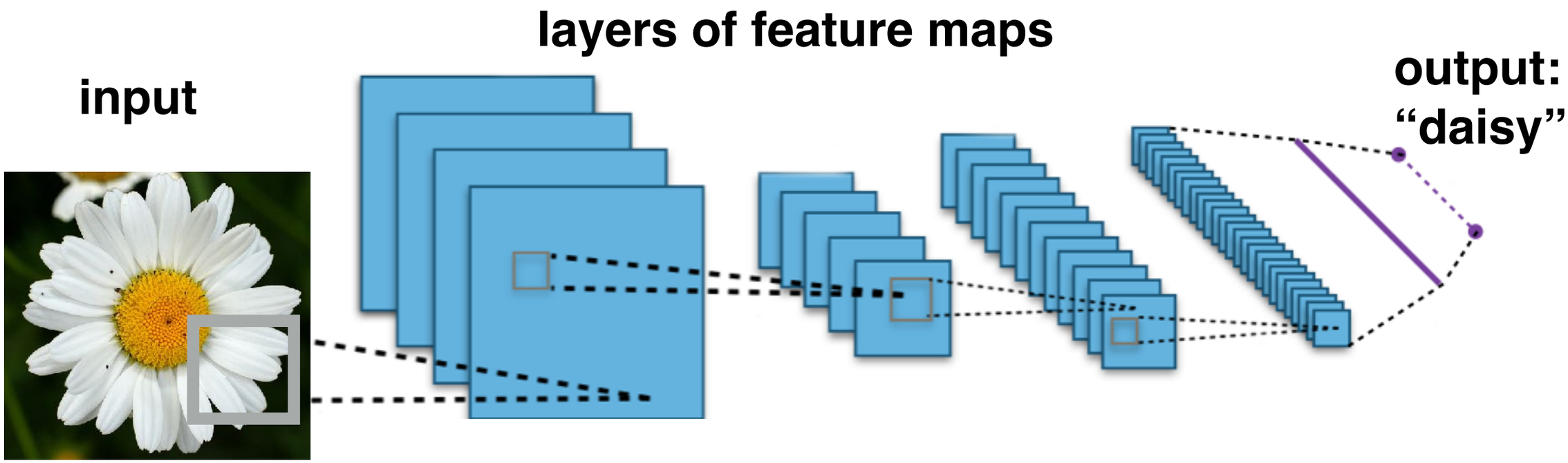


Deep convolutional neural network



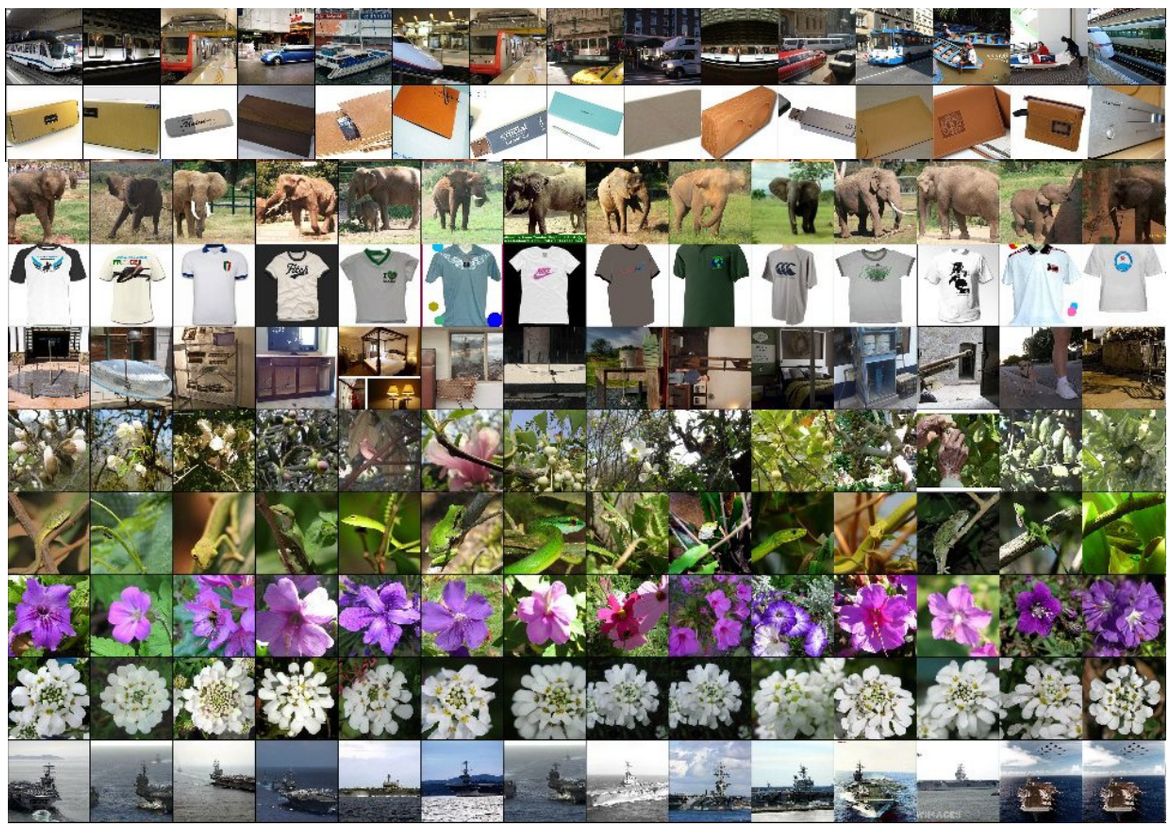
From LeCun, Bengio, & Hinton (2015).

How do we train it? Backpropagation



Training data (ImageNet)

- 1.2 million images
- 1000 categories
- ~1200 images per category



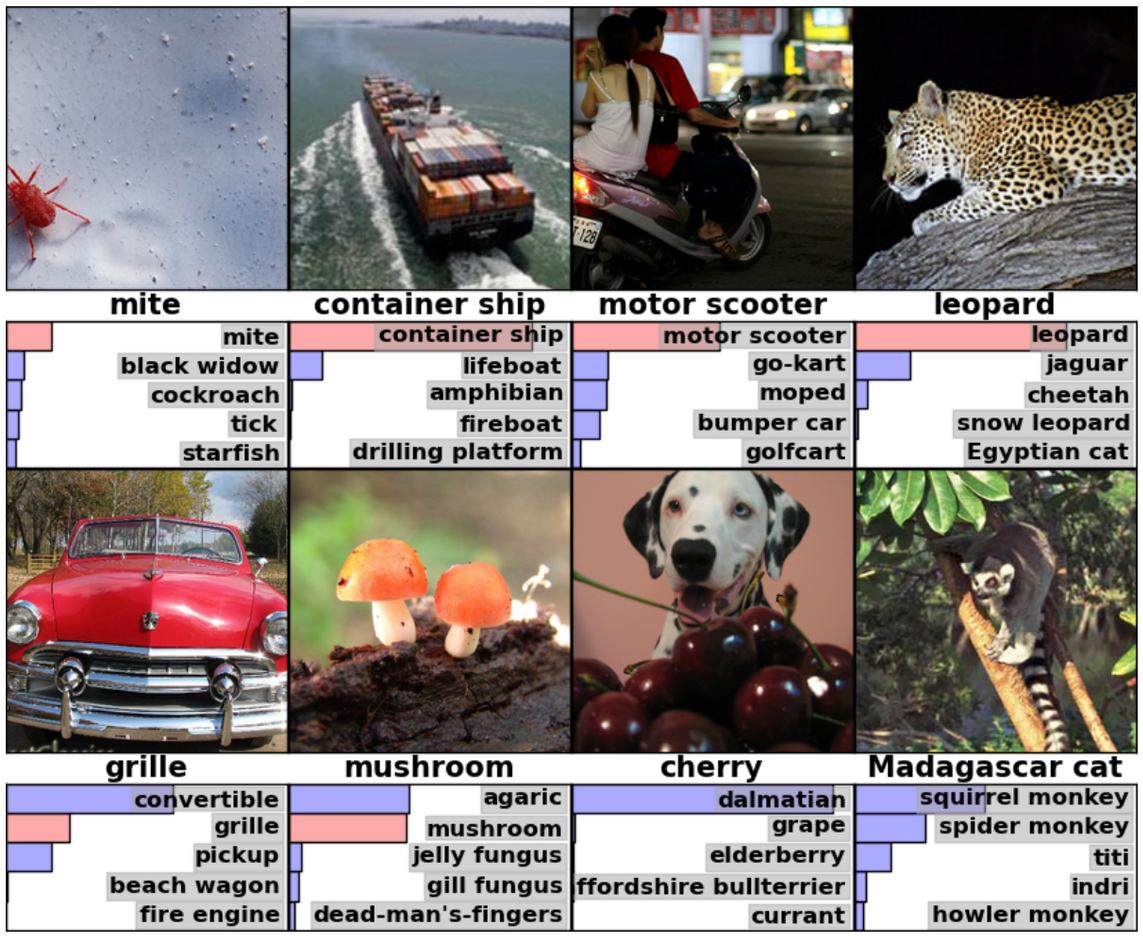
<u>AlexNet</u>
8 layers
~60 million parameters

<u>GoogLeNet</u>
22 layers
~5 million parameters

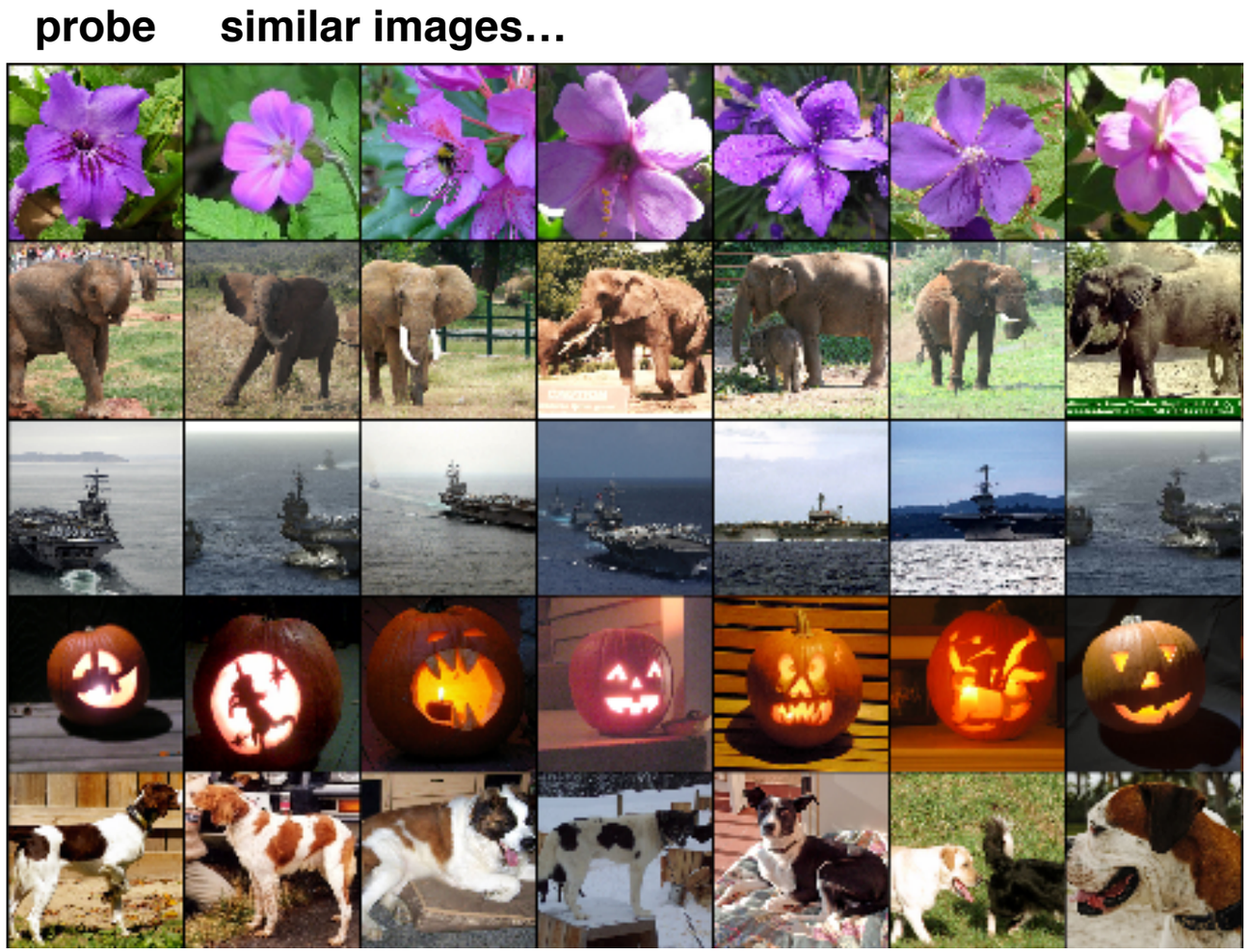
<u>Residual networks</u>
hundreds of layers…

Example results on test images

object recognition



finding similar images



(some filters learned by Krizhevsky et al., 2012)

Deep convnets for understanding psychological and neural representations

- Peterson, Abbott, and Griffiths (2016) explored convents for **predicting similarity ratings** between images.
- Lake, Zaremba, Fergus, and Gureckis (2015) showed that deep convnets can **predict human typicality ratings** for some classes of natural images
- Yamins et al. (2014) showed that deep convnets can **predict neural response in high-level visual areas**

Predicting human similarity ratings with a neural network

Peterson, J., Abbott, J., & Griffiths, T. (2016). Adapting Deep Network Features to Capture Psychological Representations.

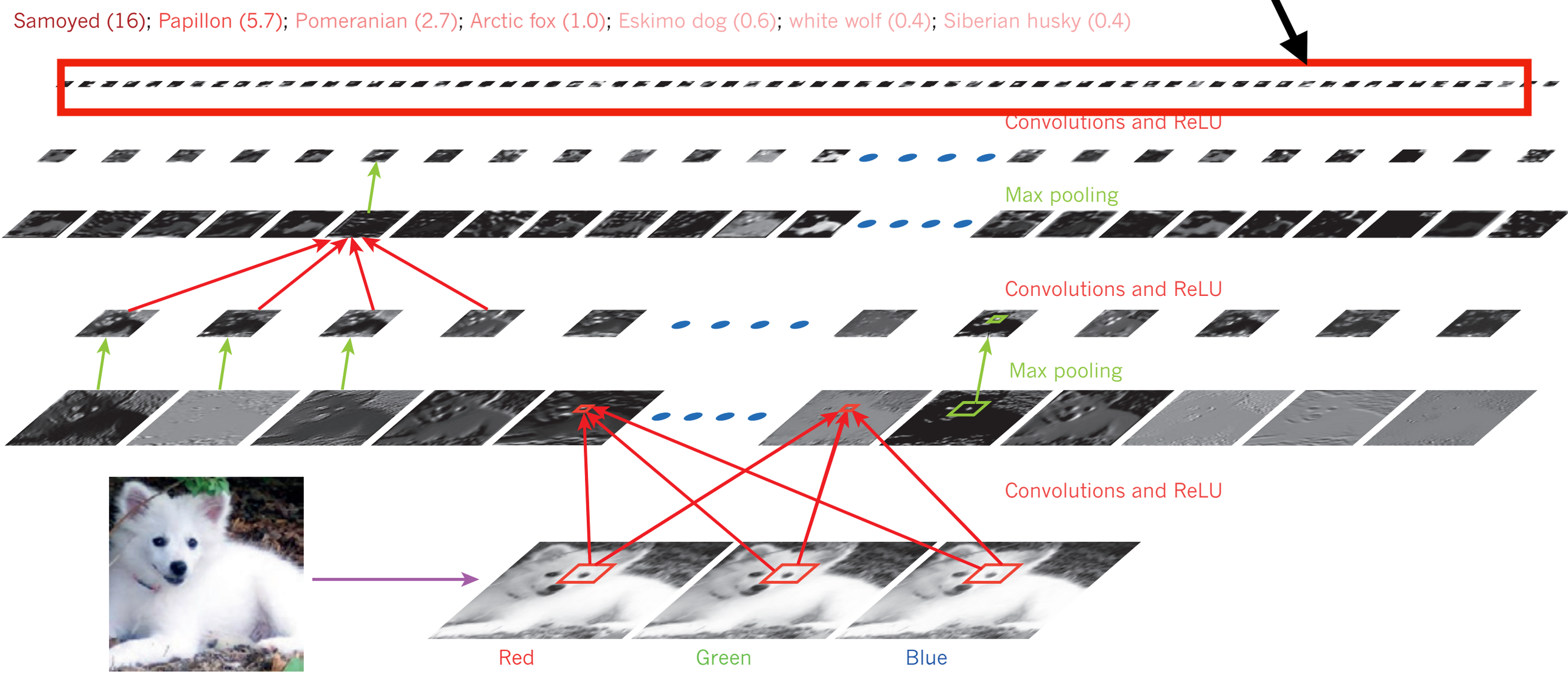
Some images look more similar to us than others. Can a neural network trained for classification help to explain similarity judgments from humans?

example animal images (they collected 120 x 120 pairwise ratings)

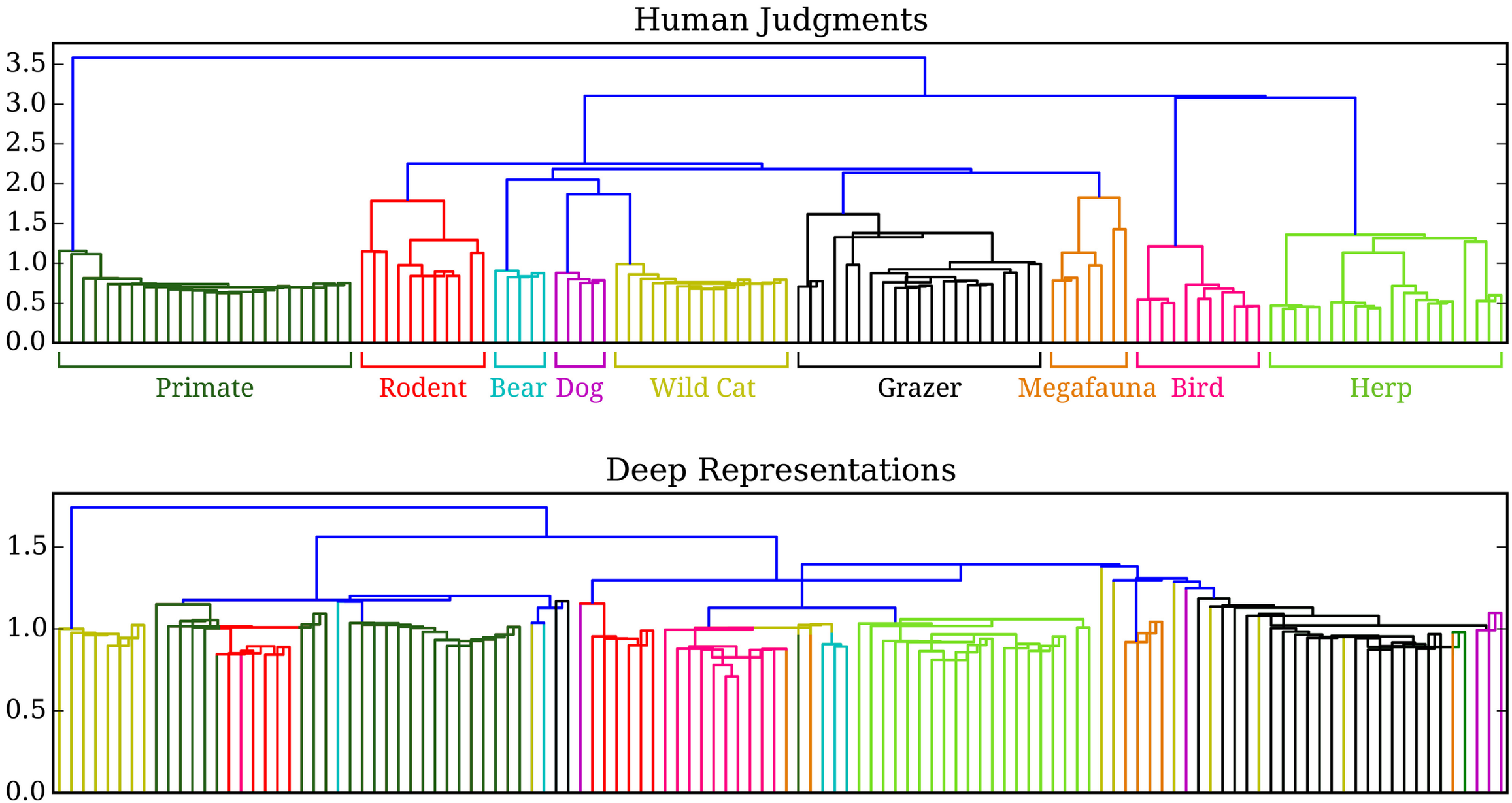
computing image-to-image similarity

$$sim(i,j) = \sum_k f_{ik} f_{jk}$$ similarity computed as dot product

summarizing images as high-level feature vector f

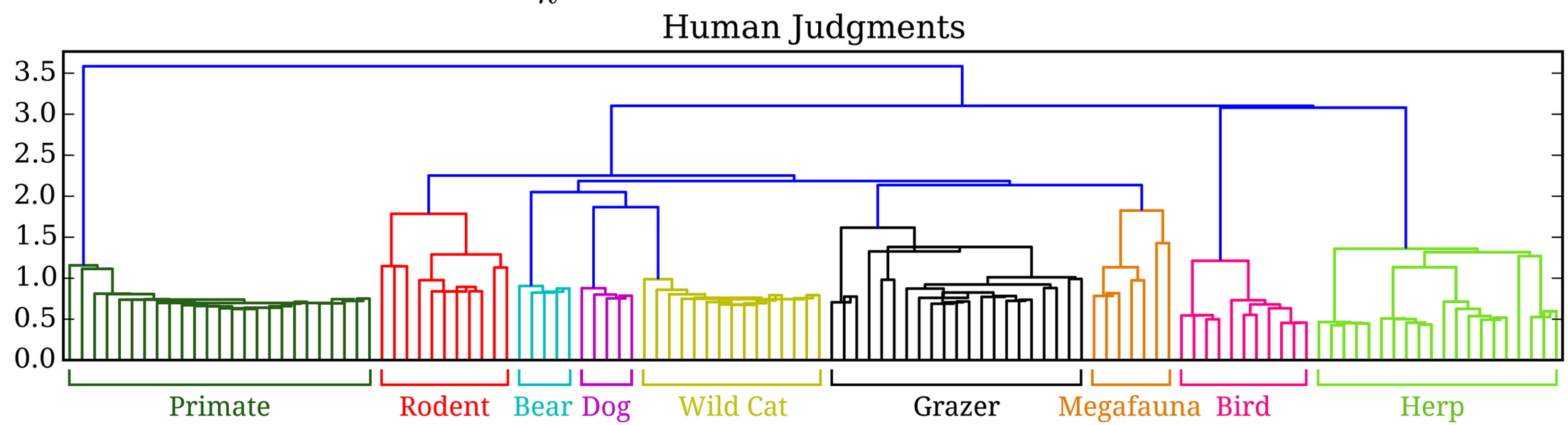


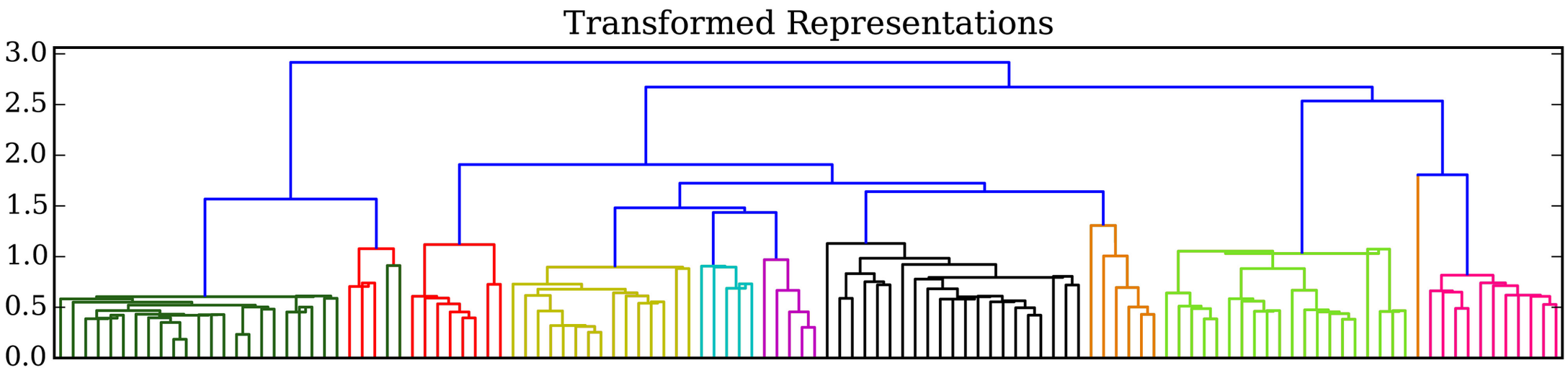
Hierarchical clustering reveals substantial differences in representation (best network explains about 40% of the variance in human judgments)



When fitting weights that allow the network features to be re-scaled, the network fits much better (best network explains about 84% of the variance in human judgments using out-of-sample predictions)

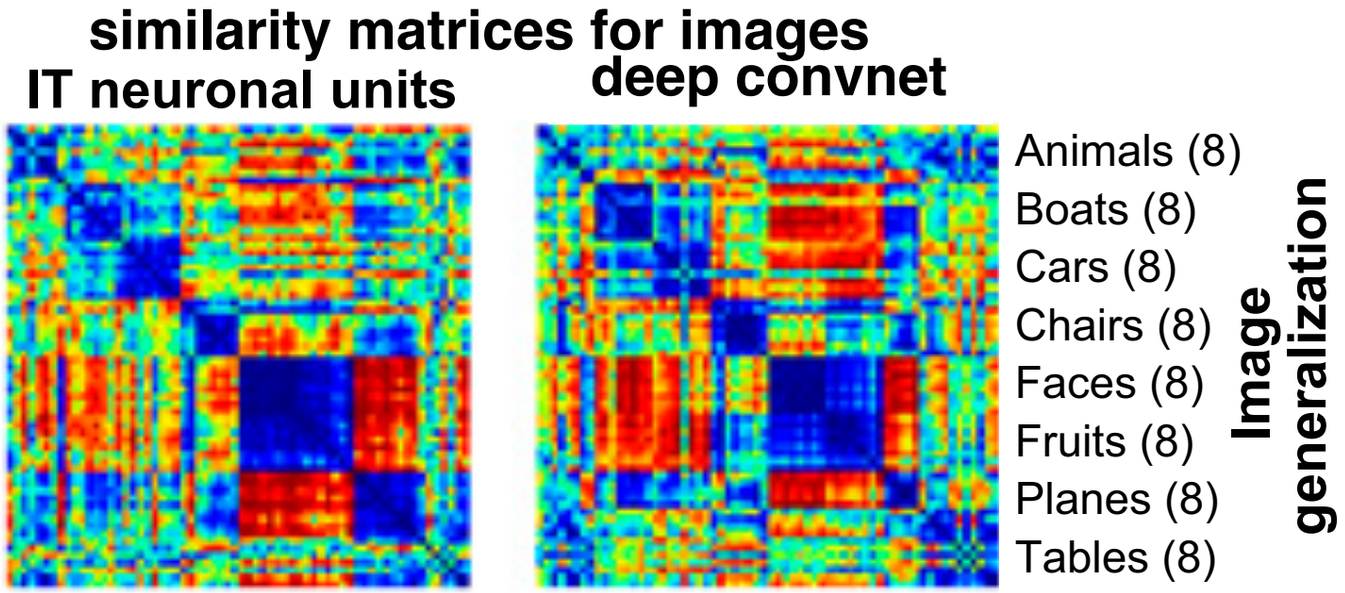
$$sim(i,j) = \sum_k w_k f_{ik} f_{jk} \qquad w_k : \text{weight for feature k}$$

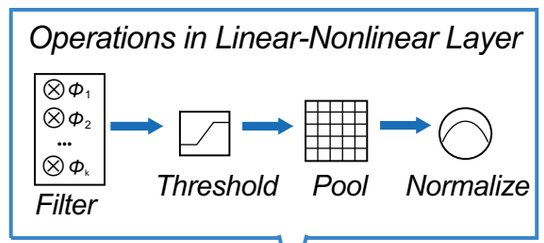




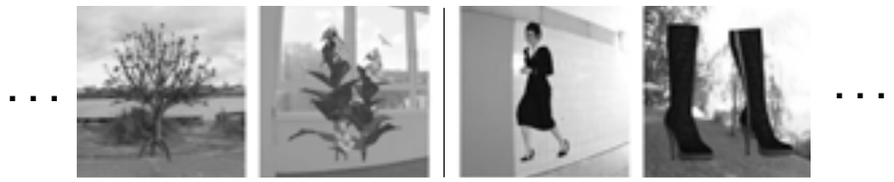
Predicting neural recordings with a deep convnet

(Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619-8624.)

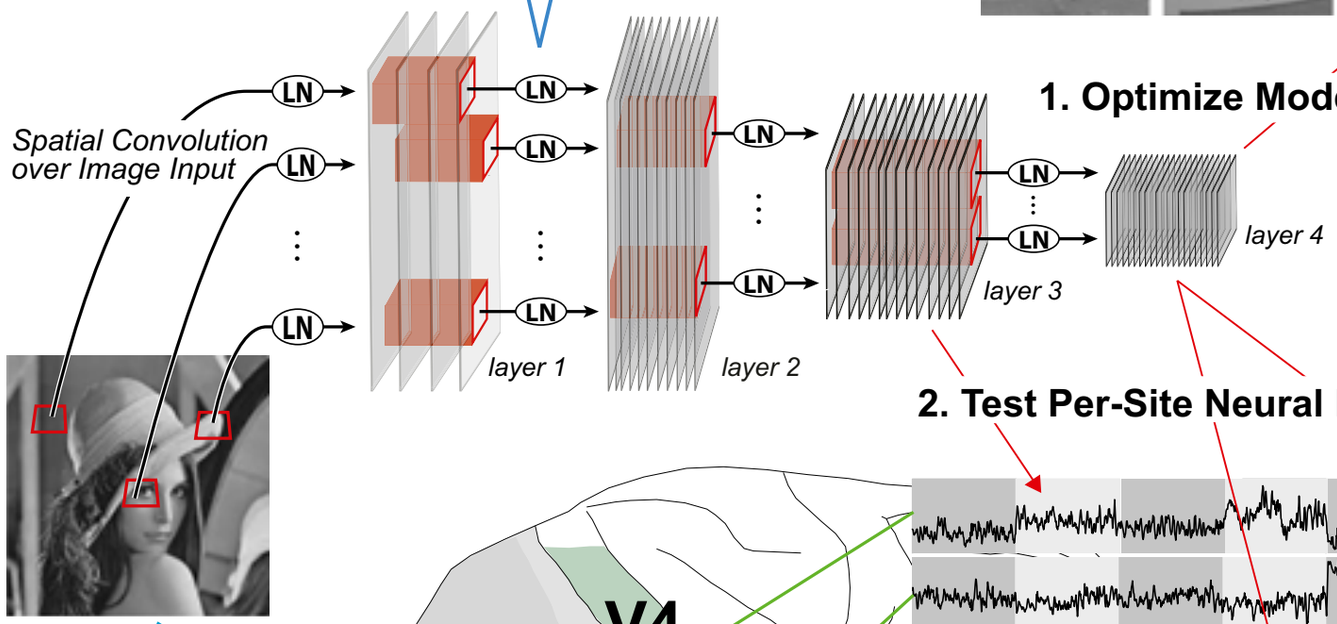




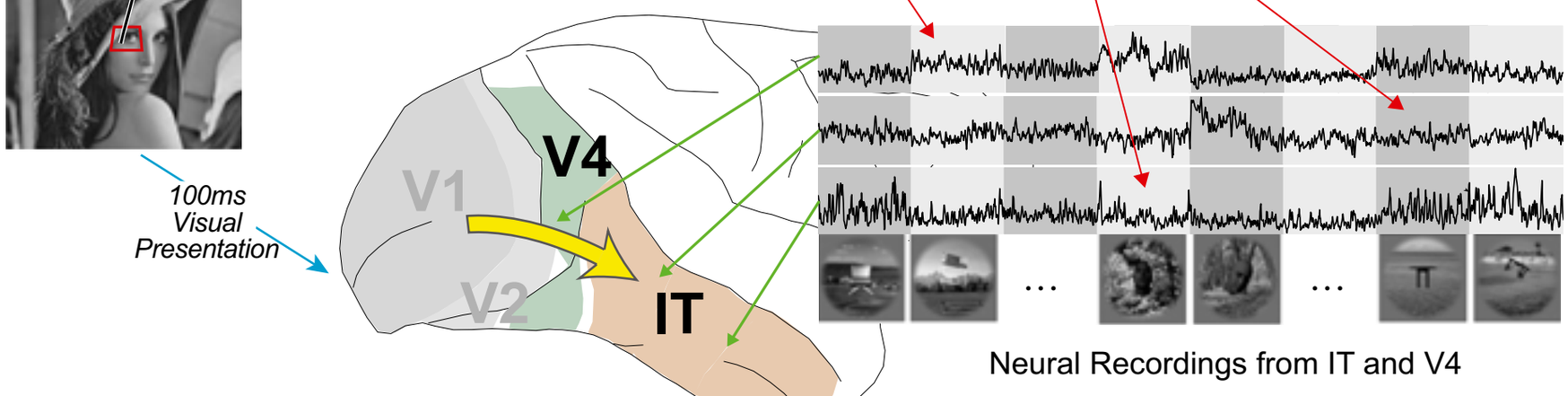
Behavioral Tasks
e.g. Trees vs non-Trees



1. Optimize Model for Task Performance



2. Test Per-Site Neural Predictions



Neural Recordings from IT and V4

Understanding category typicality with deep convnets

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep Neural Networks Predict Category Typicality Ratings for Images.

typical dog

atypical dog

• For people, typicality influences performance in practically any category-related task

- speed of categorization
- ease of production
- ease of learning
- usefulness for inductive inference

• No known model successfully predicts typicality ratings from raw images -- How do convnets perform?

Category: Banana (ρ=0.82)

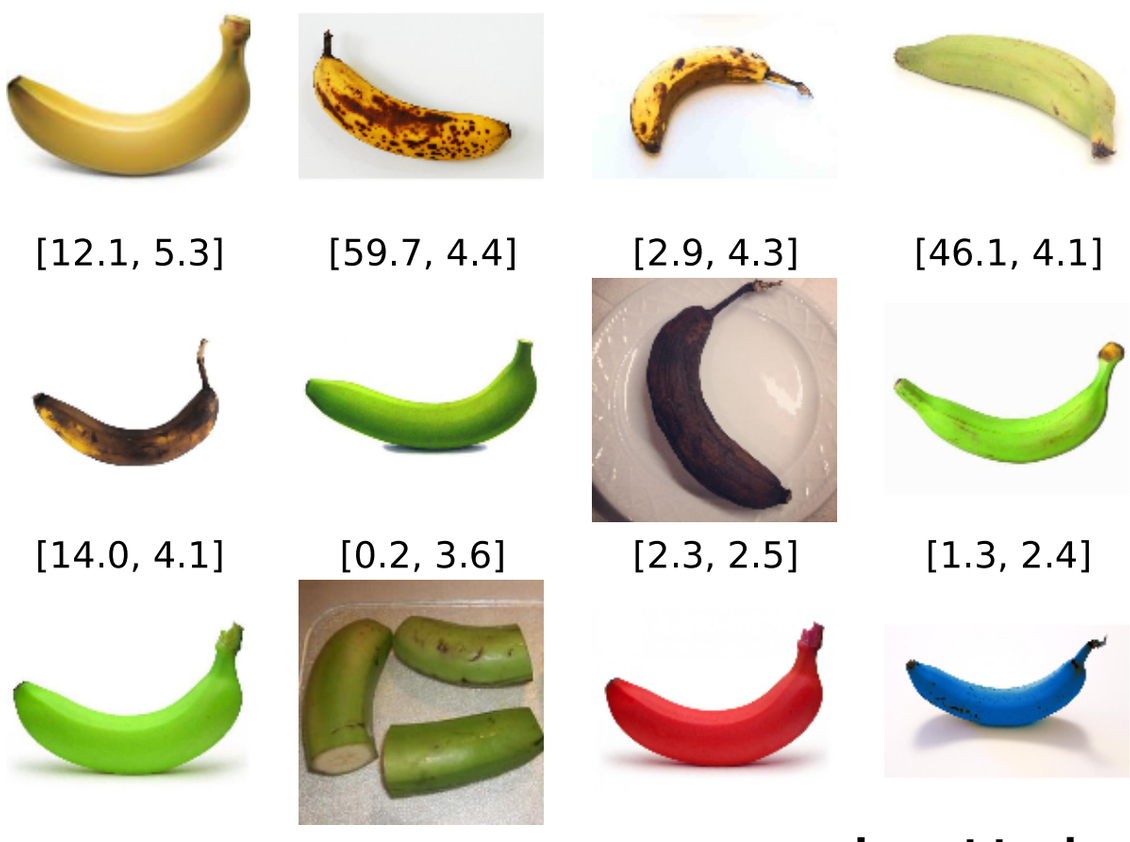
How well does this picture fit your idea or image of the category? (rated on 1-7 scale)

Human typicality ratings

Most typical →

[97.8, 6.8] [98.0, 6.8] [96.6, 6.8] [99.7, 6.6]

[96.9, 6.6] [99.3, 6.0] [78.6, 5.8] [99.5, 5.5]



Least typical

rating key: [machine (0-100), human (1-7)]

Category: Bathtub (ρ=0.68)

Human typicality ratings

Most typical →

Least typical

Convnet typicality ratings

rating key: [machine (0-100), human (1-7)]

Category: Envelope (ρ=0.79)

Human typicality ratings

Most typical →

[91.5, 6.7] [75.2, 6.6] [96.4, 6.6] [98.5, 6.6]

[97.7, 6.6] [82.8, 6.2] [69.5, 5.3] [59.7, 5.2]

[31.2, 5.1] [32.5, 5.1] [10.8, 4.8] [5.8, 4.2]

[10.4, 4.1] [50.6, 3.8] [41.9, 3.4] [24.9, 3.2]

Least typical

Convnet typicality ratings

[98.5, 6.6] [97.7, 6.6] [96.4, 6.6] [91.5, 6.7]

[82.8, 6.2] [75.2, 6.6] [69.5, 5.3] [59.7, 5.2]

[50.6, 3.8] [41.9, 3.4] [32.5, 5.1] [31.2, 5.1]

[24.9, 3.2] [10.8, 4.8] [10.4, 4.1] [5.8, 4.2]

rating key: [machine (0-100), human (1-7)]

Category: Teapots (ρ=0.38)

Human typicality ratings

Most typical →

Least typical

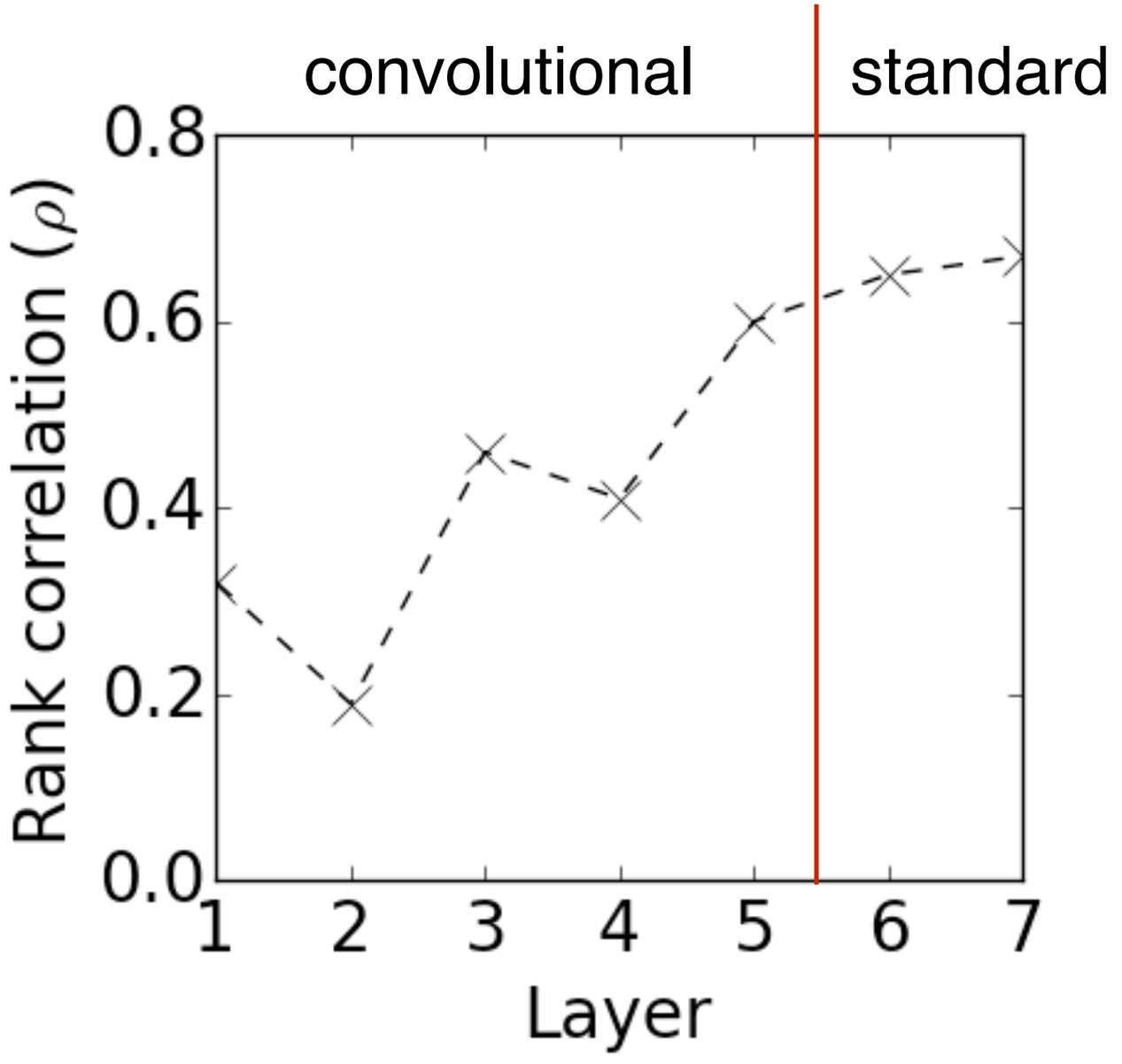
Convnet typicality ratings

rating key: [machine (0-100), human (1-7)]

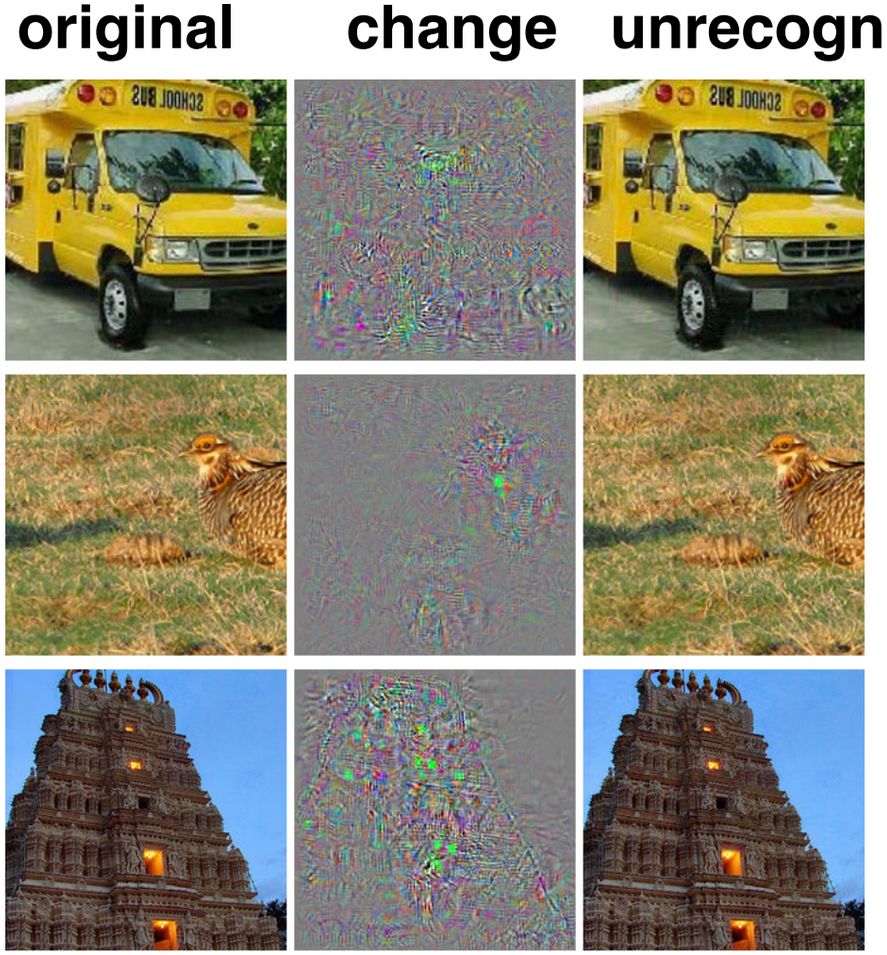
Summary of typicality predictions

	Rank Correlation
Banana	0.82
Bathtub	0.68
Coffee Mug	0.62
Envelope	0.79
Pillow	0.67
Soap dispenser	0.74
Table lamp	0.69
Teapot	0.38
Average	**0.67**

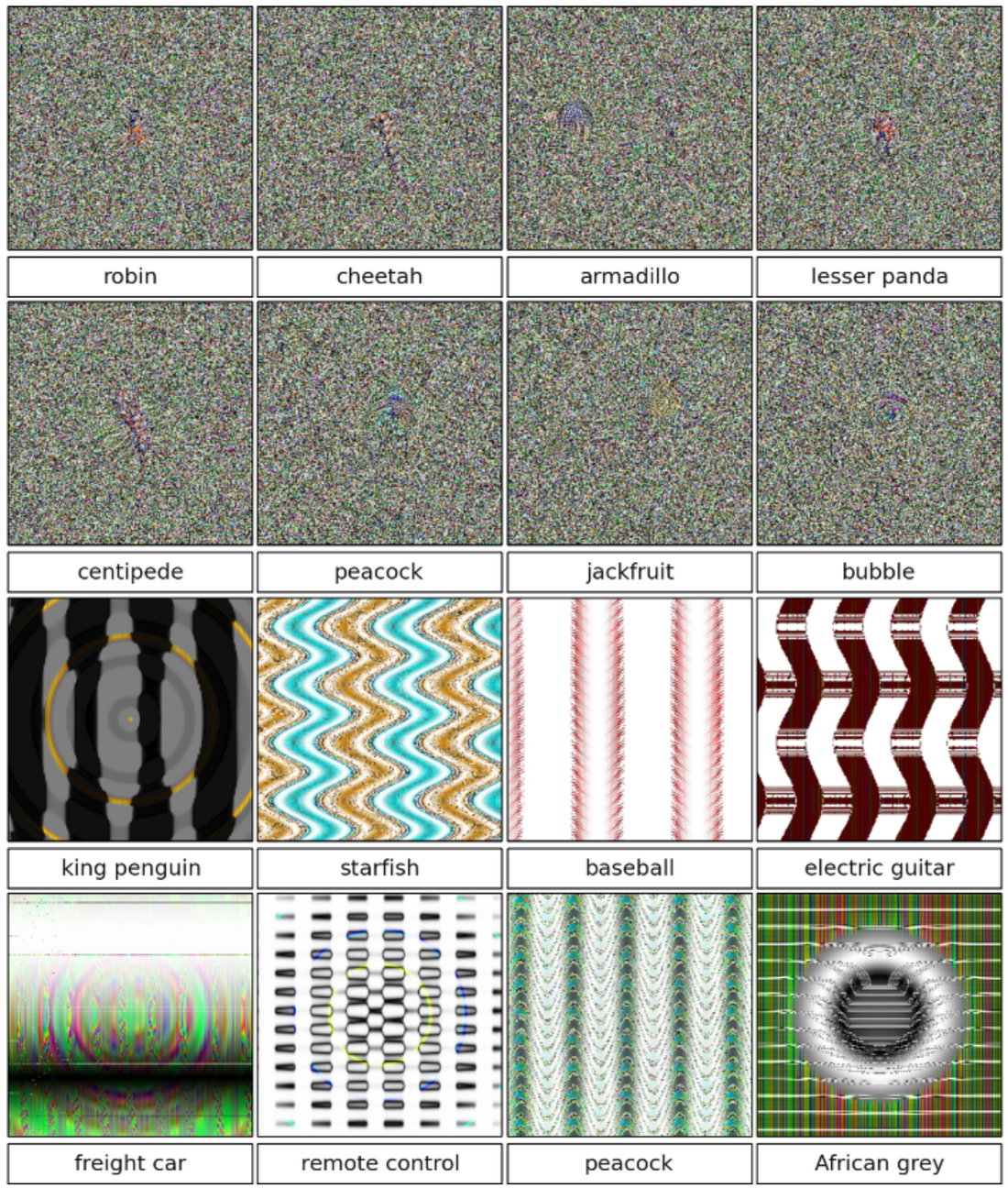
Prediction quality varies as a function of network depth.



Critiques of deep convolutional networks



(Szegedy et al., 2013)



(Nguyen et al., 2015)

Critiques of deep convolutional networks

Compare to deep convnets, people can learn much richer concepts from less data.

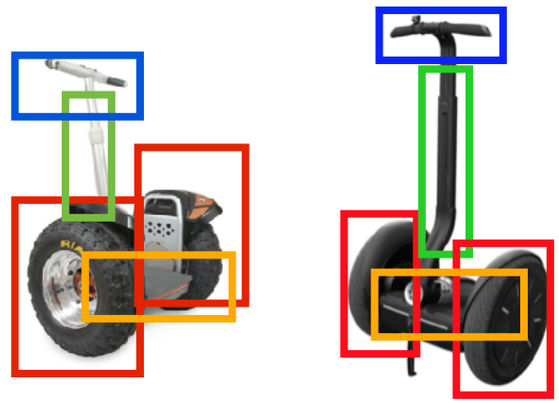
People learn from less data

"one-shot learning"

where are the others?

People learn richer concepts

parsing



generating new concepts

generating new examples