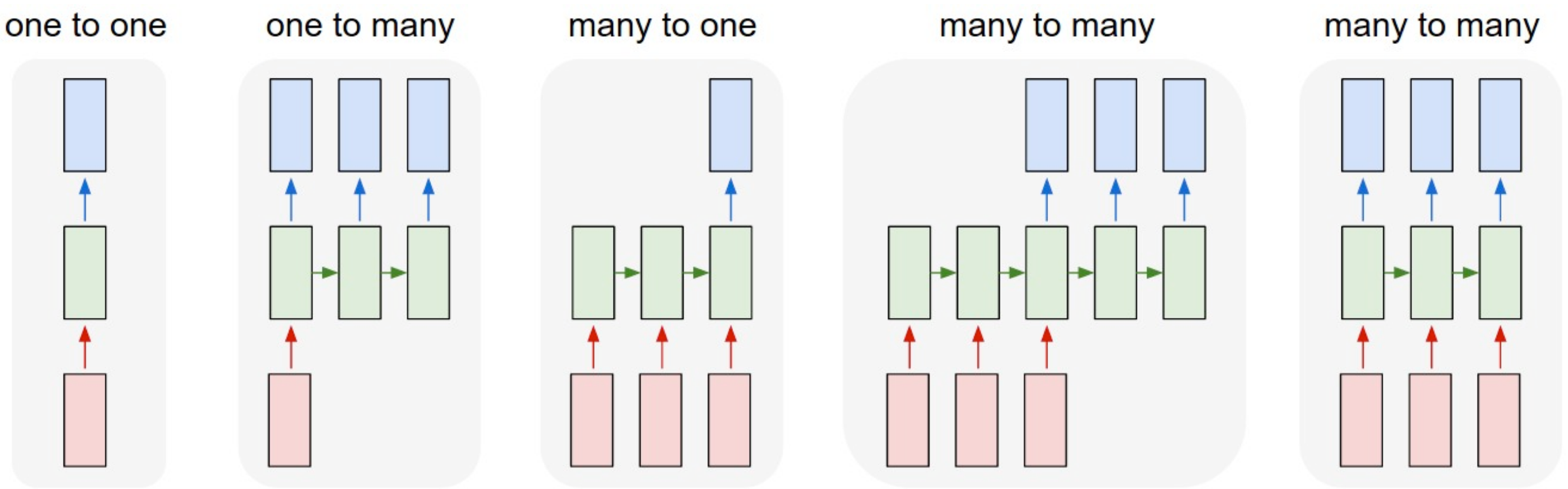
Recurrent Neural Nets
&
Visual Captioning

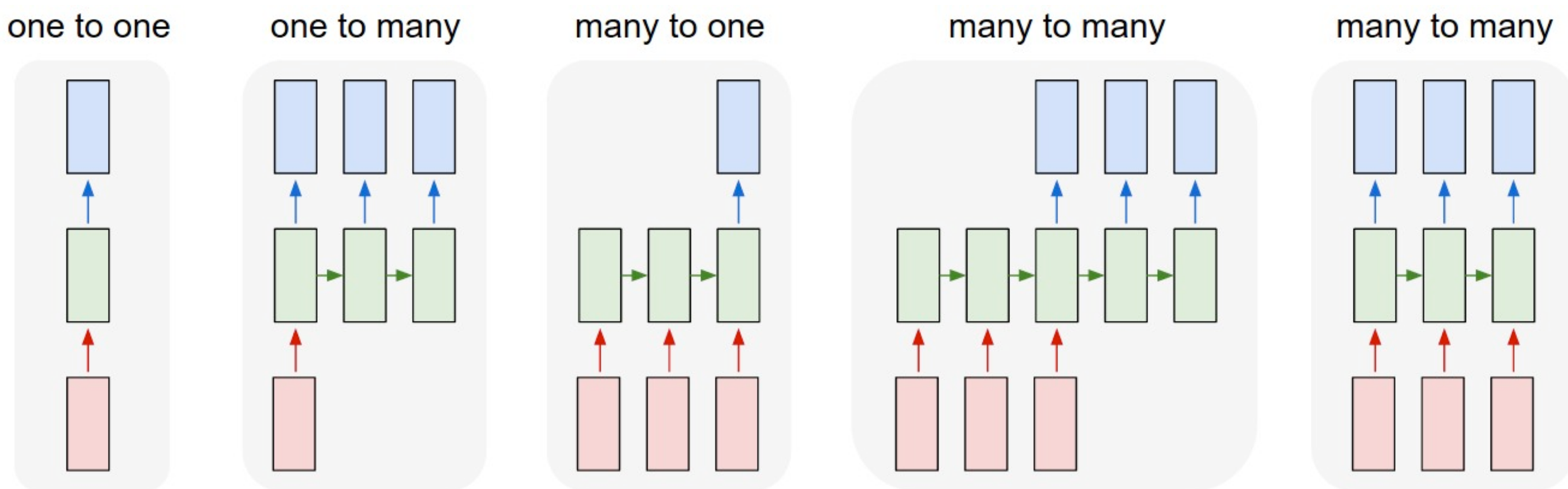
Lecture 6

Slides from: Dhruv Bhatra, Fei-Fei Li, Justin Johnson,
Serena Yeung, Andrej Karpathy

Recurrent Neural Nets



Recurrent Neural Nets



Input: No sequence

Output: No sequence

Example: “standard” classification / regression problems

Input: No sequence

Output: Sequence

Example: Im2Caption

Input: Sequence

Output: No sequence

Example: sentence classification, multiple-choice question answering

Input: Sequence

Output: Sequence

Example: machine translation, video captioning, open-ended question answering, video question answering

Image Credit: Andrej Karpathy

Synonyms

- Recurrent Neural Networks (RNNs)

- Types:
  - “Vanilla” RNNs
  - Long Short Term Memory (LSTMs)
  - Gated Recurrent Units (GRUs)
  - …

- Algorithms
  - BackProp Through Time (BPTT)

What's wrong with MLPs/ConvNets?

- Problem 1: Can't model sequences
  - Fixed-sized Inputs & Outputs
  - No temporal structure

- Problem 2: Pure feed-forward processing
  - No "memory", no feedback

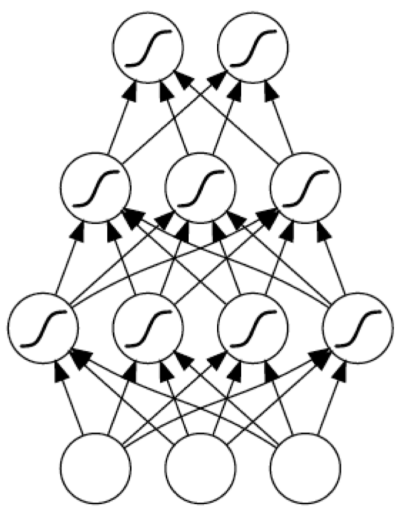


Image Credit: Alex Graves, book

Sequences are everywhere…

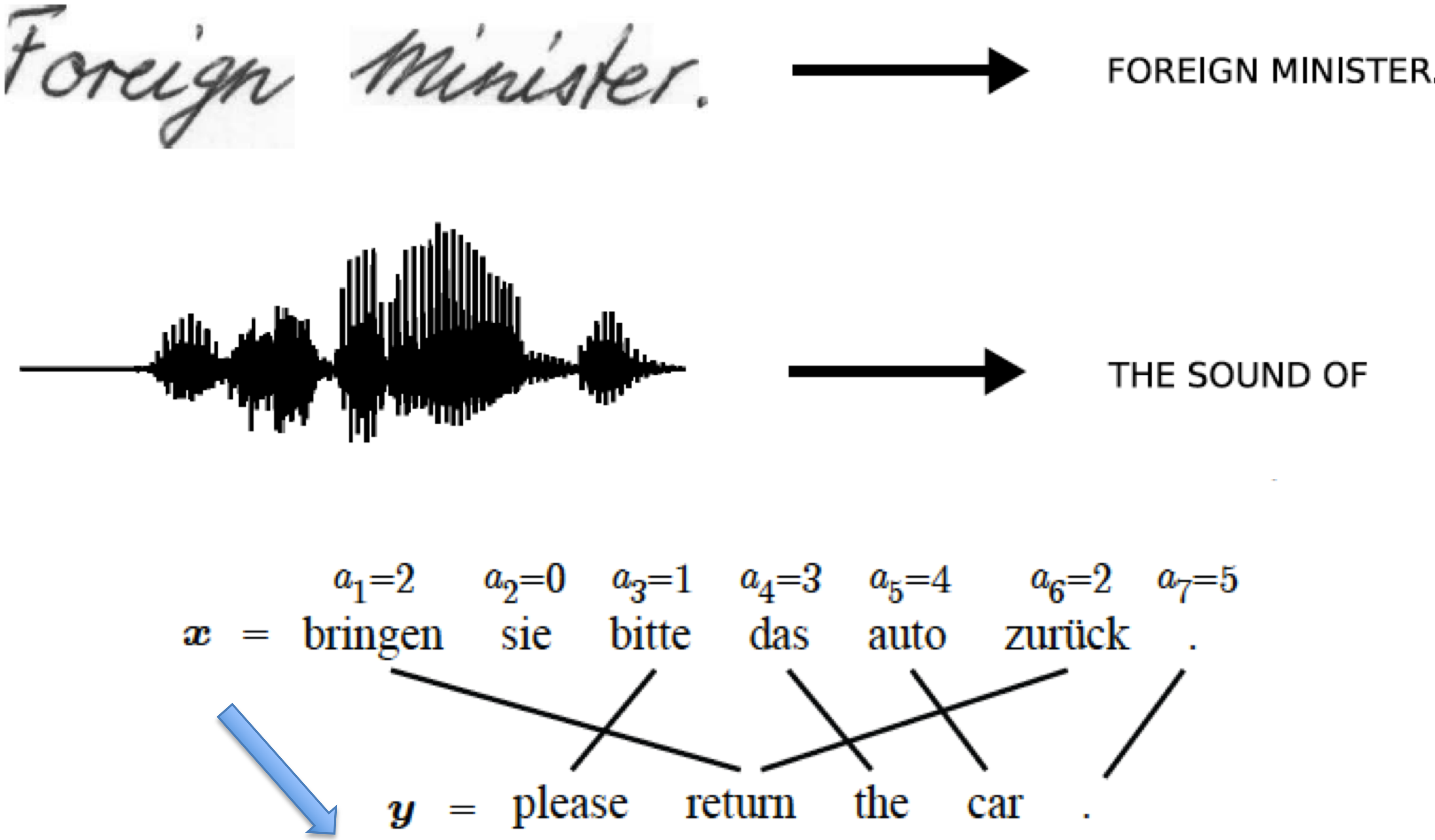
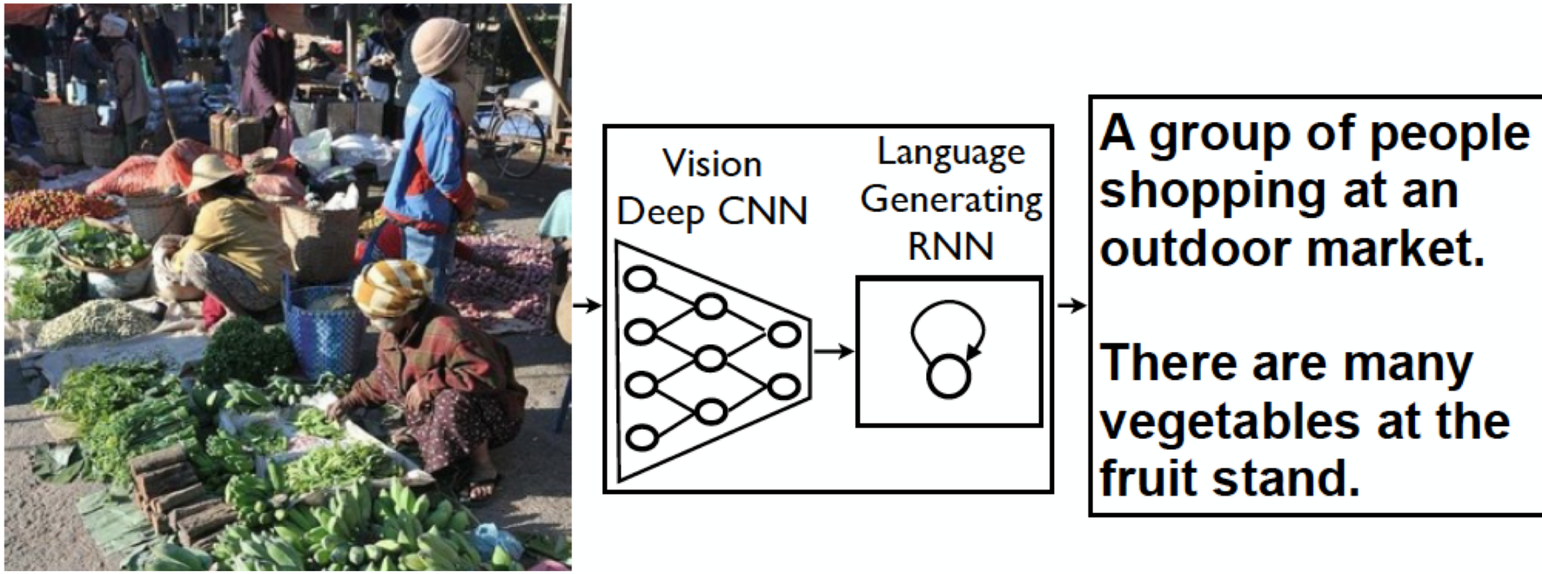
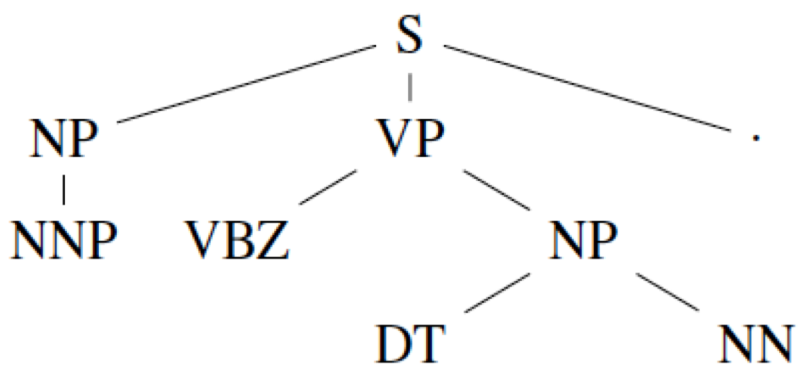


Image Credit: Alex Graves and Kevin Gimpel

Even where you might not expect a sequence…

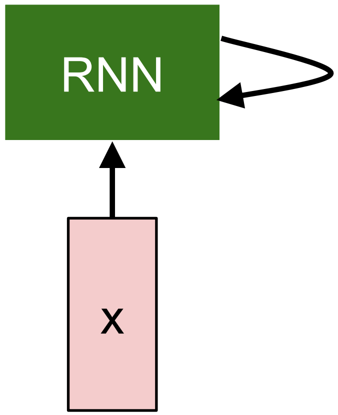


John has a dog . →



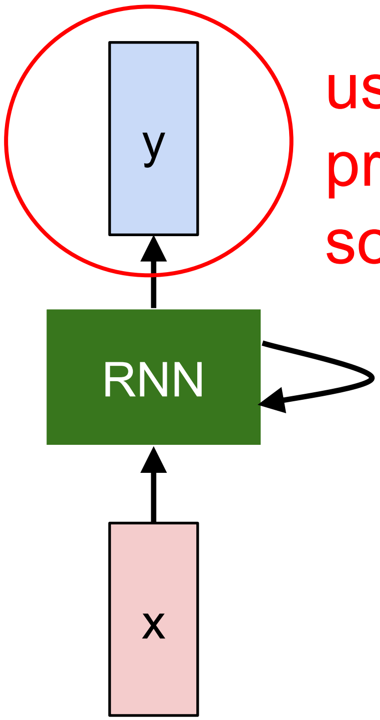
John has a dog . → (S (NP NNP $)_{NP}$ (VP VBZ (NP DT NN $)_{NP}$ $)_{VP}$. $)_{S}$

Recurrent Neural Network



Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n

Recurrent Neural Network



usually want to predict a vector at some time steps

Recurrent Neural Network

We can process a sequence of vectors **x** by applying a **recurrence formula** at every time step:

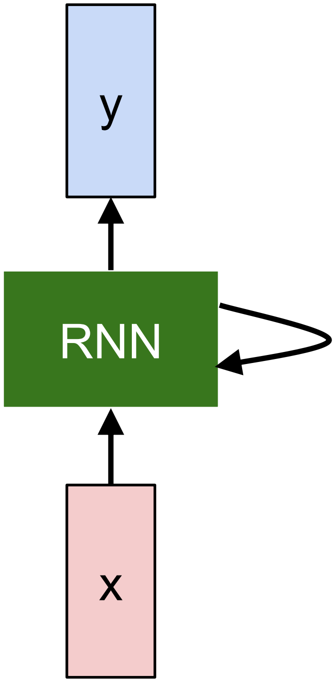
$$\boxed{h_t} = \boxed{f_W}(\boxed{h_{t-1}}, \boxed{x_t})$$

new state

old state

input vector at some time step

some function with parameters W

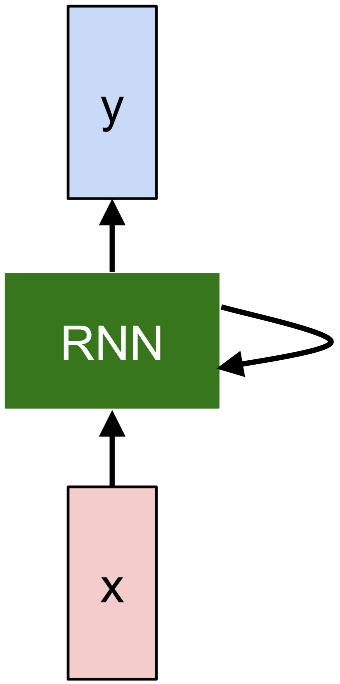


Recurrent Neural Network

We can process a sequence of vectors **x** by applying a **recurrence formula** at every time step:

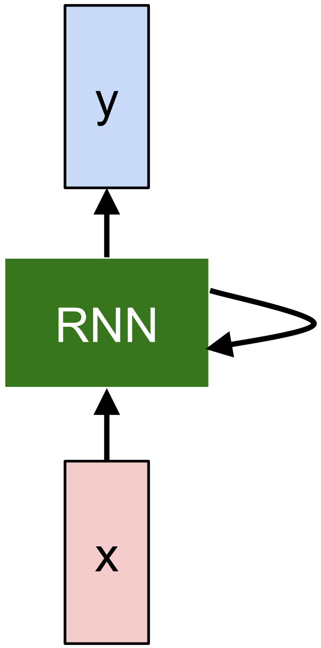
$$h_t = f_W(h_{t-1}, x_t)$$

Notice: the same function and the same set of parameters are used at every time step.



(Vanilla) Recurrent Neural Network

The state consists of a single *“hidden”* vector **h**:

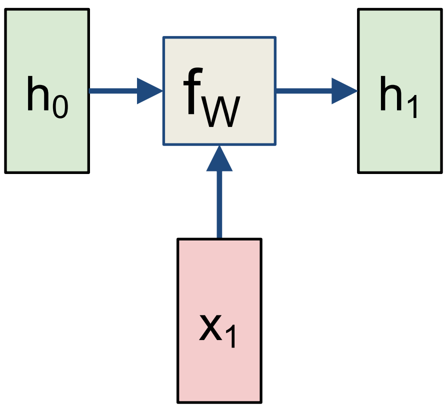


$$y_t = W_{hy}h_t + b_y$$

$$h_t = f_W(h_{t-1}, x_t)$$

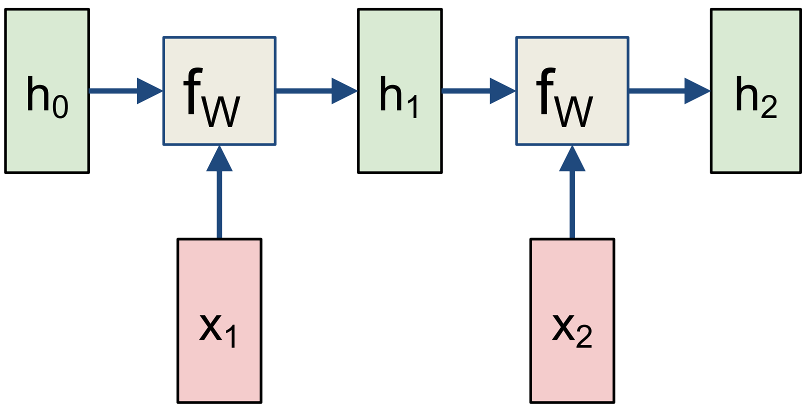
$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t + b_h)$$

RNN: Computational Graph

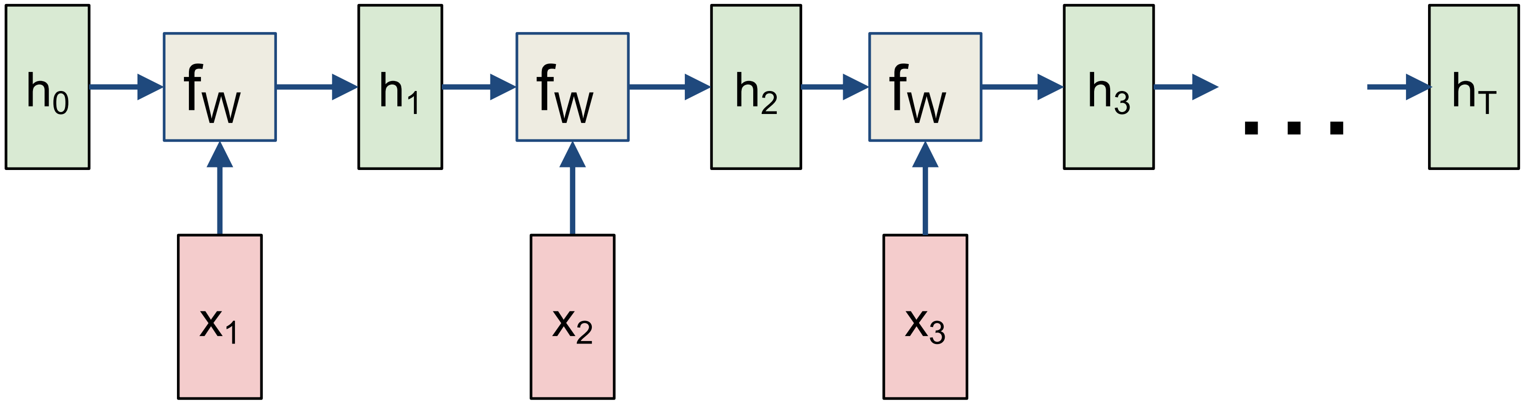


Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n

RNN: Computational Graph

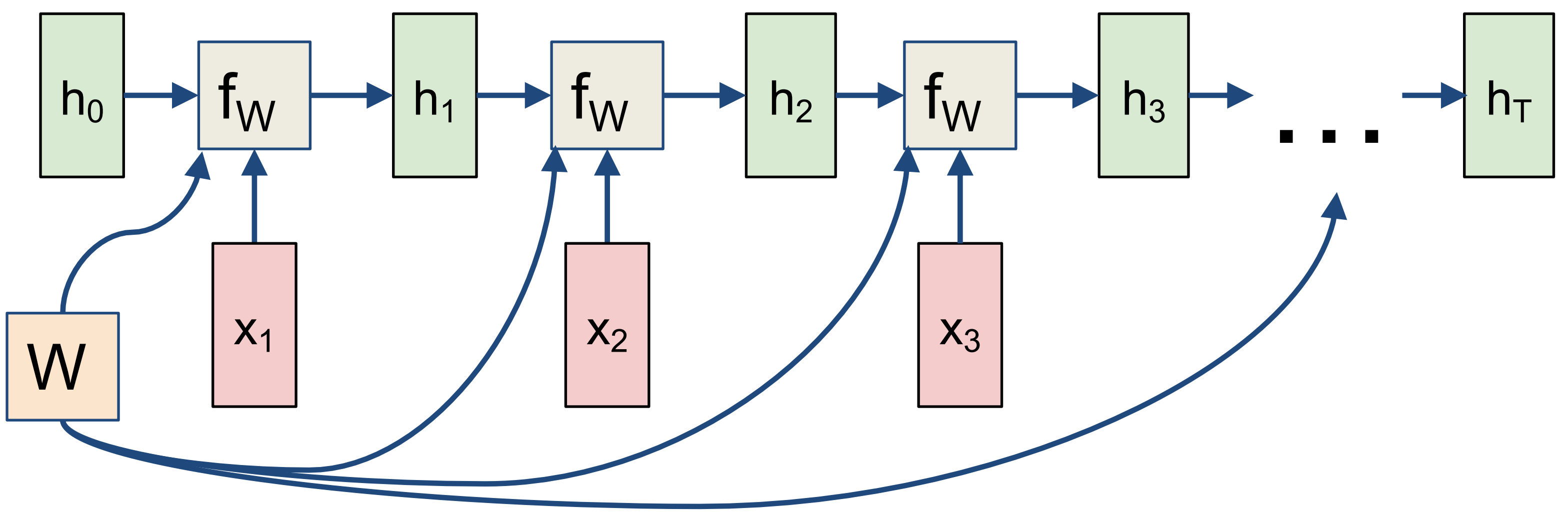


RNN: Computational Graph



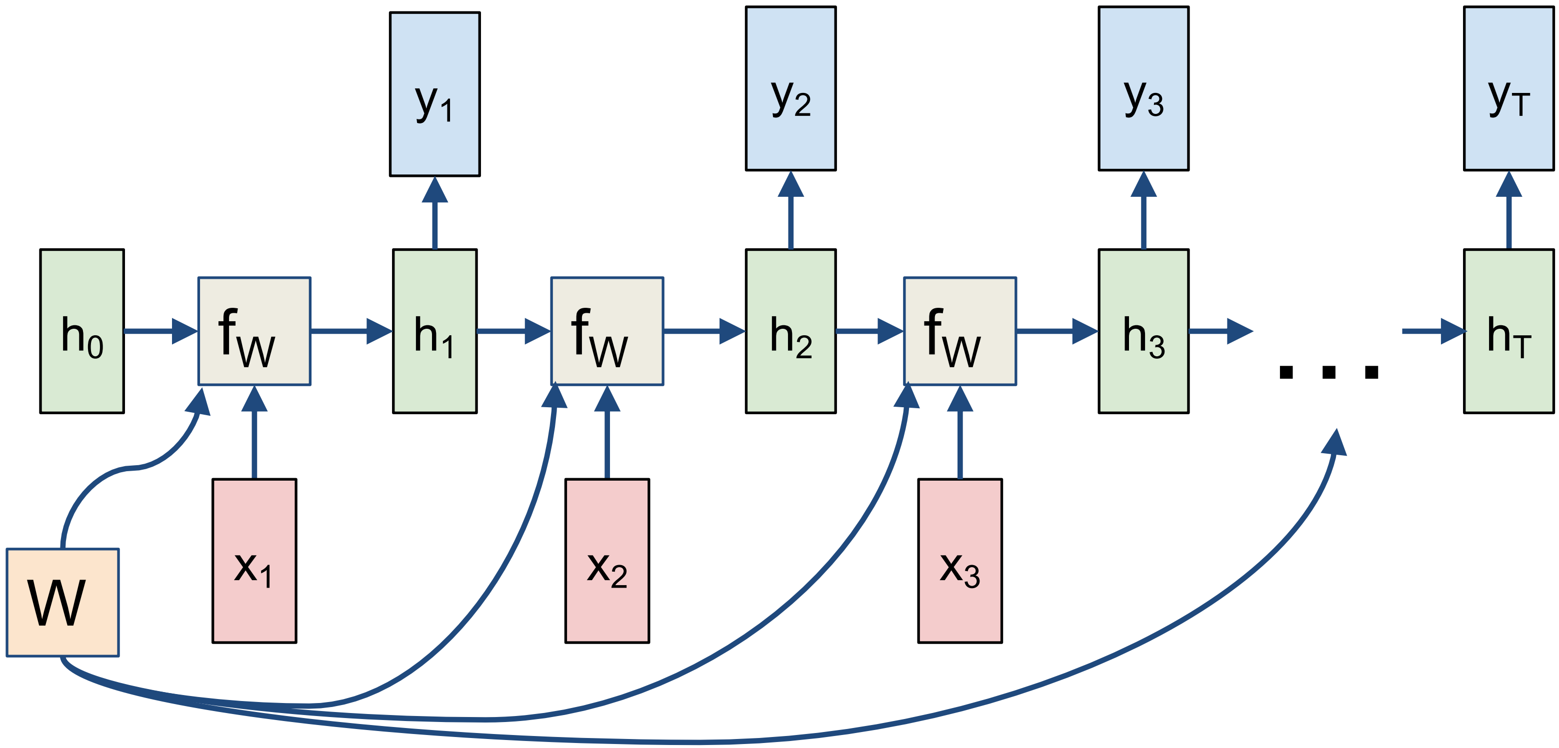
RNN: Computational Graph

Re-use the same weight matrix at every time-step

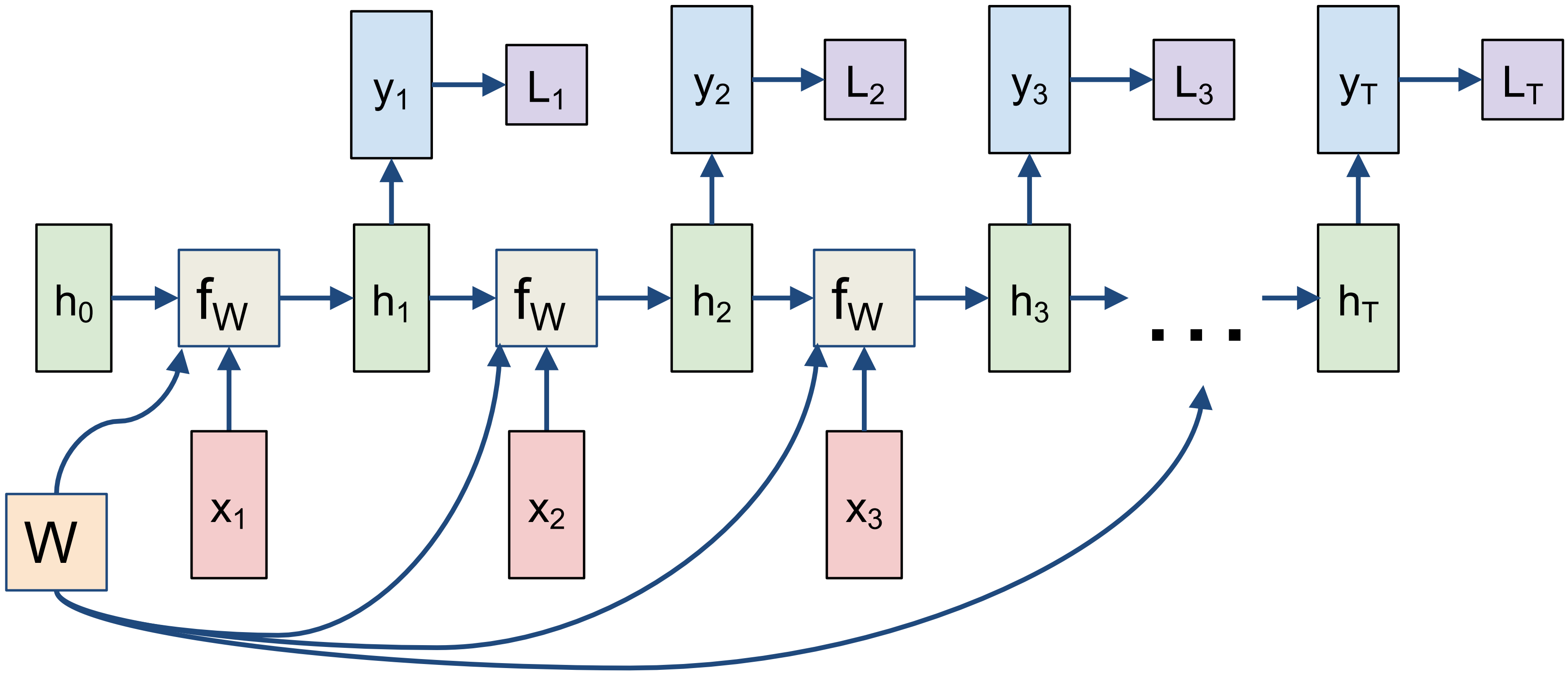


Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n

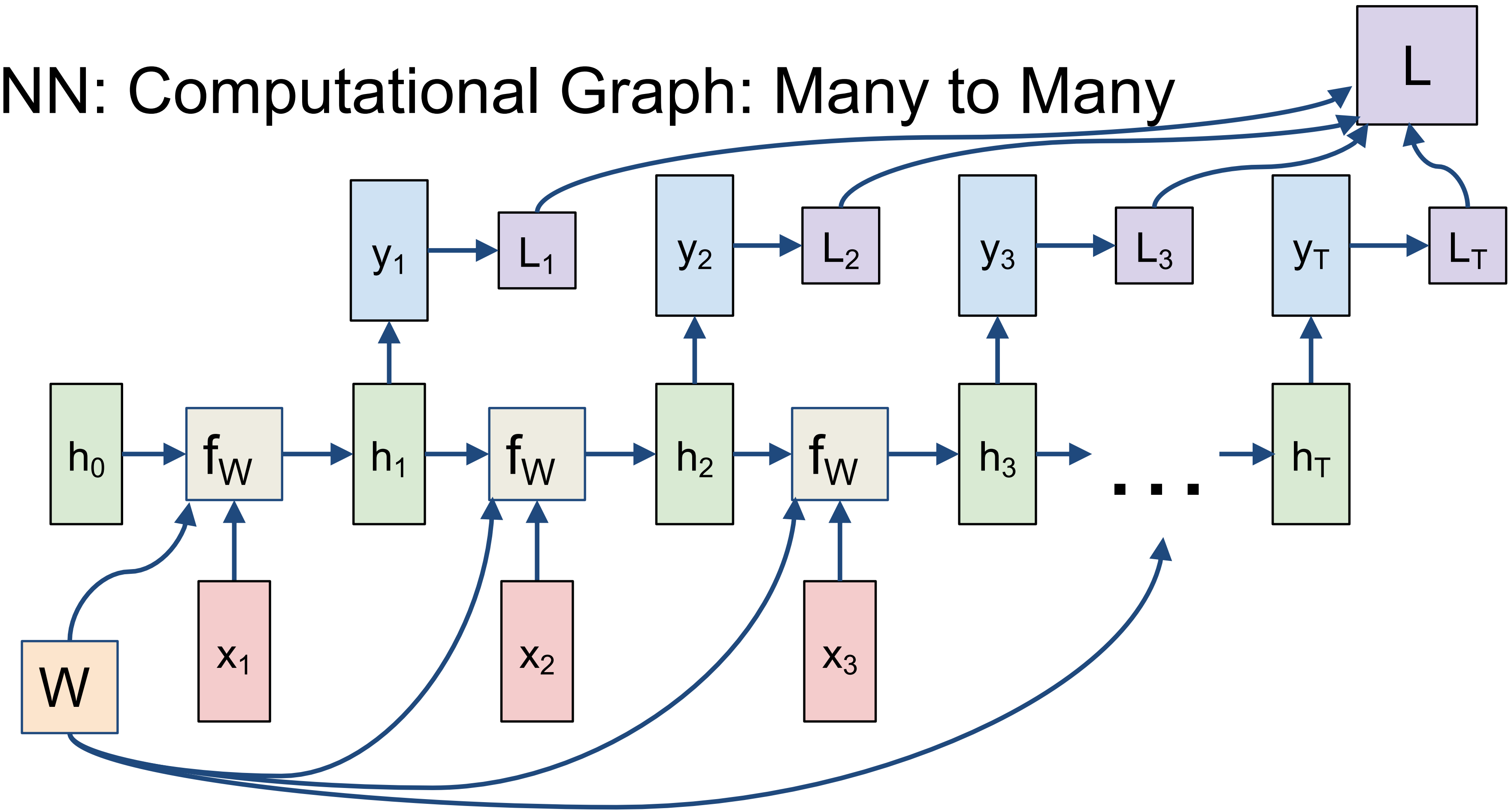
RNN: Computational Graph: Many to Many



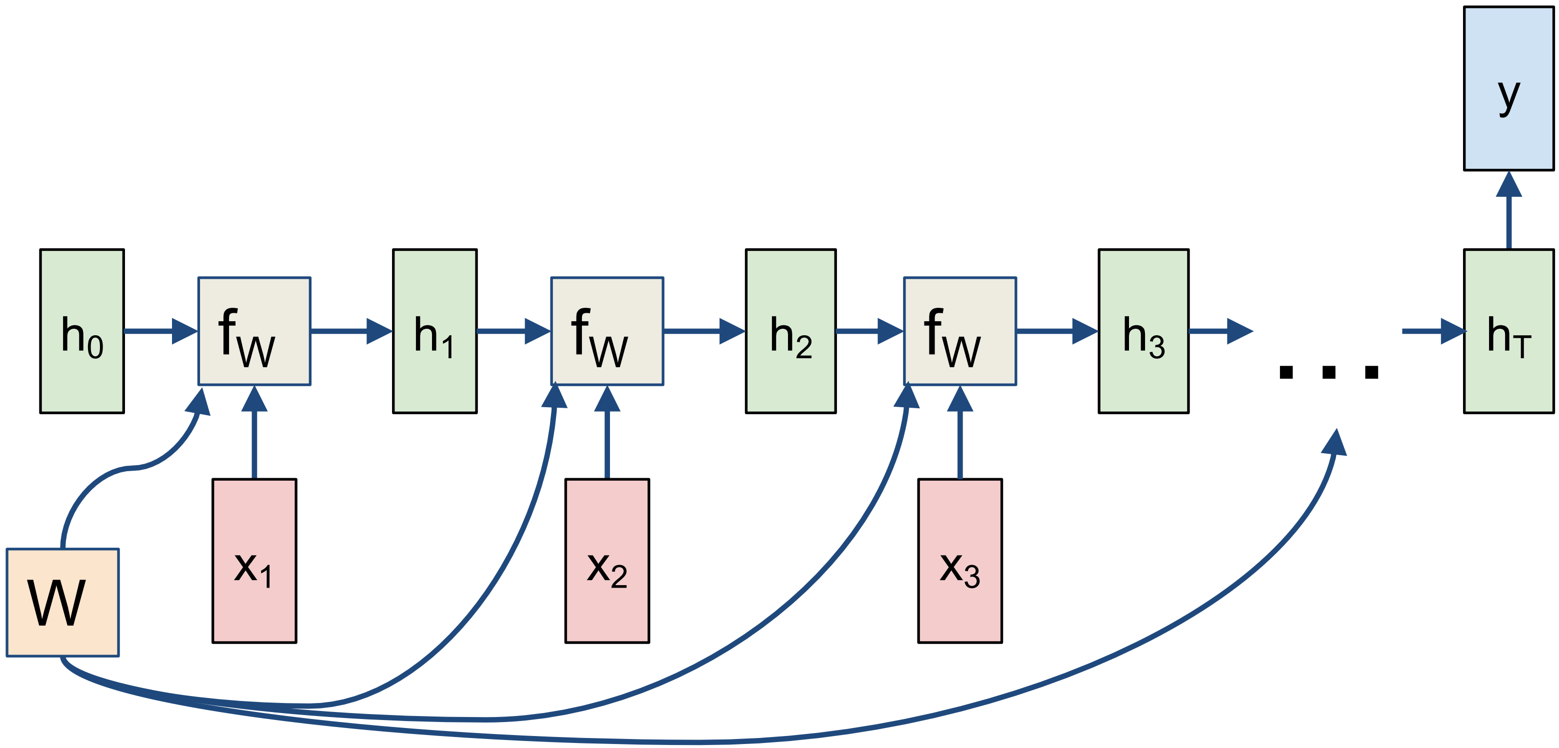
RNN: Computational Graph: Many to Many



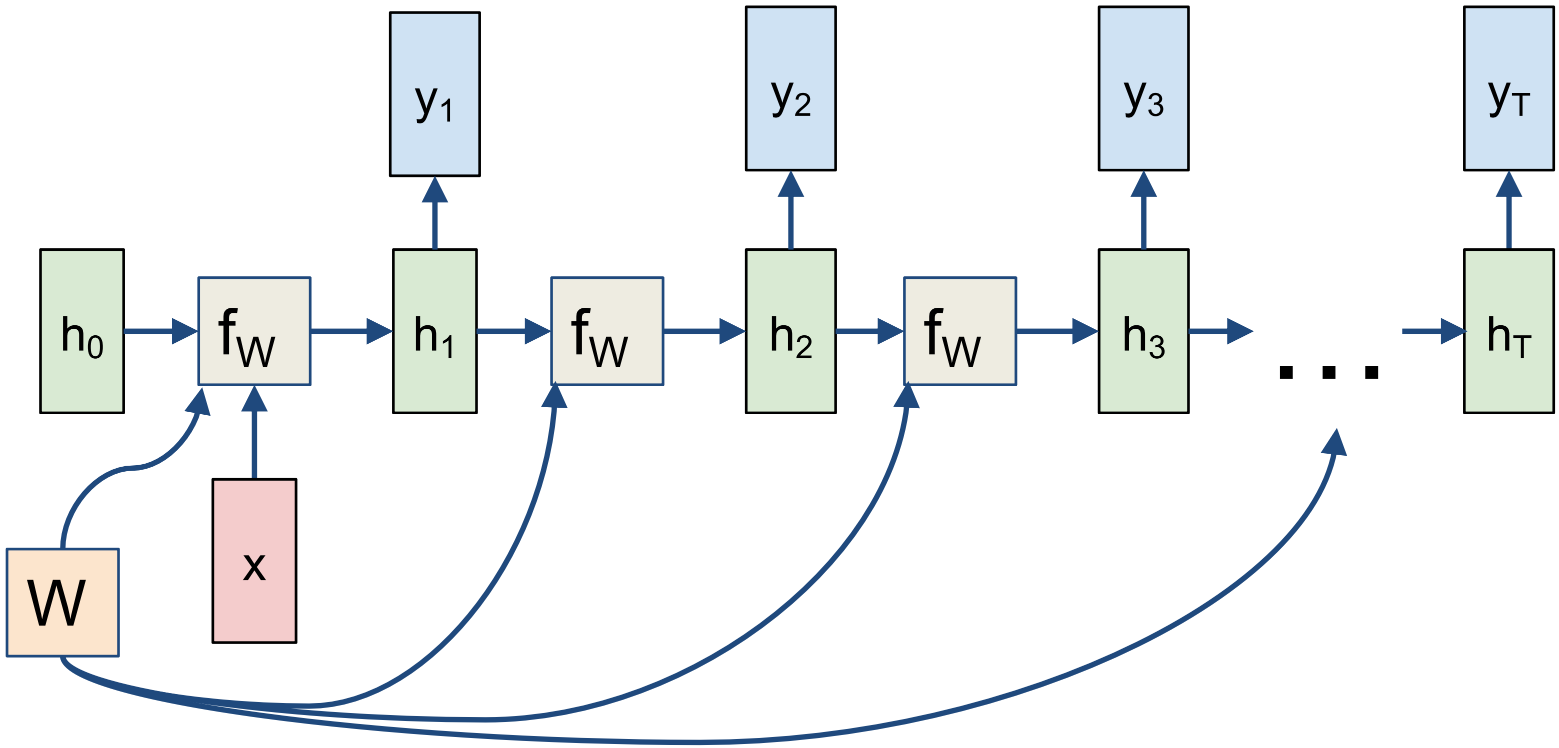
RNN: Computational Graph: Many to Many



RNN: Computational Graph: Many to One

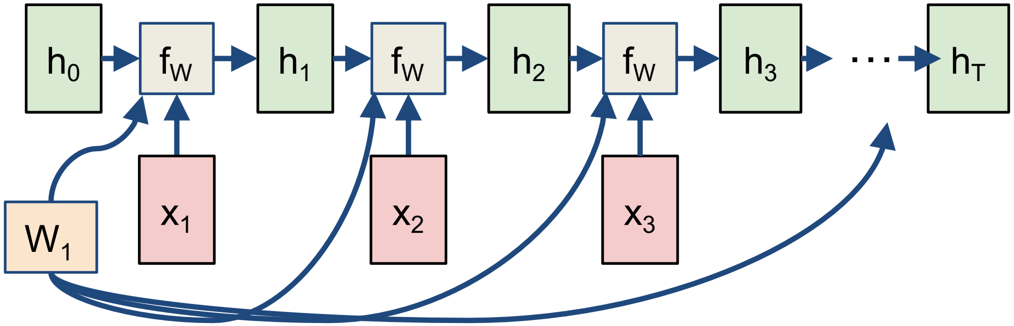


RNN: Computational Graph: One to Many



Sequence to Sequence: Many-to-one + one-to-many

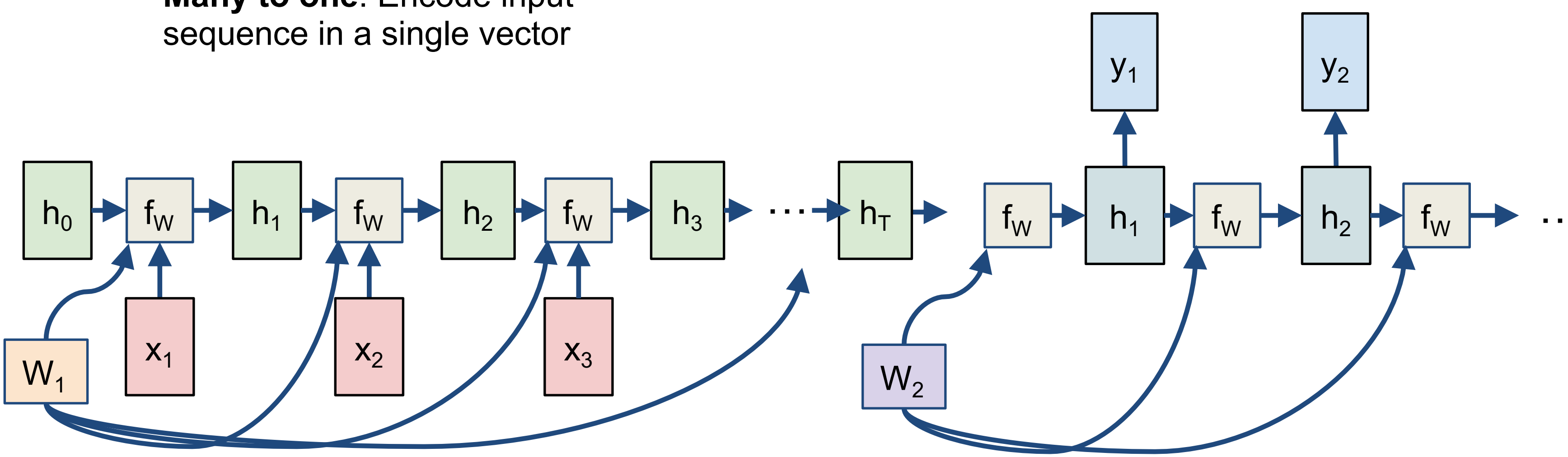
Many to one: Encode input sequence in a single vector



Sequence to Sequence: Many-to-one + one-to-many

One to many: Produce output sequence from single input vector

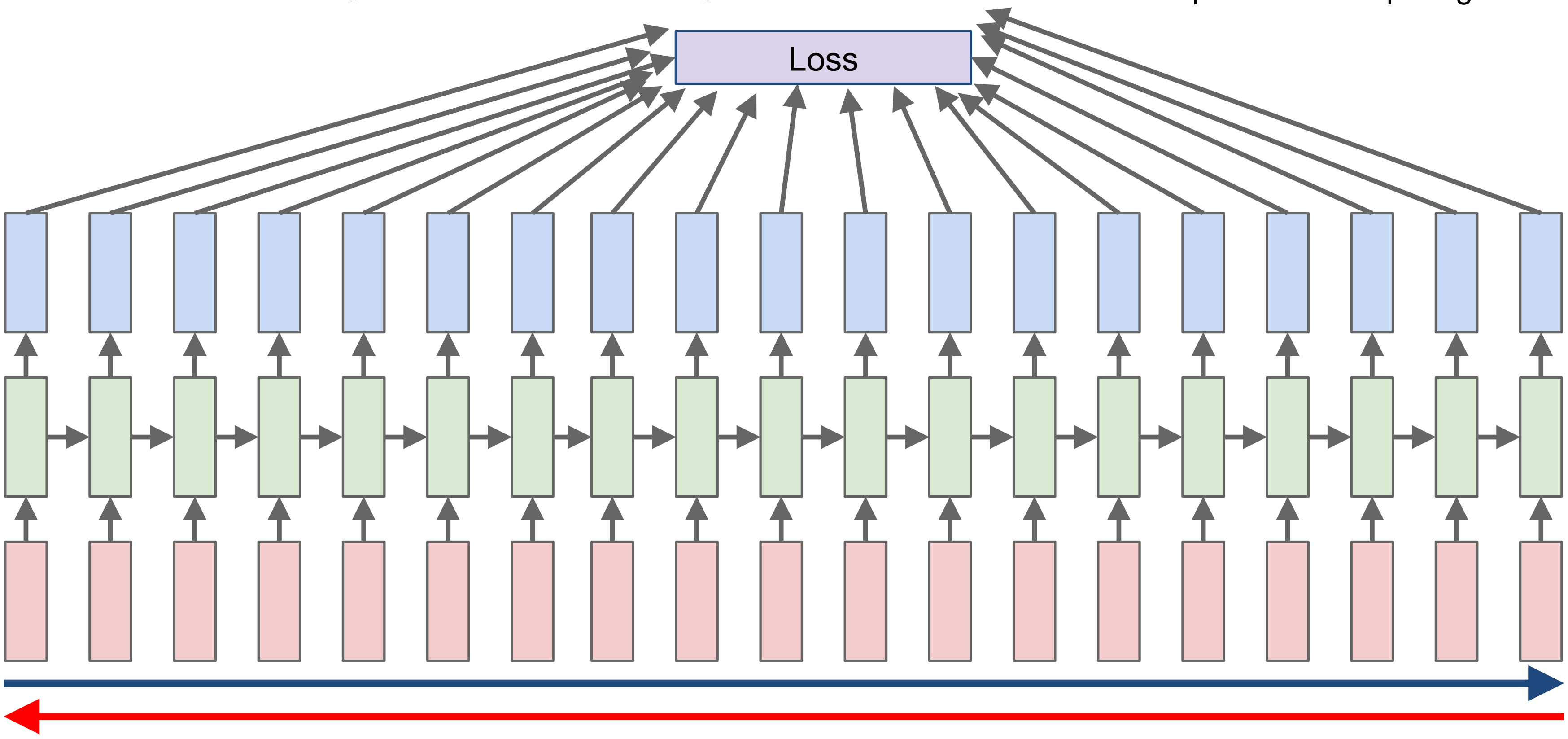
Many to one: Encode input sequence in a single vector



Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n

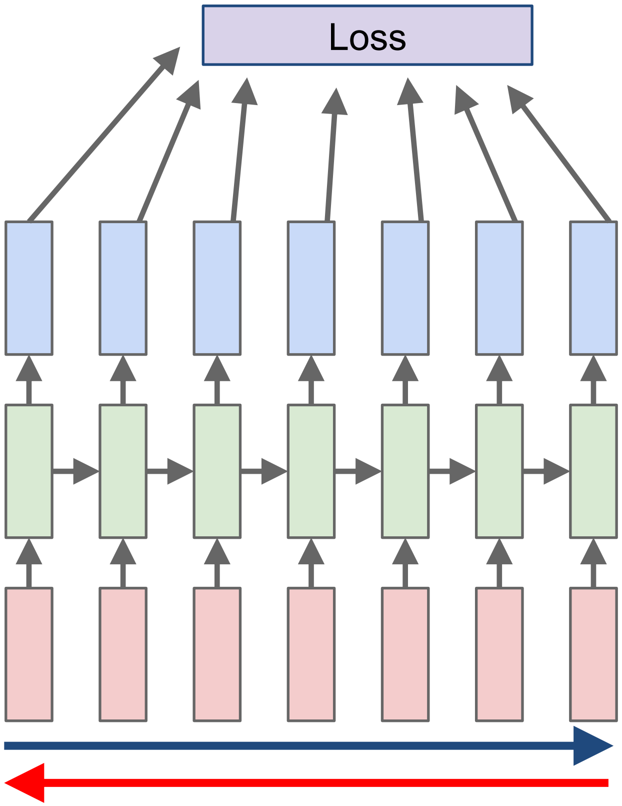
Backpropagation through time

Forward through entire sequence to compute loss, then backward through entire sequence to compute gradient



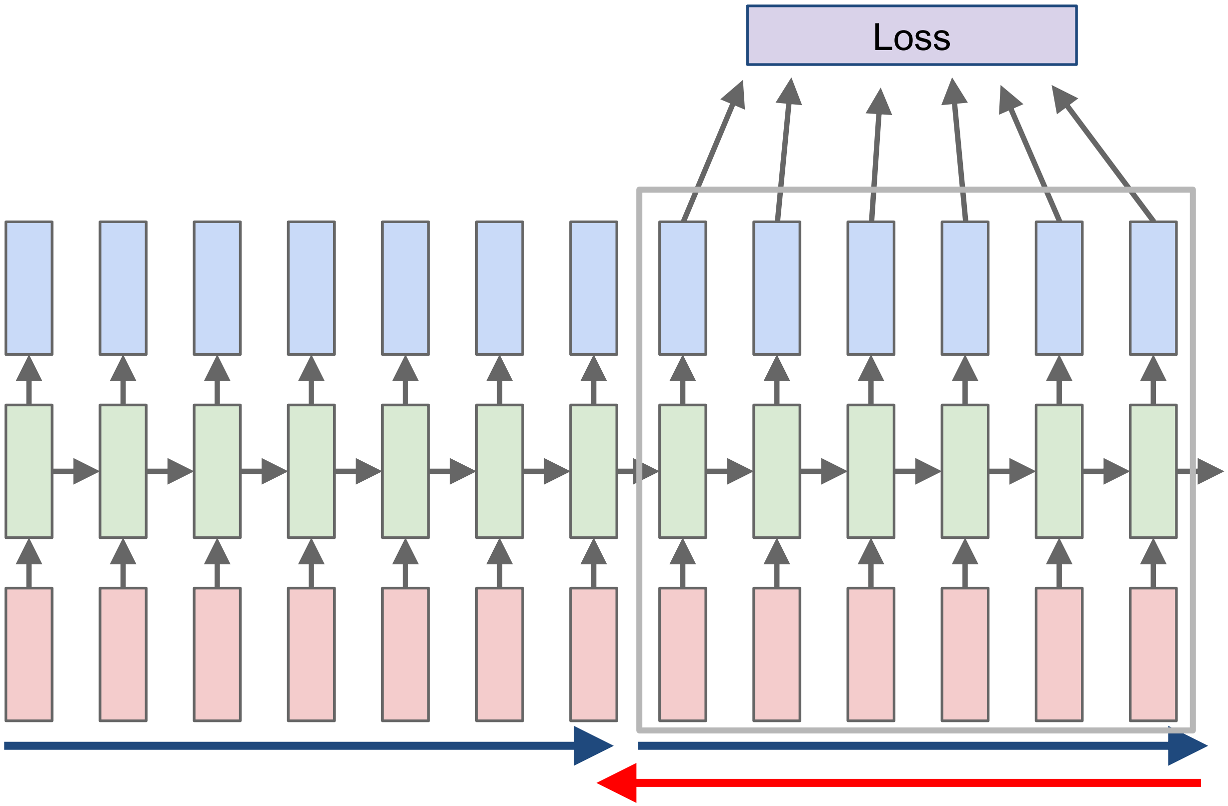
Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n

Truncated Backpropagation through time



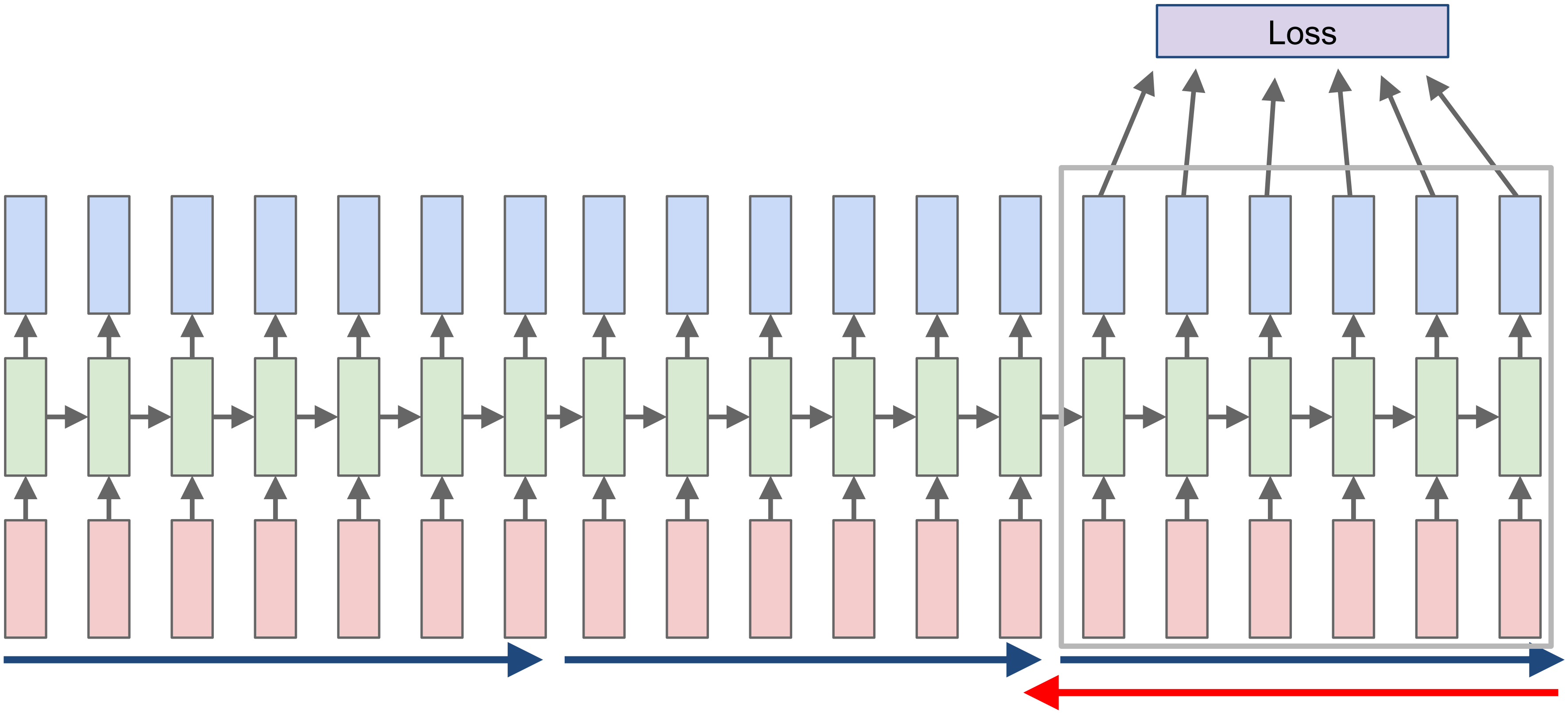
Run forward and backward through chunks of the sequence instead of whole sequence

Truncated Backpropagation through time



Carry hidden states forward in time forever, but only backpropagate for some smaller number of steps

Truncated Backpropagation through time

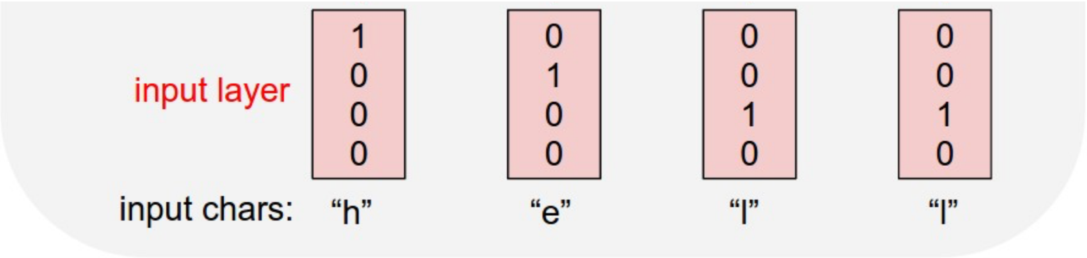


Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n

Example: Character-level Language Model

Vocabulary:
[h,e,l,o]

Example training sequence:
“hello”

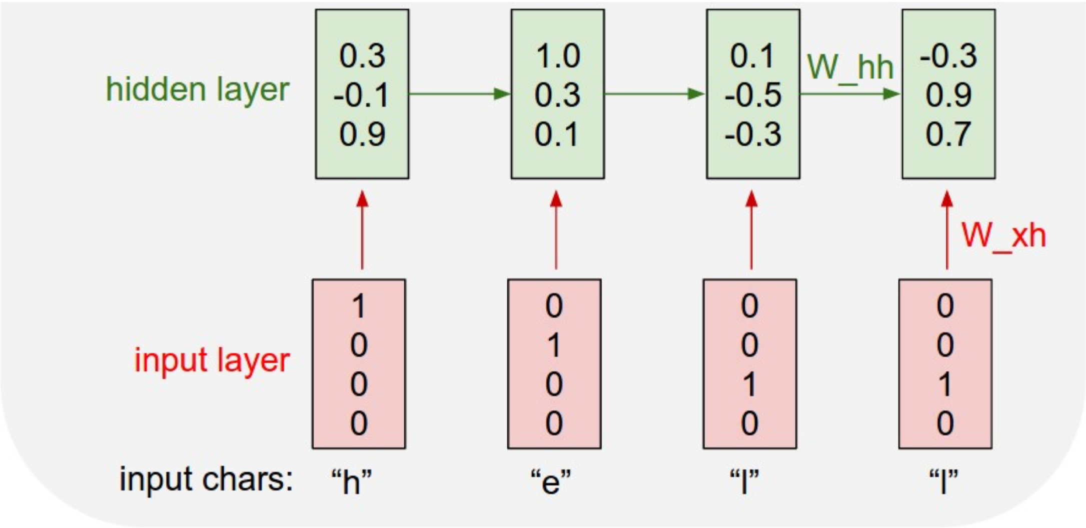


Example: Character-level Language Model

$$\boxed{h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t + b_h)}$$

Vocabulary: [h,e,l,o]

Example training sequence: **“hello”**

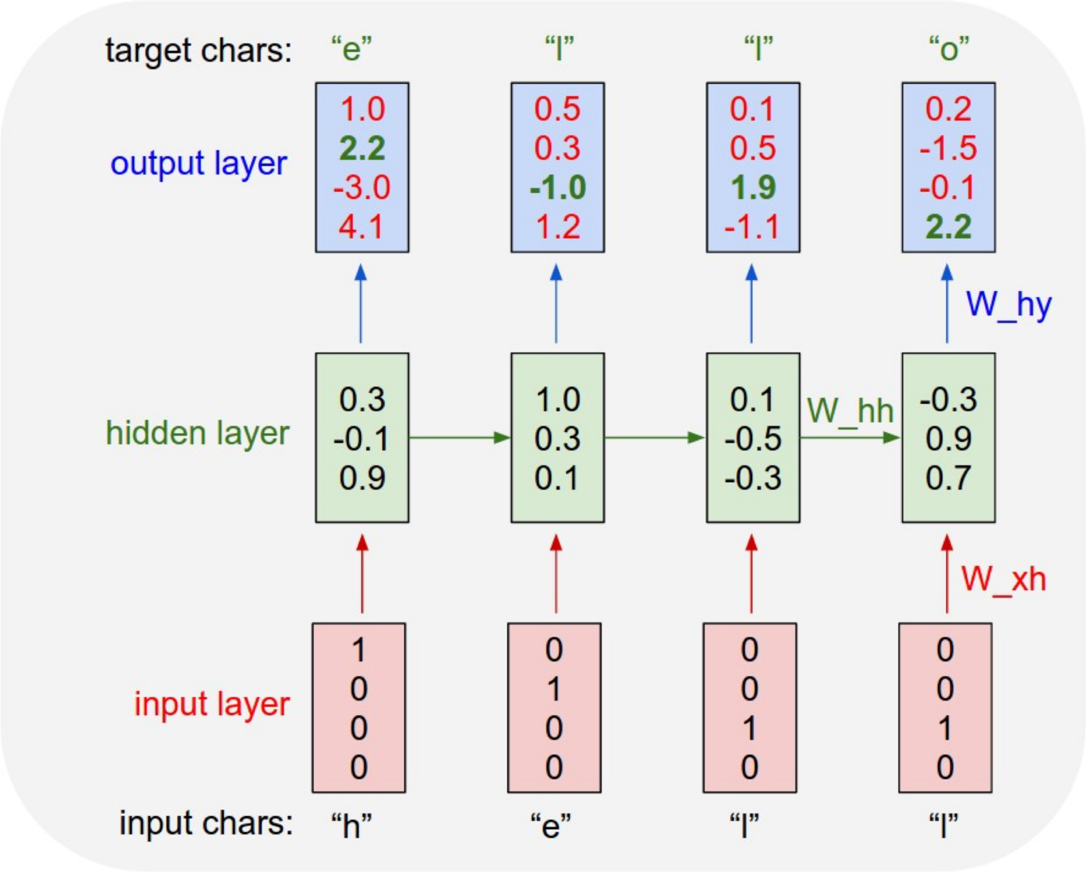


Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n

Example: Character-level Language Model

Vocabulary: [h,e,l,o]

Example training sequence: **“hello”**

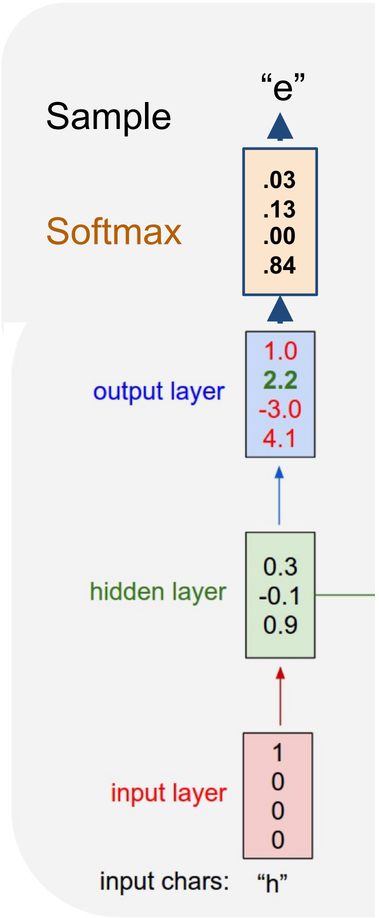


Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n

Example: Character-level Language Model Sampling

Vocabulary: [h,e,l,o]

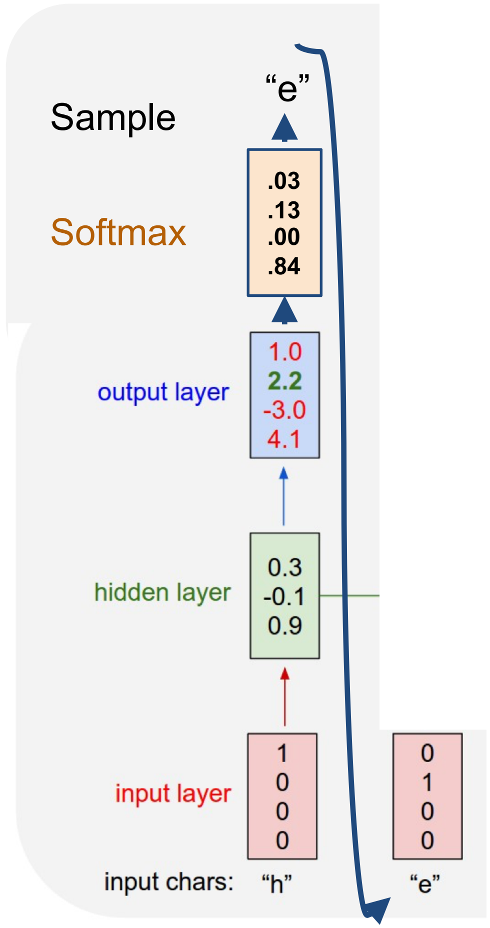
At test-time sample characters one at a time, feed back to model



Example: Character-level Language Model Sampling

Vocabulary: [h,e,l,o]

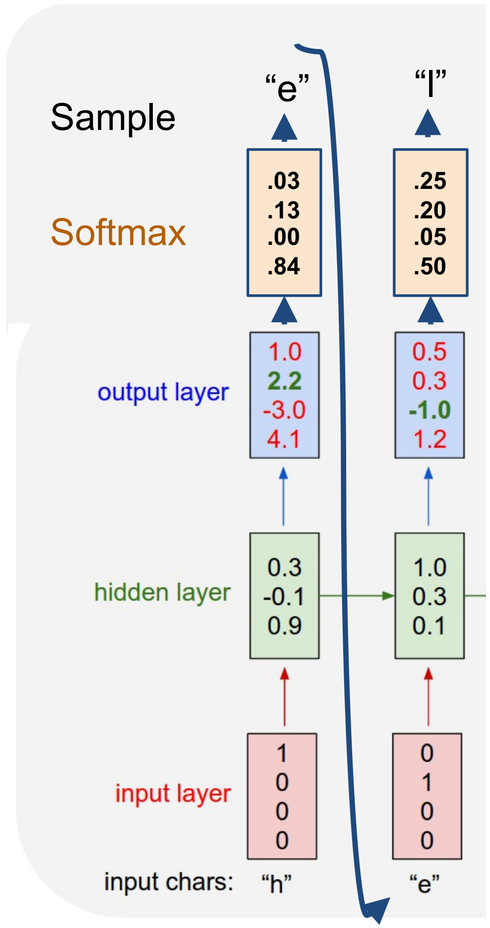
At test-time sample characters one at a time, feed back to model



Example: Character-level Language Model Sampling

Vocabulary: [h,e,l,o]

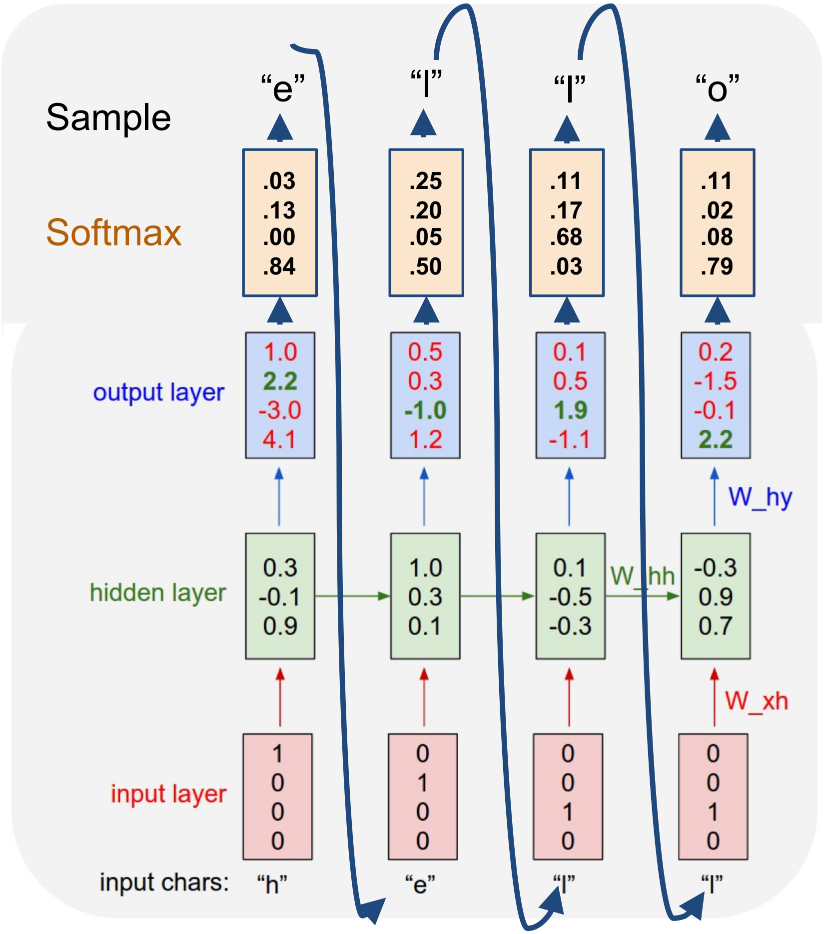
At test-time sample characters one at a time, feed back to model



Example: Character-level Language Model Sampling

Vocabulary:
[h,e,l,o]

At test-time sample characters one at a time, feed back to model



[min-char-rnn.py gist: 112 lines of Python](https://gist.github.com/karpathy/d4dee566867f8291f086)

```python
"""
Minimal character-level Vanilla RNN model. Written by Andrej Karpathy (@karpathy)
BSD License
"""
import numpy as np

# data I/O
data = open('input.txt', 'r').read() # should be simple plain text file
chars = list(set(data))
data_size, vocab_size = len(data), len(chars)
print 'data has %d characters, %d unique.' % (data_size, vocab_size)
char_to_ix = { ch:i for i,ch in enumerate(chars) }
ix_to_char = { i:ch for i,ch in enumerate(chars) }

# hyperparameters
hidden_size = 100 # size of hidden layer of neurons
seq_length = 25 # number of steps to unroll the RNN for
learning_rate = 1e-1

# model parameters
Wxh = np.random.randn(hidden_size, vocab_size)*0.01 # input to hidden
Whh = np.random.randn(hidden_size, hidden_size)*0.01 # hidden to hidden
Why = np.random.randn(vocab_size, hidden_size)*0.01 # hidden to output
bh = np.zeros((hidden_size, 1)) # hidden bias
by = np.zeros((vocab_size, 1)) # output bias

def lossFun(inputs, targets, hprev):
  """
  inputs,targets are both list of integers.
  hprev is Hx1 array of initial hidden state
  returns the loss, gradients on model parameters, and last hidden state
  """
  xs, hs, ys, ps = {}, {}, {}, {}
  hs[-1] = np.copy(hprev)
  loss = 0
  # forward pass
  for t in xrange(len(inputs)):
    xs[t] = np.zeros((vocab_size,1)) # encode in 1-of-k representation
    xs[t][inputs[t]] = 1
    hs[t] = np.tanh(np.dot(Wxh, xs[t]) + np.dot(Whh, hs[t-1]) + bh) # hidden state
    ys[t] = np.dot(Why, hs[t]) + by # unnormalized log probabilities for next chars
    ps[t] = np.exp(ys[t]) / np.sum(np.exp(ys[t])) # probabilities for next chars
    loss += -np.log(ps[t][targets[t],0]) # softmax (cross-entropy loss)
  # backward pass: compute gradients going backwards
  dWxh, dWhh, dWhy = np.zeros_like(Wxh), np.zeros_like(Whh), np.zeros_like(Why)
  dbh, dby = np.zeros_like(bh), np.zeros_like(by)
  dhnext = np.zeros_like(hs[0])
  for t in reversed(xrange(len(inputs))):
    dy = np.copy(ps[t])
    dy[targets[t]] -= 1 # backprop into y
    dWhy += np.dot(dy, hs[t].T)
    dby += dy
    dh = np.dot(Why.T, dy) + dhnext # backprop into h
    dhraw = (1 - hs[t] * hs[t]) * dh # backprop through tanh nonlinearity
    dbh += dhraw
    dWxh += np.dot(dhraw, xs[t].T)
    dWhh += np.dot(dhraw, hs[t-1].T)
    dhnext = np.dot(Whh.T, dhraw)
  for dparam in [dWxh, dWhh, dWhy, dbh, dby]:
    np.clip(dparam, -5, 5, out=dparam) # clip to mitigate exploding gradients
  return loss, dWxh, dWhh, dWhy, dbh, dby, hs[len(inputs)-1]

def sample(h, seed_ix, n):
  """
  sample a sequence of integers from the model
  h is memory state, seed_ix is seed letter for first time step
  """
  x = np.zeros((vocab_size, 1))
  x[seed_ix] = 1
  ixes = []
  for t in xrange(n):
    h = np.tanh(np.dot(Wxh, x) + np.dot(Whh, h) + bh)
    y = np.dot(Why, h) + by
    p = np.exp(y) / np.sum(np.exp(y))
    ix = np.random.choice(range(vocab_size), p=p.ravel())
    x = np.zeros((vocab_size, 1))
    x[ix] = 1
    ixes.append(ix)
  return ixes

n, p = 0, 0
mWxh, mWhh, mWhy = np.zeros_like(Wxh), np.zeros_like(Whh), np.zeros_like(Why)
mbh, mby = np.zeros_like(bh), np.zeros_like(by) # memory variables for Adagrad
smooth_loss = -np.log(1.0/vocab_size)*seq_length # loss at iteration 0
while True:
  # prepare inputs (we're sweeping from left to right in steps seq_length long)
  if p+seq_length+1 >= len(data) or n == 0:
    hprev = np.zeros((hidden_size,1)) # reset RNN memory
    p = 0 # go from start of data
  inputs = [char_to_ix[ch] for ch in data[p:p+seq_length]]
  targets = [char_to_ix[ch] for ch in data[p+1:p+seq_length+1]]

  # sample from the model now and then
  if n % 100 == 0:
    sample_ix = sample(hprev, inputs[0], 200)
    txt = ''.join(ix_to_char[ix] for ix in sample_ix)
    print '----\n %s \n----' % (txt, )

  # forward seq_length characters through the net and fetch gradient
  loss, dWxh, dWhh, dWhy, dbh, dby, hprev = lossFun(inputs, targets, hprev)
  smooth_loss = smooth_loss * 0.999 + loss * 0.001
  if n % 100 == 0: print 'iter %d, loss: %f' % (n, smooth_loss) # print progress

  # perform parameter update with Adagrad
  for param, dparam, mem in zip([Wxh, Whh, Why, bh, by],
                                [dWxh, dWhh, dWhy, dbh, dby],
                                [mWxh, mWhh, mWhy, mbh, mby]):
    mem += dparam * dparam
    param += -learning_rate * dparam / np.sqrt(mem + 1e-8) # adagrad update

  p += seq_length # move data pointer
  n += 1 # iteration counter
```

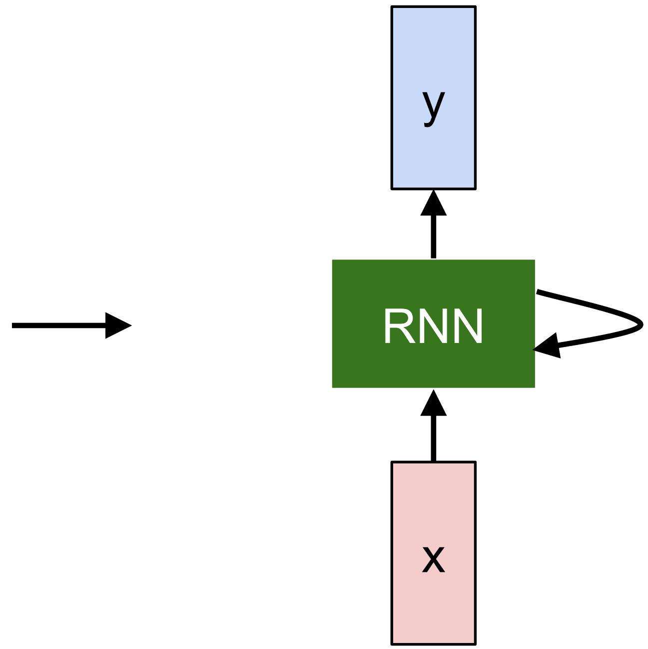
(https://gist.github.com/karpathy/d4dee566867f8291f086)

THE SONNETS

by William Shakespeare

From fairest creatures we desire increase,
That thereby beauty's rose might never die,
But as the riper should by time decease,
His tender heir might bear his memory:
But thou, contracted to thine own bright eyes,
Feed'st thy light's flame with self-substantial fuel,
Making a famine where abundance lies,
Thyself thy foe, to thy sweet self too cruel:
Thou that art now the world's fresh ornament,
And only herald to the gaudy spring,
Within thine own bud buriest thy content,
And tender churl mak'st waste in niggarding:
    Pity the world, or else this glutton be,
    To eat the world's due, by the grave and thee.

When forty winters shall besiege thy brow,
And dig deep trenches in thy beauty's field,
Thy youth's proud livery so gazed on now,
Will be a tatter'd weed of small worth held:
Then being asked, where all thy beauty lies,
Where all the treasure of thy lusty days;
To say, within thine own deep sunken eyes,
Were an all-eating shame, and thriftless praise.
How much more praise deserv'd thy beauty's use,
If thou couldst answer 'This fair child of mine
Shall sum my count, and make my old excuse,'
Proving his beauty by succession thine!
    This were to be new made when thou art old,
    And see thy blood warm when thou feel'st it cold.



Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n

at first:

```
tyntd-iafhatawiaoihrdemot  lytdws  e ,tfti, astai f ogoh eoase rrranbyne 'nhthnee e
plia tklrgd t o idoe ns,smtt   h ne etie h,hregtrs nigtike,aoaenns lng
```

train more

```
"Tmont thithey" fomesscerliund
Keushey. Thom here
sheulke, anmerenith ol sivh I lalterthend Bleipile shuwy fil on aseterlome
coaniogennc Phe lism thond hon at. MeiDimorotion in ther thize."
```

train more

```
Aftair fall unsuch that the hall for Prince Velzonski's that me of
her hearly, and behs to so arwage fiving were to it beloge, pavu say falling misfort
how, and Gogition is so overelical and ofter.
```

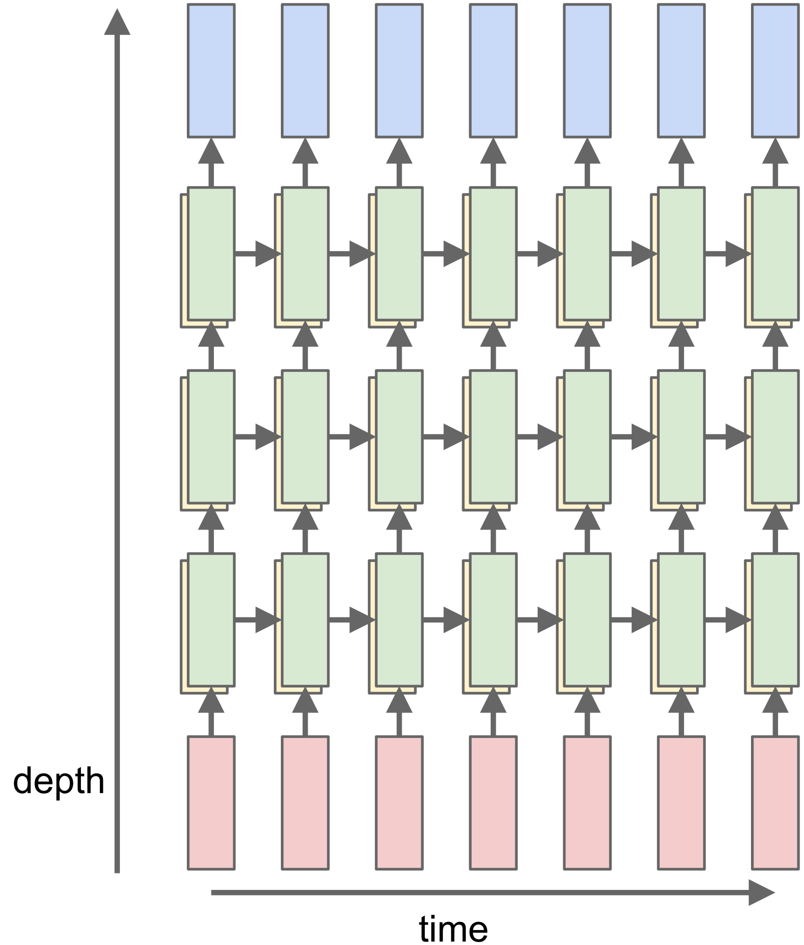
train more

```
"Why do what that day," replied Natasha, and wishing to himself the fact the
princess, Princess Mary was easier, fed in had oftened him.
Pierre aking his soul came to the packs and drove up his father-in-law women.
```

Multilayer RNNs

$$\boxed{h_t^l = \tanh W^l \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}}$$

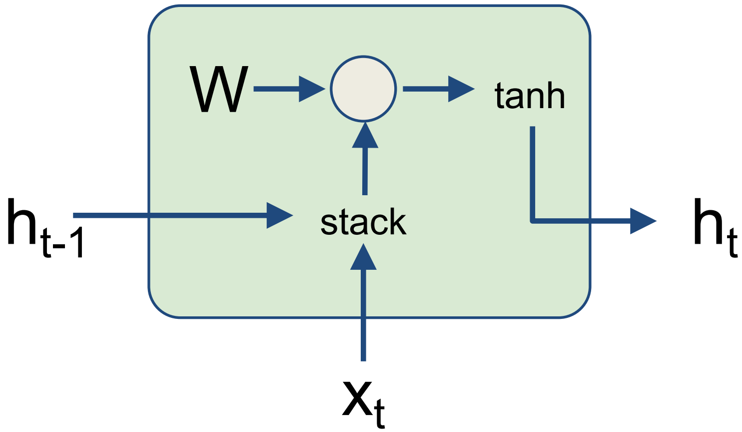
$h \in \mathbb{R}^n$ $\qquad W^l \ [n \times 2n]$



Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n

Vanilla RNN Gradient Flow

Bengio et al, "Learning long-term dependencies with gradient descent is difficult", IEEE Transactions on Neural Networks, 1994
Pascanu et al, "On the difficulty of training recurrent neural networks", ICML 2013

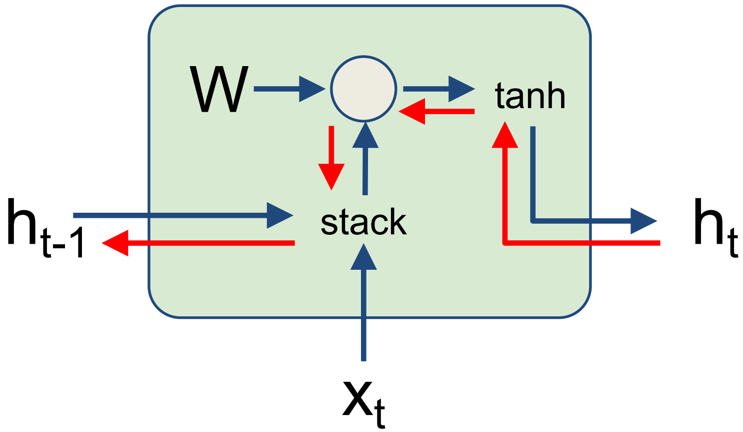


$$
\begin{aligned}
h_t &= \tanh(W_{hh}h_{t-1} + W_{xh}x_t) \\
&= \tanh\left(\begin{pmatrix} W_{hh} & W_{hx} \end{pmatrix} \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}\right) \\
&= \tanh\left(W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}\right)
\end{aligned}
$$

Vanilla RNN Gradient Flow

Bengio et al, "Learning long-term dependencies with gradient descent is difficult", IEEE Transactions on Neural Networks, 1994
Pascanu et al, "On the difficulty of training recurrent neural networks", ICML 2013

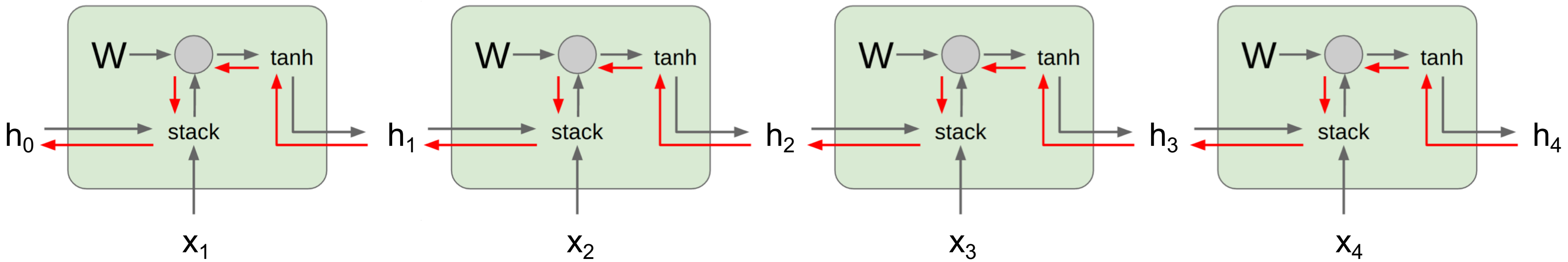
Backpropagation from h_t to h_{t-1} multiplies by W (actually W_{hh}^T)



$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$
$$= \tanh\left(\begin{pmatrix} W_{hh} & W_{hx} \end{pmatrix} \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}\right)$$
$$= \tanh\left(W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}\right)$$

Vanilla RNN Gradient Flow

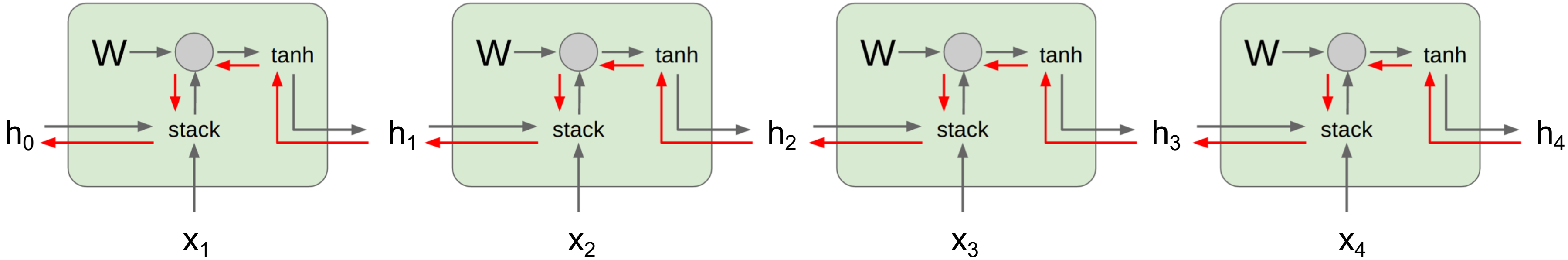
Bengio et al, "Learning long-term dependencies with gradient descent is difficult", IEEE Transactions on Neural Networks, 1994
Pascanu et al, "On the difficulty of training recurrent neural networks", ICML 2013



Computing gradient of h_0 involves many factors of W (and repeated tanh)

Vanilla RNN Gradient Flow

Bengio et al, “Learning long-term dependencies with gradient descent is difficult”, IEEE Transactions on Neural Networks, 1994
Pascanu et al, “On the difficulty of training recurrent neural networks”, ICML 2013



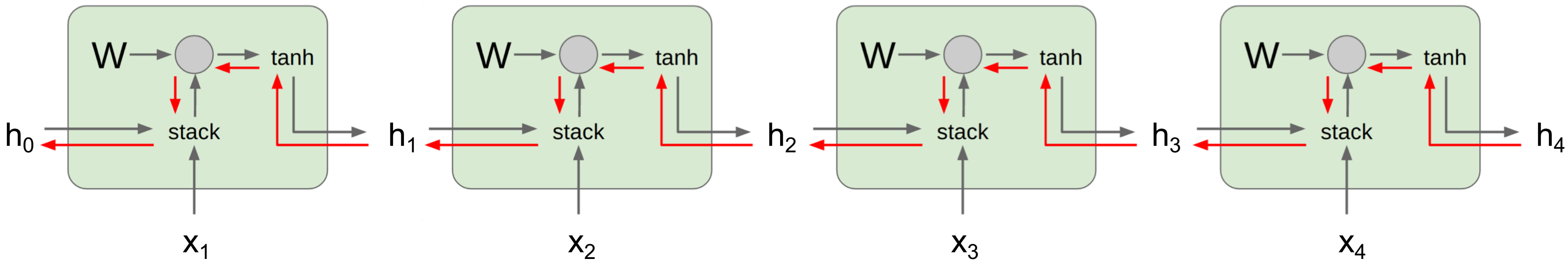
Computing gradient of h_0 involves many factors of W (and repeated tanh)

Largest singular value > 1:
Exploding gradients

Largest singular value < 1:
Vanishing gradients

Vanilla RNN Gradient Flow

Bengio et al, "Learning long-term dependencies with gradient descent is difficult", IEEE Transactions on Neural Networks, 1994
Pascanu et al, "On the difficulty of training recurrent neural networks", ICML 2013



Computing gradient of h_0 involves many factors of W (and repeated tanh)

Largest singular value > 1:
Exploding gradients

Largest singular value < 1:
Vanishing gradients

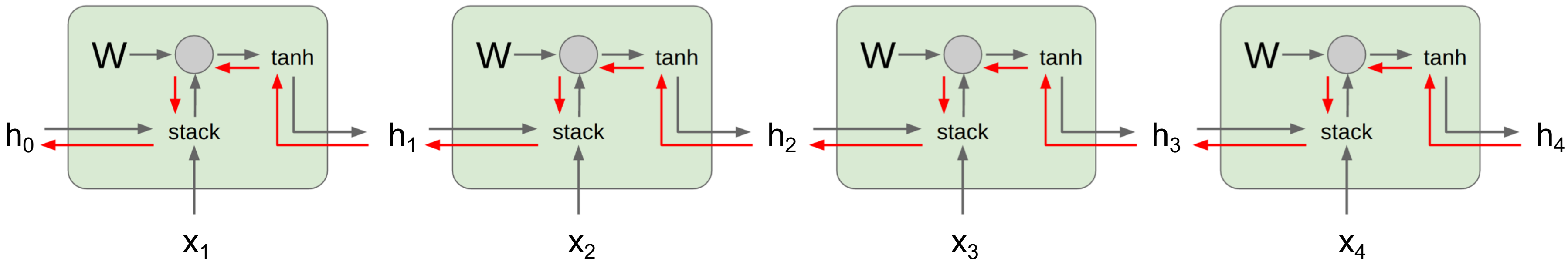
Gradient clipping: Scale gradient if its norm is too big

```
grad_norm = np.sum(grad * grad)
if grad_norm > threshold:
  grad *= (threshold / grad_norm)
```

Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n

Vanilla RNN Gradient Flow

Bengio et al, "Learning long-term dependencies with gradient descent is difficult", IEEE Transactions on Neural Networks, 1994
Pascanu et al, "On the difficulty of training recurrent neural networks", ICML 2013



Computing gradient of h_0 involves many factors of W
(and repeated tanh)

Largest singular value > 1:
Exploding gradients

Largest singular value < 1:
Vanishing gradients

→ Change RNN architecture

Long Short Term Memory (LSTM)

Vanilla RNN

$$h_t = \tanh\left(W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}\right)$$

LSTM

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}$$

$$c_t = f \odot c_{t-1} + i \odot g$$

$$h_t = o \odot \tanh(c_t)$$

Hochreiter and Schmidhuber, “Long Short Term Memory”, Neural Computation 1997

Meet LSTMs

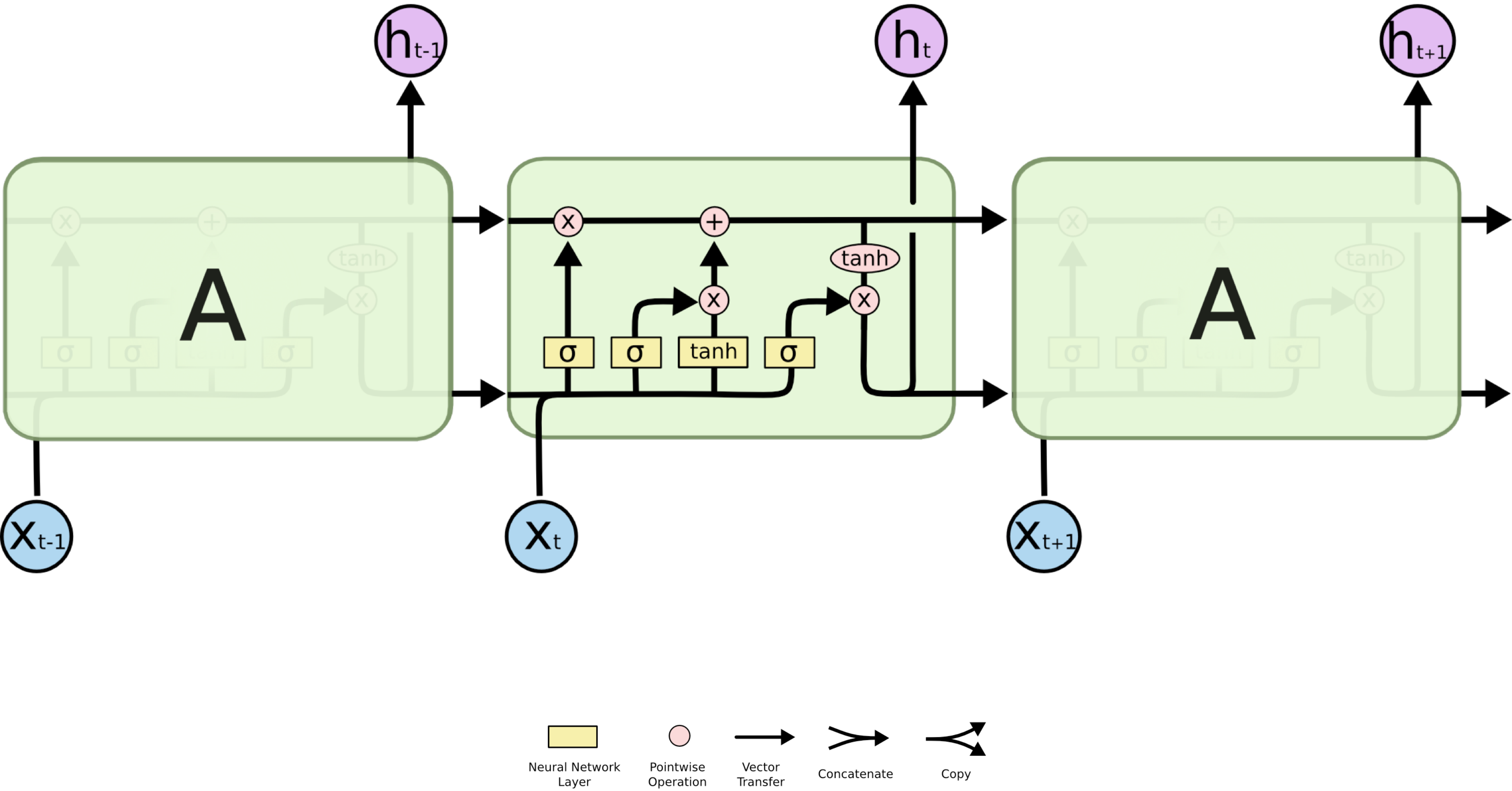
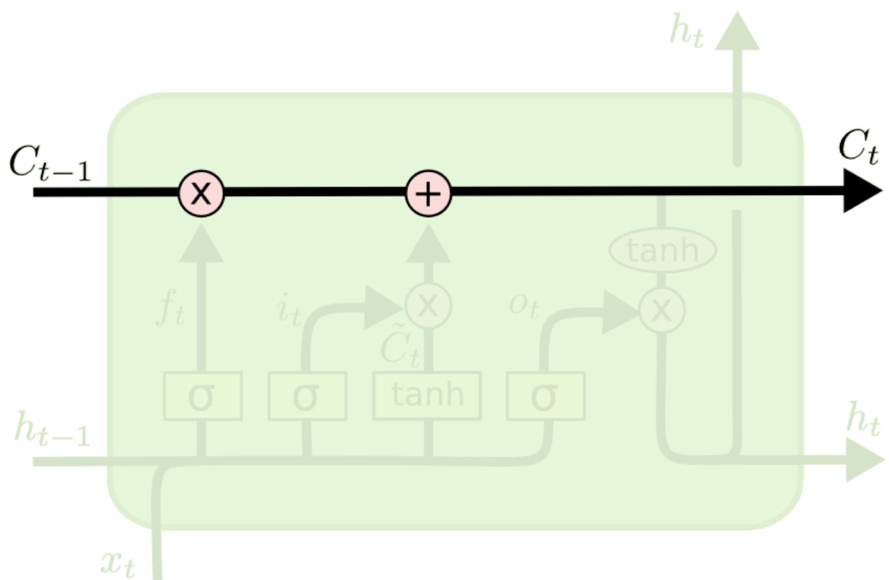


 Image Credit: Christopher Olah (http://colah.github.io/posts/2015-08-Understanding-LSTMs/)

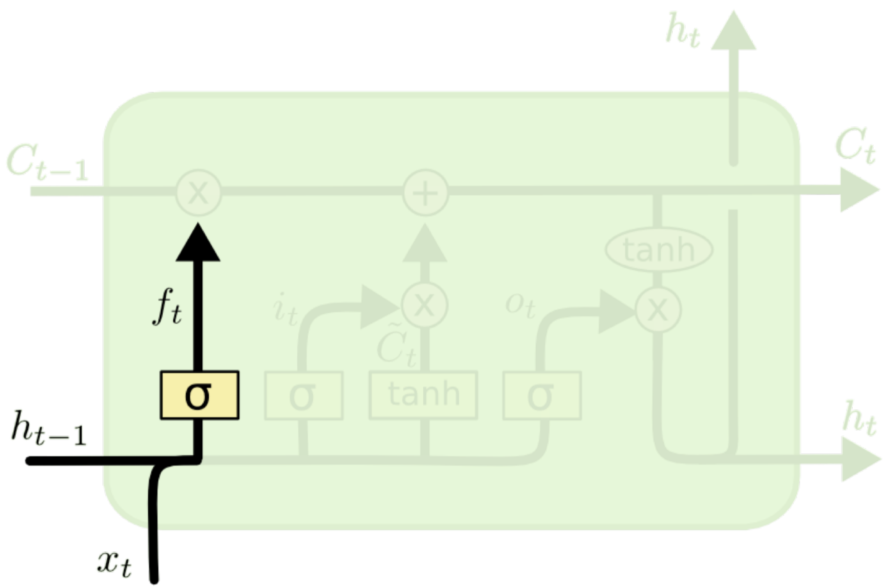
LSTMs Intuition: Memory

- Cell State / Memory



LSTMs Intuition: Forget Gate

- Should we continue to remember this "bit" of information or not?

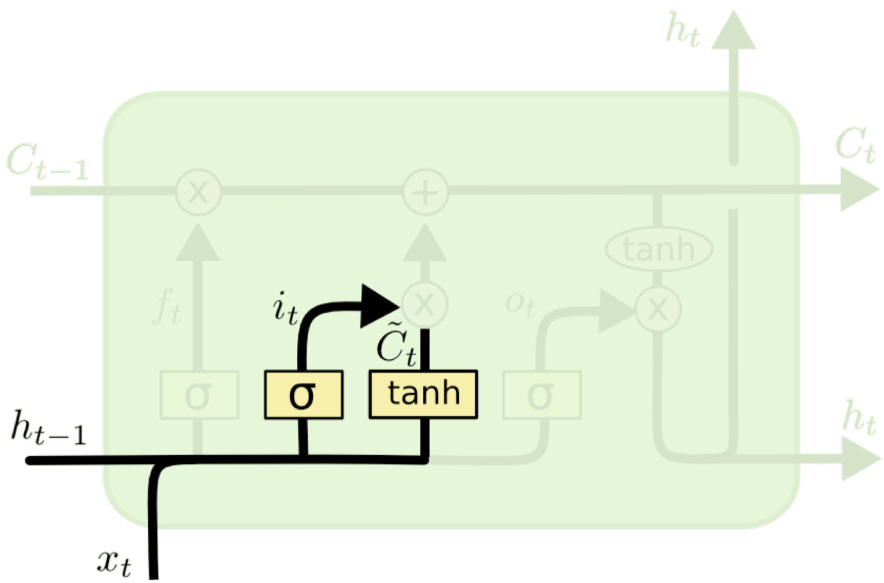


$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \ + \ b_f\right)$$

Image Credit: Christopher Olah (http://colah.github.io/posts/2015-08-Understanding-LSTMs/)

LSTMs Intuition: Input Gate

- Should we update this “bit” of information or not?
  - If so, with what?



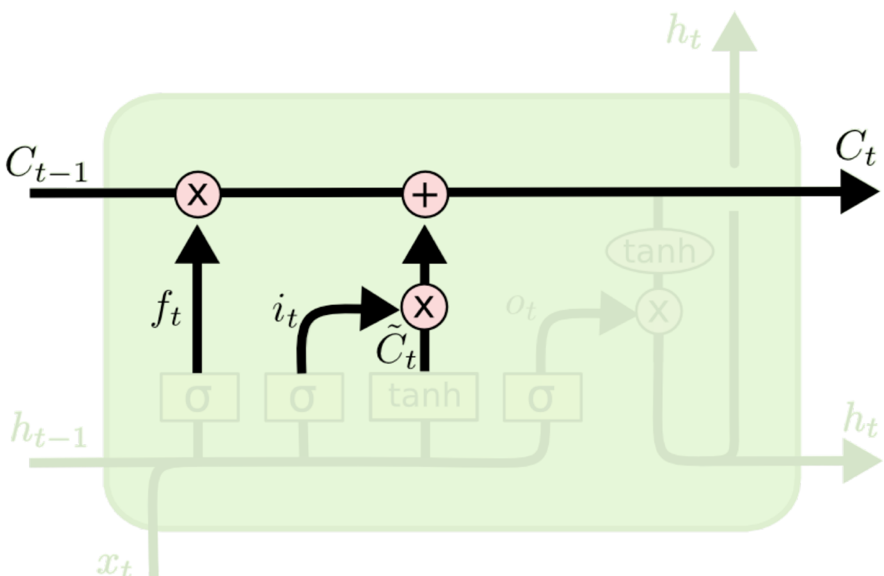
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \ + \ b_C)$$

Image Credit: Christopher Olah (http://colah.github.io/posts/2015-08-Understanding-LSTMs/)

LSTMs Intuition: Memory Update

- Forget that + memorize this

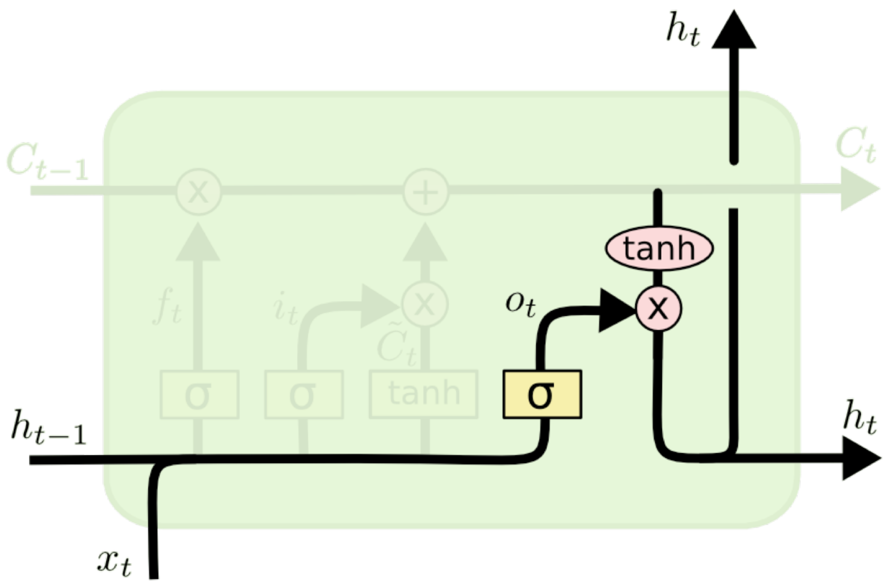


$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

 Image Credit: Christopher Olah (http://colah.github.io/posts/2015-08-Understanding-LSTMs/)

LSTMs Intuition: Output Gate

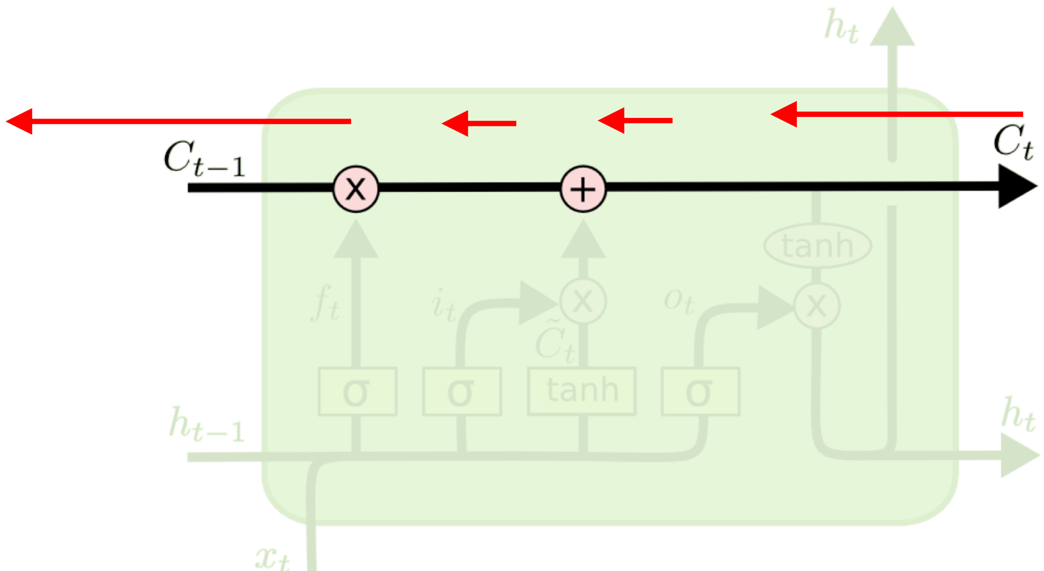
- Should we output this "bit" of information to "deeper" layers?



$$o_t = \sigma\left(W_o\ [h_{t-1}, x_t]\ +\ b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

Image Credit: Christopher Olah (http://colah.github.io/posts/2015-08-Understanding-LSTMs/)

LSTMs Intuition: Additive Updates



Backpropagation from c_t to c_{t-1} only elementwise multiplication by f, no matrix multiply by W

LSTMs Intuition: Additive Updates

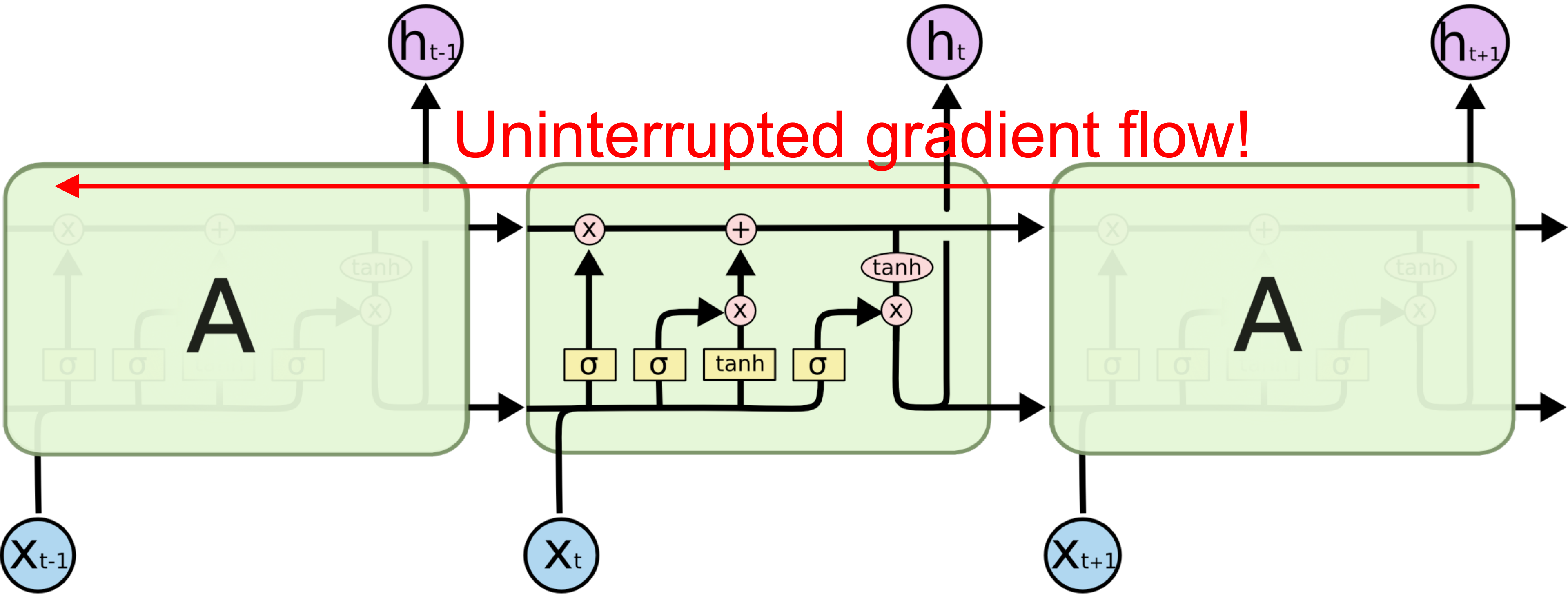
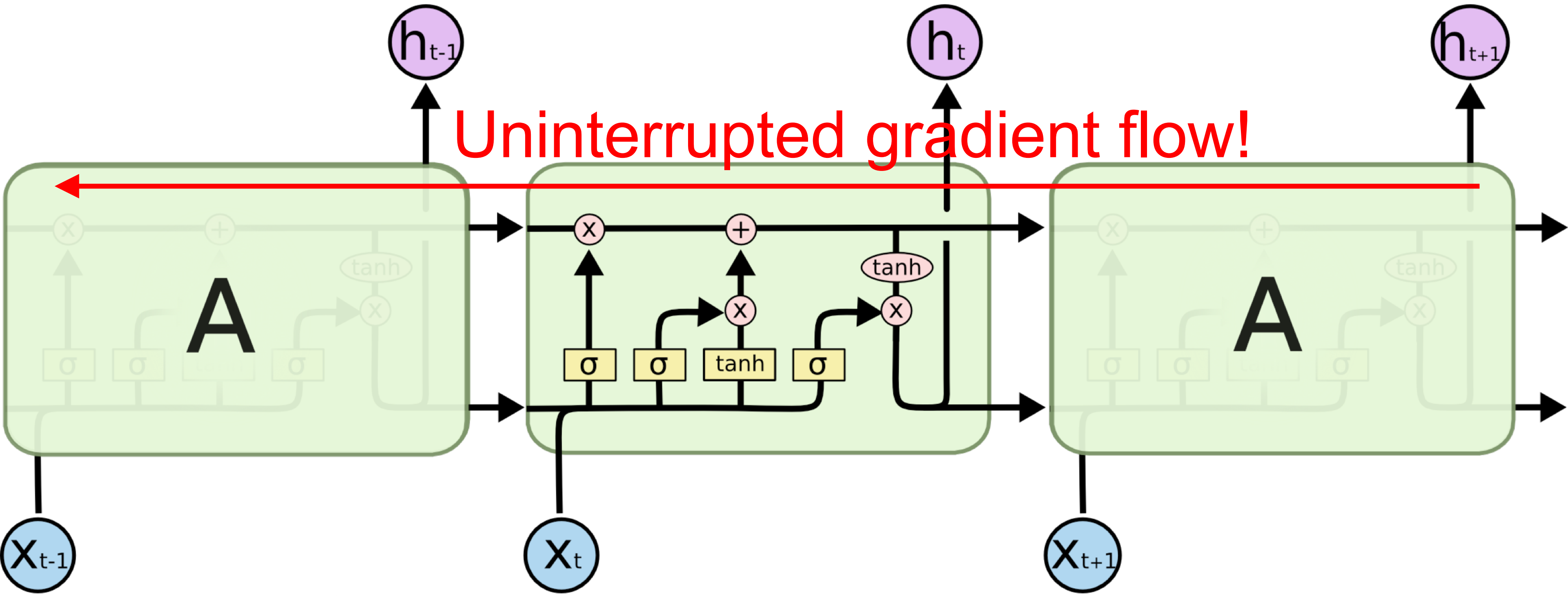
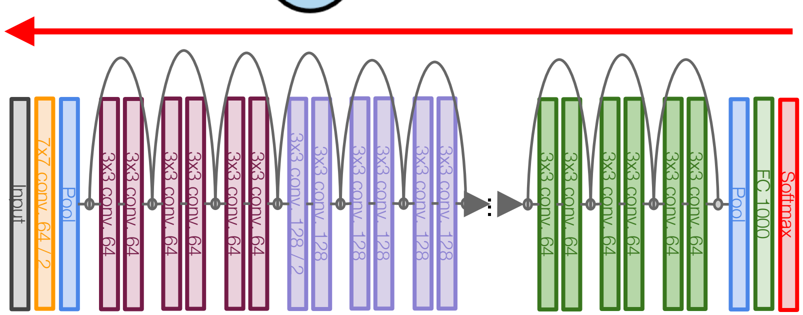


Image Credit: Christopher Olah (http://colah.github.io/posts/2015-08-Understanding-LSTMs/)

LSTMs Intuition: Additive Updates



Similar to ResNet!



LSTMs

- A pretty sophisticated cell

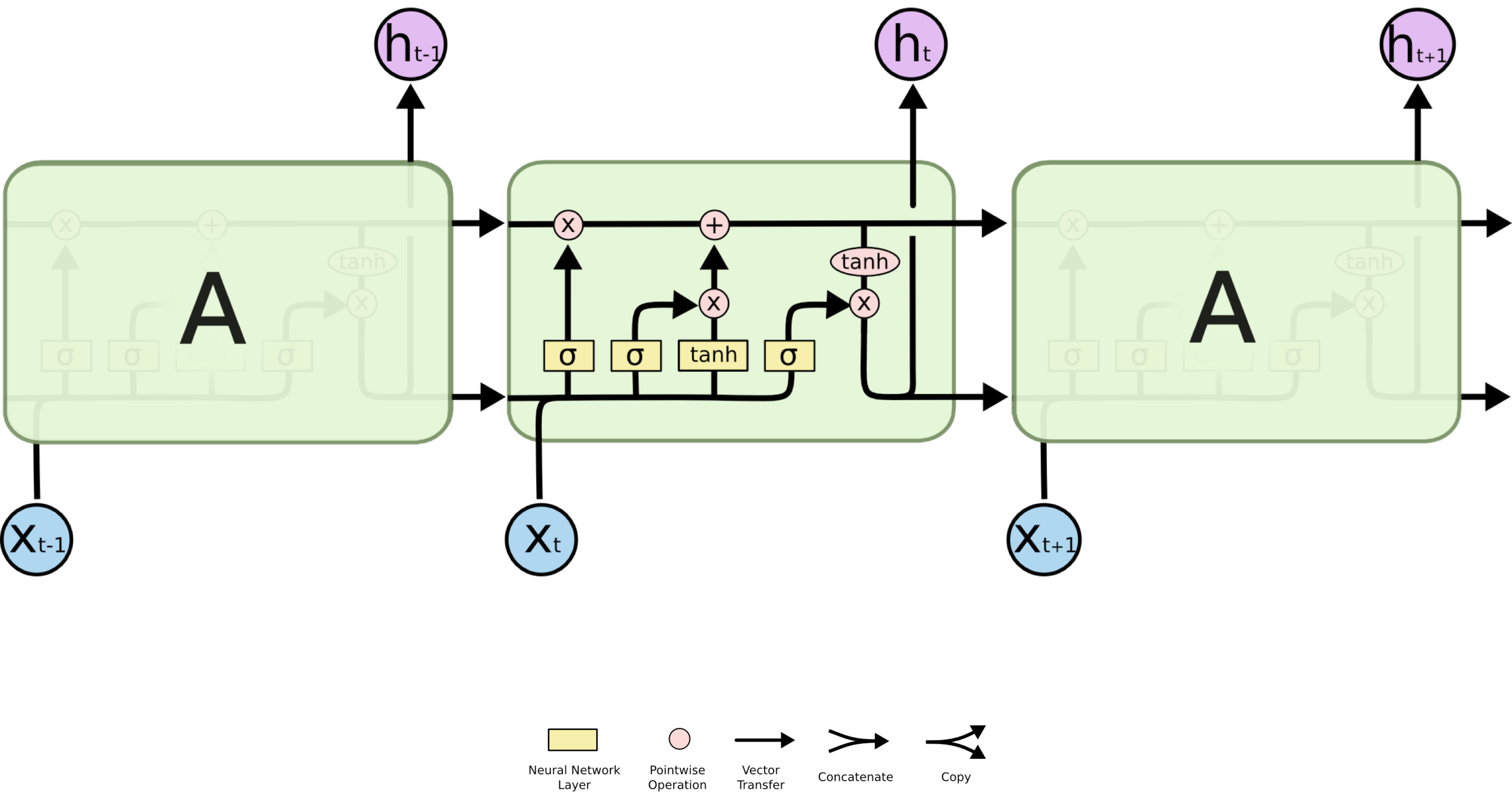
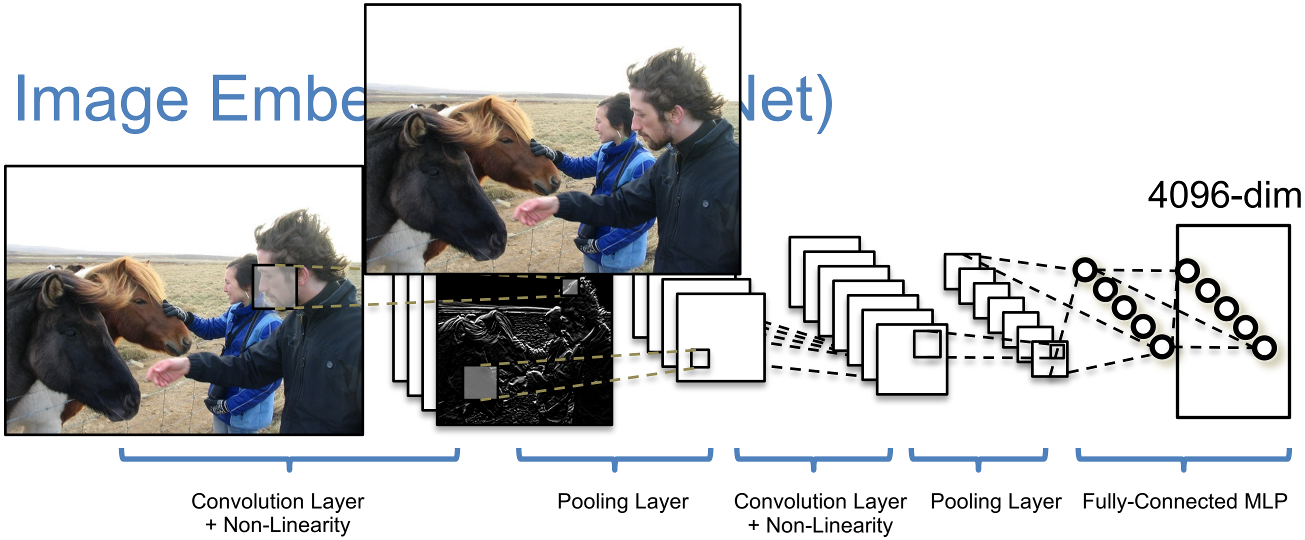


Image Credit: Christopher Olah (http://colah.github.io/posts/2015-08-Understanding-LSTMs/)

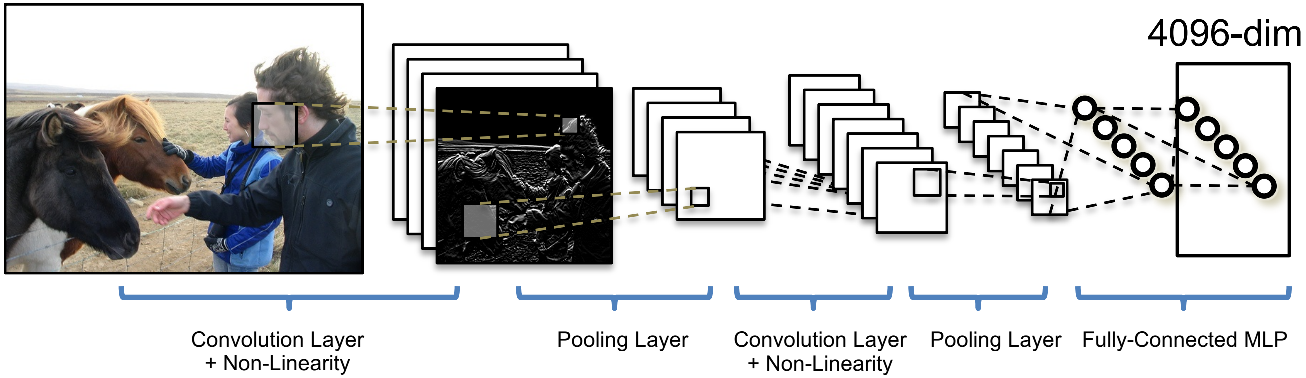
Neural Image Captioning

Image Embe Net)

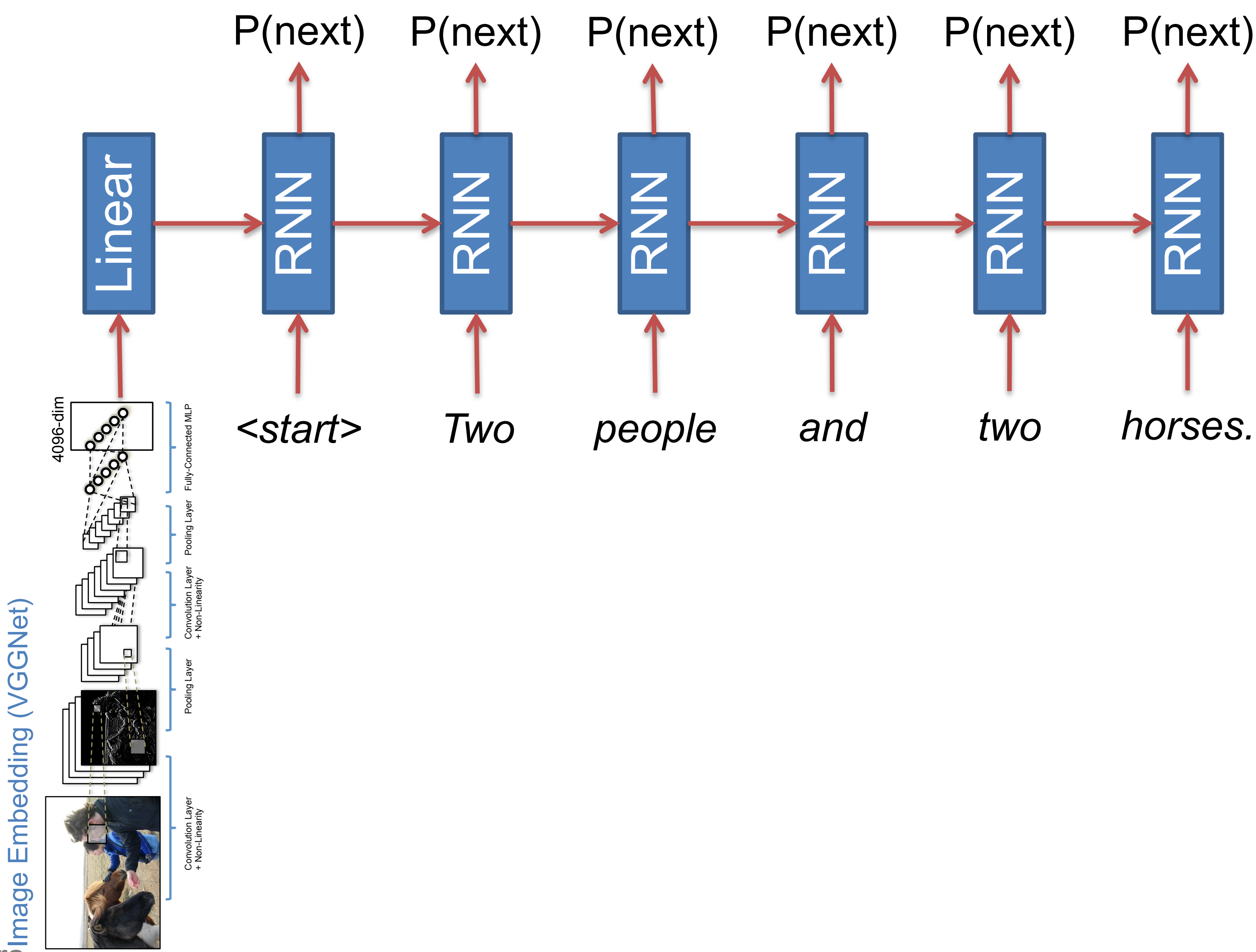


Neural Image Captioning

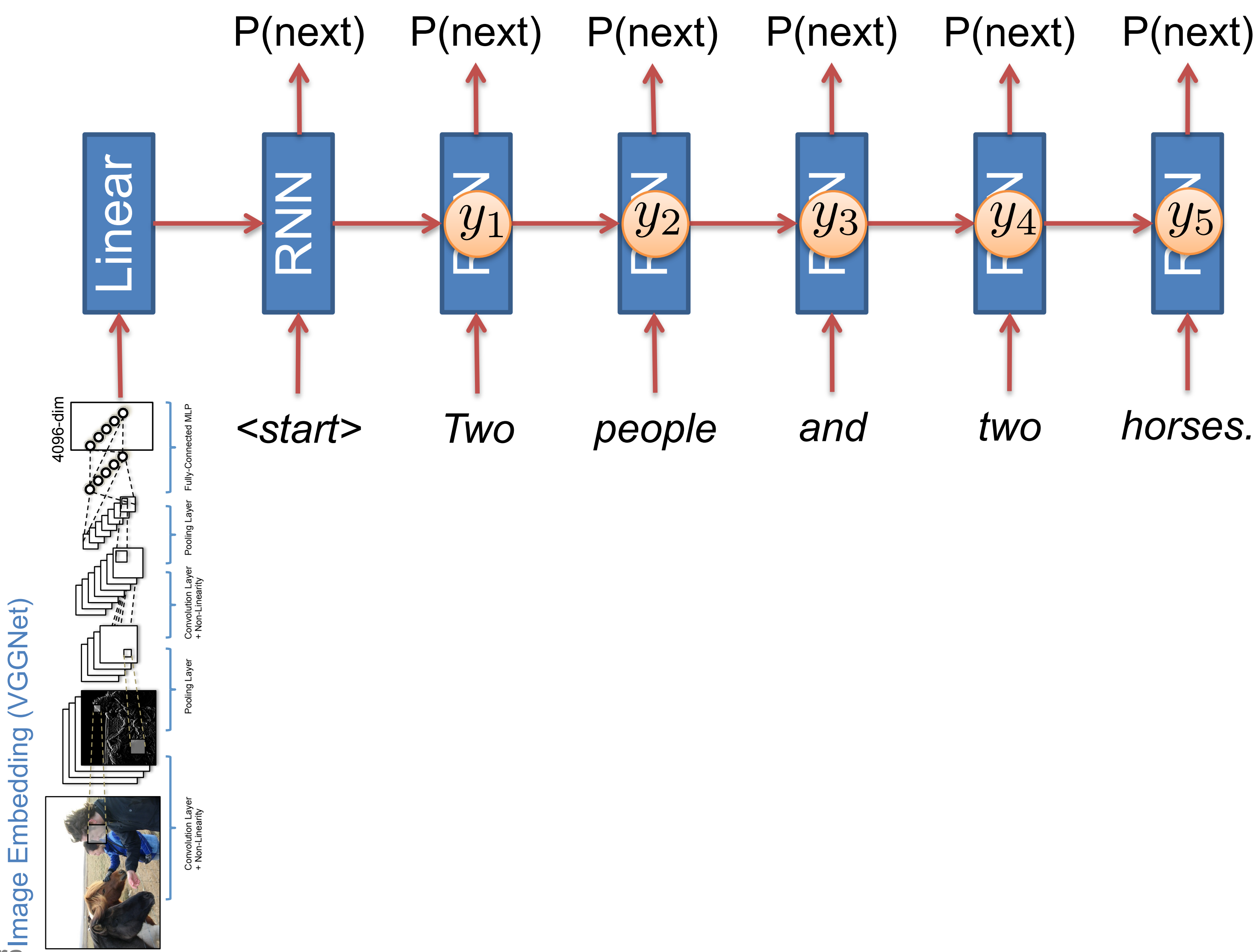
Image Embedding (VGGNet)



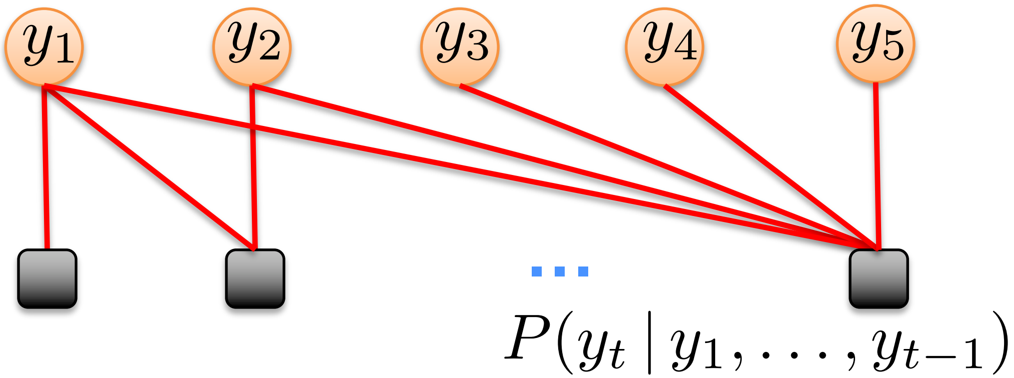
Neural Image Captioning



Neural Image Captioning



Sequence Model Factor Graph



Beam Search Demo

- http://dbs.cloudcv.org/captioning&mode=interactive

Image Captioning

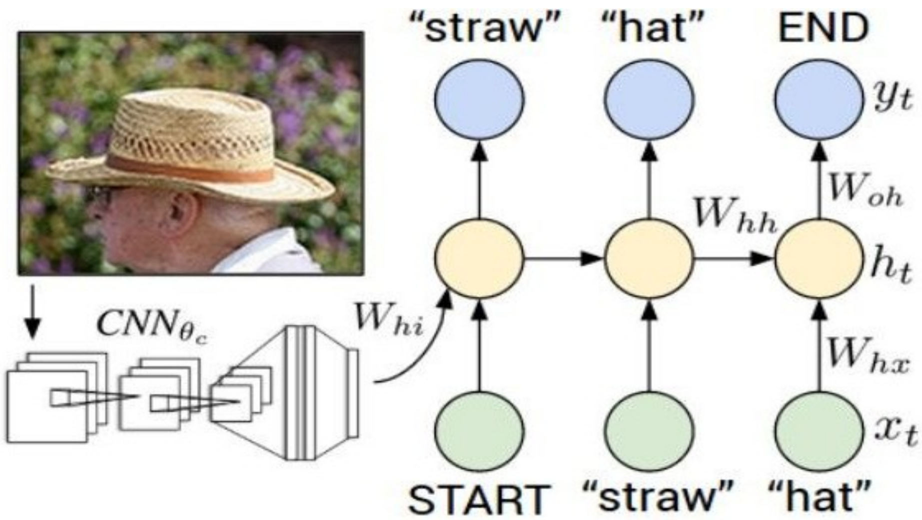
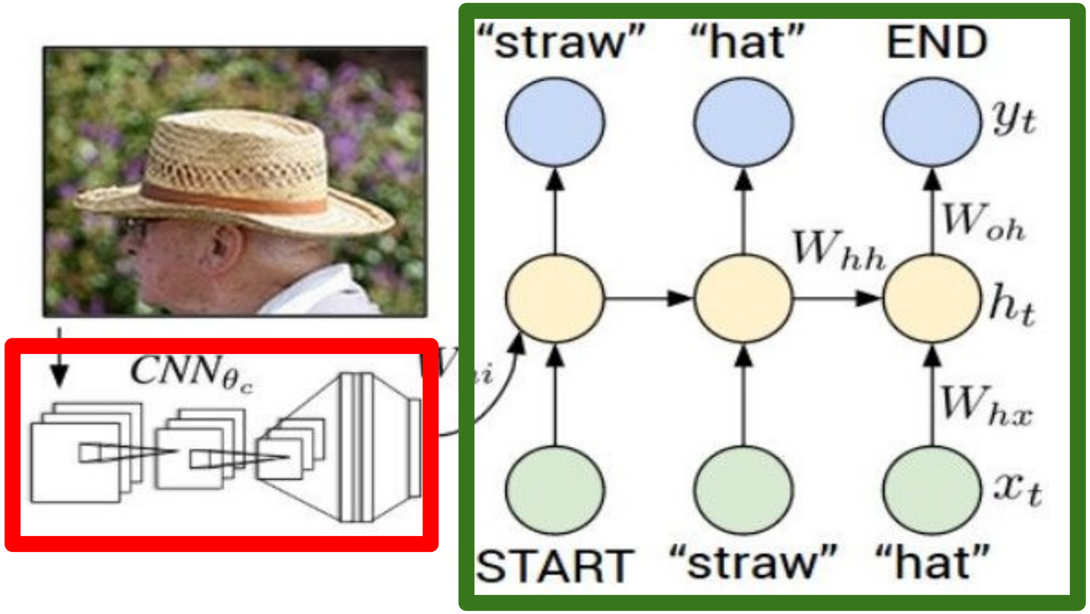


Figure from Karpathy et a, "Deep Visual-Semantic Alignments for Generating Image Descriptions", CVPR 2015; figure copyright IEEE, 2015. Reproduced for educational purposes.

- Many recent works on this:
- Baidu/UCLA: Explain Images with Multimodal Recurrent Neural Networks
- Toronto: Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models
- Berkeley: Long-term Recurrent Convolutional Networks for Visual Recognition and Description
- Google: Show and Tell: A Neural Image Caption Generator
- Stanford: Deep Visual-Semantic Alignments for Generating Image Description
- UML/UT: Translating Videos to Natural Language Using Deep Recurrent Neural Networks
- Microsoft/CMU: Learning a Recurrent Visual Representation for Image Caption Generation
- Microsoft: From Captions to Visual Concepts and Back

Recurrent Neural Network



Convolutional Neural Network

test image

image

conv-64
conv-64
maxpool

conv-128
conv-128
maxpool

conv-256
conv-256
maxpool

conv-512
conv-512
maxpool

conv-512
conv-512
maxpool

FC-4096
FC-4096
FC-1000
softmax

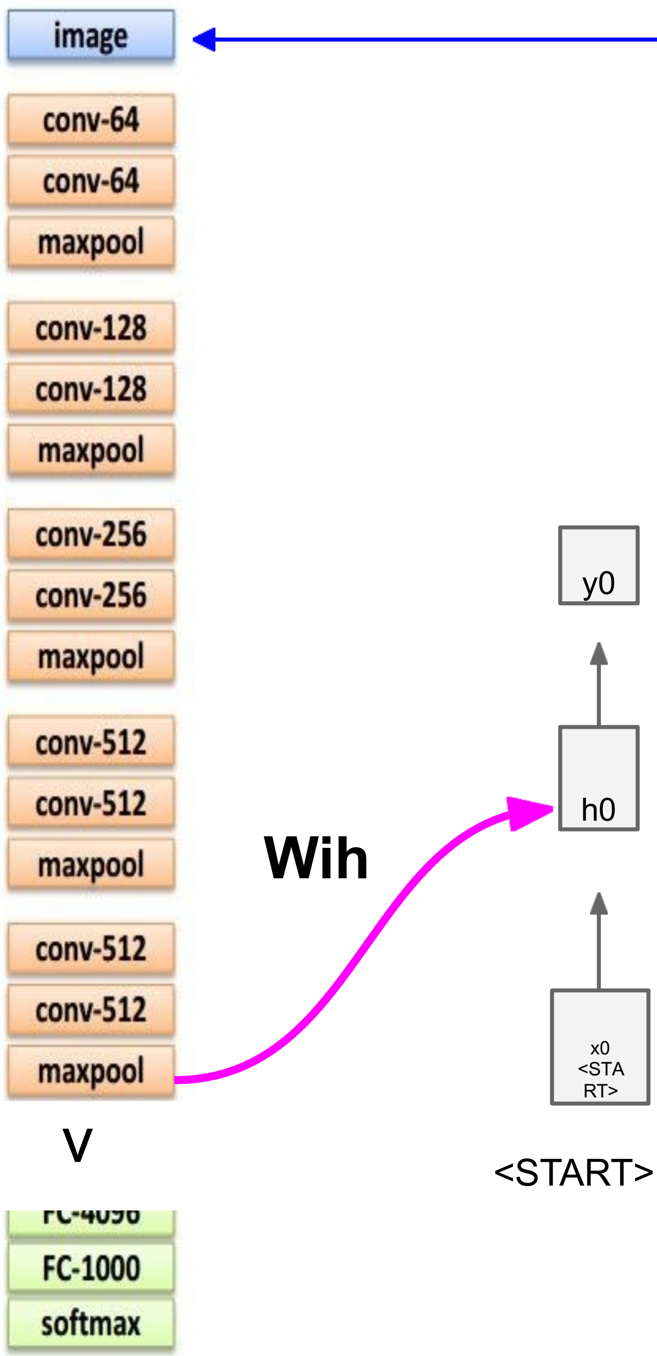
test image


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
FC-1000
softmax
test image

test image

x0
<STA
RT>

<START>



test image

before:

$$h = \tanh(W_{xh} * x + W_{hh} * h)$$

now:

$$h = \tanh(W_{xh} * x + W_{hh} * h + W_{ih} * v)$$

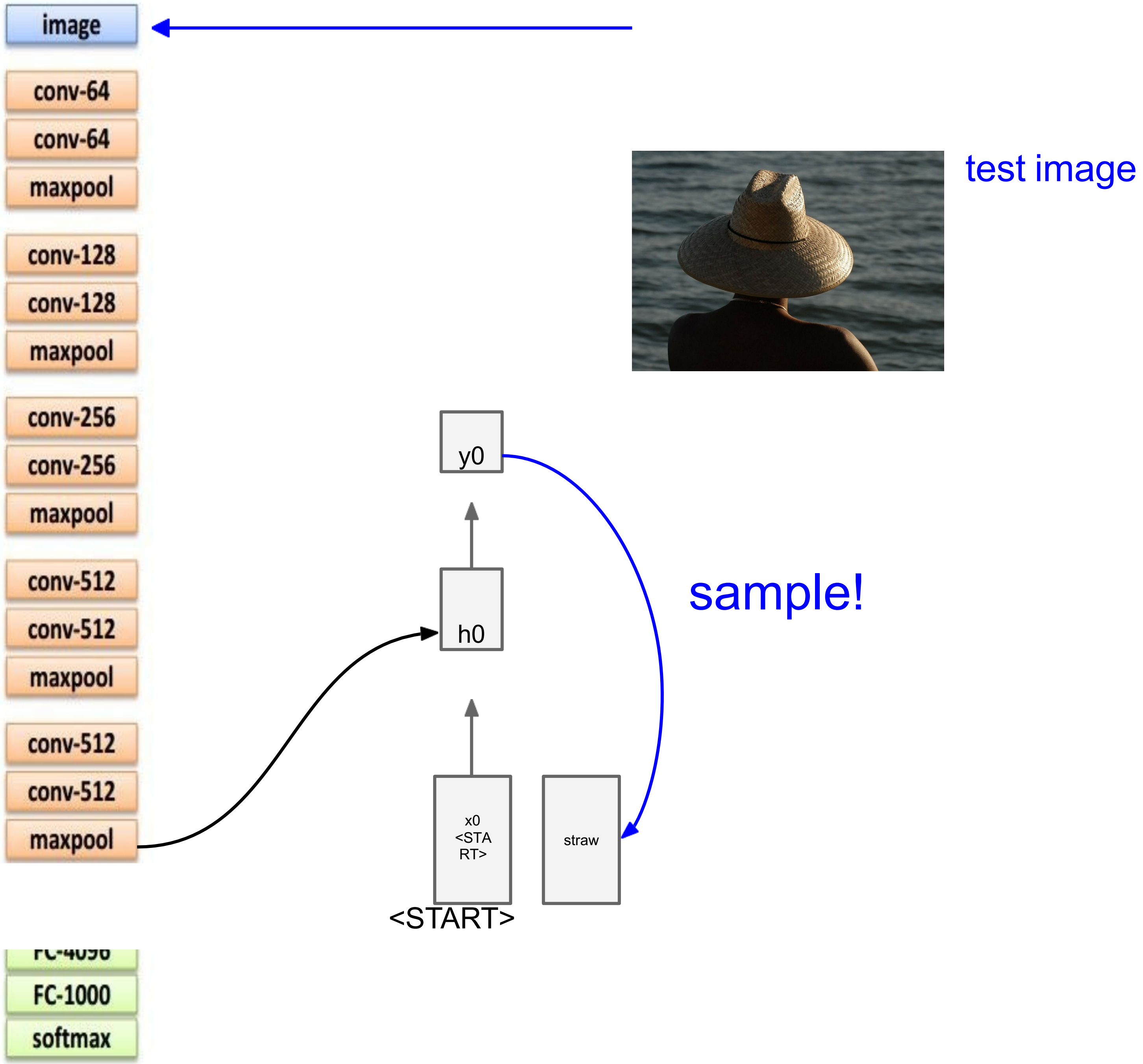


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-1000
softmax
test image
y0
h0
sample!
x0 <START>
straw
<START>

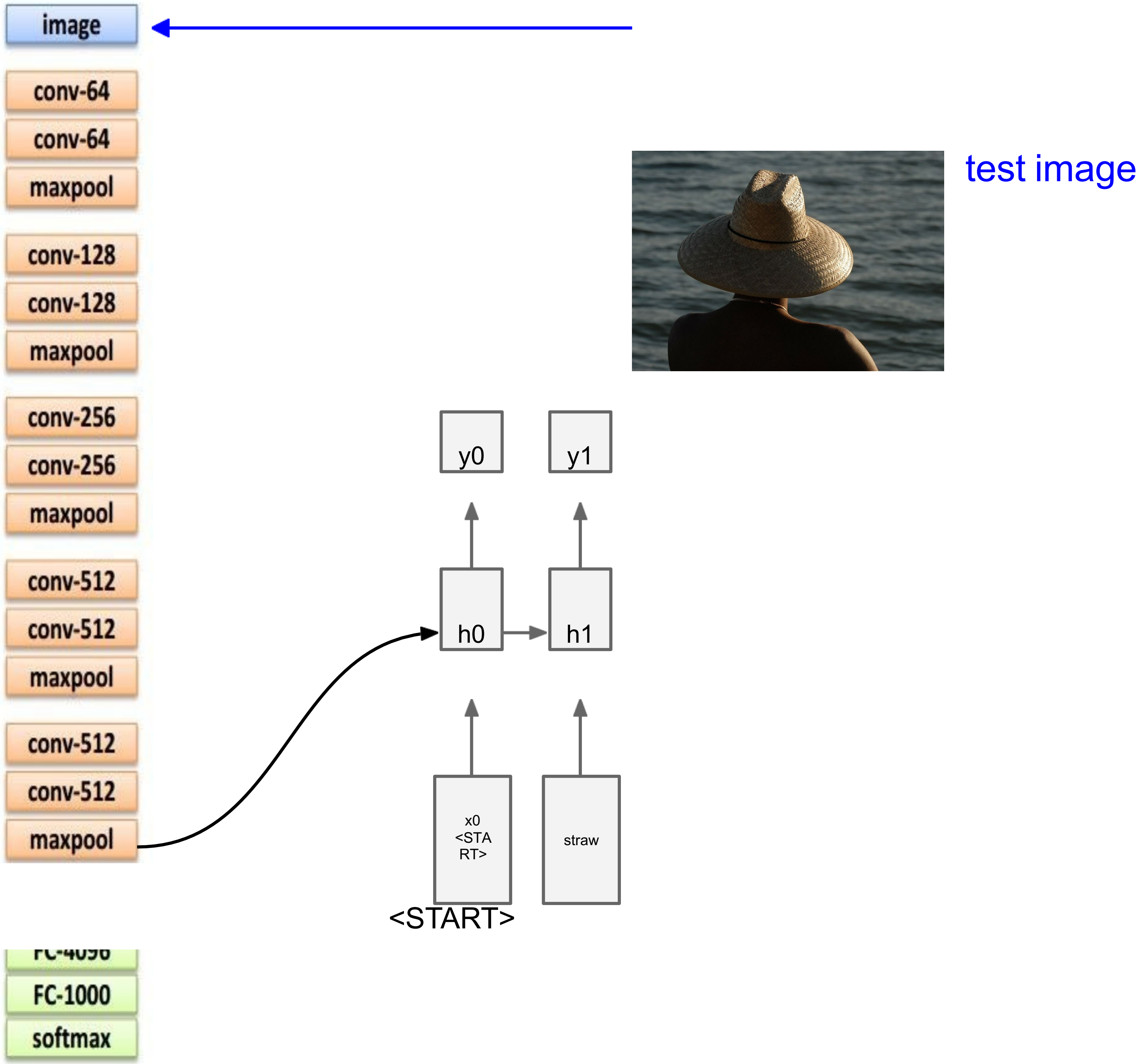


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-1000
softmax
test image
y0
y1
h0
h1
x0
<START>
straw
<START>

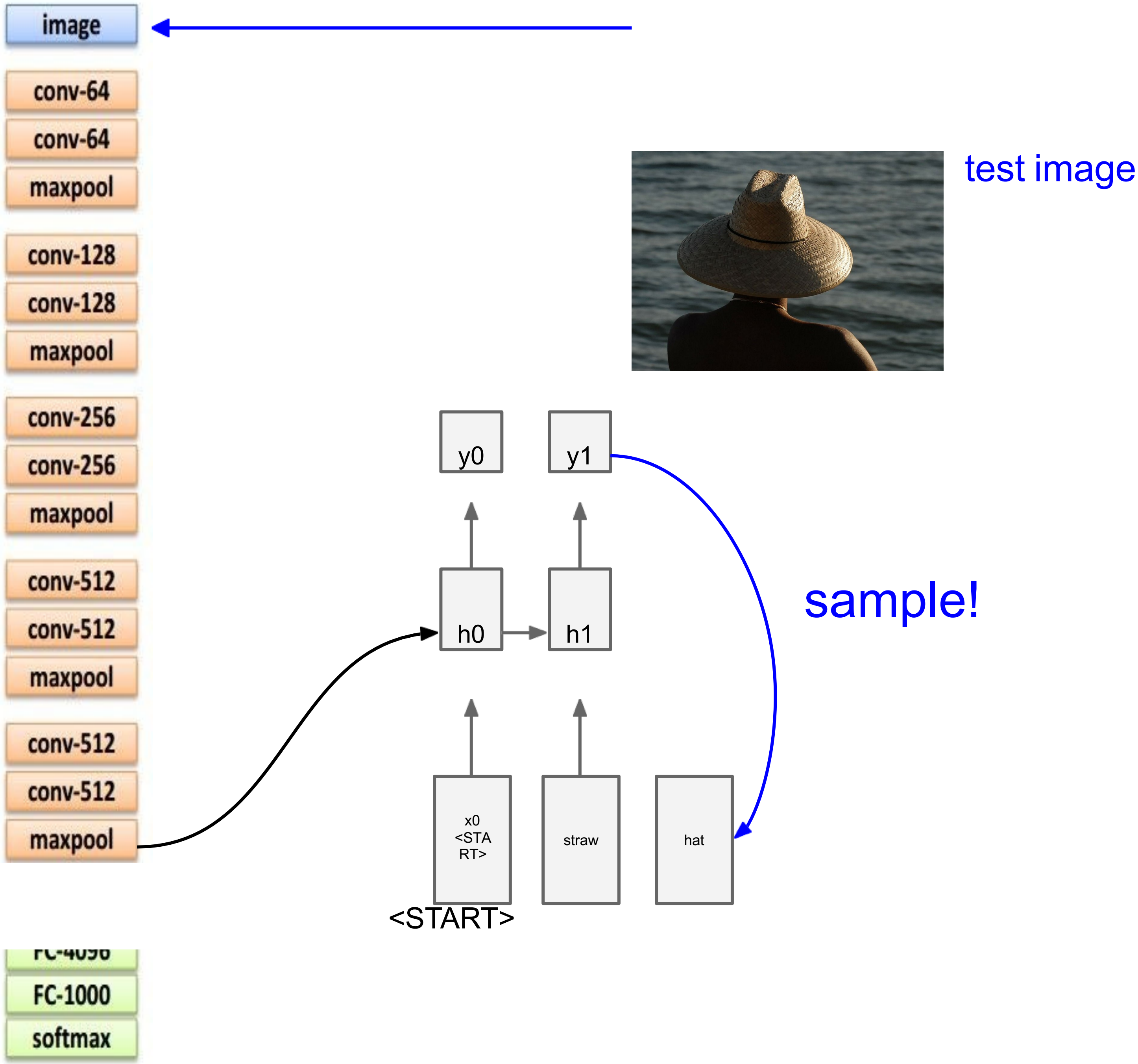


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-1000
softmax
test image
y0
y1
h0
h1
sample!
x0
<STA
RT>
straw
hat
<START>

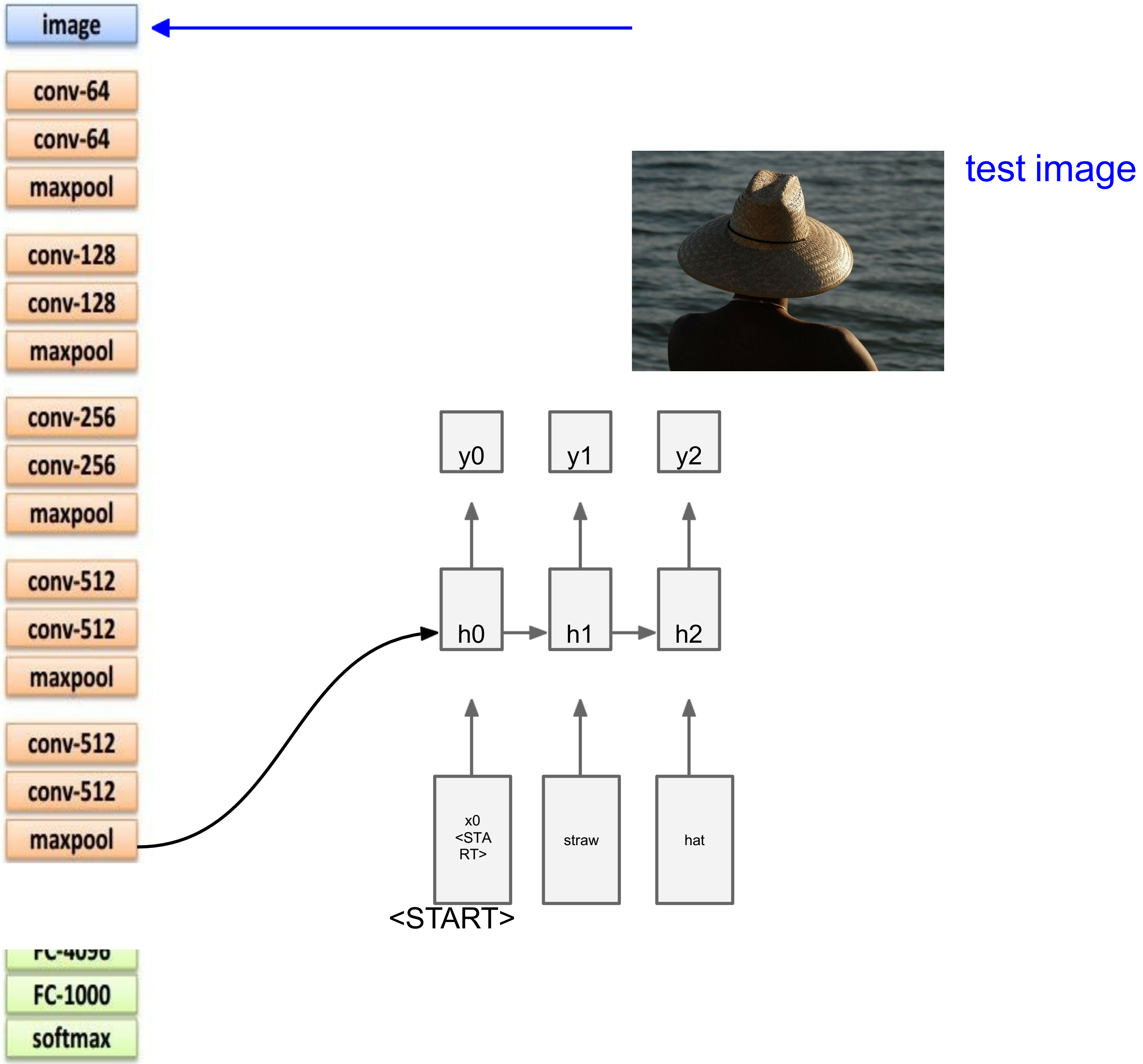


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-1000
softmax
test image
y0
y1
y2
h0
h1
h2
x0
<STA
RT>
straw
hat
<START>

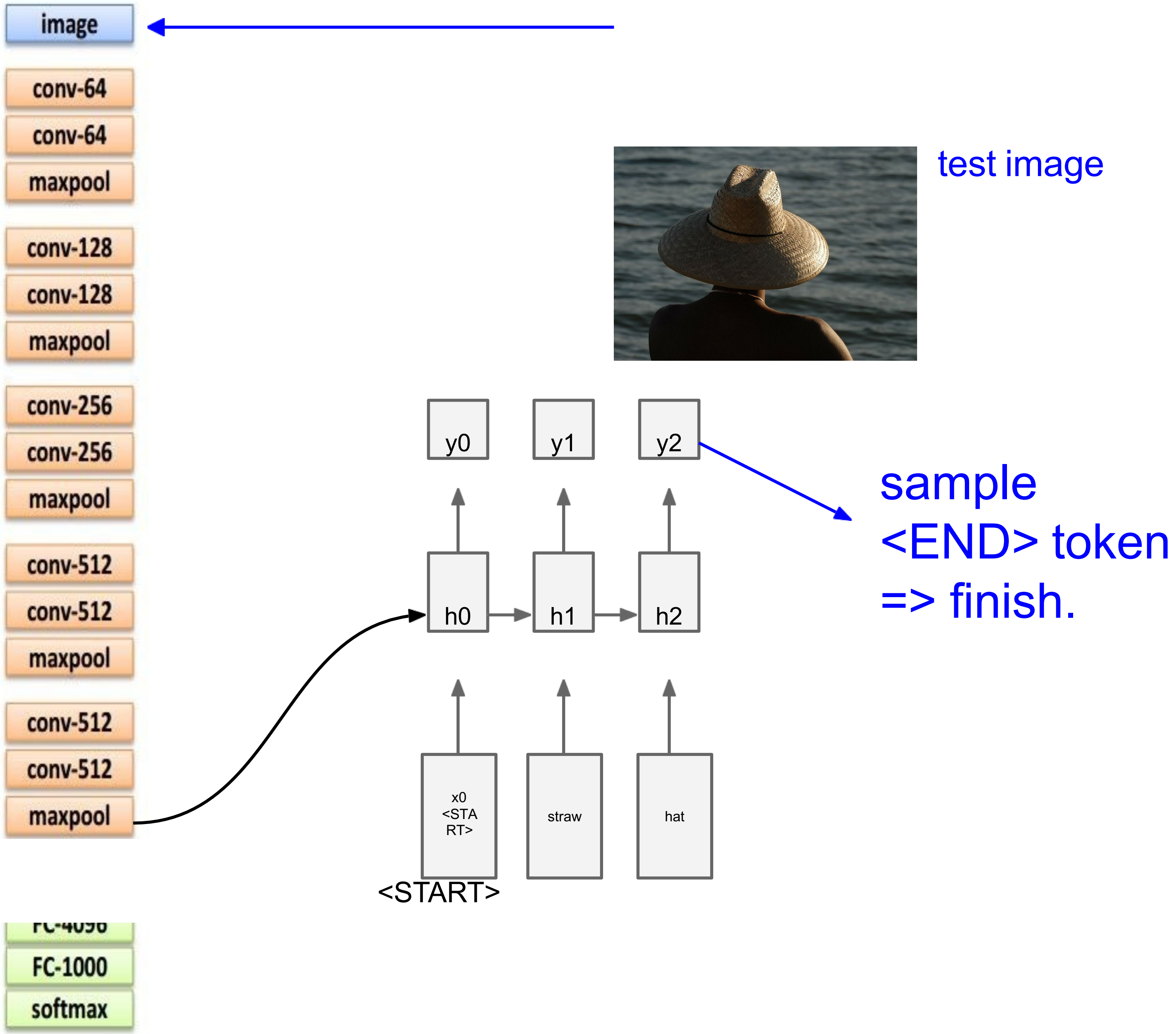


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-1000
softmax
test image
y0
y1
y2
h0
h1
h2
x0 <STA RT>
straw
hat
<START>
sample
<END> token
=> finish.

Image Captioning: Example Results

Captions generated using neuraltalk2
All images are CC0 Public domain: cat suitcase, cat tree, dog, bear, surfers, tennis, giraffe, motorcycle

A cat sitting on a suitcase on the floor

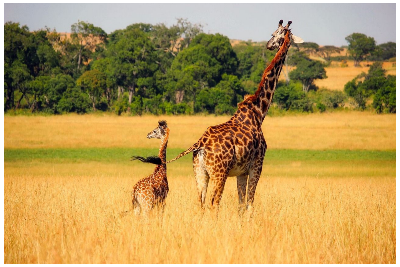
A cat is sitting on a tree branch

A dog is running in the grass with a frisbee

A white teddy bear sitting in the grass

Two people walking on the beach with surfboards

A tennis player in action on the court



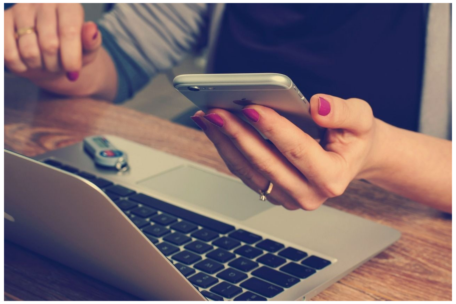
Two giraffes standing in a grassy field

A man riding a dirt bike on a dirt track

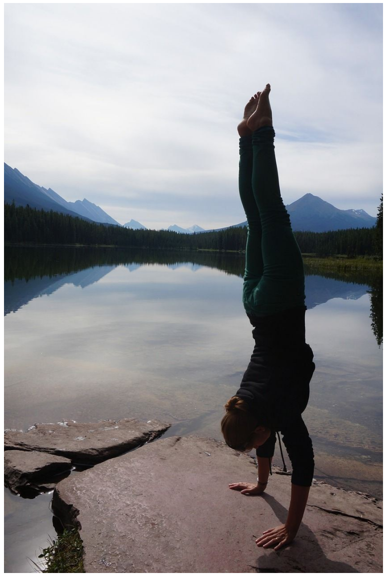
Image Captioning: Failure Cases

Captions generated using neuraltalk2
All images are CC0 Public domain: fur coat, handstand, spider web, baseball

A woman is holding a cat in her hand



A person holding a computer mouse on a desk

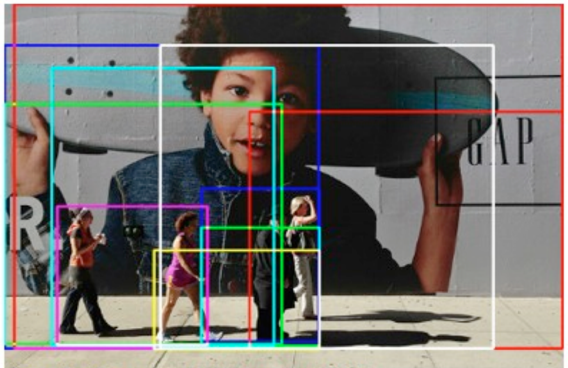


A woman standing on a beach holding a surfboard

A bird is perched on a tree branch

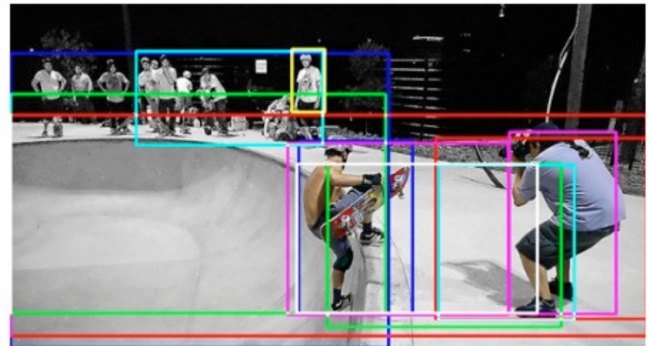
A man in a baseball uniform throwing a ball

More Image Captioning Examples



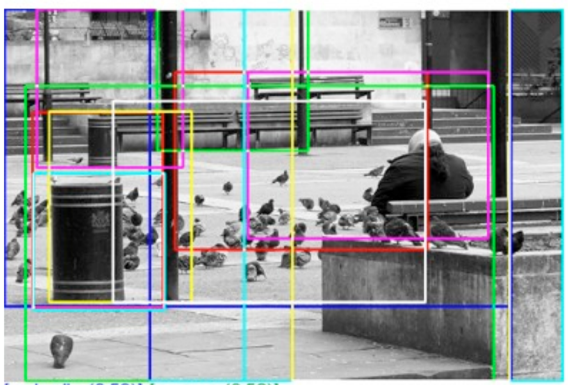
[men (0.59)] [group (0.66)] [woman (0.64)]
[people (0.89)] [holding (0.60)] [playing (0.61)] [tennis (0.69)]
[court (0.51)] [standing (0.59)] [skis (0.58)] [street (0.52)]
[man (0.77)] [skateboard (0.67)]

a group of people standing next to each other
people stand outside a large ad for gap featuring a young boy



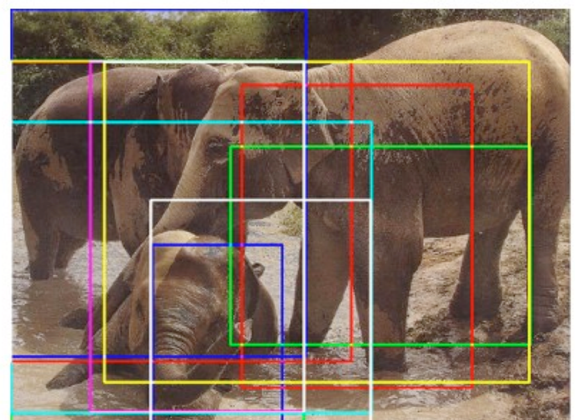
[person (0.55)] [street (0.53)] [holding (0.55)] [group (0.63)] [slope (0.51)]
[standing (0.62)] [snow (0.91)] [skis (0.74)] [player (0.54)]
[people (0.85)] [men (0.57)] [skiing (0.51)]
[skateboard (0.89)] [riding (0.75)] [tennis (0.74)] [trick (0.53)] [skate (0.52)]
[woman (0.52)] [man (0.86)] [down (0.61)]

a group of people riding skis down a snow covered slope
a guy on a skate board on the side of a ramp



[umbrella (0.59)] [woman (0.52)]
[fire (0.96)] [hydrant (0.96)] [street (0.79)] [old (0.50)]
[bench (0.81)] [building (0.75)] [standing (0.57)] [baseball (0.55)]
[white (0.82)] [sitting (0.65)] [people (0.79)] [photo (0.53)]
[black (0.84)] [kitchen (0.54)] [man (0.72)] [water (0.56)]

a black and white photo of a fire hydrant
a courtyard full of poles pigeons and garbage cans also has benches on either side of it one of which shows the back of a large person facin g in the direction of the pigeons



[horse (0.53)] [bear (0.71)] [elephant (0.99)] [elephants (0.95)]
[brown (0.68)] [baby (0.62)] [walking (0.57)] [laying (0.61)]
[man (0.57)] [standing (0.79)] [field (0.65)]
[water (0.83)] [large (0.71)] [dirt (0.65)] [river (0.58)]

a baby elephant standing next to each other on a field
elephants are playing together in a shallow watering hole

From Captions to Visual Concepts and Back, Hao Fang∗ Saurabh Gupta∗ Forrest Iandola∗ Rupesh K. Srivastava∗, Li Deng Piotr Dollar, Jianfeng Gao Xiaodong He, Margaret Mitchell John C. Platt, C. Lawrence Zitnick, Geoffrey Zweig, CVPR 2015.

Engaging Image Captioning Via Personality

Kurt Shuster, Samuel Humeau, Hexiang Hu,
Antoine Bordes, Jason Weston

Standard (COCO) Image Captioning Models

Man in black shirt is playing guitar.

Standard (COCO) Image Captioning Models

Man in black shirt is playing guitar.

A plate with a sandwich and salad on it.

Good for: testing if model understands image content

Bad for: engaging human reader

Standard (COCO) Image Captioning Models

Man in black shirt is playing guitar.

A plate with a sandwich and salad on it.

Good for: testing if model understands image content
Bad for: engaging human reader

Want to be good at both of these!!!

What makes an utterance engaging? One answer: personality, emotion & style traits

(not always just neutral, factual tone)

Existing Work

Neutral, factual captions:

- COCO (Chen et al., 2015) and Flickr30k (Young et al., 2014)
- Many models developed for them (discussed later).

Funny captions:

- wordplay (puns) (Chandrasekaran et al., 2017)
- or training on data from humour websites (Yoshida et al., 2018).

Using user features:

- location and age (Denton et al., 2015)
- or knowledge of the reader's active vocabulary (Park et al., 2017).

Style transfer:

- unsupervised (Mathews et al., 2018).
- Small datasets, e.g. Senticap (800 examples), (Mathews et al., 2016)
- romantic and humorous only - FlickrStyle10K , 10k examples - Gan et al. (2017)

MIT Personality List - 638 Traits

638 Primary Personality Traits

Positive Traits (234 = 37%)

1. Accessible
2. Active
3. Adaptable
4. Admirable
5. Adventurous
6. Agreeable
7. Alert
8. Allocentric
9. Amiable
10. Anticipative
11. Appreciative
12. Articulate
13. Aspiring
14. Athletic
15. Attractive
16. Balanced
17. Benevolent
18. Brilliant
19. Calm
20. Capable
21. Captivating
22. Caring
23. Challenging
24. Charismatic
25. Charming
26. Cheerful

215. Tidy
216. Tolerant
217. Tractable
218. Trusting
219. Uncomplaining
220. Understanding
221. Undogmatic
222. Unfoolable
223. Upright
224. Urbane
225. Venturesome
226. Vivacious
227. Warm
228. Well-bred
229. Well-read
230. Well-rounded
231. Winning
232. Wise
233. Witty
234. Youthful

Neutral Traits (292 = 18%)

1. Absentminded
2. Aggressive
3. Ambitious
4. Amusing
5. Artful
6. Ascetic
7. Authoritarian
8. Big-thinking
9. Boyish
10. Breezy
11. Businesslike
12. Busy
13. Casual

Negative Traits (292 = 46%)

1. Abrasive
2. Abrupt
3. Agonizing
4. Aimless
5. Airy
6. Aloof
7. Amoral
8. Angry
9. Anxious
10. Apathetic
11. Arbitrary
12. Argumentative
13. Arrogantt
14. Artificial
15. Asocial
16. Assertive
17. Astigmatic
18. Barbaric
19. Bewildered
20. Bizarre
21. Bland
22. Blunt
23. Biosterous

Step 1: build a dataset

Your personality: **Sarcastic**

Your comment:

Can this island get any smaller?

- Selected 215 personality traits
- Images from YFFC100M
- Collect captions via annotators

Examples from the dataset

Sarcastic
Yes please sit by me

Mellow
Look at that smooth easy catch of the ball. like ballet.

Zany
I wish I could just run down this shore!

Contradictory
Love what you did with the place!

Contemptible
I can't believe no one has been taking care of this plant. Terrible

Energetic
About to play the best tune you've ever heard in your life. Get ready!

Examples from the dataset

Get ready!

Kind
they left me a parking spot

Spirited
That is one motor cycle enthusi-ast!!!

Creative
Falck alarm, everyone. Just a Falck alarm.

Crazy
I drove down this road backwards at 90 miles per hour three times

Morbid
I hope this car doesn't get into a wreck.

Questioning
Why do people think its cool to smoke cigarettes?

Step 1: Collect a large supervised dataset

Table 1: PERSONALITY-CAPTIONS dataset statistics.

Split	train	valid	test
Number of Examples	186,858	5,000	10,000
Number of Personality Types	215	215	215
Vocabulary Size	35559	5557	8137
Average Tokens per Caption	11.6	11.2	11.4

Step 2: Build strong models

We make use of state-of-the-art in vision and language domains to build our models:

Image Encoder:

- ResNeXt (Xie et al., 2016) trained on 3.5 billion Instagram pictures following Mahajan et al. (2018), which we call *ResNeXt-IG-3.5B*.
- *Shown to work very well on ImageNet classification (but not captioning).*

Text Encoder:

- Transformer (Vaswani et al., 2017) trained on 1.7 billion Reddit dialogue examples, following (Mazaré et al., 2018).
- *Shown to work very well for PersonaChat dialogue (but not captioning).*

Models: we consider both generative and retrieval models.

- Generative: *consider three widely used architectures:*
  - ShowTell (Vinyals et al., 2015)
  - ShowAttTell (Xu et al., 2015)
  - UpDown (Anderson et al., 2018)

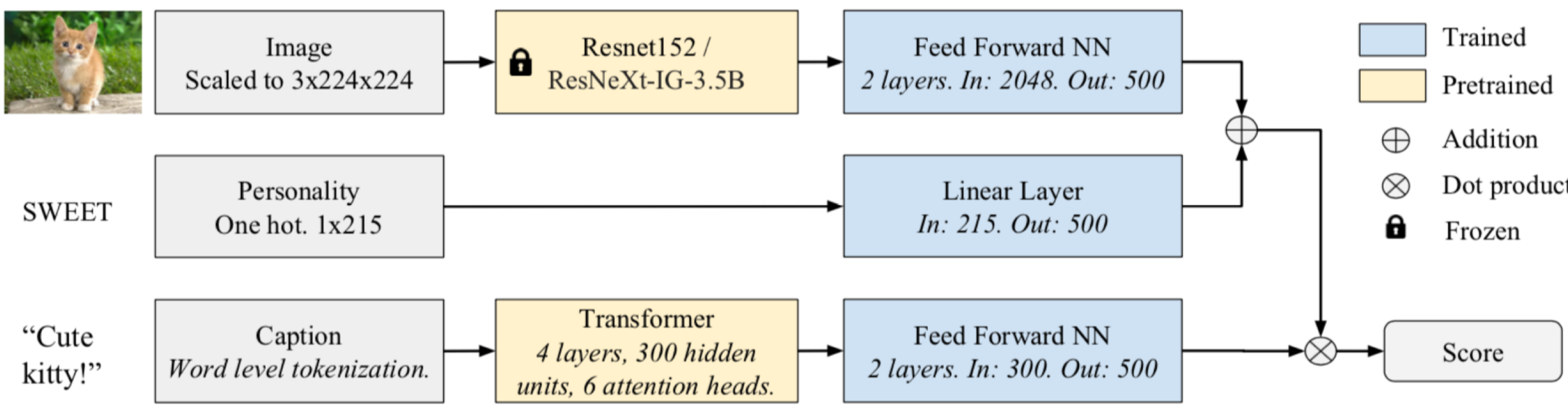
Use ResNeXt-IG-3.5B and add learnt personality features to each decoder step

Models: we consider both generative and retrieval models.

- Generative: *consider three recent best architectures:*
  - ShowTell (Vinyals et al., 2015)
  - ShowAttTell (Xu et al., 2015)
  - UpDown (Anderson et al., 2018)

Use ResNeXt-IG-3.5B and add learnt personality features to each decoder step

- Retrieval: *TransResNet*



Our generative models are good at **understanding image content**.

Table 3: Generative model performance on COCO caption using the test split of (Karpathy & Fei-Fei, 2015)

Method	Image Encoder	BLEU1	BLEU4	ROUGE-L	CIDEr	SPICE
Adaptive (Lu et al., 2017)	ResNet	74.2	32.5	-	108.5	19.5
Att2in (Rennie et al., 2017)	ResNet	-	33.3	55.3	111.4	-
NBT (Lu et al., 2018)	ResNet	75.5	34.7	-	107.2	20.1
UpDown (Anderson et al., 2018)	ResNet FRCNN	**79.8**	36.3	56.9	120.1	**21.4**
ShowTell (Our)	ResNet152	75.2	31.5	54.2	103.9	18.4
ShowAttTell (Our)	ResNet152	76.5	32.4	55.1	109.7	19.2
UpDown (Our)	ResNet152	77.0	33.9	55.6	112.7	19.6
ShowTell (Our)	ResNeXt-IG-3.5B	78.2	35.0	56.6	119.9	20.8
ShowAttTell (Our)	ResNeXt-IG-3.5B	78.8	35.6	57.1	121.8	20.6
UpDown (Our)	ResNeXt-IG-3.5B	79.3	**36.4**	**57.5**	**124.0**	21.2

Our retrieval models are good at **understanding image content**.

Table 4: Retrieval model performance on Flickr30k and COCO caption using the splits of (Karpathy & Fei-Fei, 2015). COCO caption performance is measured on the 1k image test split.

Model	Text Pre-training	Flickr30k R@1	Flickr30k R@5	Flickr30k R@10	COCO R@1	COCO R@5	COCO R@10
UVS (Kiros et al., 2014)	-	23.0	50.7	62.9	43.4	75.7	85.8
Embedding Net (Wang et al., 2018)	-	40.7	69.7	79.2	50.4	79.3	69.4
sm-LSTM (Huang et al., 2016)	-	42.5	71.9	81.5	53.2	83.1	91.5
VSE++ (ResNet, FT) (Faghri et al., 2017)	-	52.9	80.5	87.2	64.6	90.0	95.7
GXN (i2t+t2i) (Gu et al., 2017)	-	56.8	-	89.6	**68.5**	-	**97.9**
TransResNet model variants:							
Transformer, ResNet152	Full	10.3	27.3	38.8	21.7	45.6	58.9
Bag of words ResNeXt-IG-3.5B	None	50.0	81.1	90.0	51.6	85.3	93.4
Transformer ResNeXt-IG-3.5B	None	55.6	83.2	90.5	64.0	90.6	96.3
Bag of words ResNeXt-IG-3.5B	Word	58.6	87.2	92.9	54.7	87.1	94.5
Transformer ResNeXt-IG-3.5B	Word	**68.4**	**90.6**	**95.3**	67.3	**91.7**	96.5

Our generative models are good at using **personality**

Table 5: Generative model caption performance on the PERSONALITY-CAPTIONS test set.

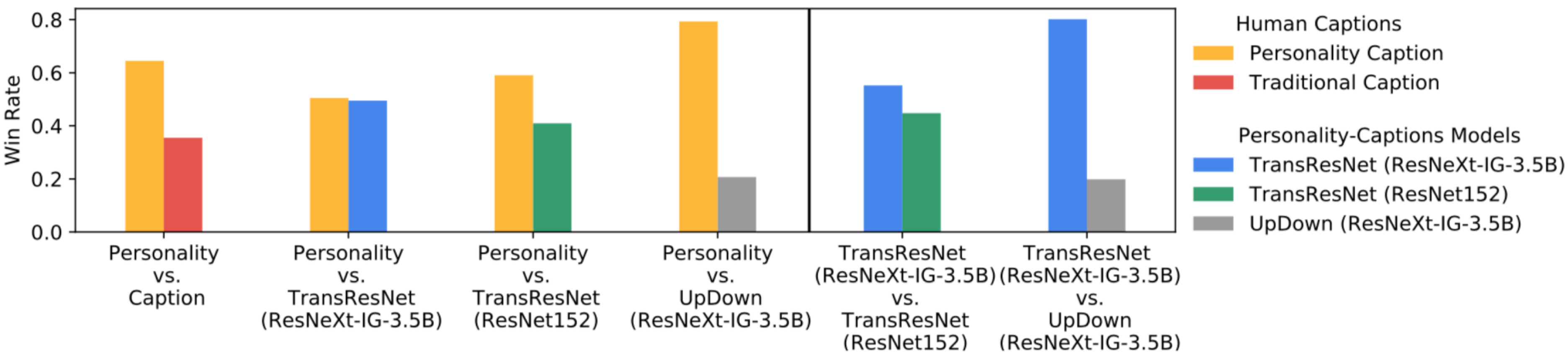
Method	Image Encoder	Personality Encoder	BLEU1	BLEU4	ROUGE-L	CIDEr	SPICE
SHOWTELL	ResNet152	Yes	12.4	1.4	13.2	14.5	1.6
SHOWATTTELL	ResNet152	Yes	15.3	1.3	13.1	15.2	3.4
UPDOWN	ResNet152	Yes	15.4	1.4	14.6	16.9	4.9
SHOWTELL	ResNeXt-IG-3.5B	No	15.2	0.9	13.3	14.4	4.6
SHOWATTTELL	ResNeXt-IG-3.5B	No	13.8	0.9	13.1	17.6	5.4
UPDOWN	ResNeXt-IG-3.5B	No	14.3	1.0	13.5	18.0	7.0
SHOWTELL	ResNeXt-IG-3.5B	Yes	14.2	1.2	14.5	15.4	2.2
SHOWATTTELL	ResNeXt-IG-3.5B	Yes	15.0	1.4	14.6	18.8	5.9
UPDOWN	ResNeXt-IG-3.5B	Yes	**15.6**	**1.6**	**15.0**	**22.0**	**7.3**

Our retrieval models are good at using **personality**

Table 6: Results for TransResNet retrieval variants on the PERSONALITY-CAPTIONS test set.

Text Encoder	Pre-training	Image Encoder	Personality Encoder	R@1
Transformer	Full	ResNet152	No	16.6
Bag of Words	None	ResNet152	Yes	24.2
Transformer	None	ResNet152	Yes	26.8
Bag of Words	Word	ResNet152	Yes	28.5
Transformer	Full	ResNet152	Yes	34.4
Transformer	Full	ResNeXt-IG-3.5B	No	38.5
Bag of Words	None	ResNeXt-IG-3.5B	Yes	38.6
Transformer	None	ResNeXt-IG-3.5B	Yes	42.9
Bag of Words	Word	ResNeXt-IG-3.5B	Yes	45.7
Transformer	Full	ResNeXt-IG-3.5B	Yes	**53.5**

Human evaluation studies: *our best model is close to matching human performance*



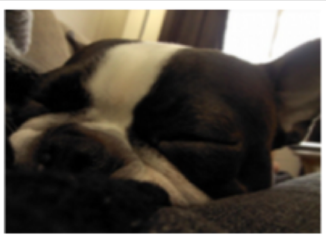
Standard captioning output: A plate with a sandwich and salad on it.
Our model with different personality traits:

Sweet	That is a lovely sandwich.
Dramatic	This sandwich looks so delicious! My goodness!
Anxious	I'm afraid this might make me sick if I eat it.
Sympathetic	I feel so bad for that carrot, about to be consumed.
Arrogant	I make better food than this
Optimistic	It will taste positively wonderful!
Money-minded	I would totally pay $100 for this plate.

Standard Captioning Model: man in black shirt is playing guitar

***Our model with different personas*:**
Artful: He has the most perfect technique of any solo artist
Overimaginative: I'm thinking he could grab that guitar and bust out Stairway to Heaven right now. In doing so, he could summon all the long-gone rock greats.
Romantic: This guitarist is so cute I want to take him on a date!
Arrogant: He holds the guitar wrong. I would do a much better job if I was in the photo.
Absentminded: Okay guys. What do I do now? Is this a banjo?

More examples of our best model

Image	Personality	Generated comment
	Sweet	I love, love, love these chairs! I want the big one in my house!
	Vague	This chair is either covered in snow or the snow is covered in the chair.
	Cultured	These chairs remind me of the Swedish interior design revolution of the 70's.
	Paranoid	What if someone fell off those chairs.
	Overimaginative	Those chairs look like they could be in a doll house.
	Arrogant	I've seen better sunsets elsewhere.
	Overimaginative	that sunset is so orange it could be a fruit
	Vague	It's the sunset.
	Optimistic	The sunset makes look forward to a happy tomorrow.
	Charming	The way the sun is hitting the water makes for a romantic evening.
	Sweet	What a cute puppy, reminds me of my friends.
	Skeptical	I don't think this dog will bite me.
	Sympathetic	poor dog! It looks so hungry :c
	Vague	it's a dog
	Wishful	I wish that I had a dog as cute as him.

More examples of our best model

	Cultured	I love a cultural celebration.
	Skeptical	I’m not sure if these are guys in costumes or time travelers.
	Sweet	I love that they are celebrating their traditions and culture.
	Overimaginative	They look like they could be dancers in a fantasy movie with dragons!
	Sympathetic	I feel sorry for him having to wear that
	Romantic	If I was an insect, I would definitely make this my mate.
	Humble	I am grateful that spiders eat these disgusting bugs.
	Paranoid	What is going on? Are these insects dangerous?
	Creative	I made something like this from colored toothpicks once
	Money-minded	how much are those? those looks expensive
	Happy	That is so cool! I I love street art!
	Optimistic	The future is bright for people who can dream in artistic ways.
	Critical	I do believe this taggers verbage is a tad junvenile
	Charming	What a charming wall.
	Adventurous	I think I could create art like that, I will go learn and take action.

More examples of our best model

Image	Personality	Generated comment
	Adventurous	This biking event looks like something that I would try!
	Vague	Those people are riding a bike.
	Charming	I bet a wonderful couple uses this bike to tour the countryside together.
	Optimistic	A hopeful cyclist trying to catch up to the pack
	Paranoid	What if all those bikes just tipped over!
	Adventurous	I am so ready for the conference.
	Cultured	This conference is one of the most important ones in the country.
	Vague	The organization on that table is uncertain.
	Dramatic	OMG!! This ceremony is frightening!
	Sympathetic	I feel bad for these people being so cramped in this room.
	Old-fashioned	Such old fashioned script, a true lost art.
	Charming	I could use these to write to my loved ones.
	Argumentative	Can you even read this through all the jpeg artifacts?
	Anxious	I hope this paper doesnt tear, history will be destroyed.
	Dramatic	Some of the most profound things ever written have been on linen.

More examples of our best model

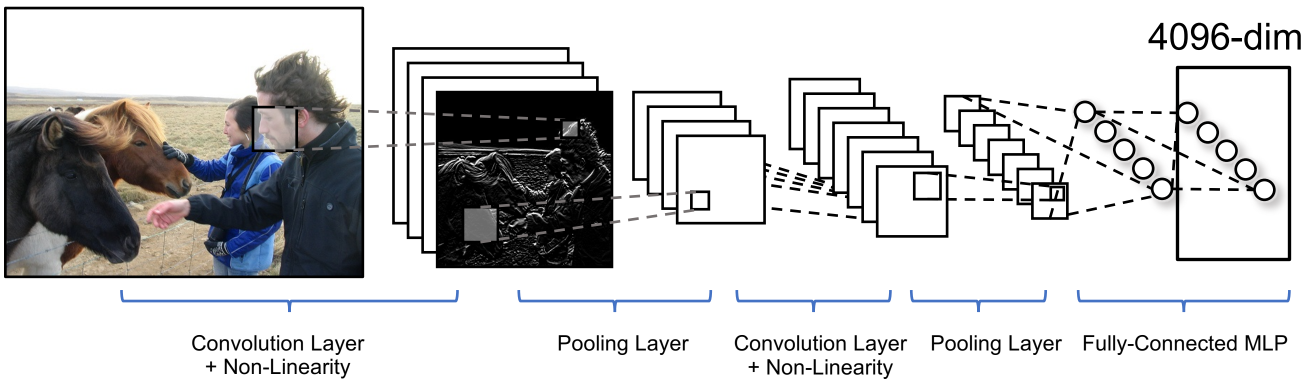
	Happy	It finally snowed, it makes me feel awesome
	Wishful	I wish there was enough for snow angels.
	Boyish	Can I go sledding now?
	Romantic	What a beautiful frost! Looks like the perfect place to fall in love!
	Cultured	The white of the snow provides a glistening contrast to the dead trees.
	Wishful	I wish I could have a life as easy as a plant.
	Money-minded	This plant is probably worth a lot of money
	Critical	the leaf is ruining the picture
	Humble	This plant is a symbol of life in humble opinion. Just gorgeous!
	Paranoid	If you eat this leaf it definetly will not poison you. Or will it...
	Romantic	This valentine concert is for lovers.
	Boyish	It's always fun to get down and jam with the boys!
	Creative	musician performing a song of theirs
	Sweet	oh what lovely young musicians
	Money-minded	I wonder how much the musicians have in student loan debt.

Human Evaluation Examples

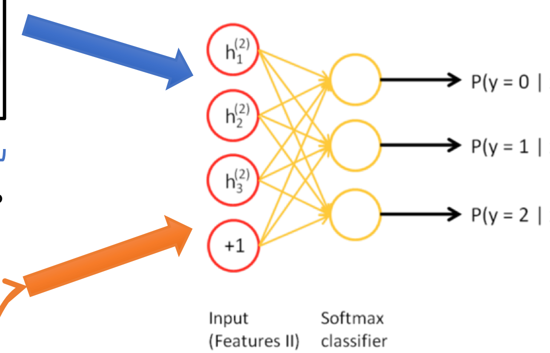
Image and Pers.	Use pers.	Captioning	Caption
	No	Standard	A city on the background, a lake on the front, during a sunset.
	No	Engaging	Talk about summer fun! Can I join? :)
	Yes	Human	i feel moved by the sunset
	Yes	TransResNet	The water at night is a beautiful sight.
Spirited	Yes	UPDOWN	This is a beautiful sunset!
	No	Standard	Rose colored soft yarn.
	No	Engaging	I really want to untangle that yarn.
	Yes	Human	I cannot believe how yummy that looks.
	Yes	TransResNet	What is up with all the knitting on my feed
Ridiculous	Yes	UPDOWN	I would love to be a of that fruit!
	No	Standard	A beautiful mesa town built into the cliffs.
	No	Engaging	That is a strange cave
	Yes	Human	It must be very dangerous if children play there
	Yes	TransResNet	I hope my kids don't climb on this.
Maternal	Yes	UPDOWN	I hope this is a beautiful place.
	No	Standard	Hockey players competing for control of the hockey puck.
	No	Engaging	Great save, goalie!!
	Yes	Human	Hockey is a little too barbaric for my taste.
	Yes	TransResNet	Hockey players gracefully skate across the ice.
Sophisticated	Yes	UPDOWN	This hockey is like they are a great of the game.
	No	Standard	two people walking through a snowy forest.
	No	Engaging	Too cold for me.
	Yes	Human	I wonder what's at the finish line for these guys?
	Yes	TransResNet	I wonder why they are running.
Curious	Yes	UPDOWN	I wonder what they are a?
	No	Standard	Hollywood Tower at Night
	No	Engaging	I went to that theme park, but was too scared to get on that ride!
	Yes	Human	I am so excited to be here!
	Yes	TransResNet	I remember going to disney world, it was one of the best trips I've ever done.
Happy	Yes	UPDOWN	This looks like a beautiful view!

Typical VQA Models

Image Embedding (VGGNet)



Neural Network Softmax over top K answers



Question Embedding (LSTM)

"How many horses are in this image?"