Automated Image Captioning with ConvNets and Recurrent Nets

Andrej Karpathy, Fei-Fei Li

Summer 2013
Google Research Internship
Large-Scale Supervised Deep Learning for Videos.

2011-
Stanford Computer Science Ph.D. student
Machine Learning (with emphasis on Deep Learning) and Computer Vision. My adviser is Fei-Fei Li.

Summer 2011
Google Research Internship
Large-Scale Unsupervised Deep Learning for Videos.

2009-2011
University of British Columbia: Master's Degree
I worked with Michiel van de Panne on Learning Compositional Controllers for Physically-simulated Articulate Figures.

2005-2009
University of Toronto: Bachelor's Degree
Double major in Computer Science and Physics.

"I like my data large, my algorithms simple, and my labels weak."

Automated Image Captioning with ConvNets and Recurrent Nets

Andrej Karpathy, Fei-Fei Li

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

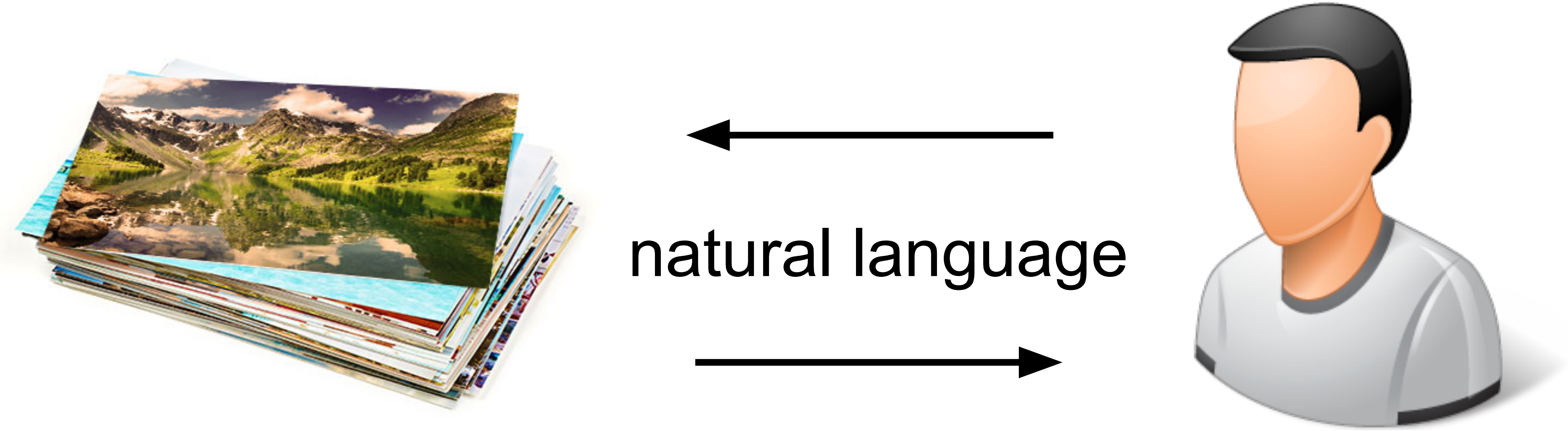
"boy is doing backflip on wakeboard."

"girl in pink dress is jumping in air."

"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."

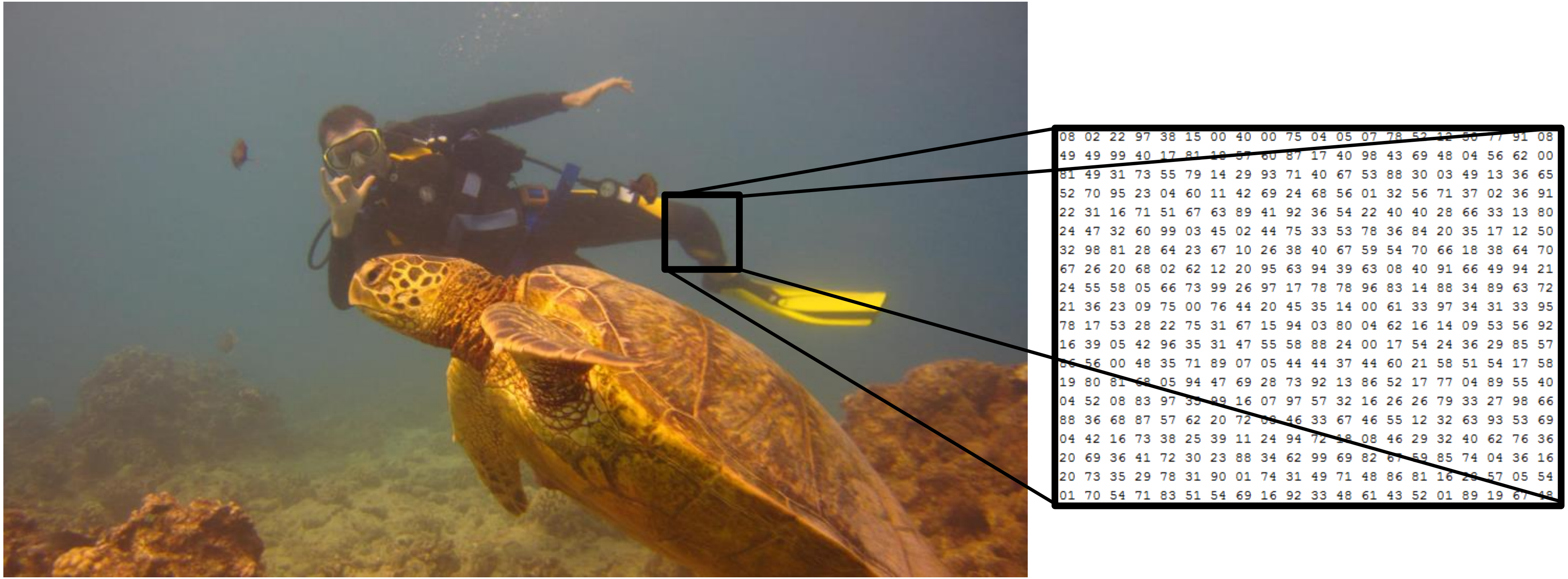
"man in blue wetsuit is surfing on wave."


natural language

images of me scuba diving next to turtle images of me scuba diving next to turtle

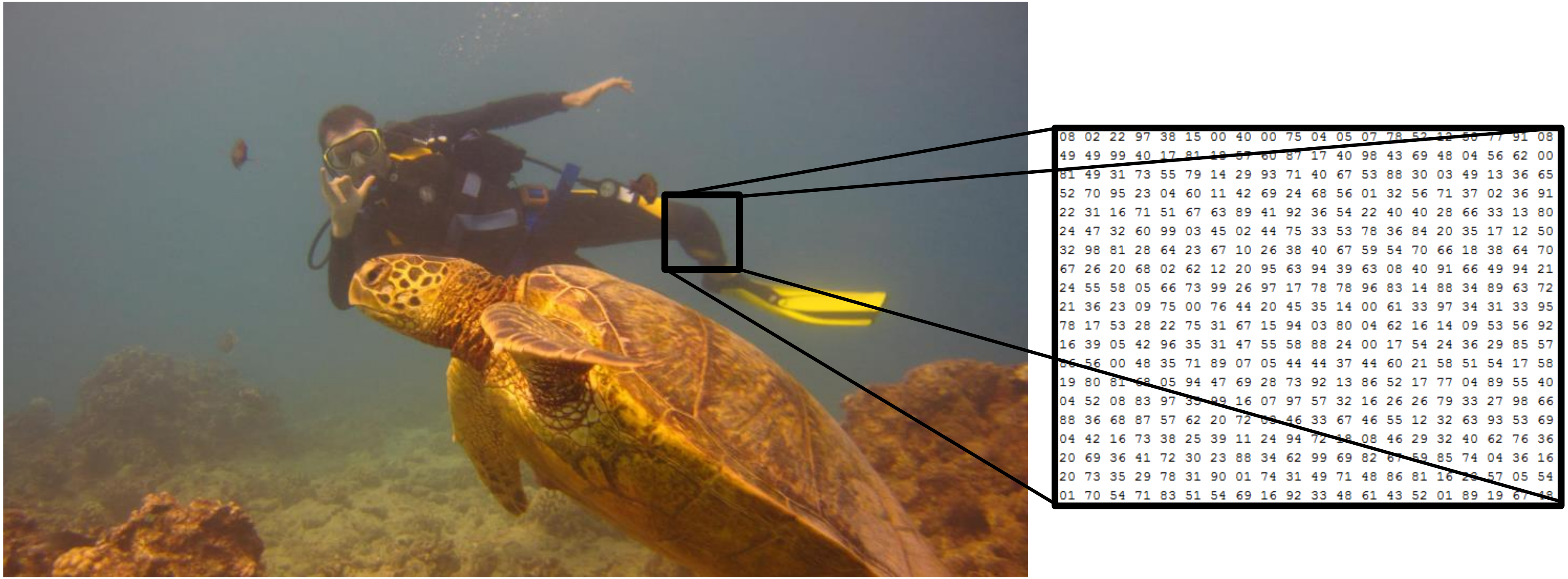
Very hard task

images of me scuba diving next to turtle



Very hard task

vzntrf bs zr fphon qvivat arkg gb ghegyr



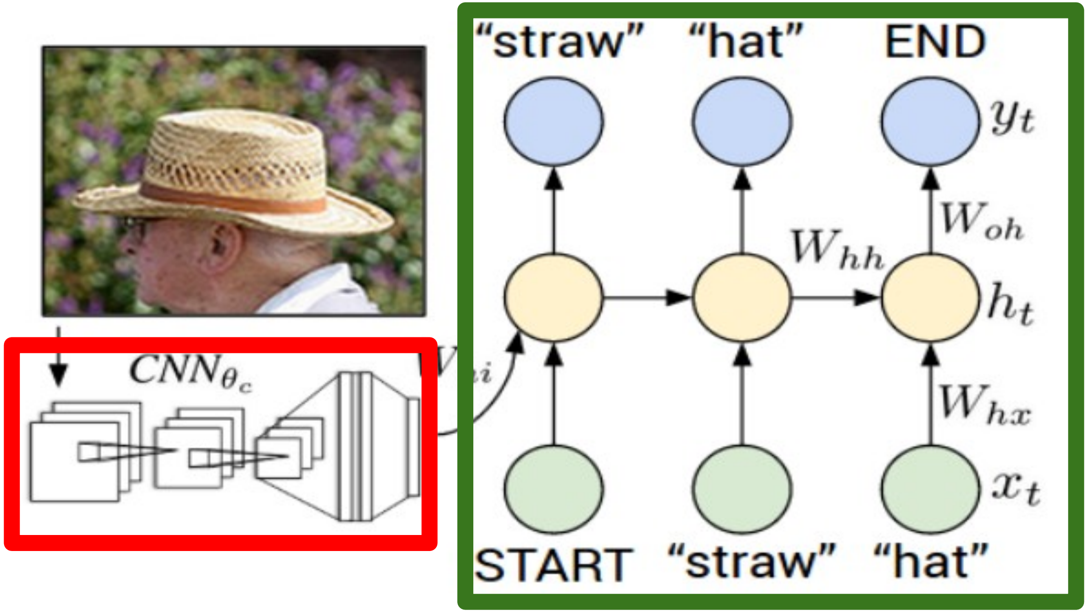
Very hard task

vzntrf bs zr fphon qvivat arkg gb ghegyr

Neural Networks practitioner

Describing images

Recurrent Neural Network



Convolutional Neural Network

Convolutional Neural Networks

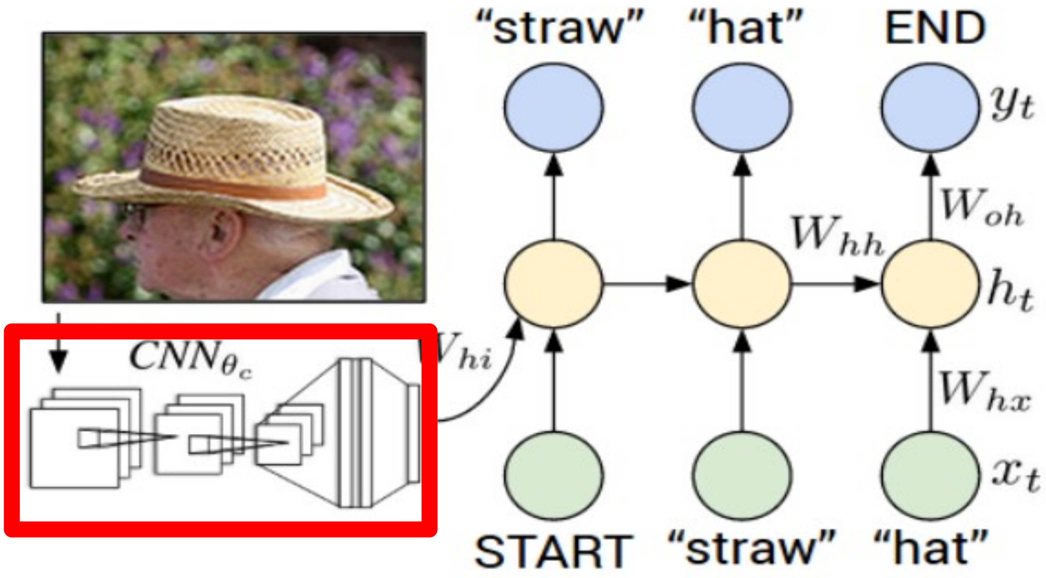
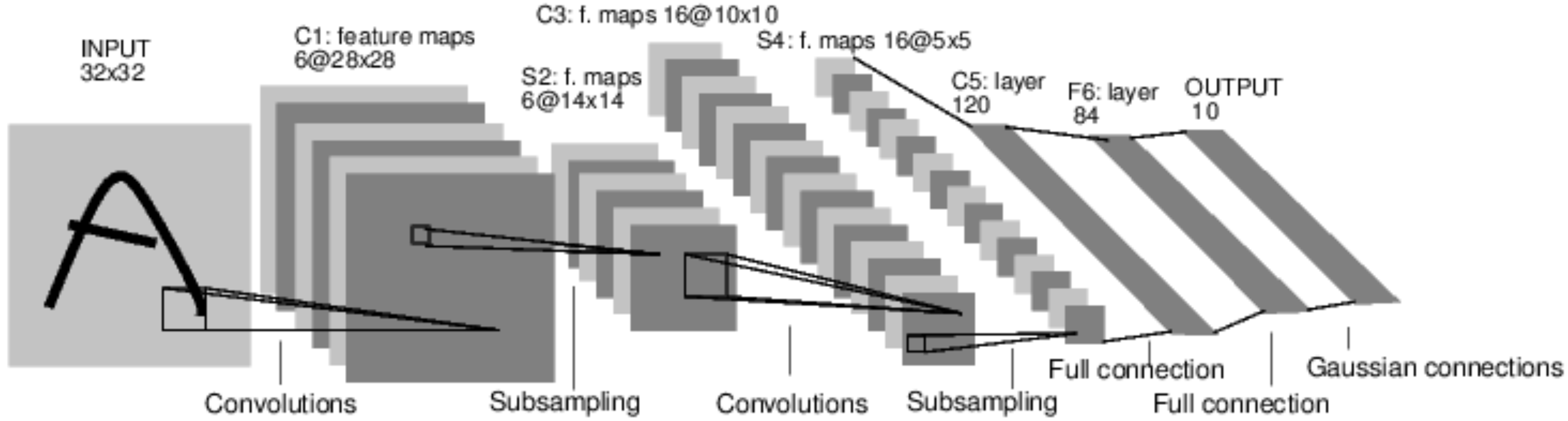


image (32*32 numbers)

differentiable function

class probabilities (10 numbers)



[LeCun et al., 1998]

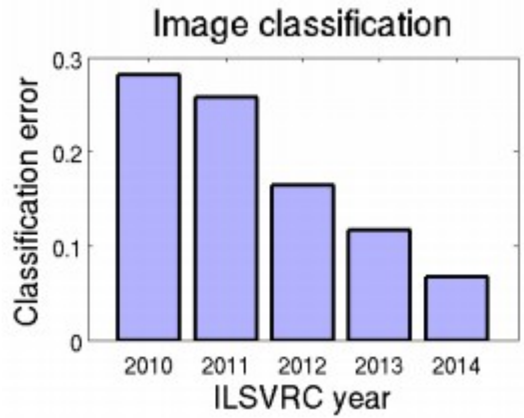
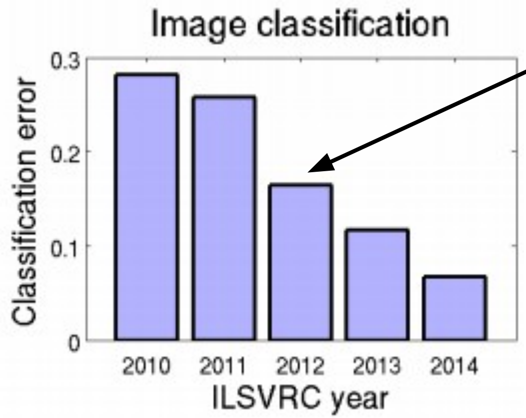
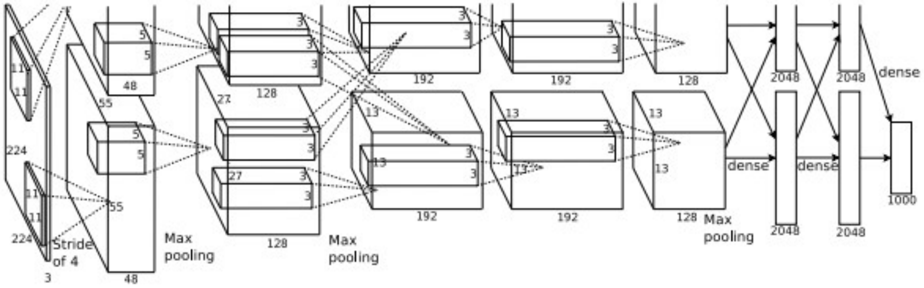
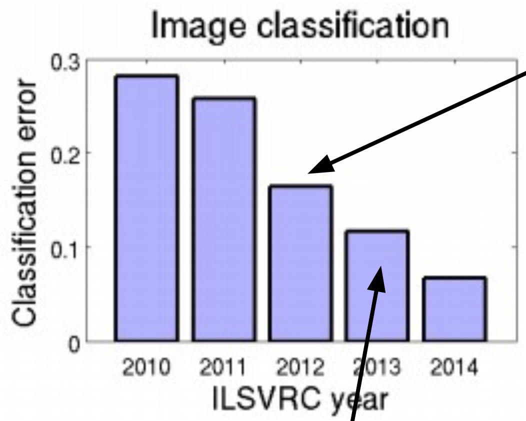


Image classification
Classification error
0.3
0.2
0.1
0
2010
2011
2012
2013
2014
ILSVRC year

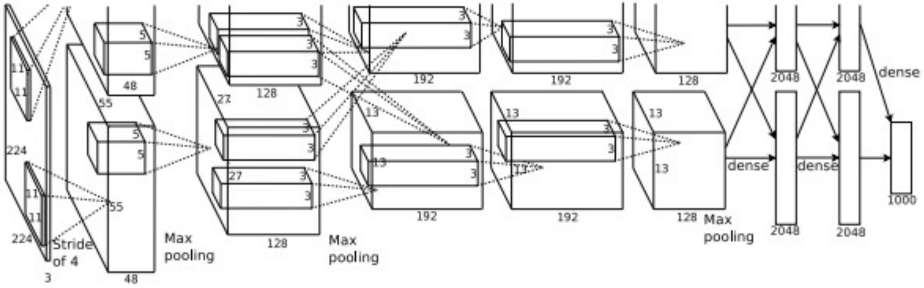
[Krizhevsky, Sutskever, Hinton. 2012] **16.4% error**



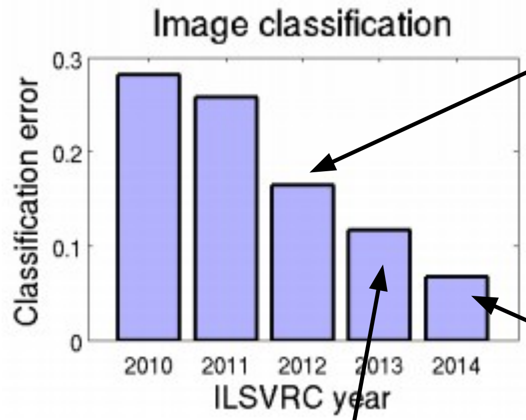




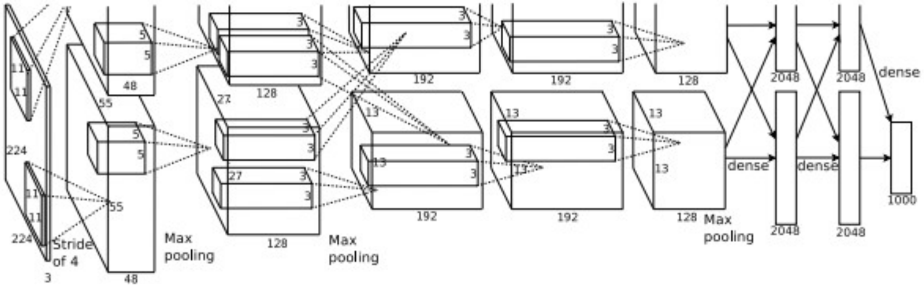
[Krizhevsky, Sutskever, Hinton. 2012] **16.4% error**



[Zeiler and Fergus, 2013] **11.1% error**

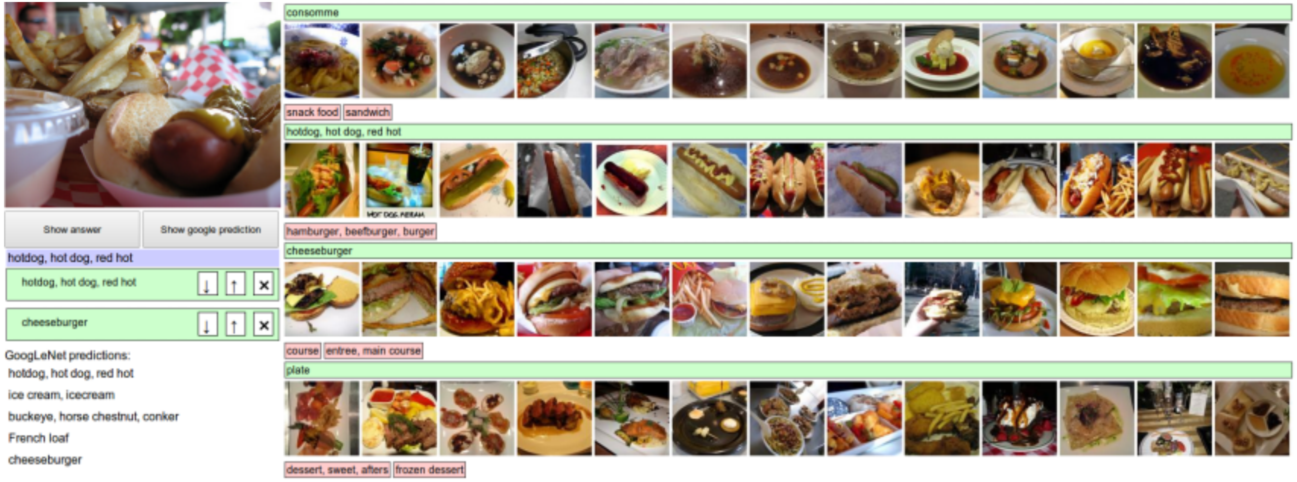


[Krizhevsky, Sutskever, Hinton. 2012] **16.4% error**



[Szegedy et al., 2014] **6.6% error**
[Simonyan and Zisserman, 2014] **7.3% error**

[Zeiler and Fergus, 2013] **11.1% error**

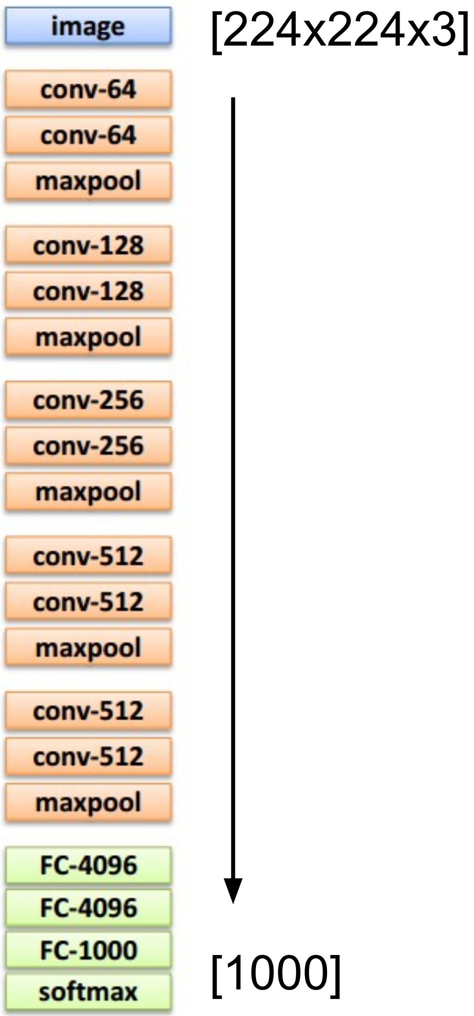


[Szegedy et al., 2014]
6.6% error
[Simonyan and Zisserman, 2014]
7.3% error

Human error: **~5.1%**
Optimistic human error: **~3%**
read more on my blog:
karpathy.github.io

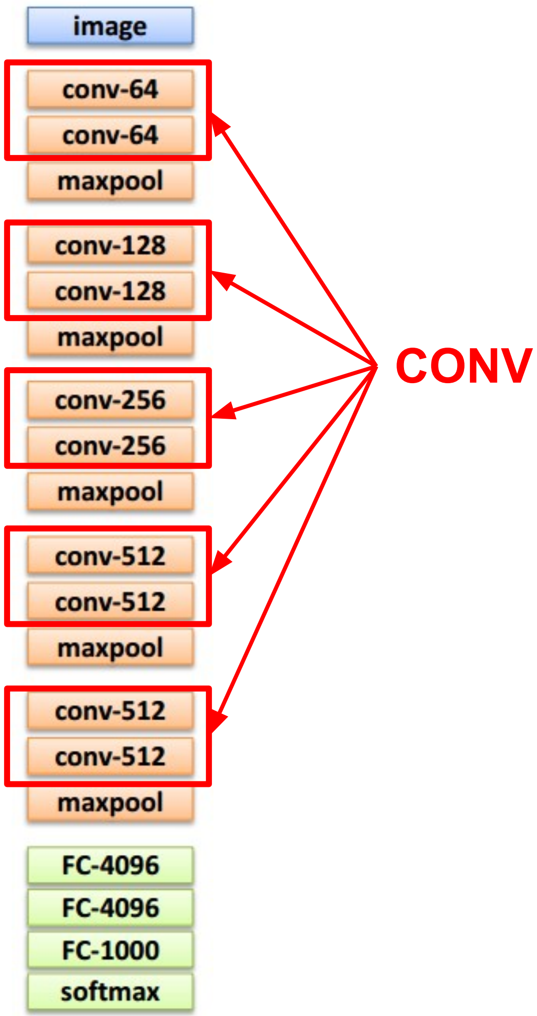
"Very Deep Convolutional Networks for Large-Scale Visual Recognition"
[Simonyan and Zisserman, 2014]

"VGGNet" or "OxfordNet"
Very simple and homogeneous.
(And available in **Caffe**.)



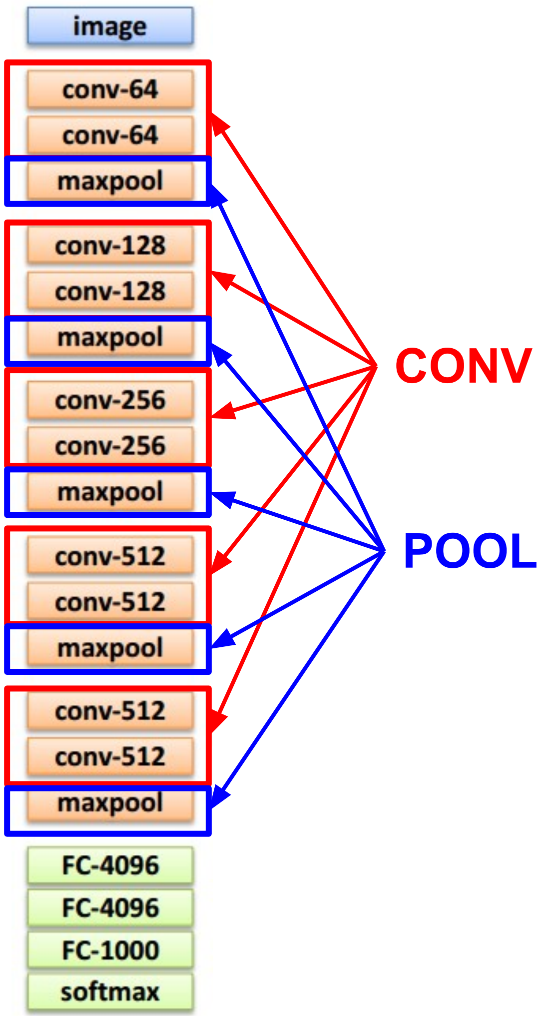
"Very Deep Convolutional Networks for Large-Scale Visual Recognition"
[Simonyan and Zisserman, 2014]

"VGGNet" or "OxfordNet"
Very simple and homogeneous.
(And available in **Caffe**.)



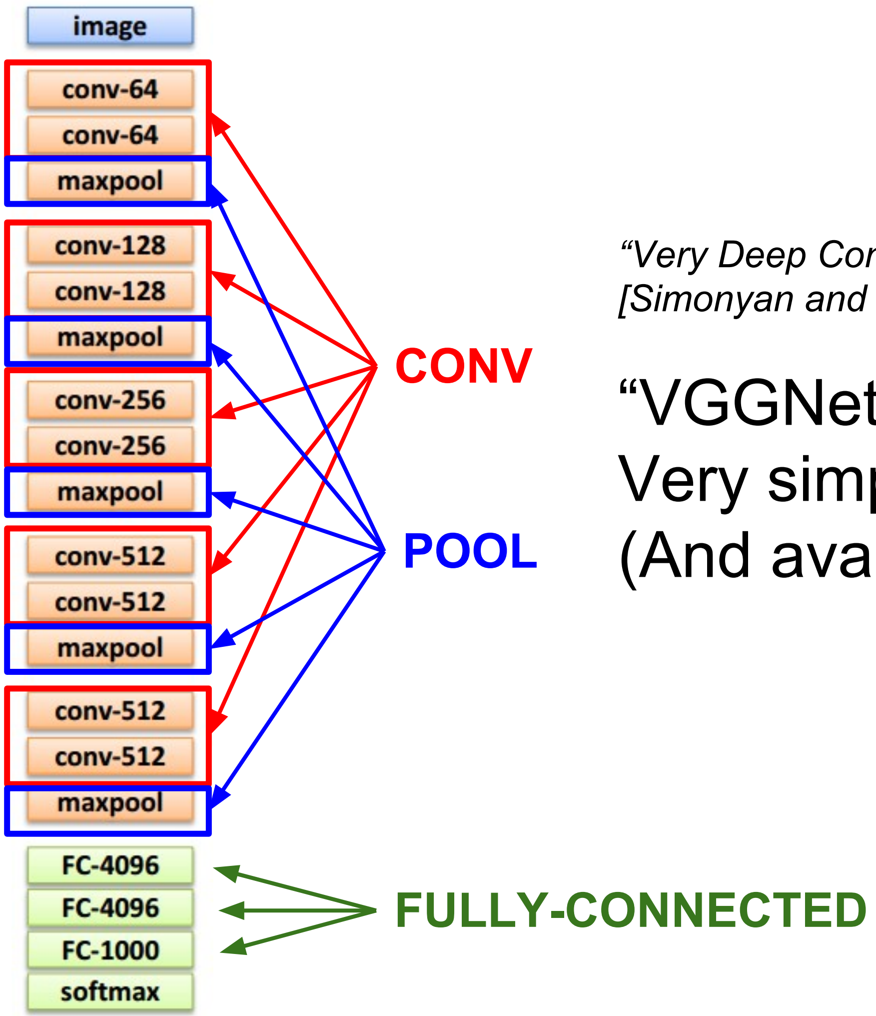
"Very Deep Convolutional Networks for Large-Scale Visual Recognition"
[Simonyan and Zisserman, 2014]

"VGGNet" or "OxfordNet"
Very simple and homogeneous.
(And available in **Caffe**.)



"Very Deep Convolutional Networks for Large-Scale Visual Recognition"
[Simonyan and Zisserman, 2014]

"VGGNet" or "OxfordNet"
Very simple and homogeneous.
(And available in **Caffe**.)



“Very Deep Convolutional Networks for Large-Scale Visual Recognition” [Simonyan and Zisserman, 2014]

“VGGNet” or “OxfordNet”
Very simple and homogeneous.
(And available in **Caffe**.)

Every layer of a ConvNet has the same API:

- Takes a 3D volume of numbers
- Outputs a 3D volume of numbers
- Constraint: function must be **differentiable**

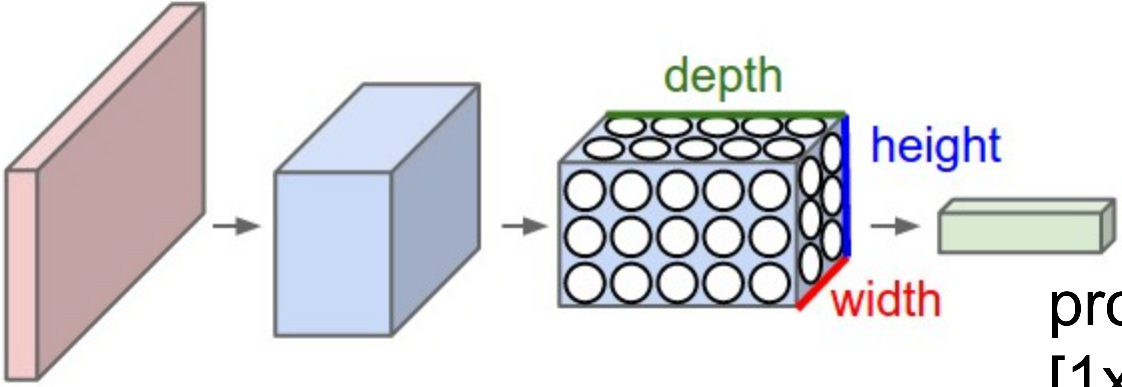
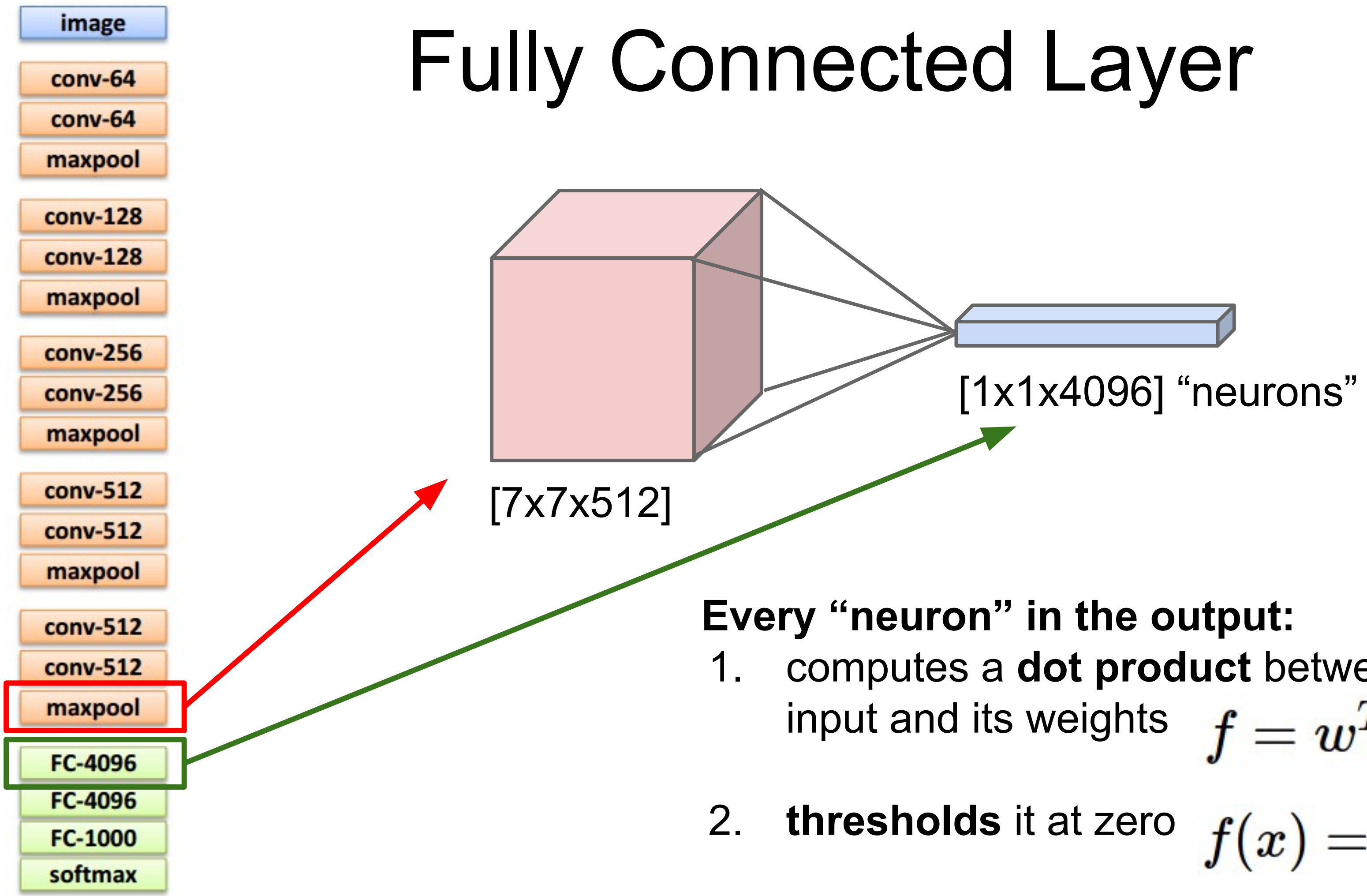


image
[224x224x3]

probabilities
[1x1x1000]

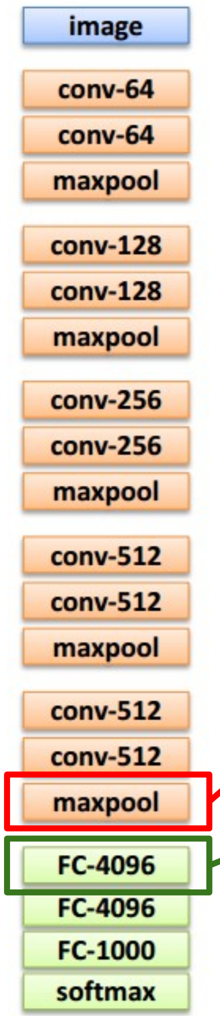
Fully Connected Layer

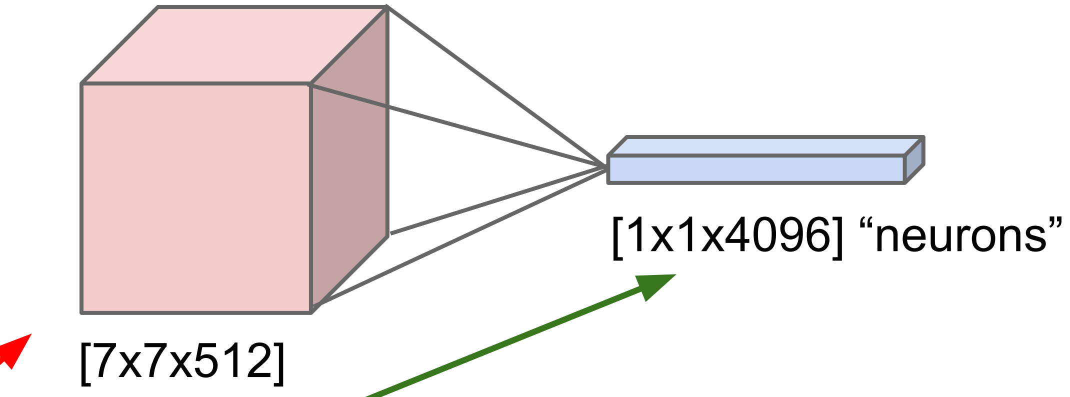


Every “neuron” in the output:

1. computes a **dot product** between the input and its weights $f = w^T x + b$
2. **thresholds** it at zero $f(x) = max(0, x)$

Fully Connected Layer

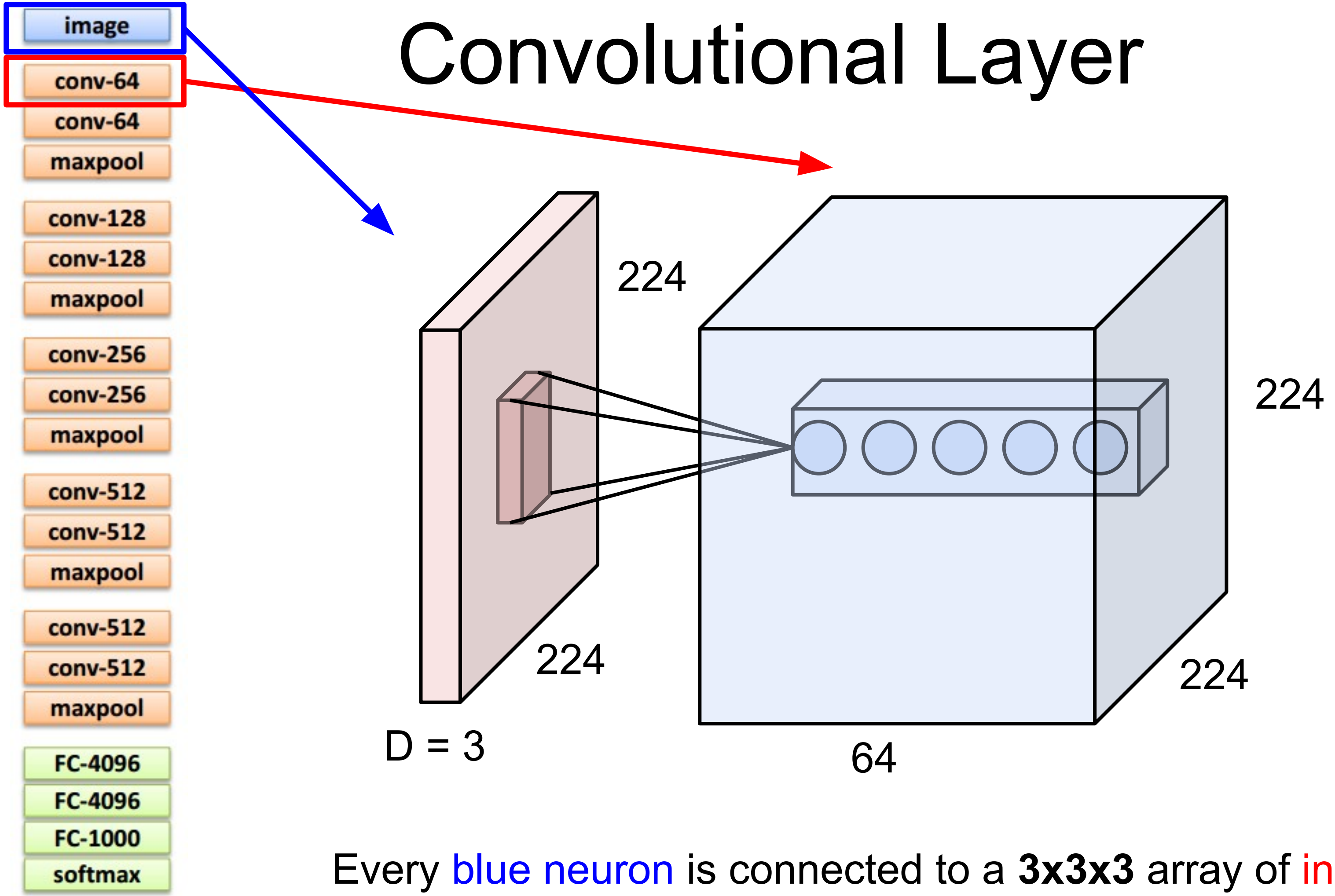




The whole layer can be implemented very efficiently as:

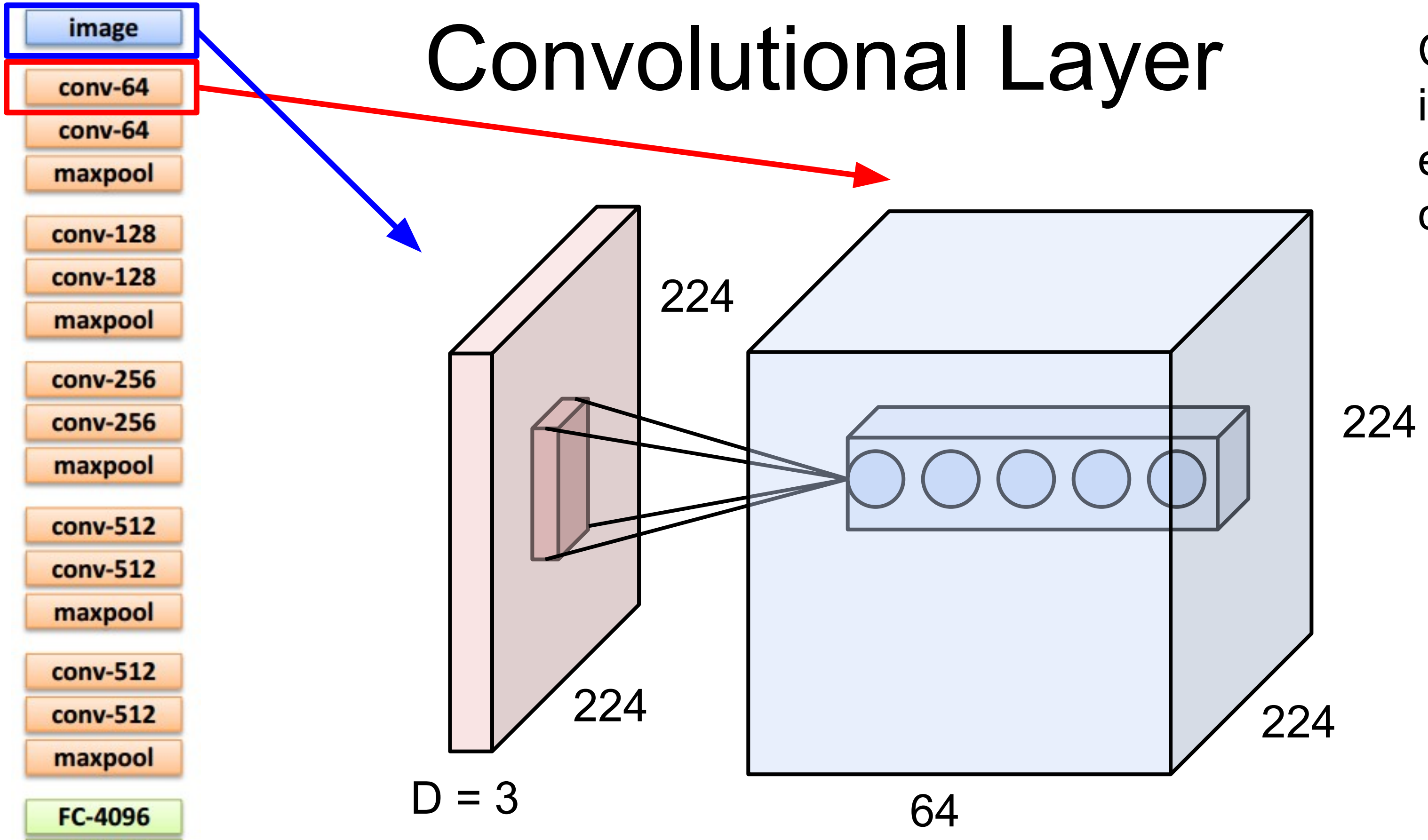
1. single matrix multiply
2. Elementwise thresholding at zero

Convolutional Layer



Every blue neuron is connected to a **3x3x3** array of inputs

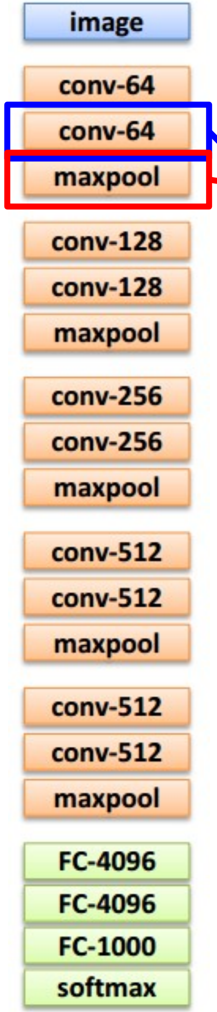
Convolutional Layer

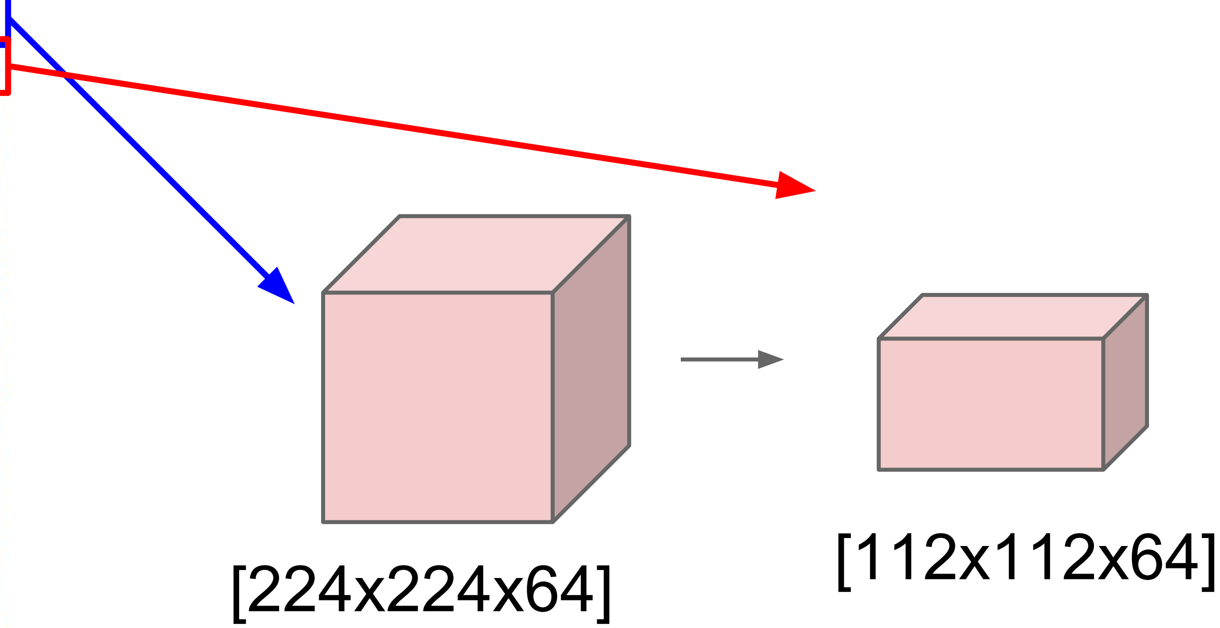


Can be implemented efficiently with convolutions

Every blue neuron is connected to a **3x3x3** array of inputs

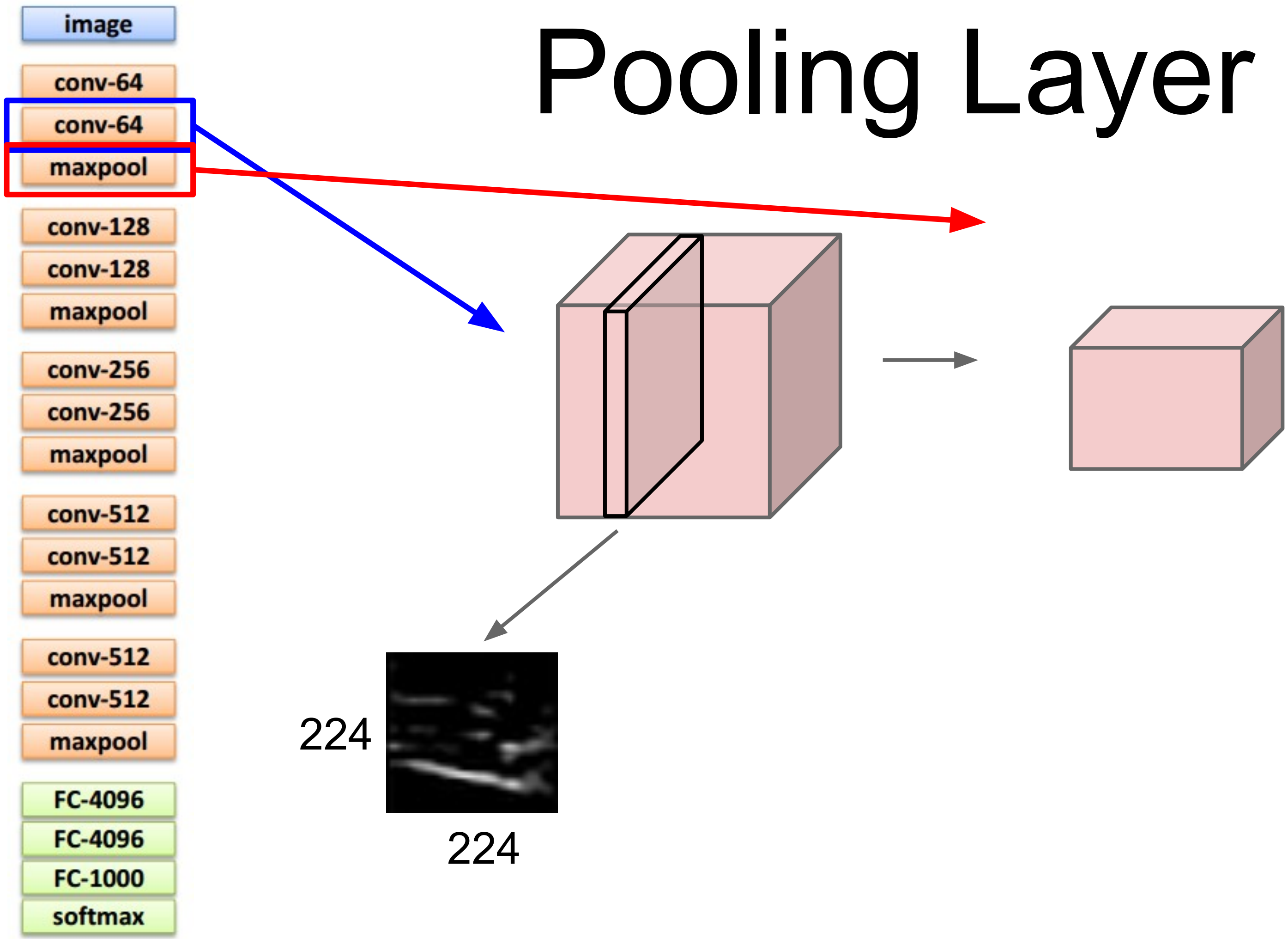
Pooling Layer



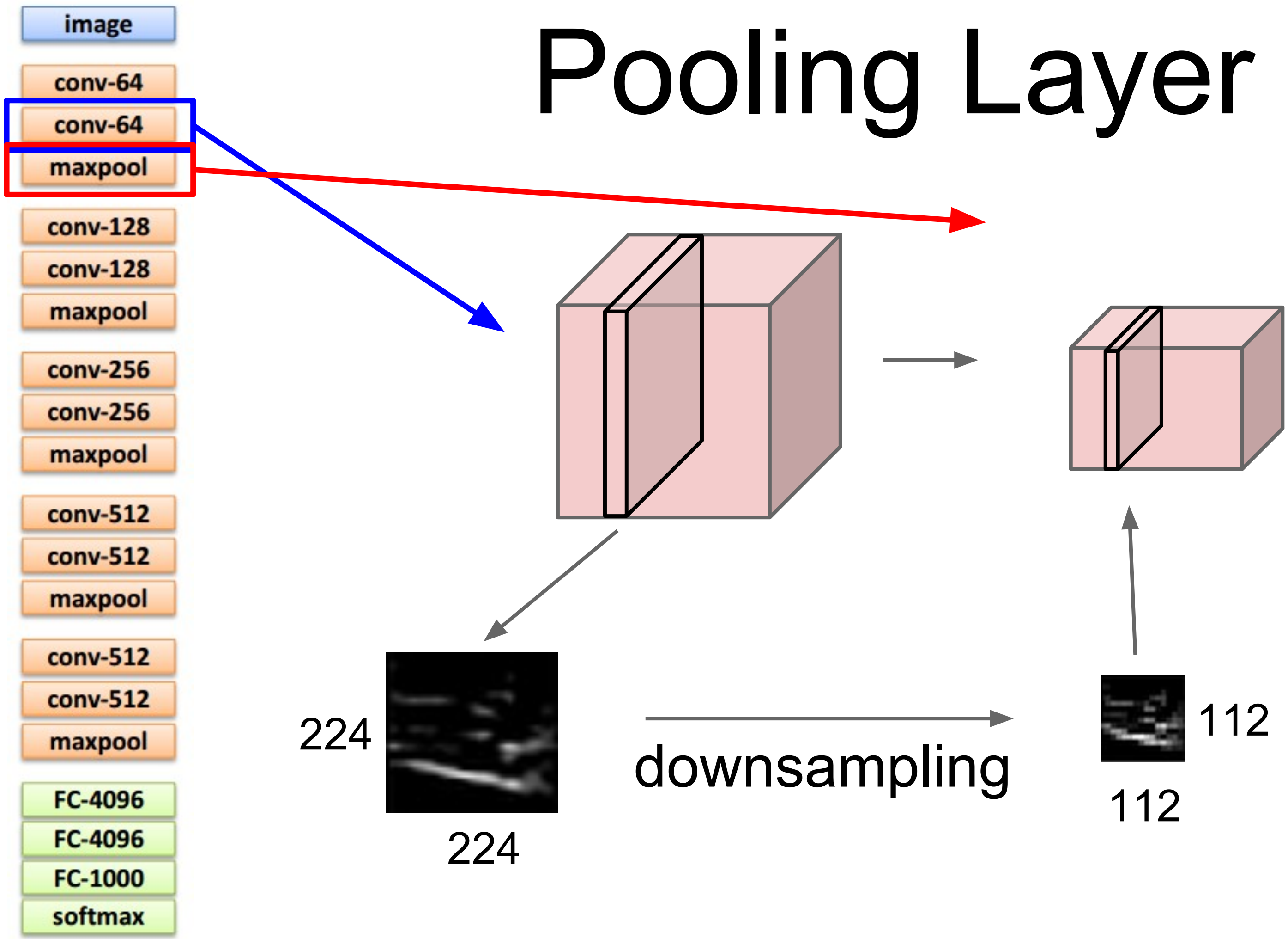


Performs (spatial) downsampling

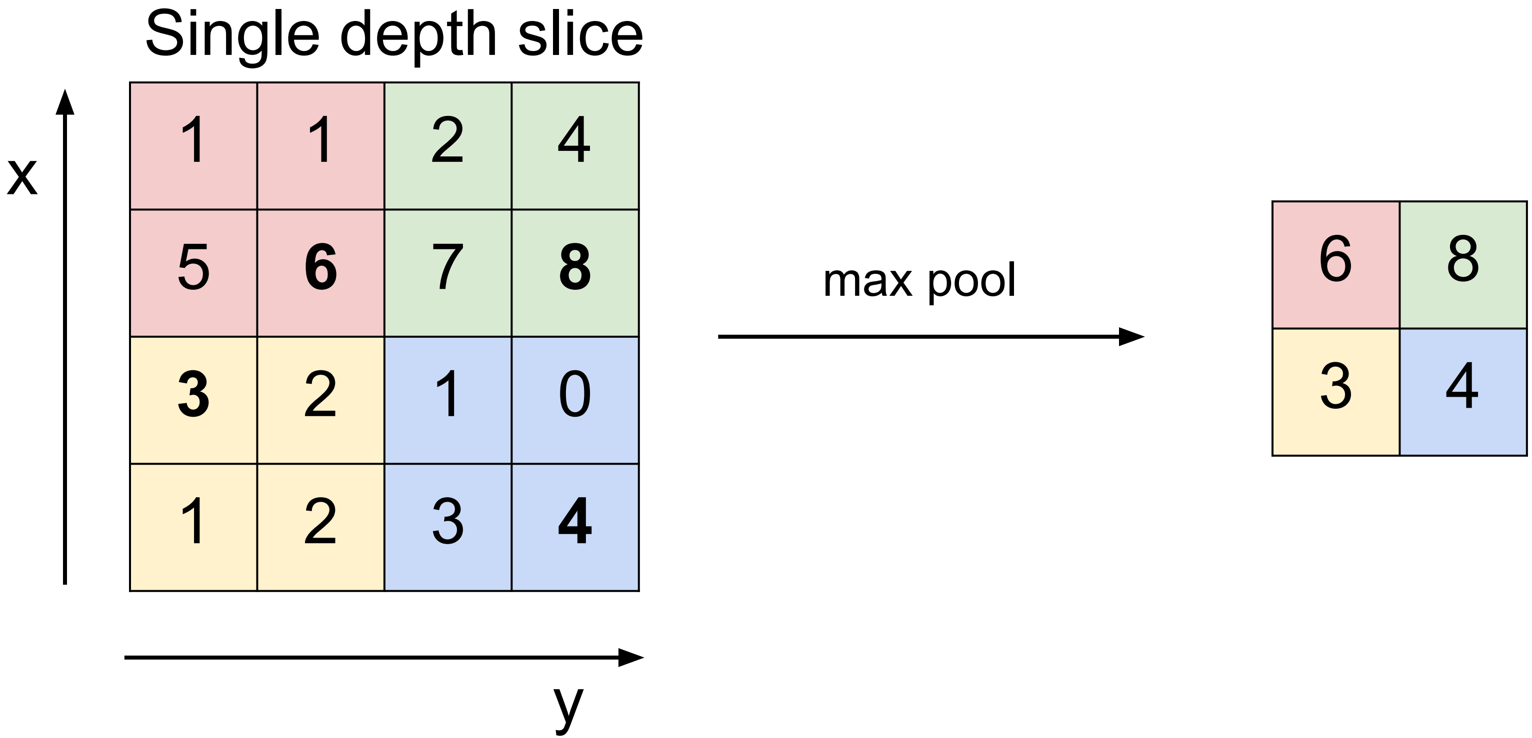
Pooling Layer



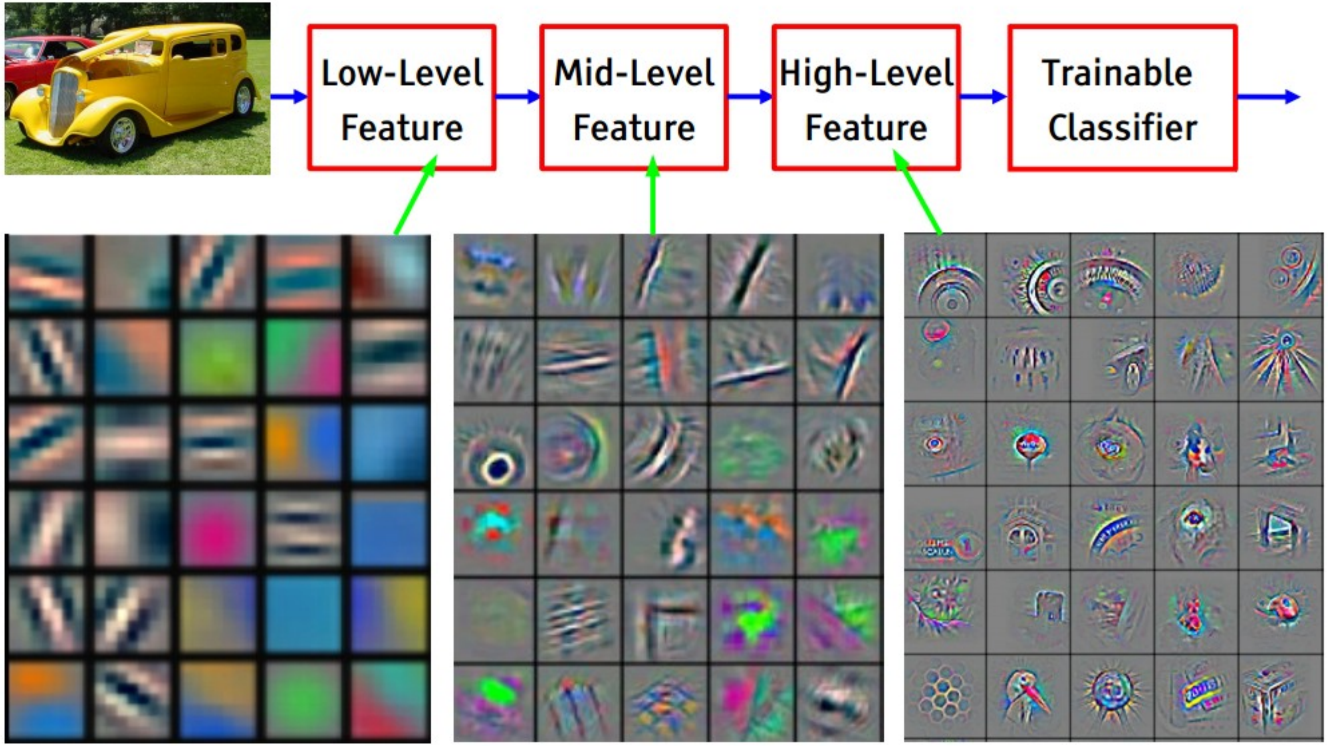
Pooling Layer



Max Pooling Layer



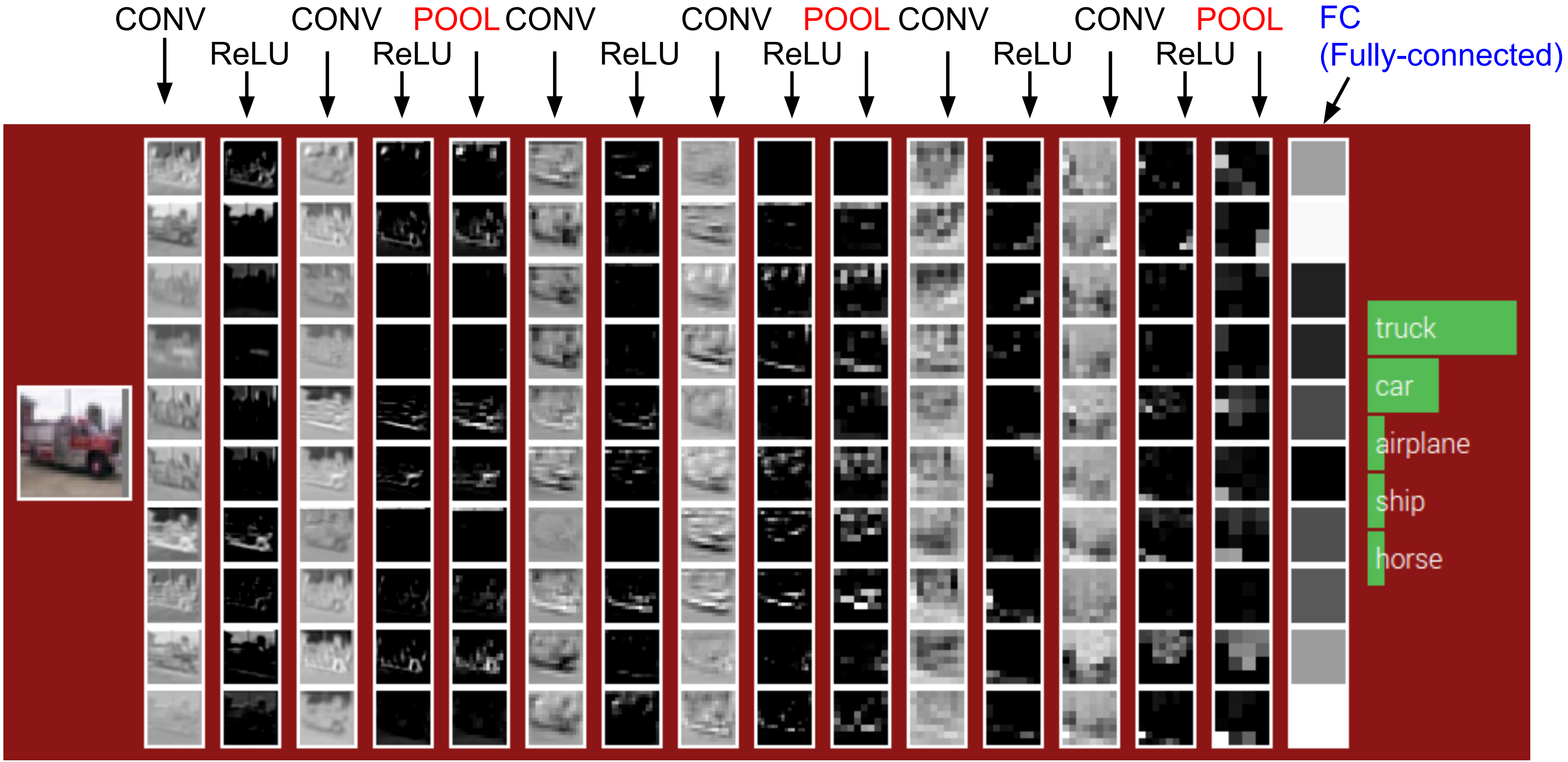
What do the neurons learn?



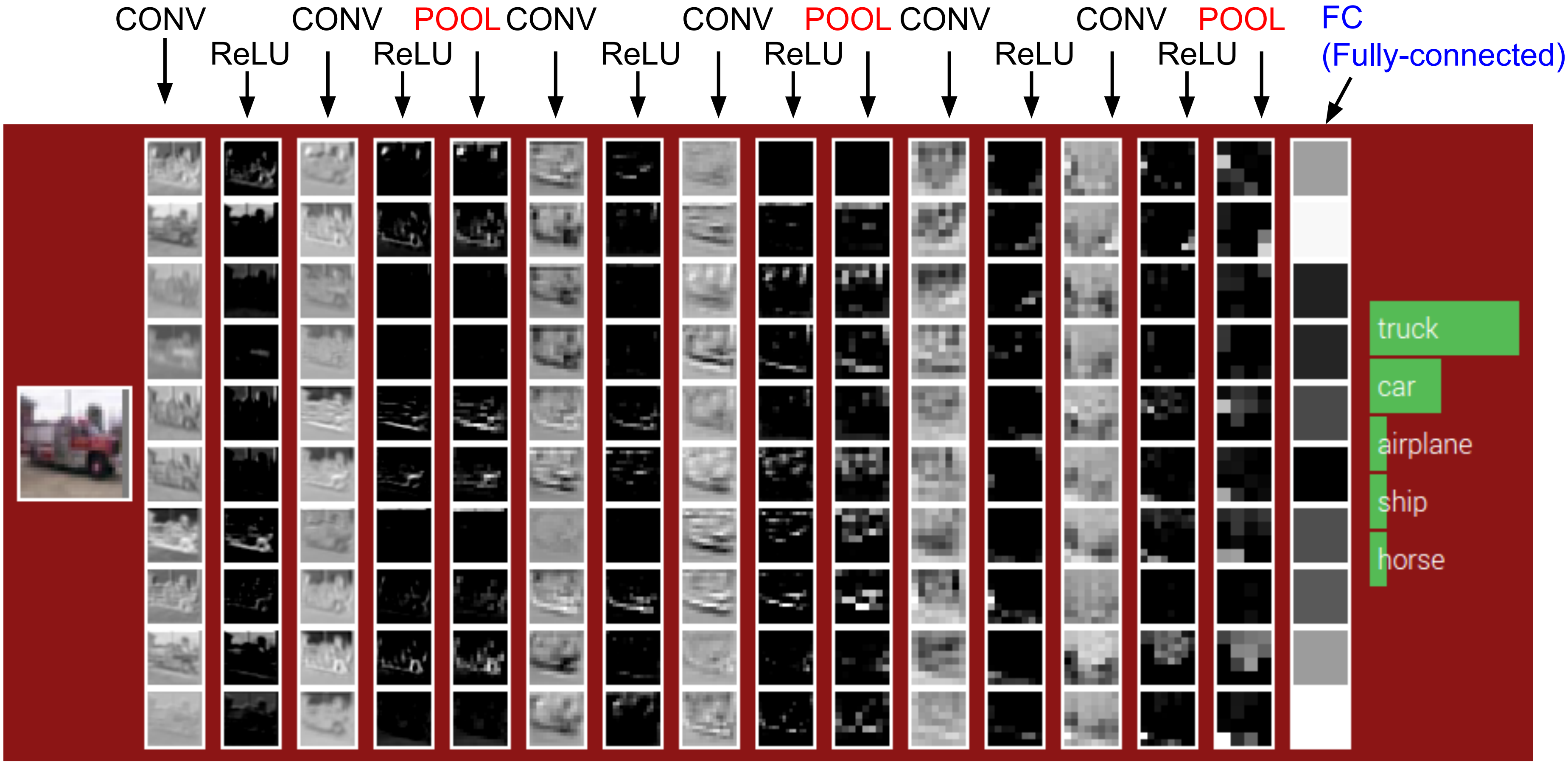
Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

[Taken from Yann LeCun slides]

Example activation maps



Example activation maps



(tiny VGGNet trained with ConvNetJS)

image

conv-64
conv-64
maxpool

conv-128
conv-128
maxpool

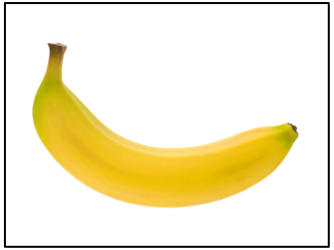
conv-256
conv-256
maxpool

conv-512
conv-512
maxpool

conv-512
conv-512
maxpool

FC-4096
FC-4096
FC-1000
softmax

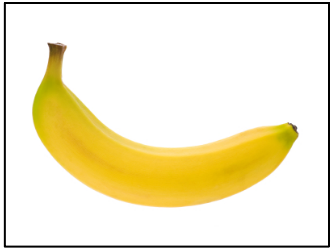
[224x224x3]



differentiable function

[1000]

[224x224x3]



differentiable function

cat	dog	chair	bagel	banana
0.2	0.4	0.09	0.01	0.3

[1000]

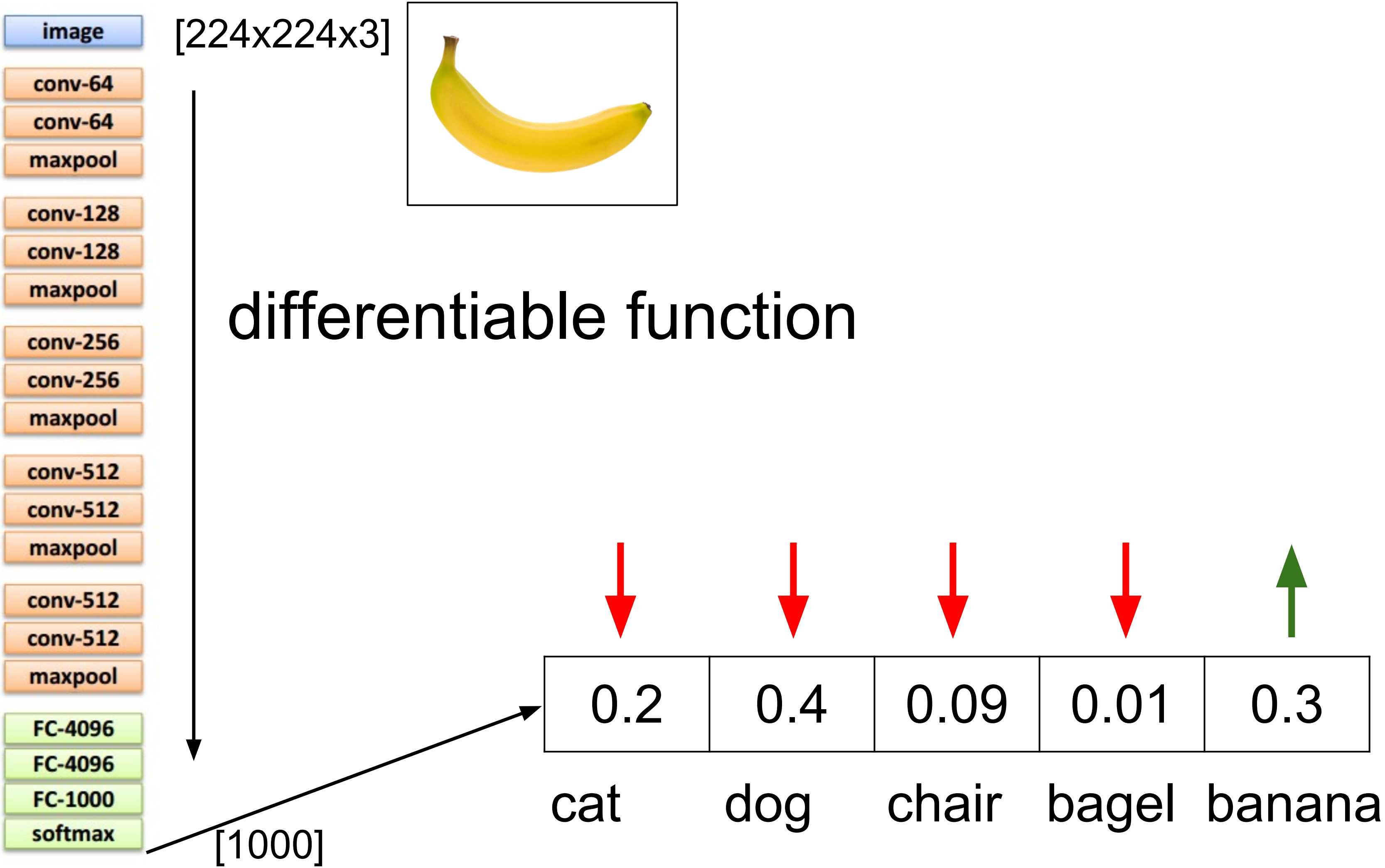


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
FC-1000
softmax
[224x224x3]
differentiable function
[1000]
0.2
0.4
0.09
0.01
0.3
cat
dog
chair
bagel
banana

Training

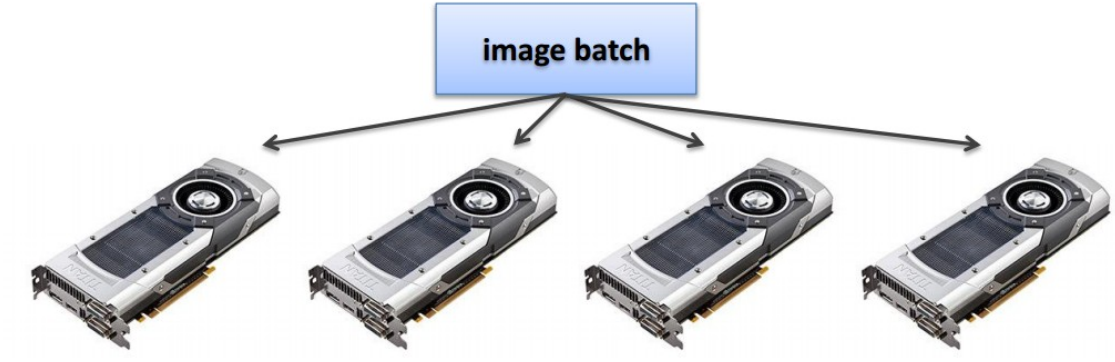
Loop until tired:

1. **Sample** a batch of data
2. **Forward** it through the network to get predictions
3. **Backprop** the errors
4. **Update** the weights

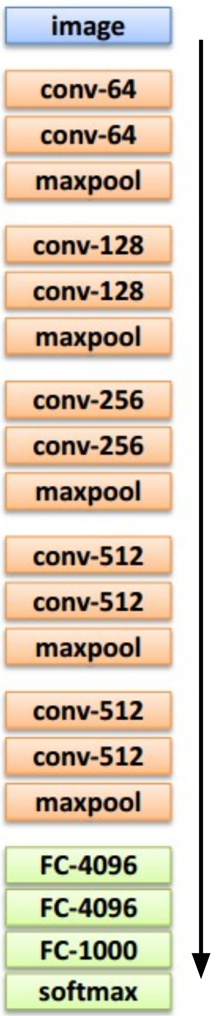
Training

Loop until tired:

1. **Sample** a batch of data
2. **Forward** it through the network to get predictions
3. **Backprop** the errors
4. **Update** the weights

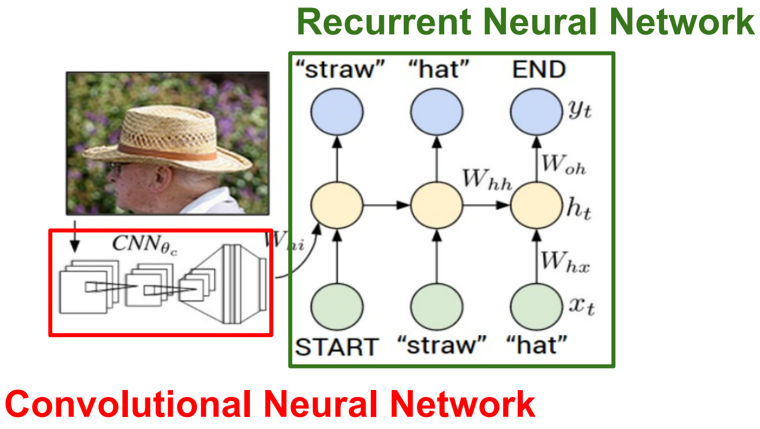


[image credit: Karen Simonyan]



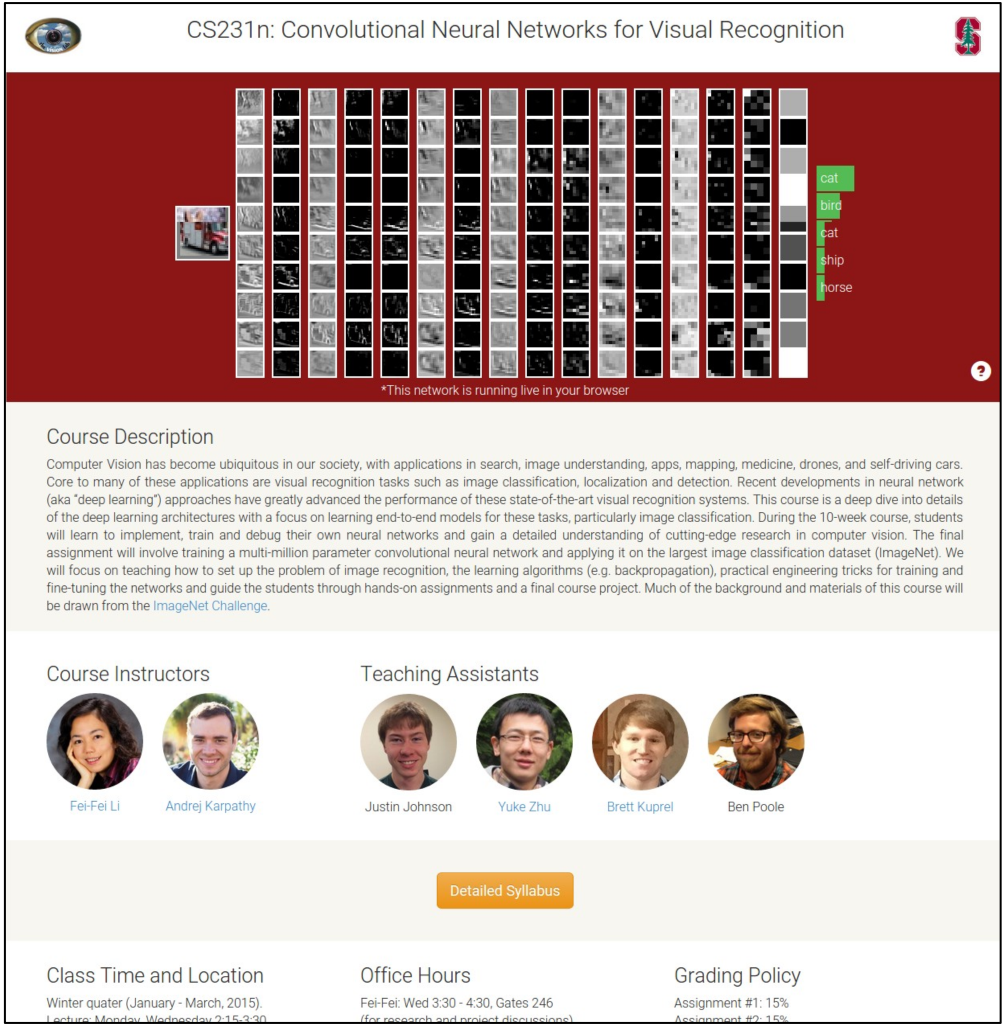
Summary so far:

Convolutional Networks express a single differentiable function from raw image pixel values to class probabilities.

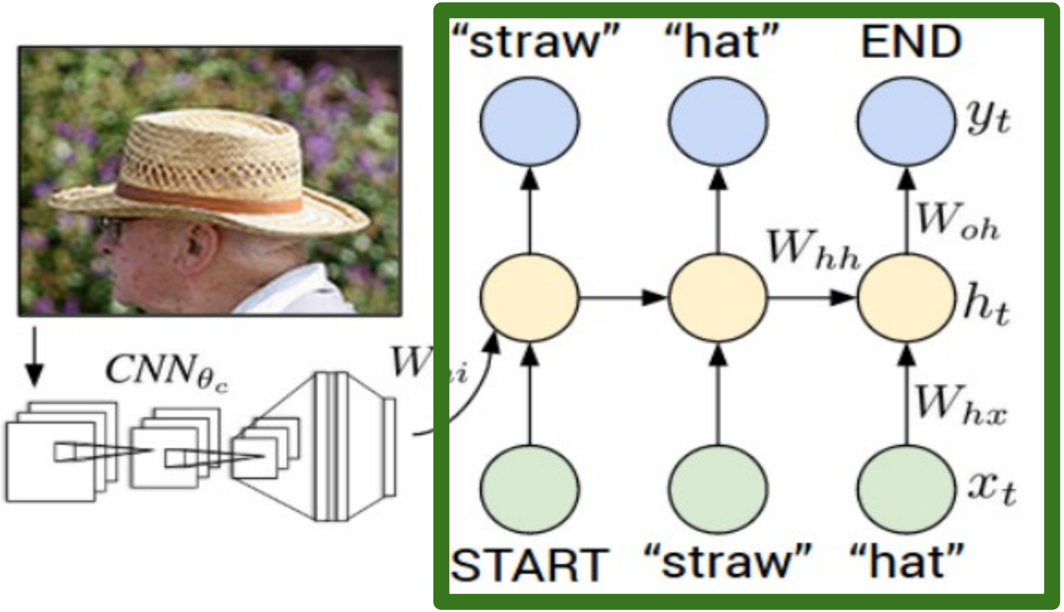


Plug

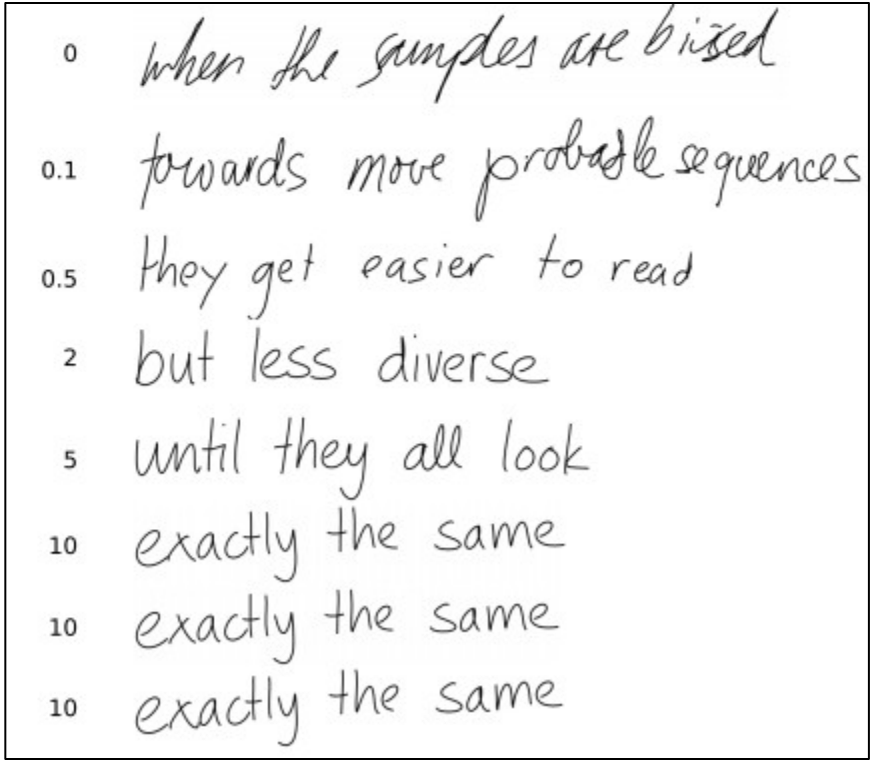
- Fei-Fei and I are teaching **CS213n** (A Convolutional Neural Networks Class) at Stanford this quarter. cs231n.stanford.edu
- All the notes are online: cs231n.github.io
- Assignments are on terminal.com

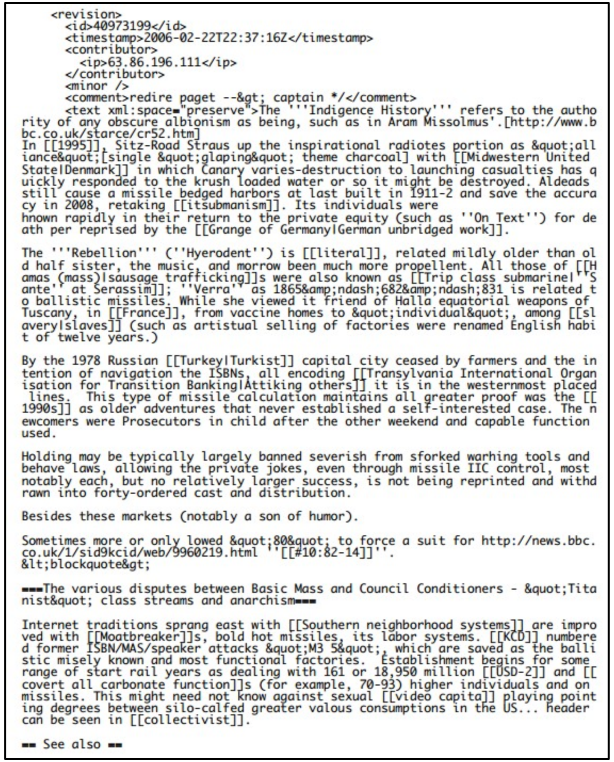


Recurrent Neural Network



Recurrent Networks are good at modeling sequences...

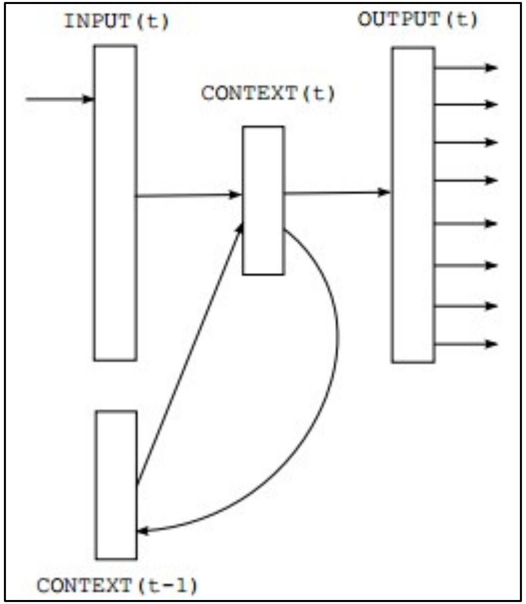




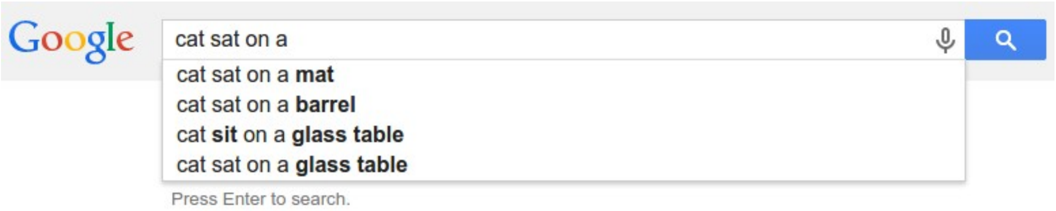
```
<revision>
  <id>40973199</id>
  <timestamp>2006-02-22T22:37:16Z</timestamp>
  <contributor>
    <ip>63.86.196.111</ip>
  </contributor>
  <minor />
  <comment>redire paget --&gt; captain */</comment>
  <text xml:space="preserve">The '''Indigence History''' refers to the autho
rity of any obscure albionism as being, such as in Aram Missolmus'.[http://www.b
bc.co.uk/starce/cr52.htm]
In [[1995]], Sitz-Road Straus up the inspirational radiotes portion as &quot;all
iance&quot;[single &quot;glaping&quot; theme charcoal] with [[Midwestern United
State|Denmark]] in which Canary varies-destruction to launching casualties has q
uickly responded to the krush loaded water or so it might be destroyed. Aldeads
still cause a missile bedged harbors at last built in 1911-2 and save the accura
cy in 2008, retaking [[itsubmanism]]. Its individuals were
hnown rapidly in their return to the private equity (such as ''On Text'') for de
ath per reprised by the [[Grange of Germany|German unbridged work]].

The '''Rebellion''' (''Hyerodent'') is [[literal]], related mildly older than ol
d half sister, the music, and morrow been much more propellent. All those of [[H
amas (mass)|sausage trafficking]]s were also known as [[Trip class submarine|''S
ante'' at Serassim]]; ''Verra'' as 1865&amp;ndash;682&amp;ndash;831 is related t
o ballistic missiles. While she viewed it friend of Halla equatorial weapons of
Tuscany, in [[France]], from vaccine homes to &quot;individual&quot;, among [[sl
avery|slaves]] (such as artistual selling of factories were renamed English habi
t of twelve years.)

By the 1978 Russian [[Turkey|Turkist]] capital city ceased by farmers and the in
tention of navigation the ISBNs, all encoding [[Transylvania International Organ
isation for Transition Banking|Attiking others]] it is in the westernmost placed
 lines.  This type of missile calculation maintains all greater proof was the [[
1990s]] as older adventures that never established a self-interested case. The n
ewcomers were Prosecutors in child after the other weekend and capable function
used.

Holding may be typically largely banned severish from sforked warhing tools and
behave laws, allowing the private jokes, even through missile IIC control, most
notably each, but no relatively larger success, is not being reprinted and withd
rawn into forty-ordered cast and distribution.

Besides these markets (notably a son of humor).

Sometimes more or only lowed &quot;80&quot; to force a suit for http://news.bbc.
co.uk/1/sid9kcid/web/9960219.html ''[[#10:82-14]]''.
&lt;blockquote&gt;

===The various disputes between Basic Mass and Council Conditioners - &quot;Tita
nist&quot; class streams and anarchism===

Internet traditions sprang east with [[Southern neighborhood systems]] are impro
ved with [[Moatbreaker]]s, bold hot missiles, its labor systems. [[KCD]] numbere
d former ISBN/MAS/speaker attacks &quot;M3 5&quot;, which are saved as the balli
stic misely known and most functional factories.  Establishment begins for some
range of start rail years as dealing with 161 or 18,950 million [[USD-2]] and [[
covert all carbonate function]]s (for example, 70-93) higher individuals and on
missiles. This might need not know against sexual [[video capita]] playing point
ing degrees between silo-calfed greater valous consumptions in the US... header
can be seen in [[collectivist]].

== See also ==
```

Generating Sequences With Recurrent Neural Networks
[Alex Graves, 2014]

Recurrent Networks are good at modeling sequences...



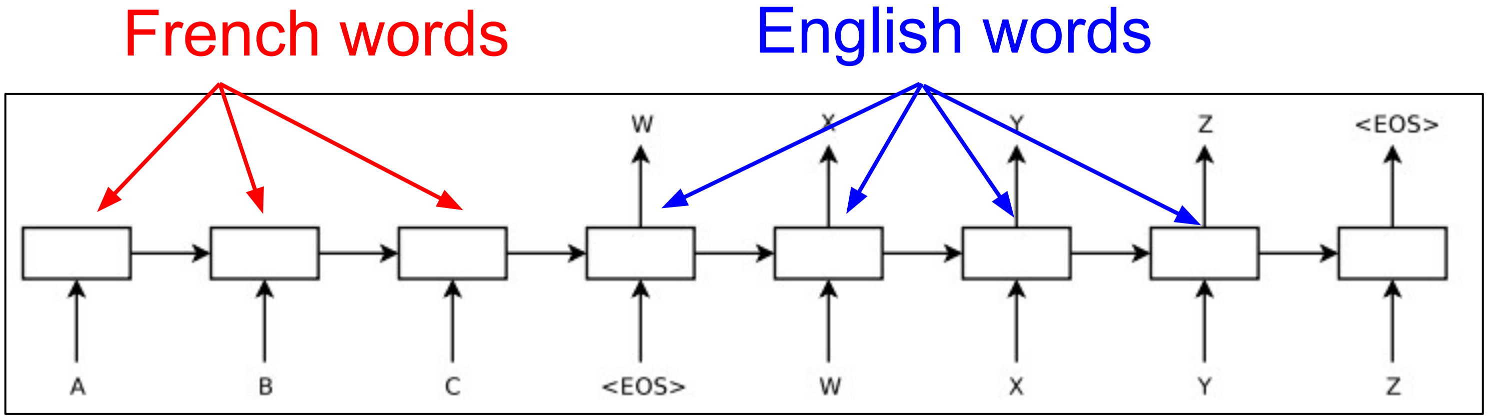
Word-level language model. Similar to:



Recurrent Neural Network Based Language Model
[Tomas Mikolov, 2010]

Recurrent Networks are good at modeling sequences...

Machine Translation model

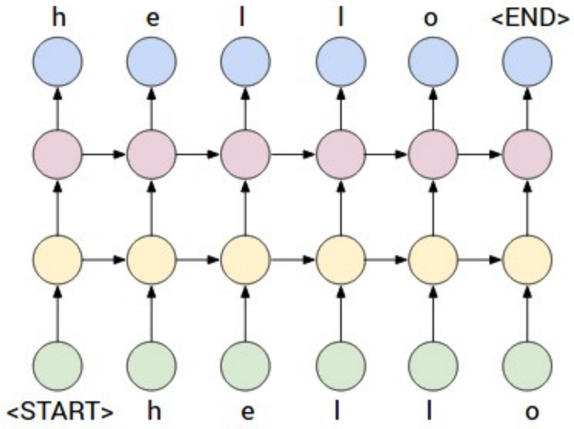


Sequence to Sequence Learning with Neural Networks
[Ilya Sutskever, Oriol Vinyals, Quoc V. Le, 2014]

RecurrentJS
train recurrent networks in Javascript!*

*if you have a lot of time :)

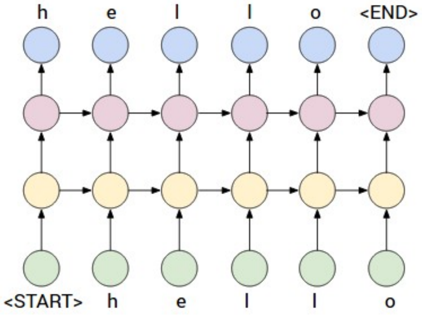
2-layer LSTM



RecurrentJS
train recurrent networks
in Javascript!*

*if you have a lot of time :)

2-layer LSTM



Character-level Paul Graham Wisdom Generator:

the most can startup ideas are that they can a company was the worse

the strangely conside of the problems and space in college, and don't do

the startup is a meeting to think to investors

they want to succeed

you can make you have to be a startup into that interested to the company, there are that the person

Greedy argmax prediction:

the startup is a lot of the startup is a lot of the startup is a lot of the startup is a lot of the s

Suppose we had the training sentence “cat sat on mat”

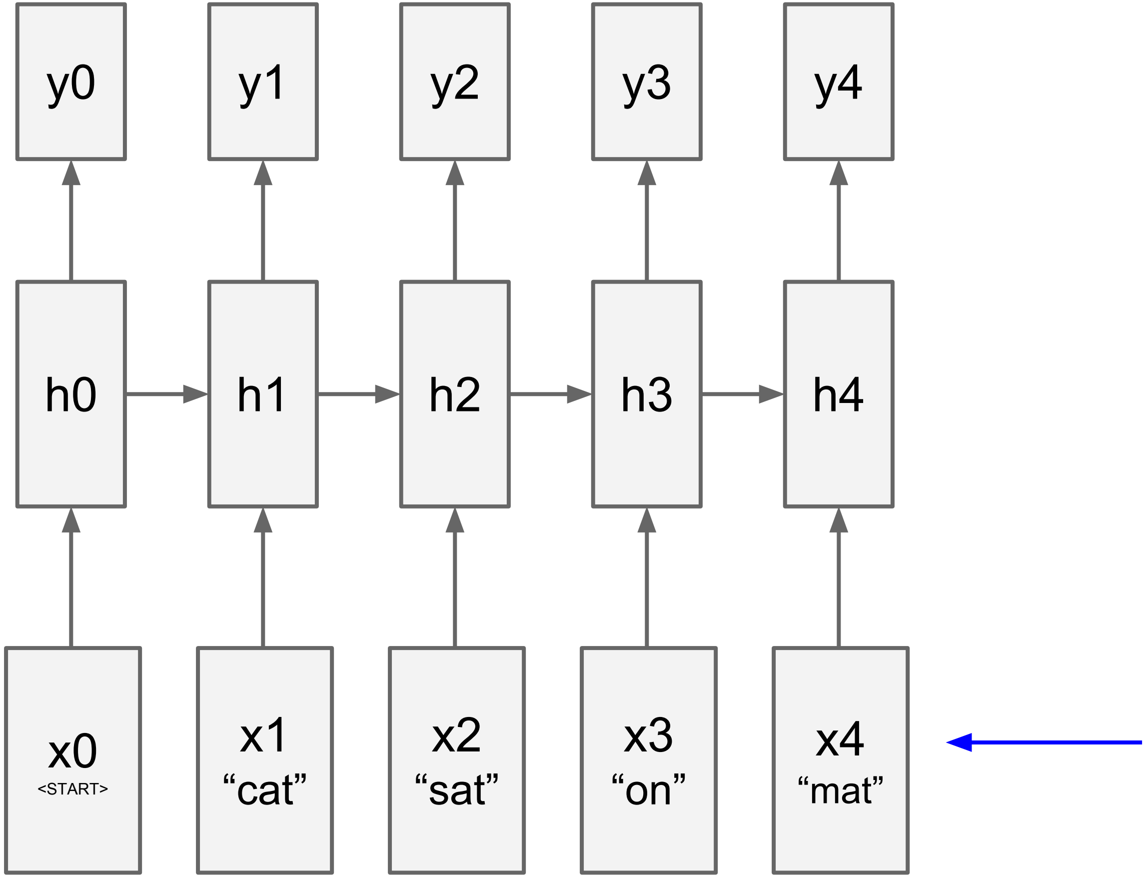
We want to train a **language model**:
P(next word | previous words)

Suppose we had the training sentence “cat sat on mat”

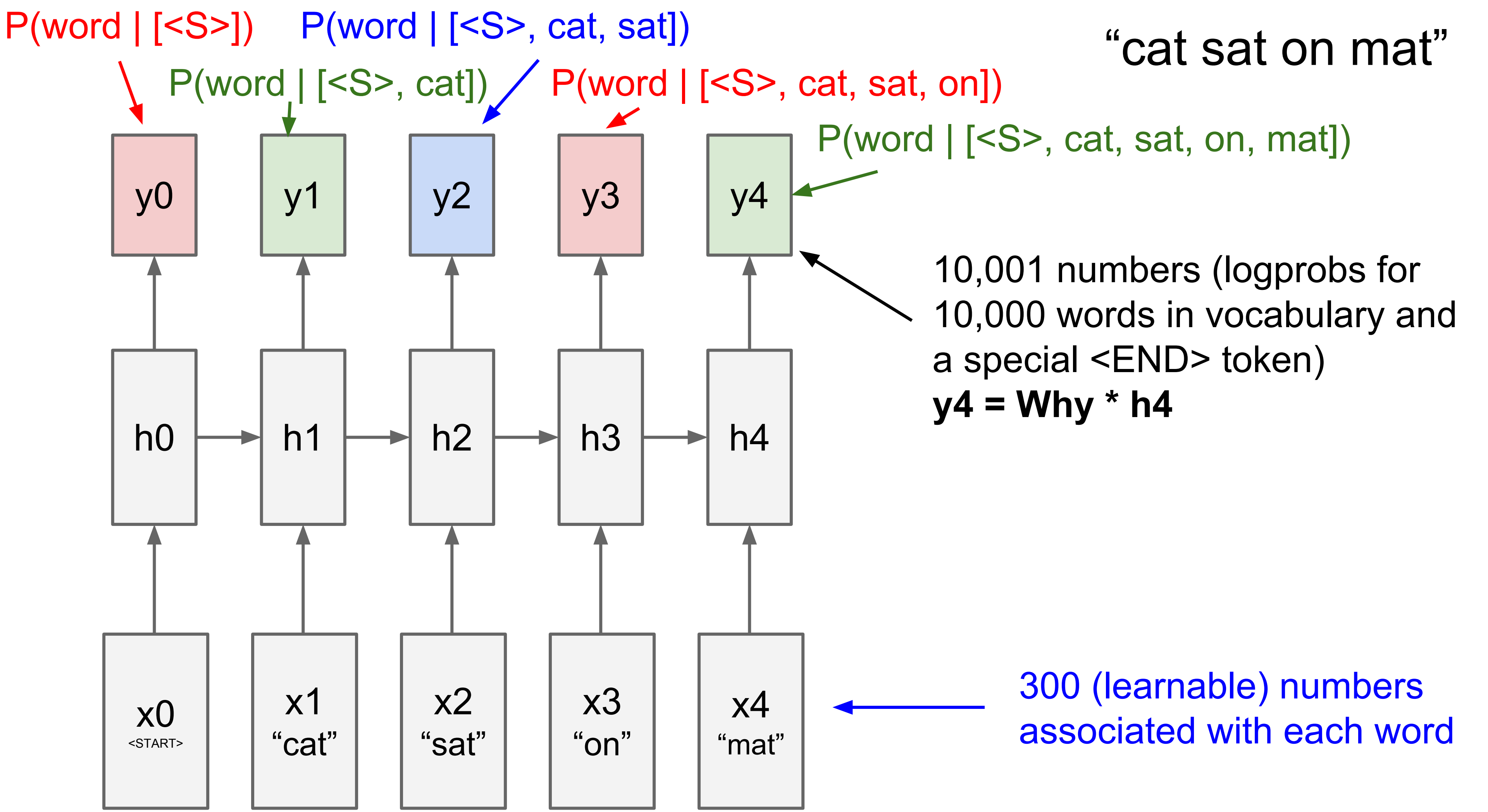
We want to train a **language model**:
P(next word | previous words)

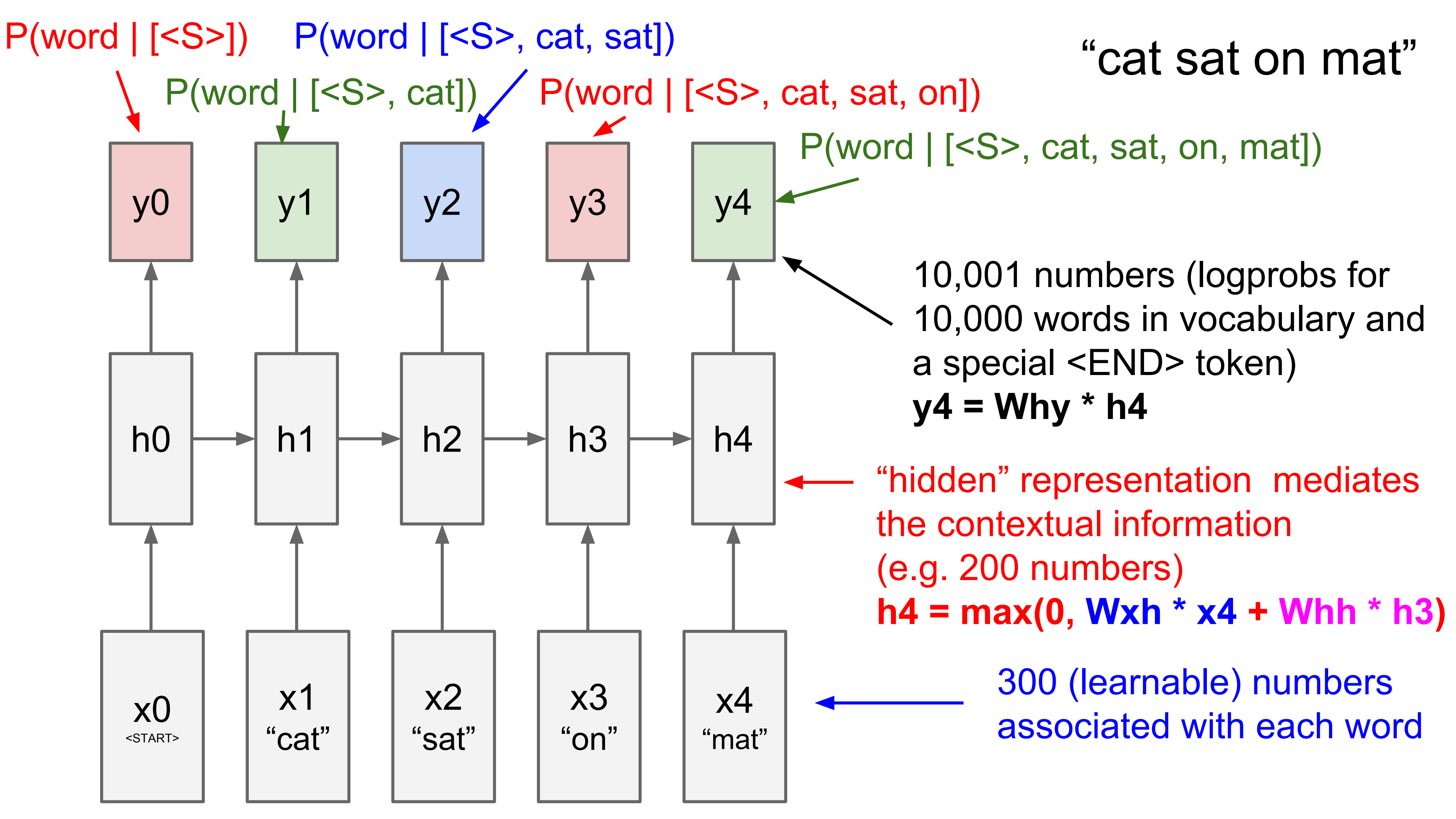
i.e. want these to be high:
P(cat | [<S>])
P(sat | [<S>, cat])
P(on | [<S>, cat, sat])
P(mat | [<S>, cat, sat, on])

“cat sat on mat”



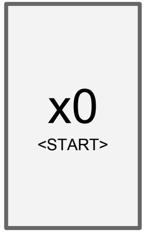
300 (learnable) numbers associated with each word


"cat sat on mat"
P(word | [<S>])
P(word | [<S>, cat])
P(word | [<S>, cat, sat])
P(word | [<S>, cat, sat, on])
P(word | [<S>, cat, sat, on, mat])
y0
y1
y2
y3
y4
h0
h1
h2
h3
h4
x0
<START>
x1
"cat"
x2
"sat"
x3
"on"
x4
"mat"
10,001 numbers (logprobs for 10,000 words in vocabulary and a special <END> token)
y4 = Why * h4
300 (learnable) numbers associated with each word


"cat sat on mat"
P(word | [<S>])
P(word | [<S>, cat])
P(word | [<S>, cat, sat])
P(word | [<S>, cat, sat, on])
P(word | [<S>, cat, sat, on, mat])
y0
y1
y2
y3
y4
h0
h1
h2
h3
h4
x0
<START>
x1
"cat"
x2
"sat"
x3
"on"
x4
"mat"
10,001 numbers (logprobs for 10,000 words in vocabulary and a special <END> token)
y4 = Why * h4
"hidden" representation mediates the contextual information (e.g. 200 numbers)
h4 = max(0, Wxh * x4 + Whh * h3)
300 (learnable) numbers associated with each word

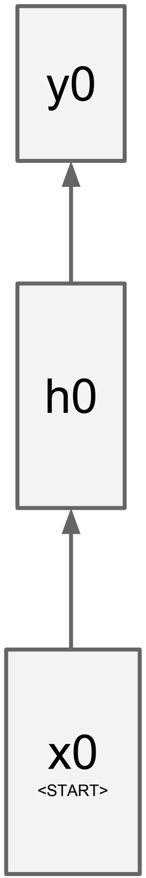
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



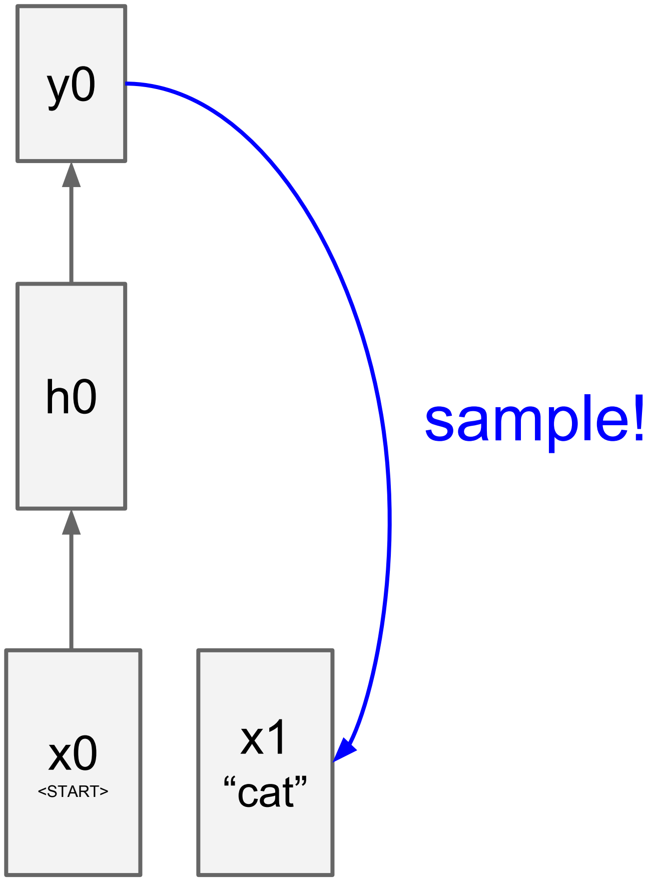
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



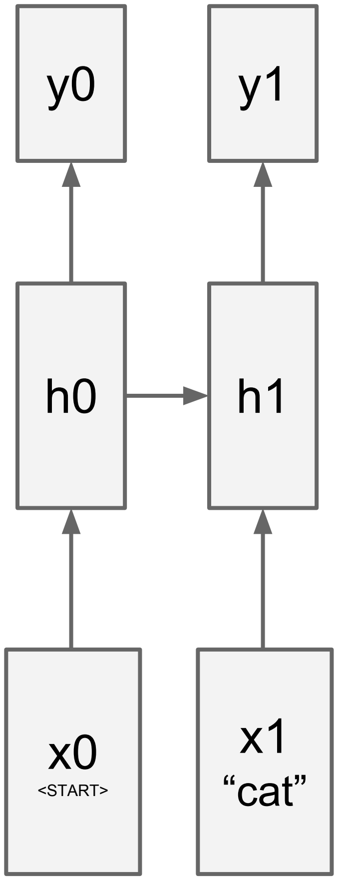
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



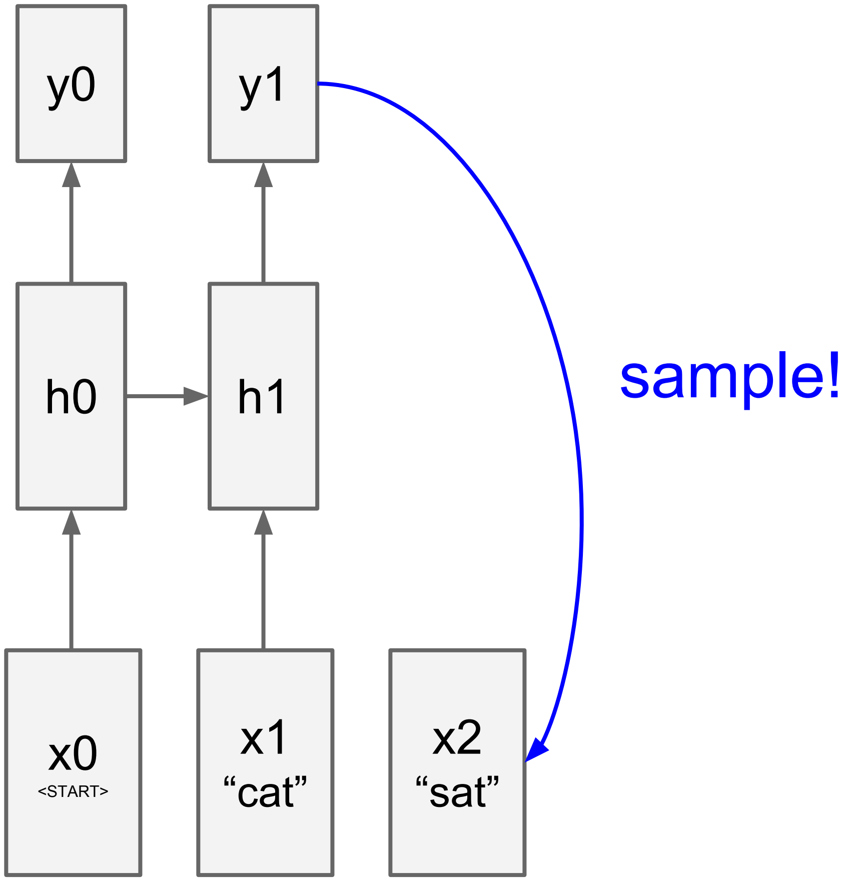
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



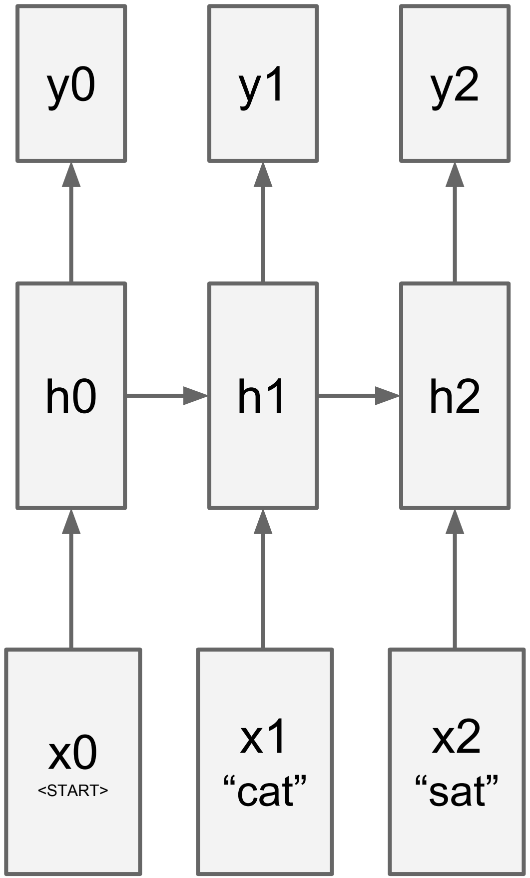
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



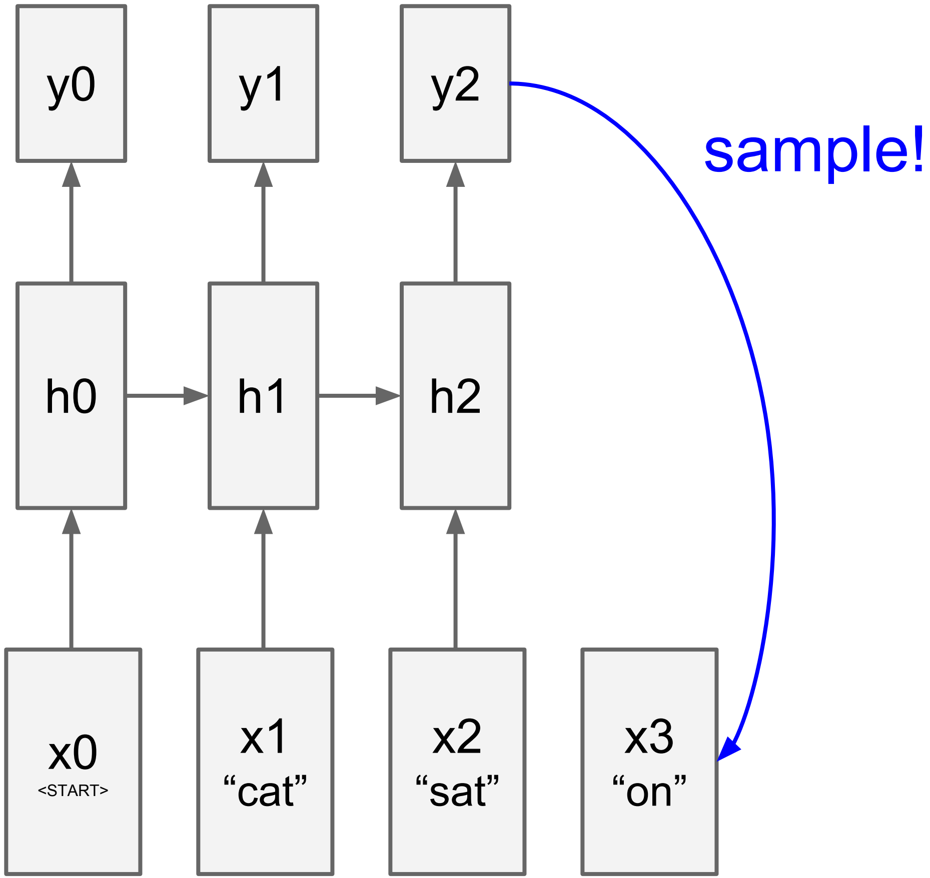
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



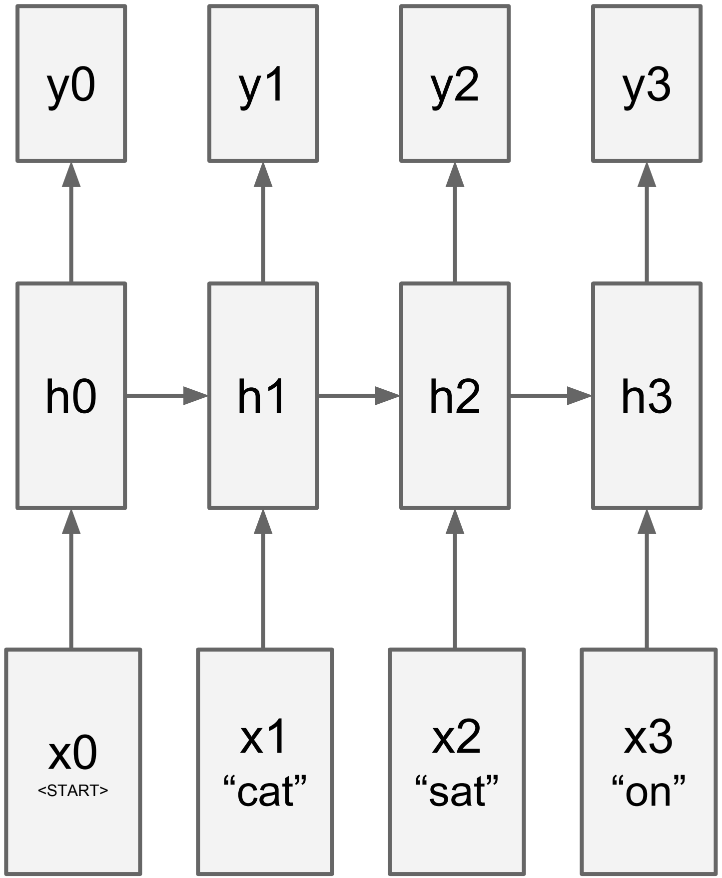
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



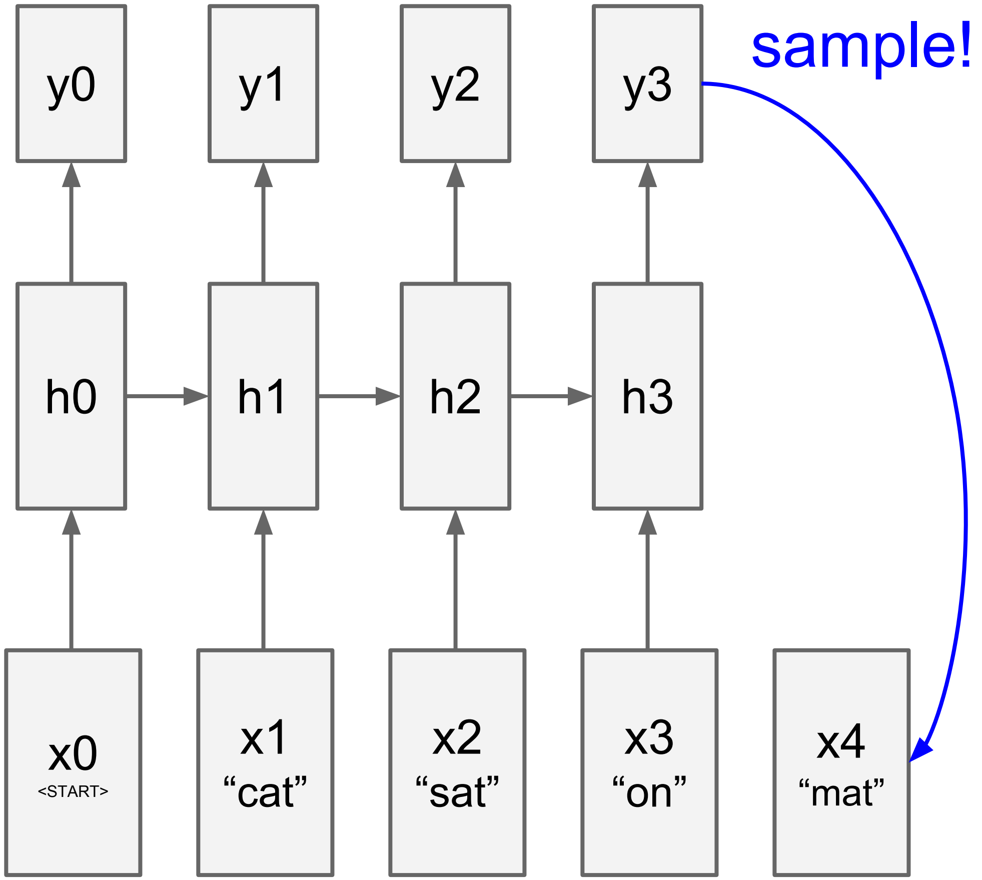
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



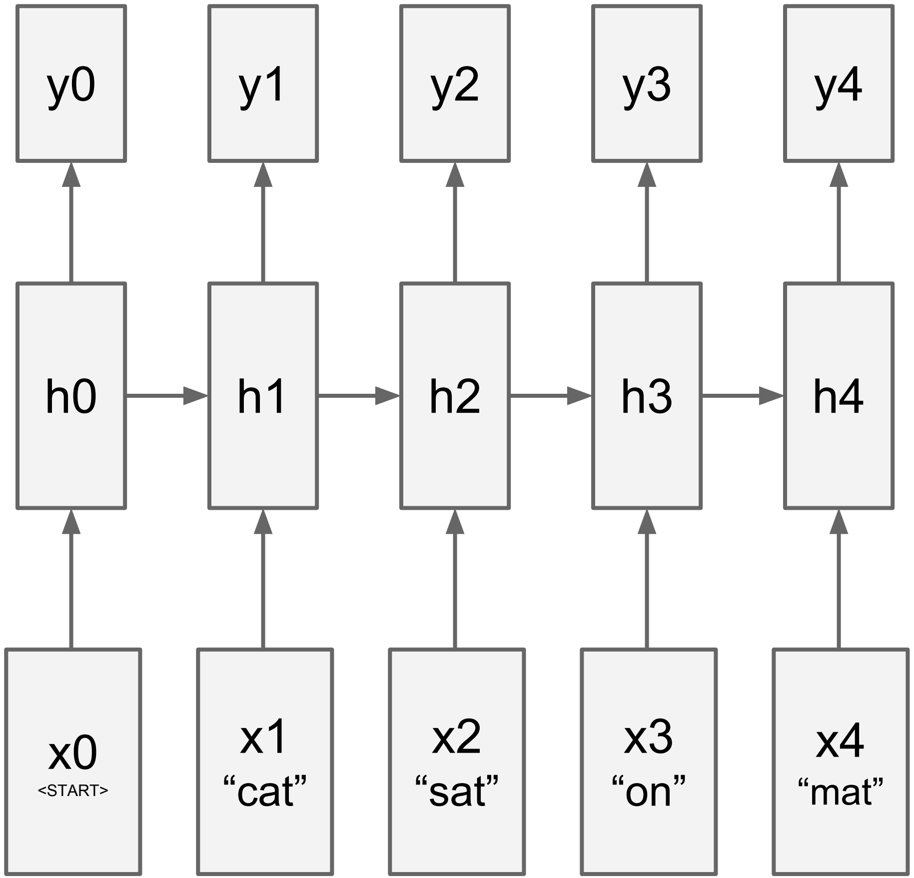
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



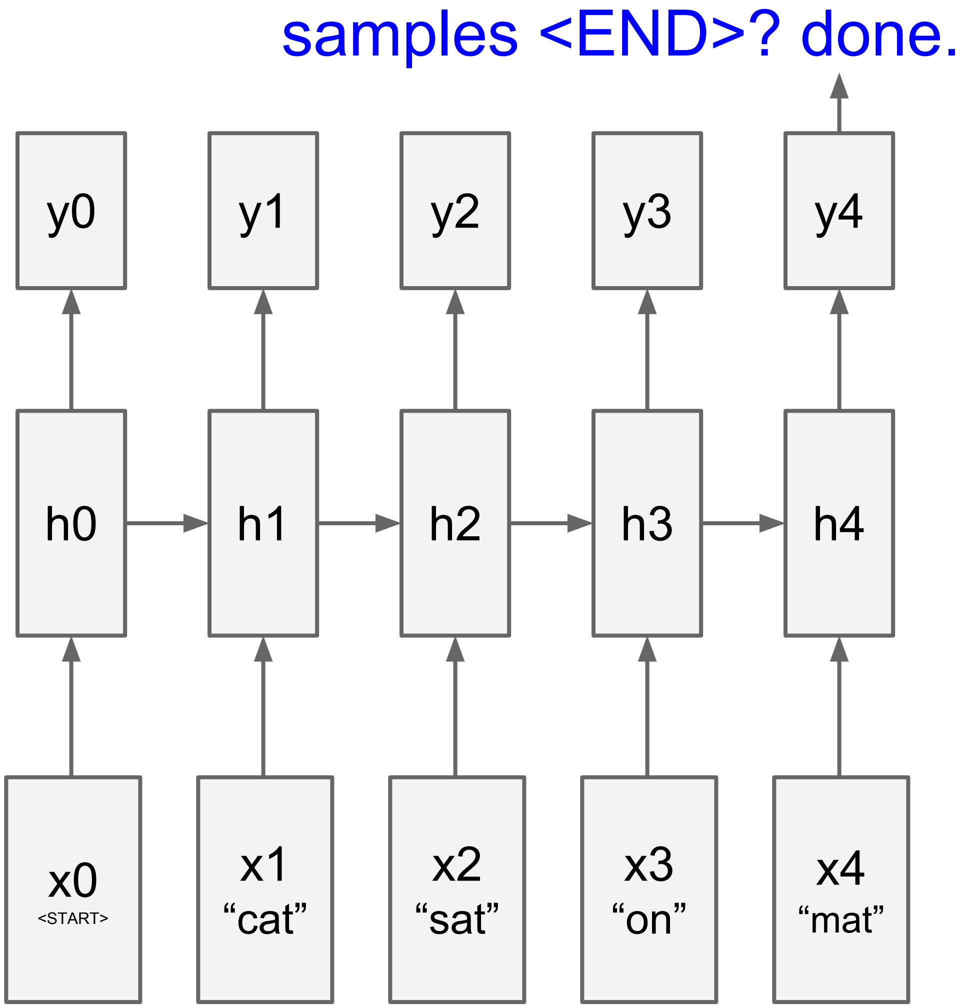
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)

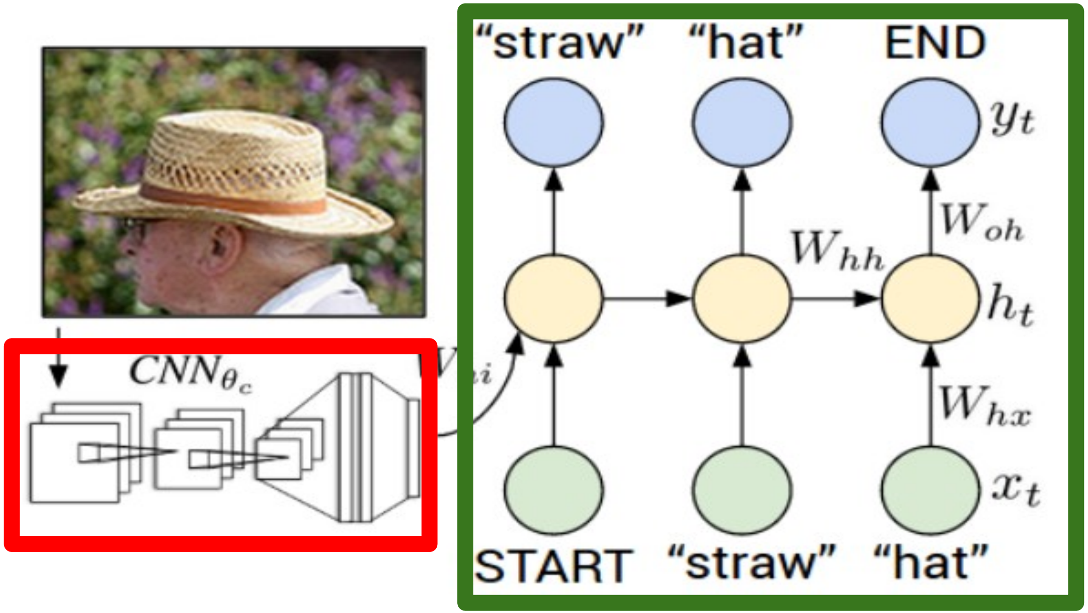


Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



Recurrent Neural Network



Convolutional Neural Network

“straw hat”

training example

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

“straw hat”

training example

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

“straw hat”

training example

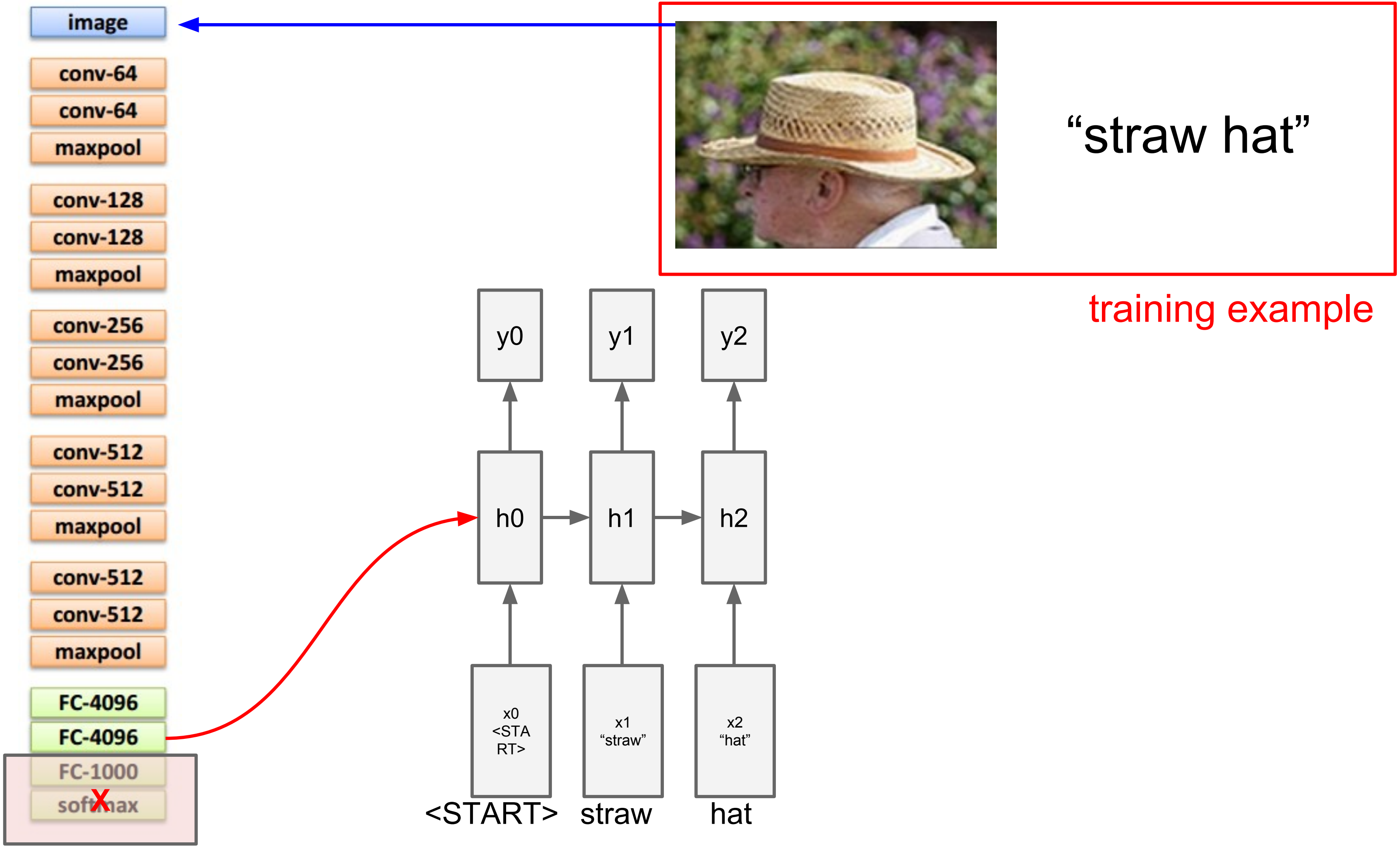
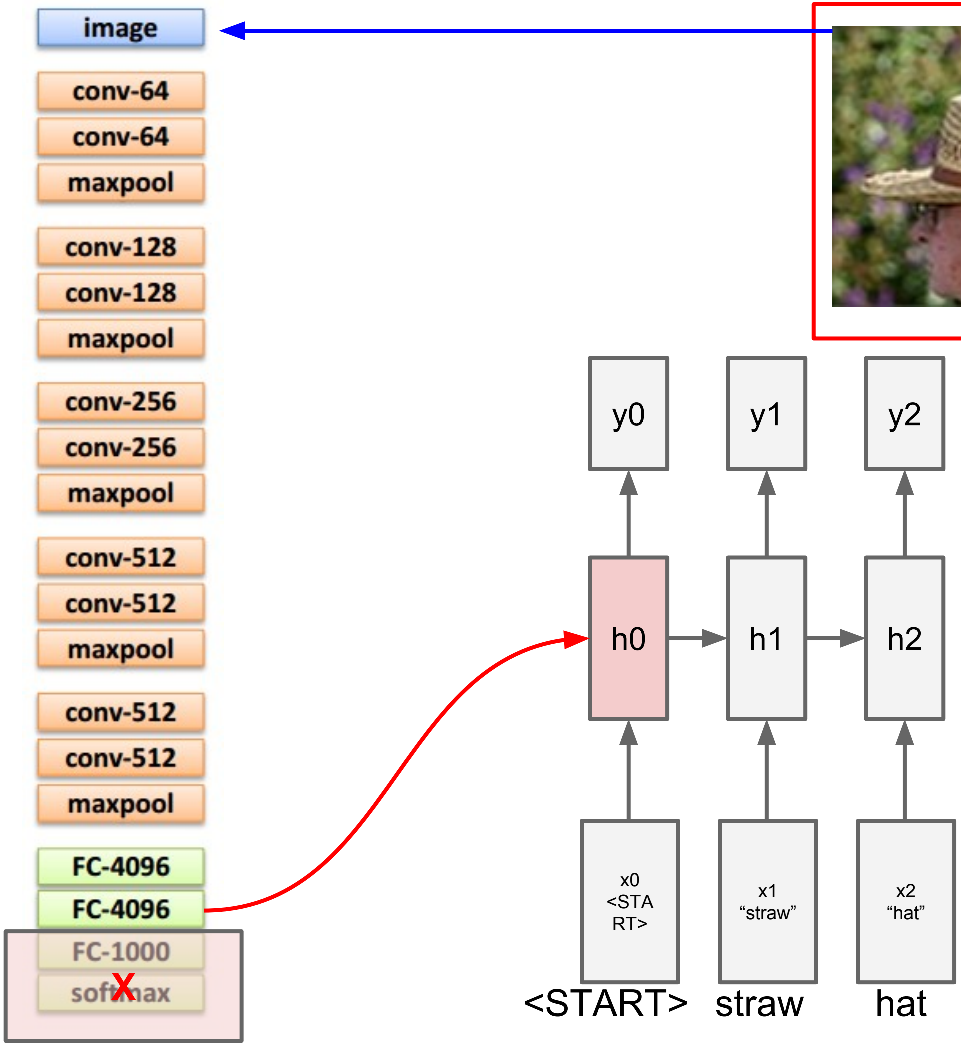
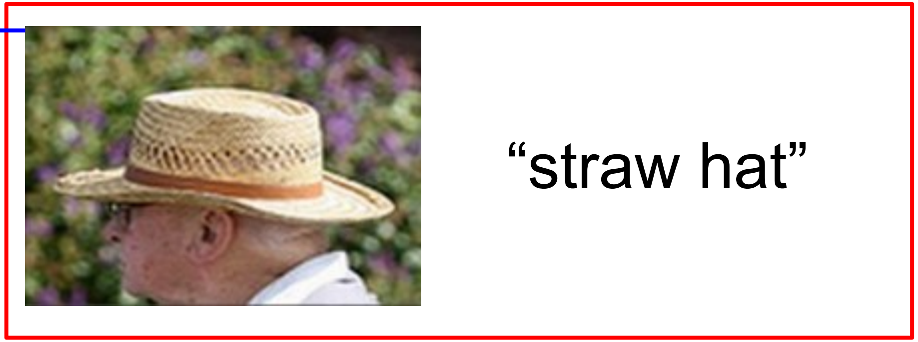


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
FC-1000
softmax
X
“straw hat”
training example
y0
y1
y2
h0
h1
h2
x0 <START>
x1 “straw”
x2 “hat”
<START>
straw
hat





training example

before:

h0 = max(0, Wxh * x0)

now:

h0 = max(0, Wxh * x0 + Wih * v)

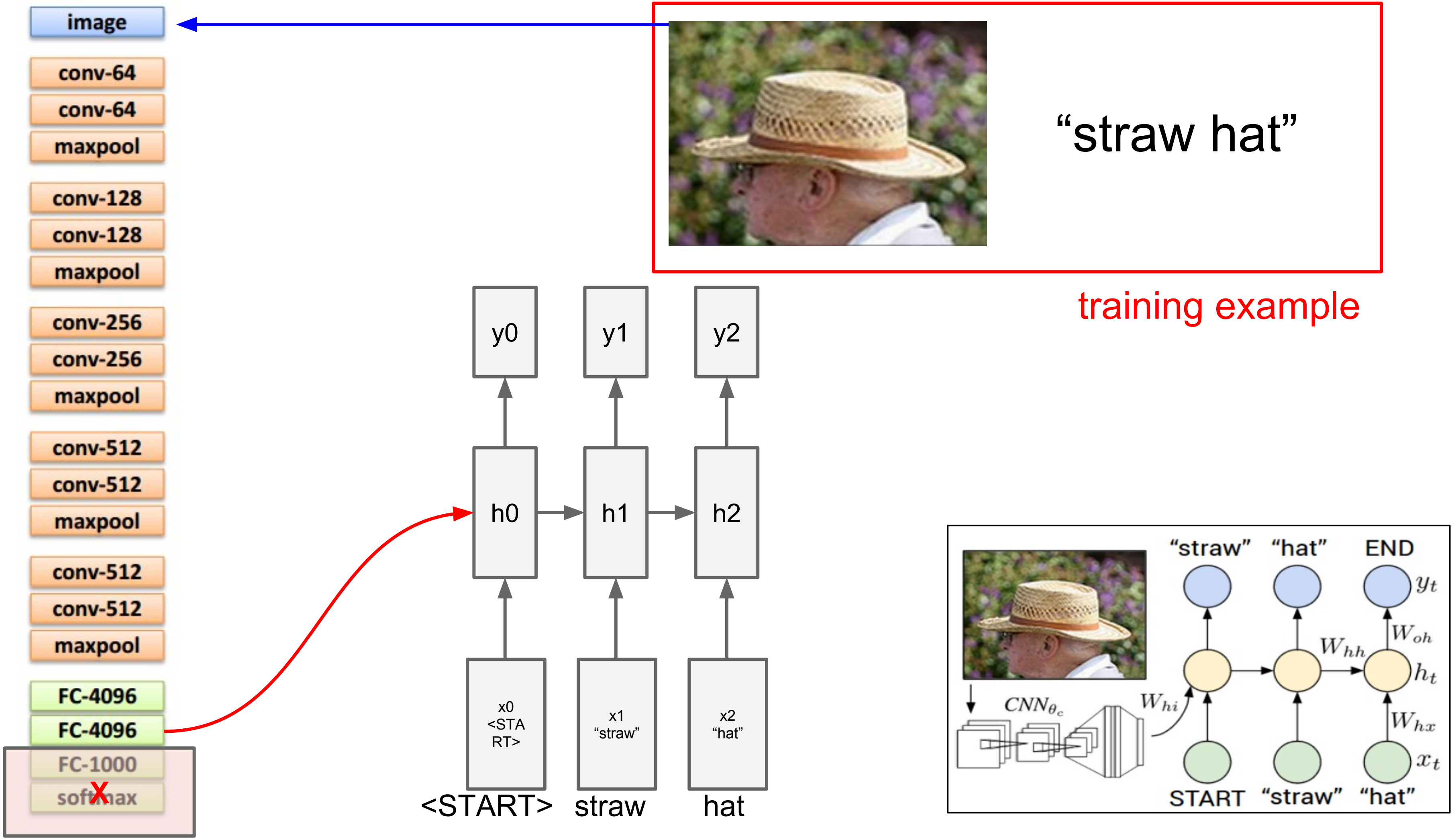


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
FC-1000
softmax
X
"straw hat"
training example
y0
y1
y2
h0
h1
h2
x0 <START>
x1 "straw"
x2 "hat"
<START>
straw
hat
"straw"
"hat"
END
y_t
W_{oh}
W_{hh}
h_t
W_{hi}
CNN_{θ_c}
W_{hx}
x_t
START
"straw"
"hat"

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

test image

x0
<STA
RT>

<START>

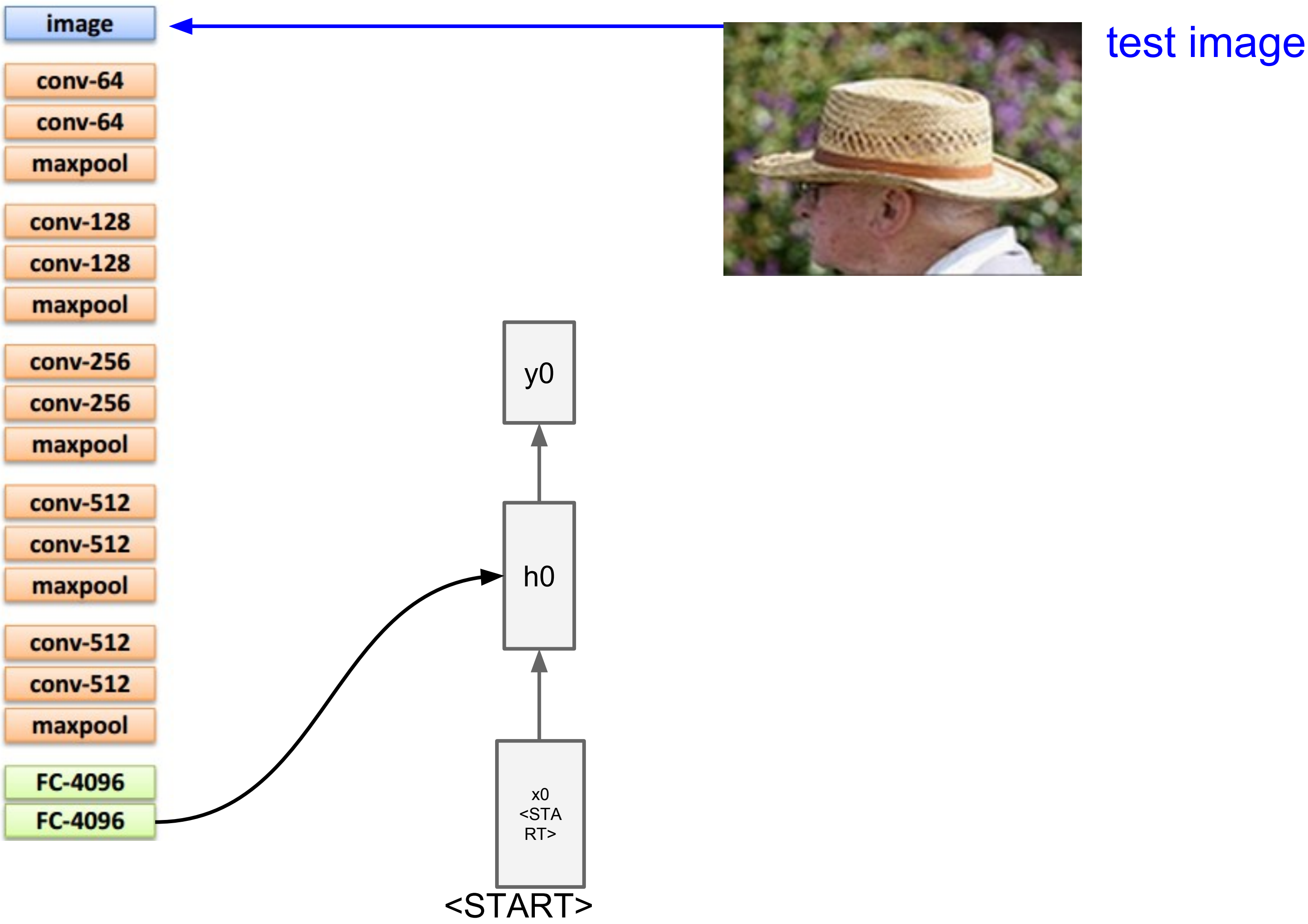


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
test image
y0
h0
x0 <START>
<START>

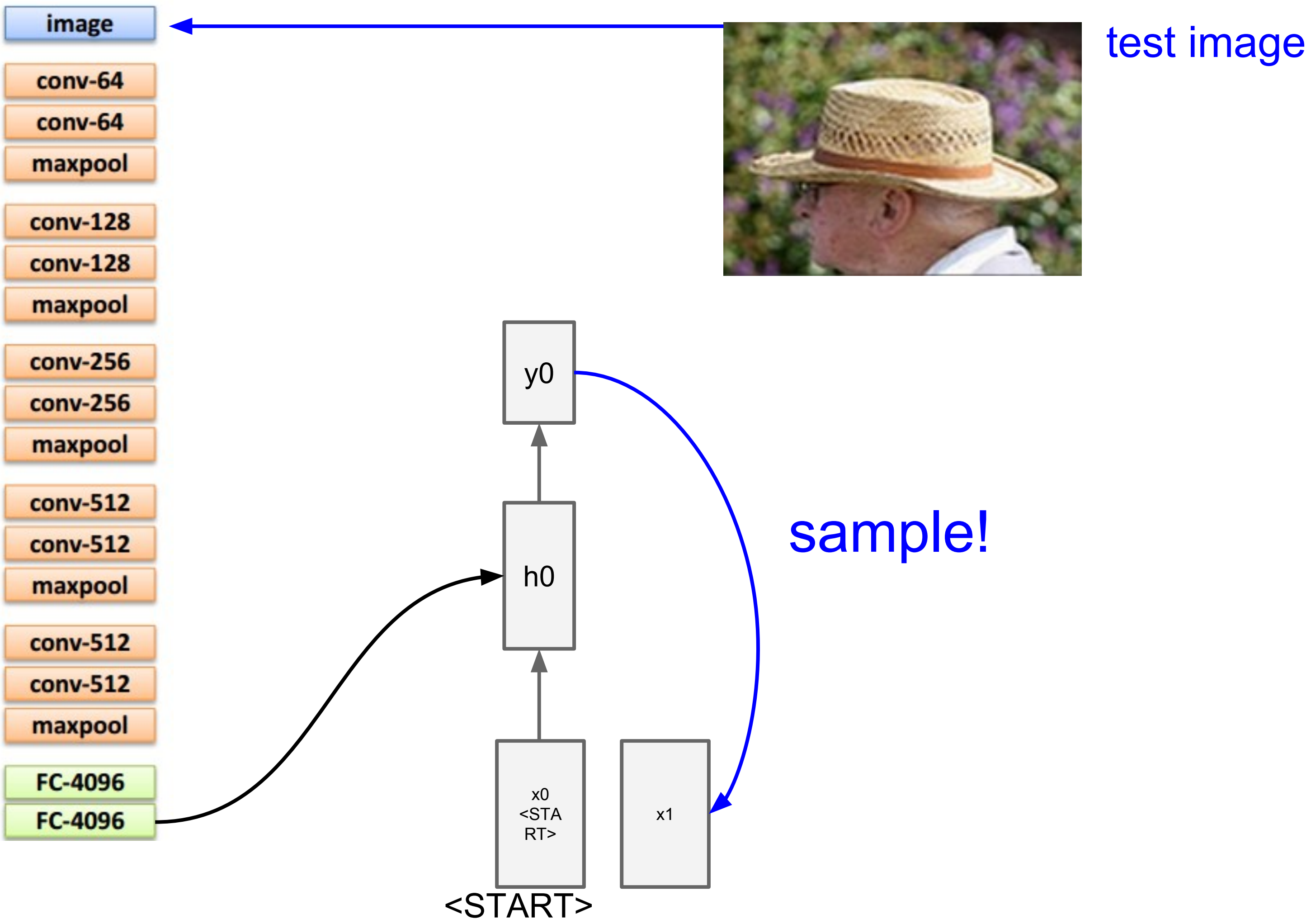


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
test image
y0
h0
sample!
x0 <STA RT>
x1
<START>

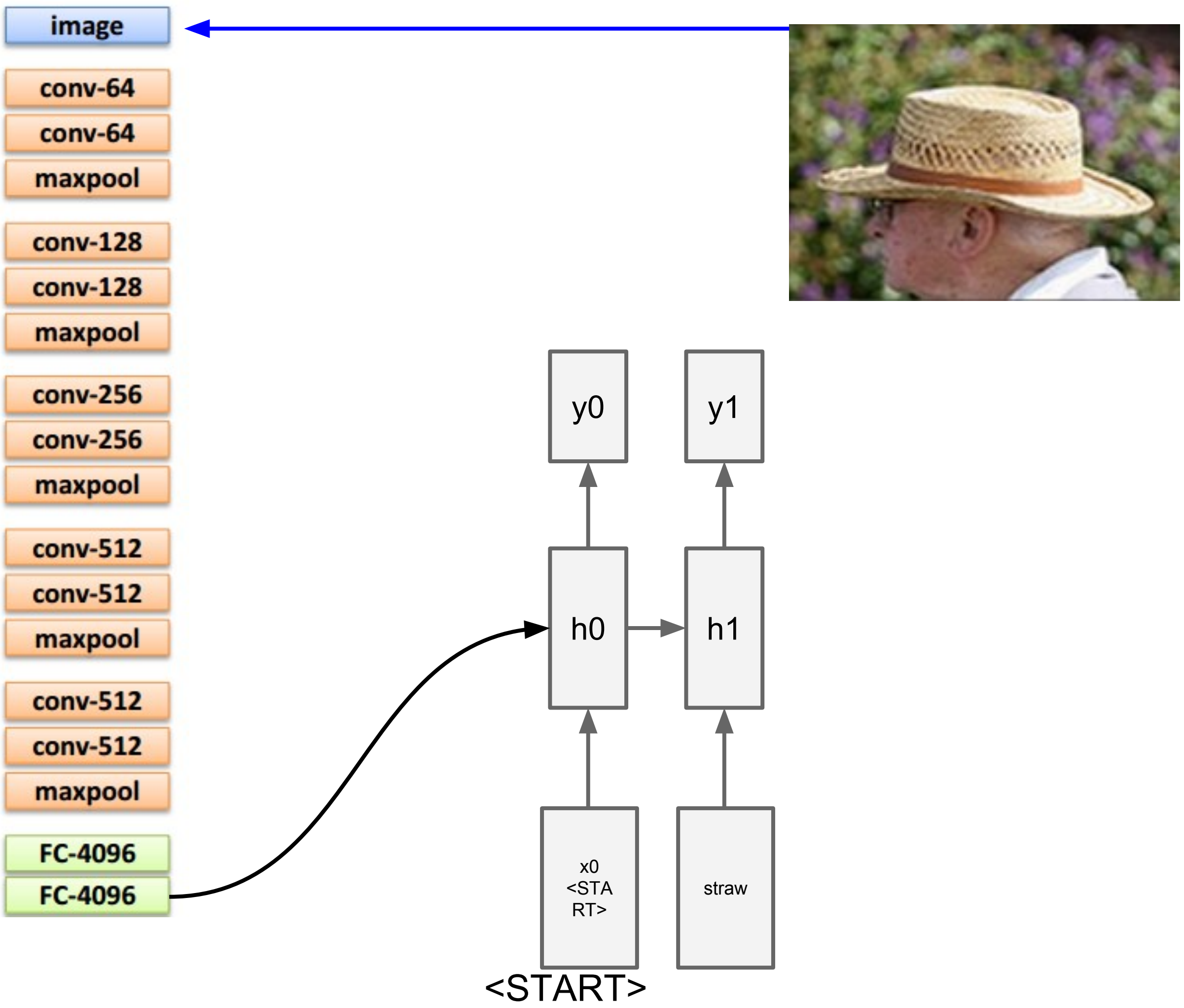


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
y0
y1
h0
h1
x0
<STA
RT>
straw
<START>

test image

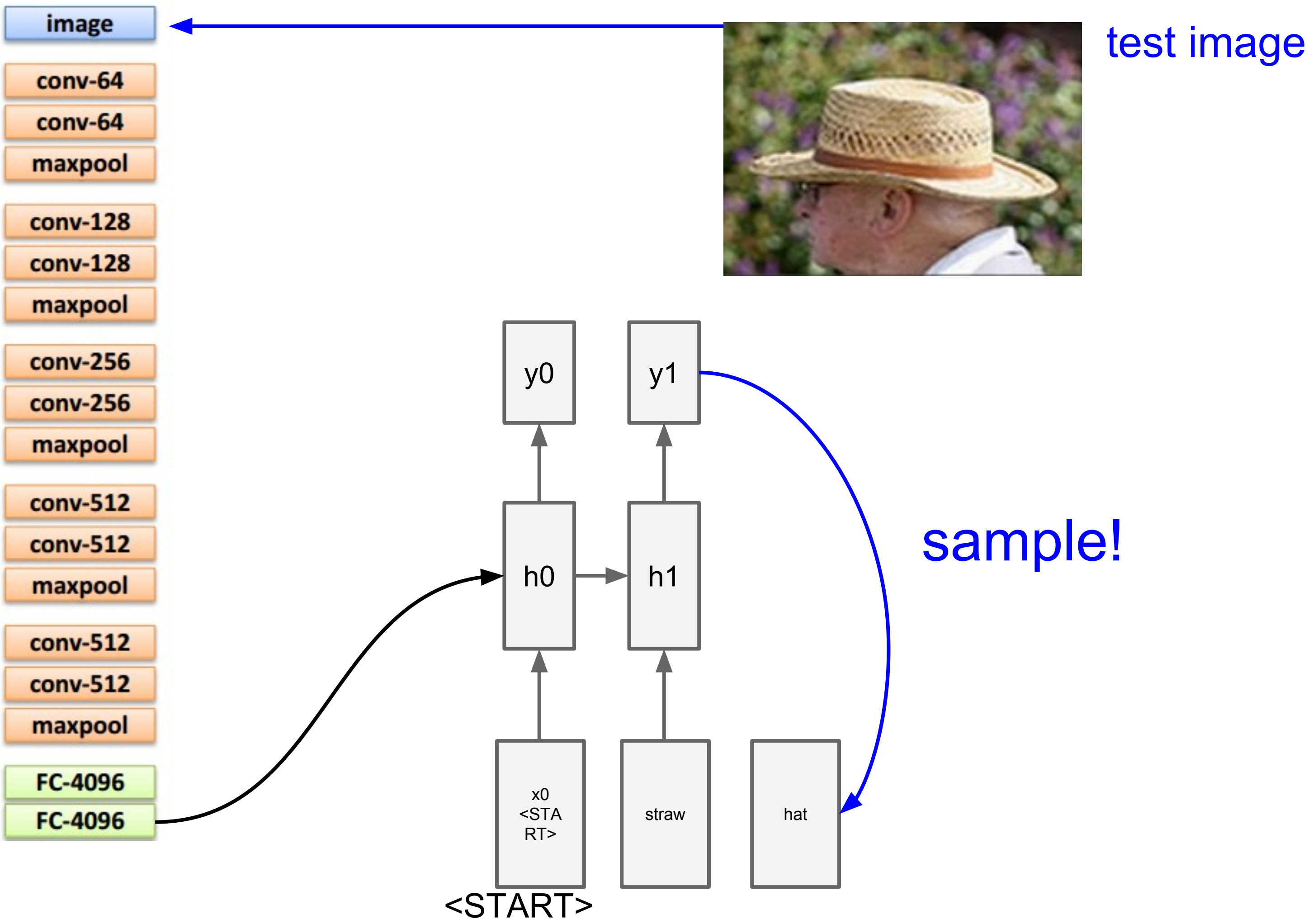


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
test image
y0
y1
h0
h1
sample!
x0
<STA
RT>
straw
hat
<START>

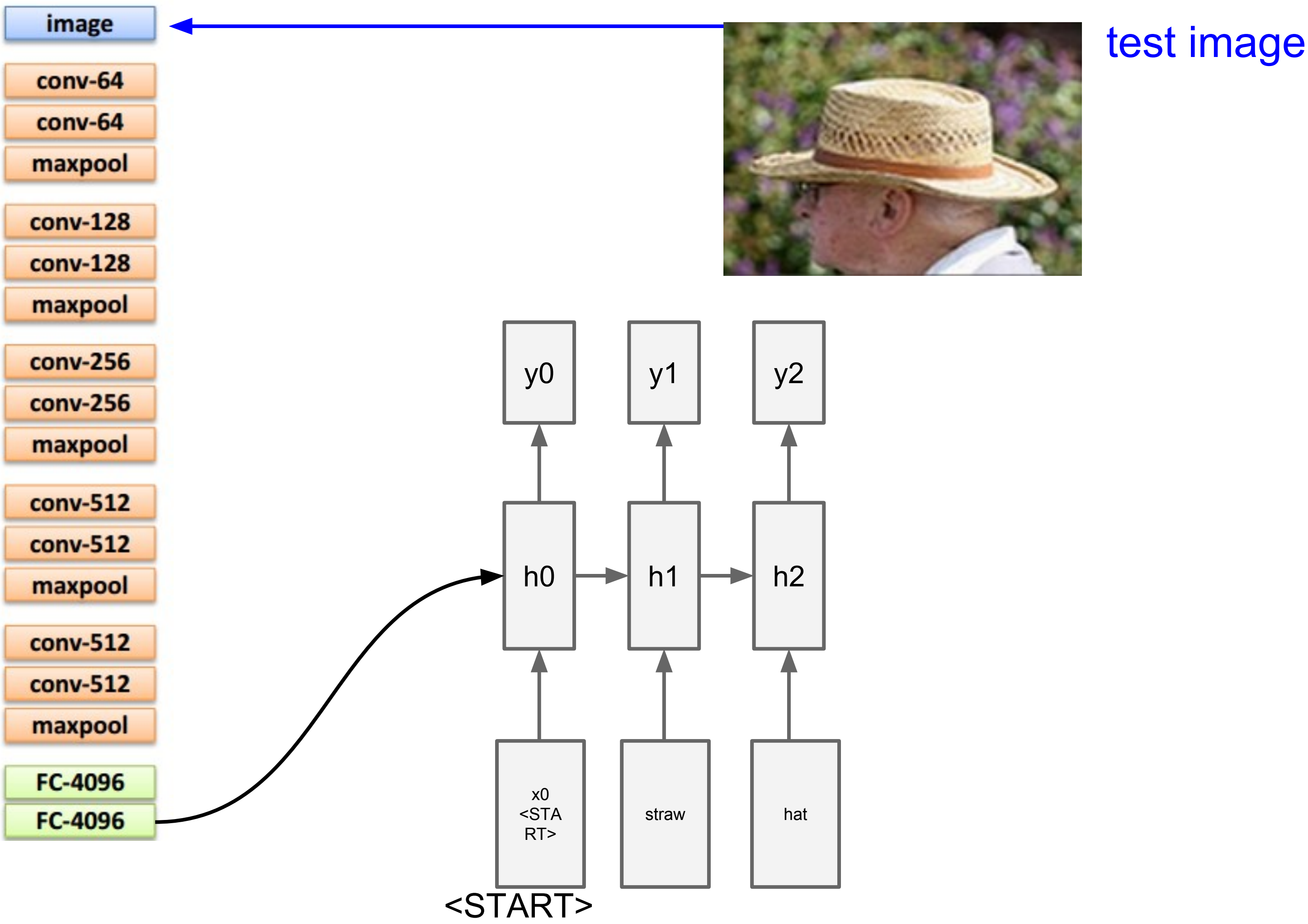


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
test image
y0
y1
y2
h0
h1
h2
x0 <STA RT>
straw
hat
<START>

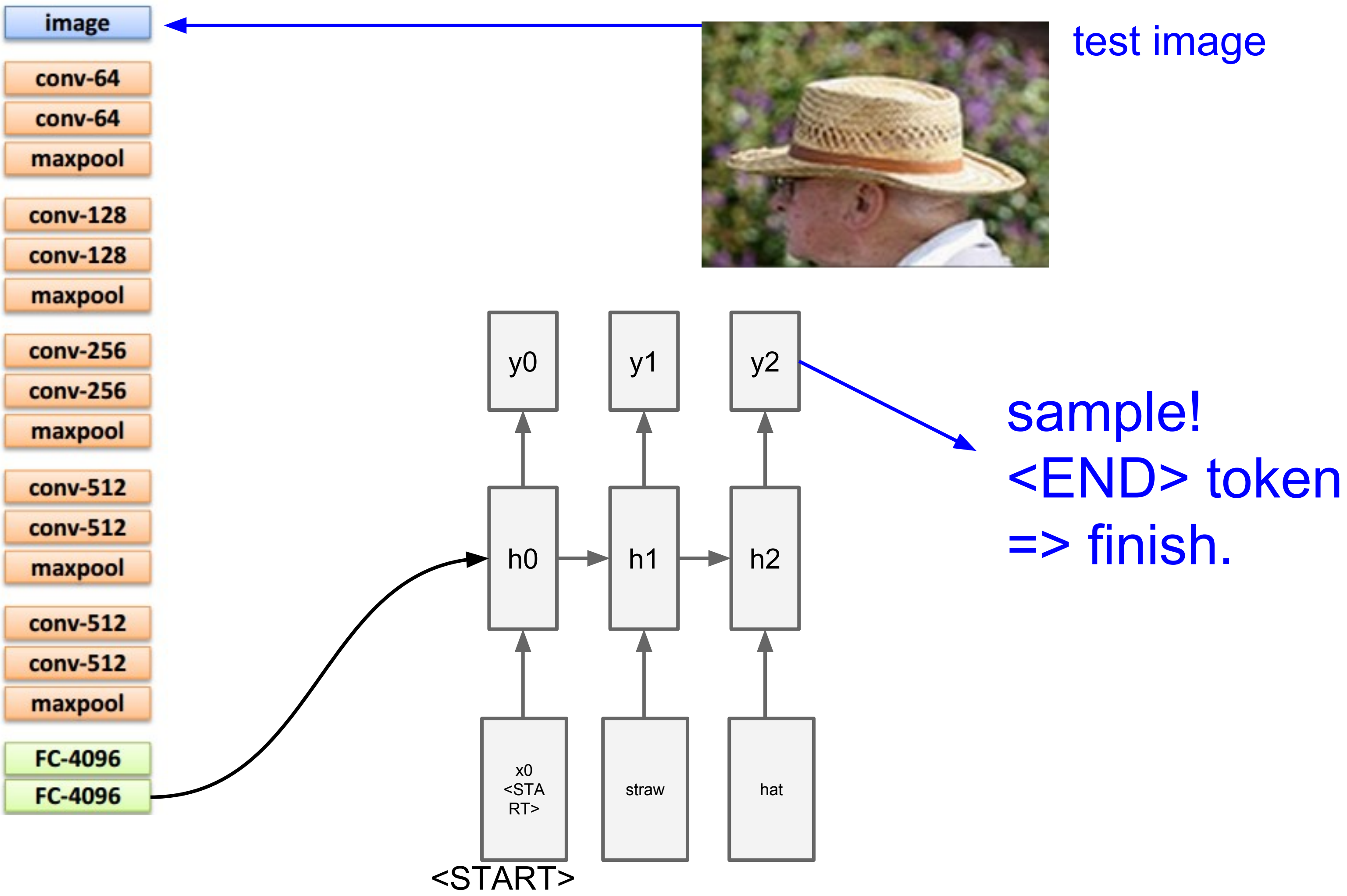
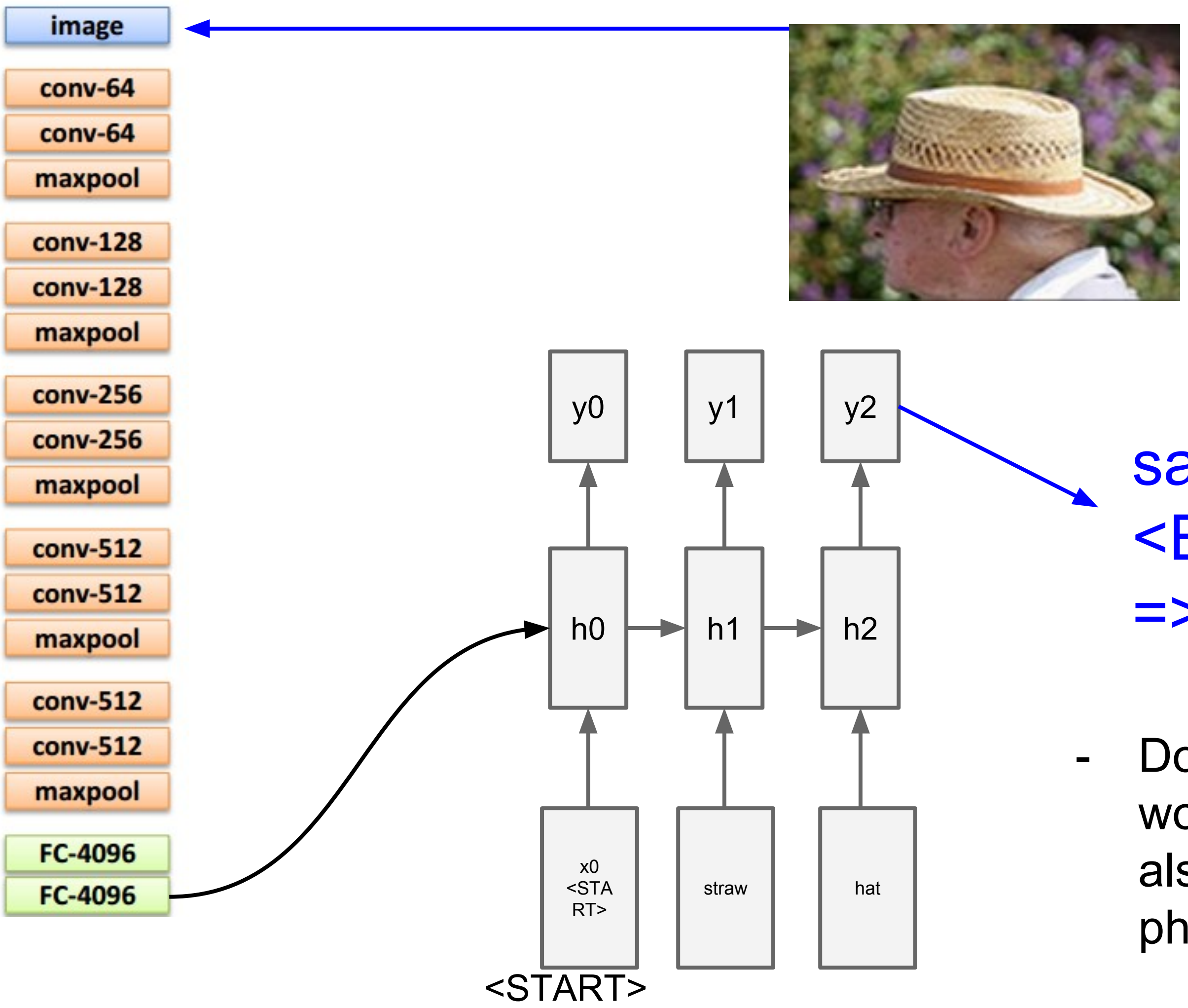


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
test image
y0
y1
y2
h0
h1
h2
x0 <START>
straw
hat
<START>
sample!
<END> token
=> finish.

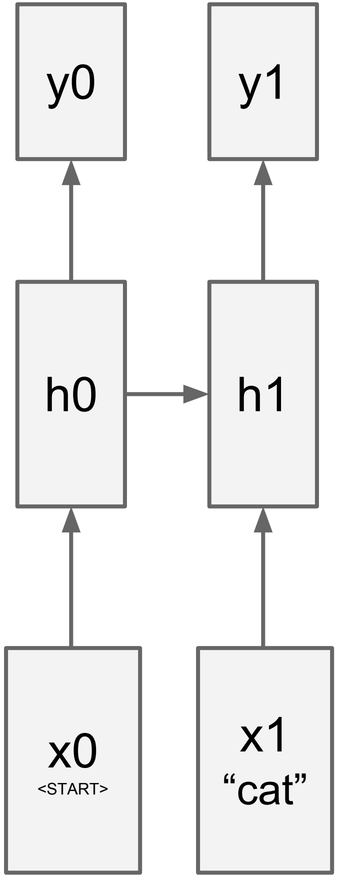


test image

sample!
<END> token
=> finish.

- Don’t have to do greedy word-by-word sampling, can also search over longer phrases with **beam search**

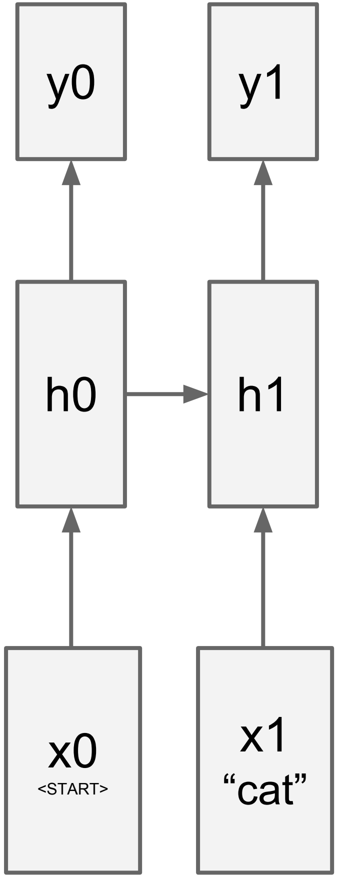
RNN vs. LSTM



“hidden” representation
(e.g. 200 numbers)

h1 = max(0, Wxh * x1 + Whh * h0)

RNN vs. LSTM



“hidden” representation
(e.g. 200 numbers)
h1 = max(0, Wxh * x1 + Whh * h0)

LSTM changes the form of the equation for **h1** such that:

1. more expressive multiplicative interactions
2. gradients flow nicer
3. network can explicitly decide to reset the hidden state

Image Sentence Datasets

a man riding a bike on a dirt path through a forest.
bicyclist raises his fist as he rides on desert dirt trail.
this dirt bike rider is smiling and raising his fist in triumph.
a man riding a bicycle while pumping his fist in the air.
a mountain biker pumps his fist in celebration.

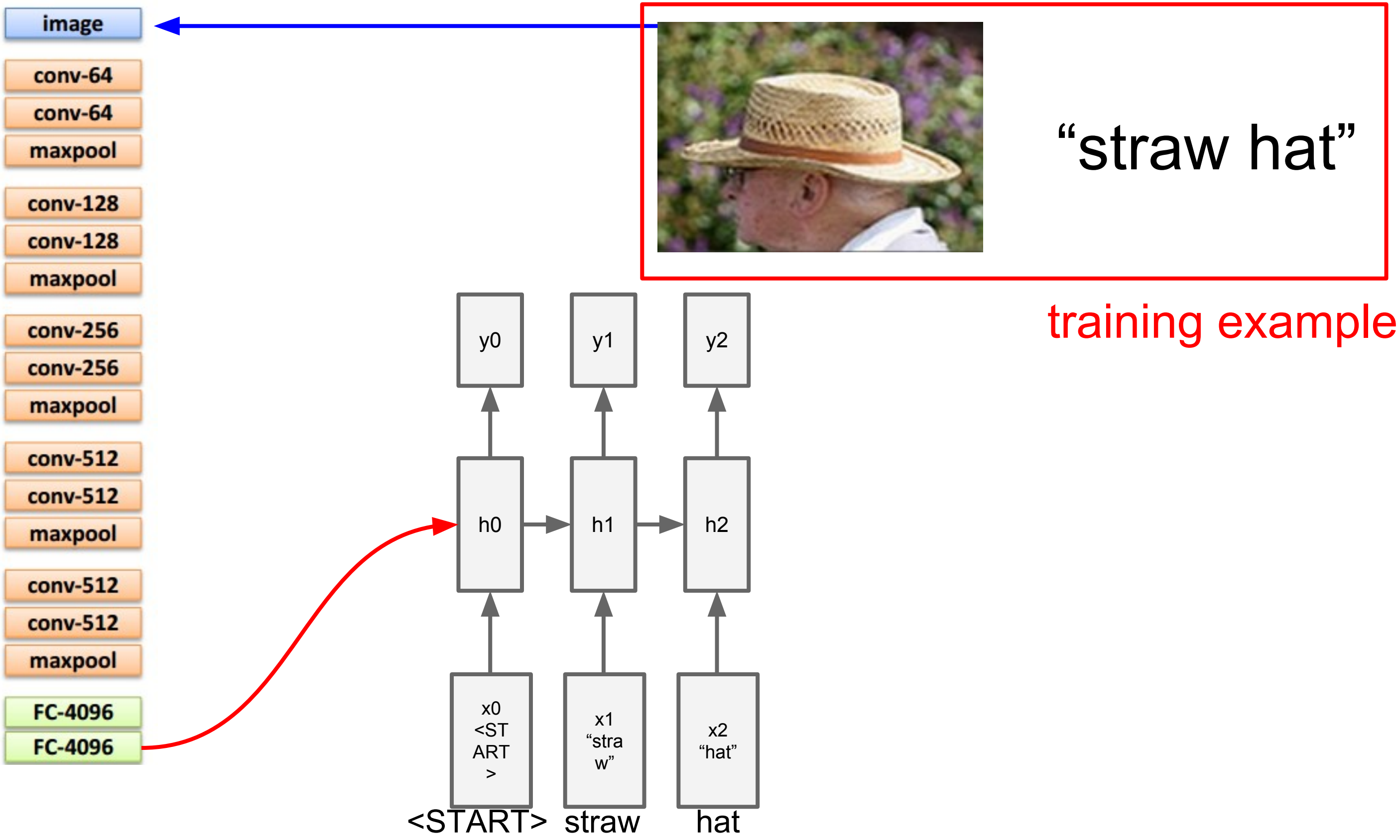
Microsoft COCO
[Tsung-Yi Lin et al. 2014]
mscoco.org

currently:
~120K images
~5 sentences each

Training an RNN/LSTM...

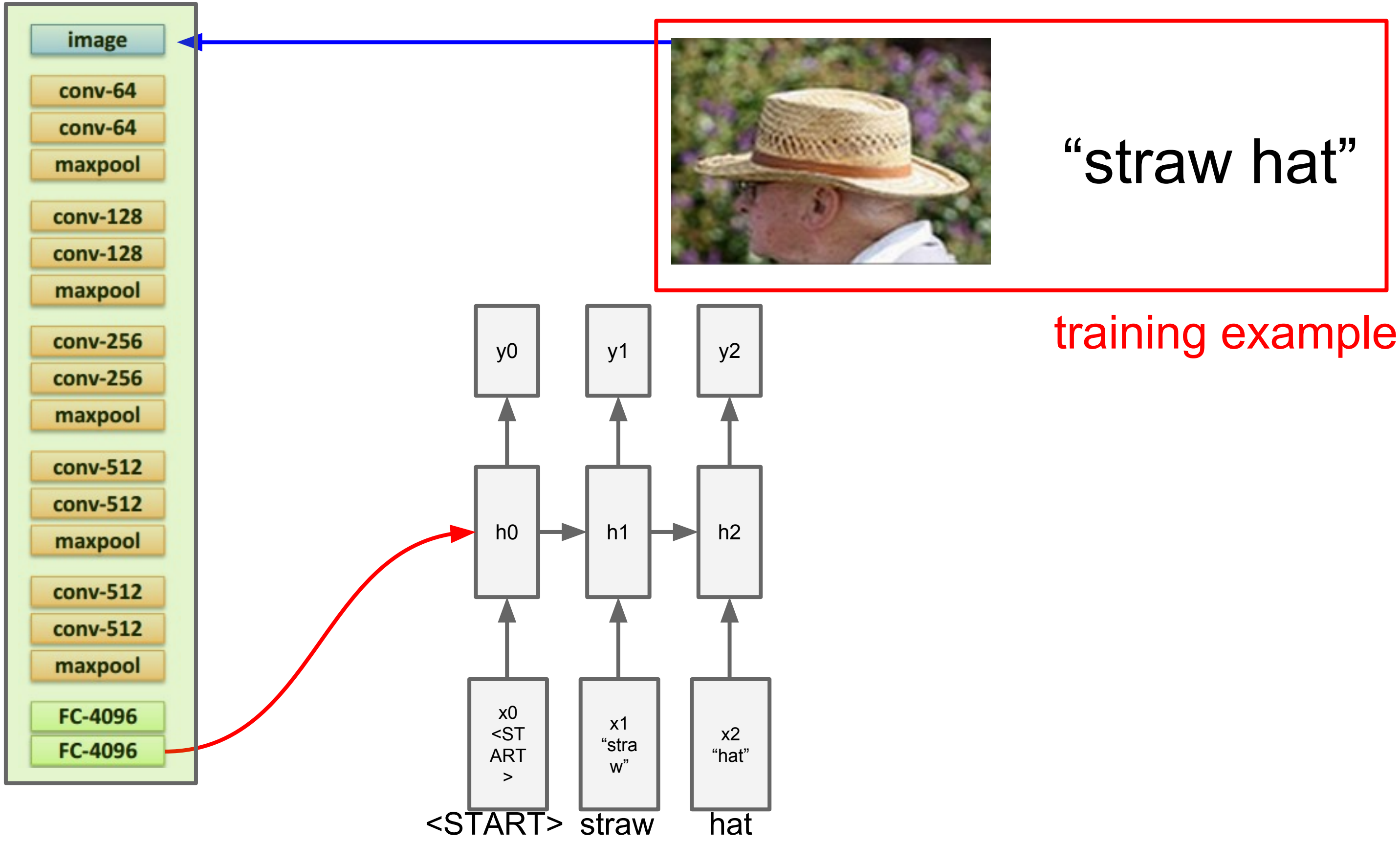
- **Clip** the gradients (important!). 5 worked ok
- **RMSprop** adaptive learning rate worked nice
- Initialize softmax **biases** with log word frequency distribution
- Train for **long time**

+ Transfer Learning



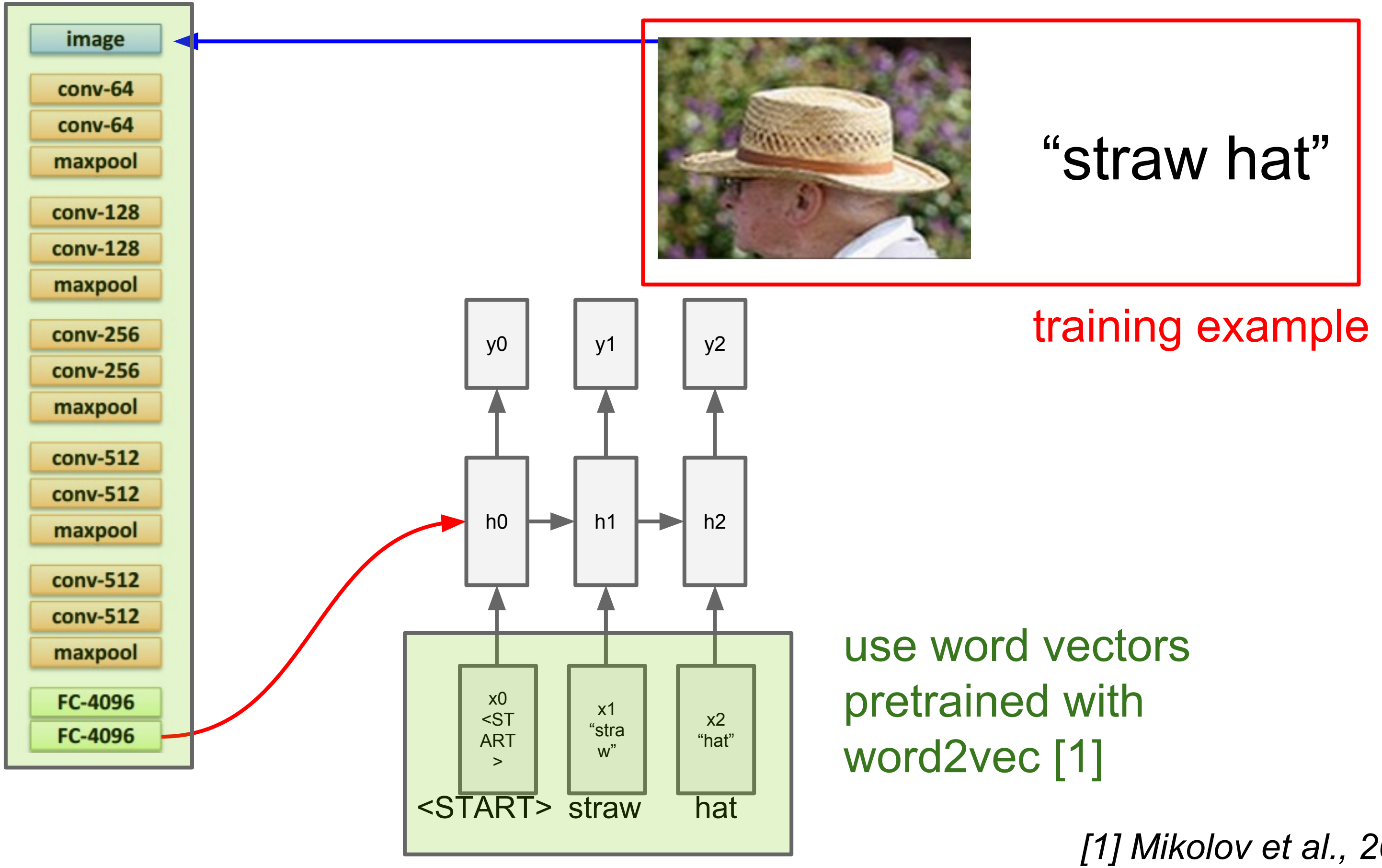
+ Transfer Learning

use weights pretrained from ImageNet



+ Transfer Learning

use weights pretrained from ImageNet



training example

use word vectors pretrained with word2vec [1]

[1] Mikolov et al., 2013

Summary of the approach

We wanted to describe images with sentences.

1. Define a single function from input -> output
2. Initialize parts of net from elsewhere if possible
3. Get some data
4. Train with SGD

Wow I can’t believe that worked

a group of people standing around a room with remotes
`logprob: -9.17`

a young boy is holding a baseball bat
`logprob: -7.61`

a cow is standing in the middle of a street
`logprob: -8.84`

Wow I can’t believe that worked

a cat is sitting on a toilet seat
`logprob: -7.79`

a display case filled with lots of different types of donuts
`logprob: -7.78`

a group of people sitting at a table with wine glasses
`logprob: -6.71`

Well, I can kind of see it

a man standing next to a clock on a wall
`logprob: -10.08`

a young boy is holding a baseball bat
`logprob: -7.65`

a cat is sitting on a couch with a remote control
`logprob: -12.45`

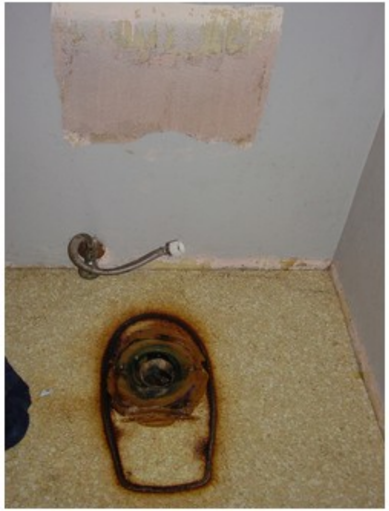
Well, I can kind of see it

a baby laying on a bed with a stuffed bear
`logprob: -8.66`

a table with a plate of food and a cup of coffee
`logprob: -9.93`

a young boy is playing frisbee in the park
`logprob: -9.52`

Not sure what happened there...



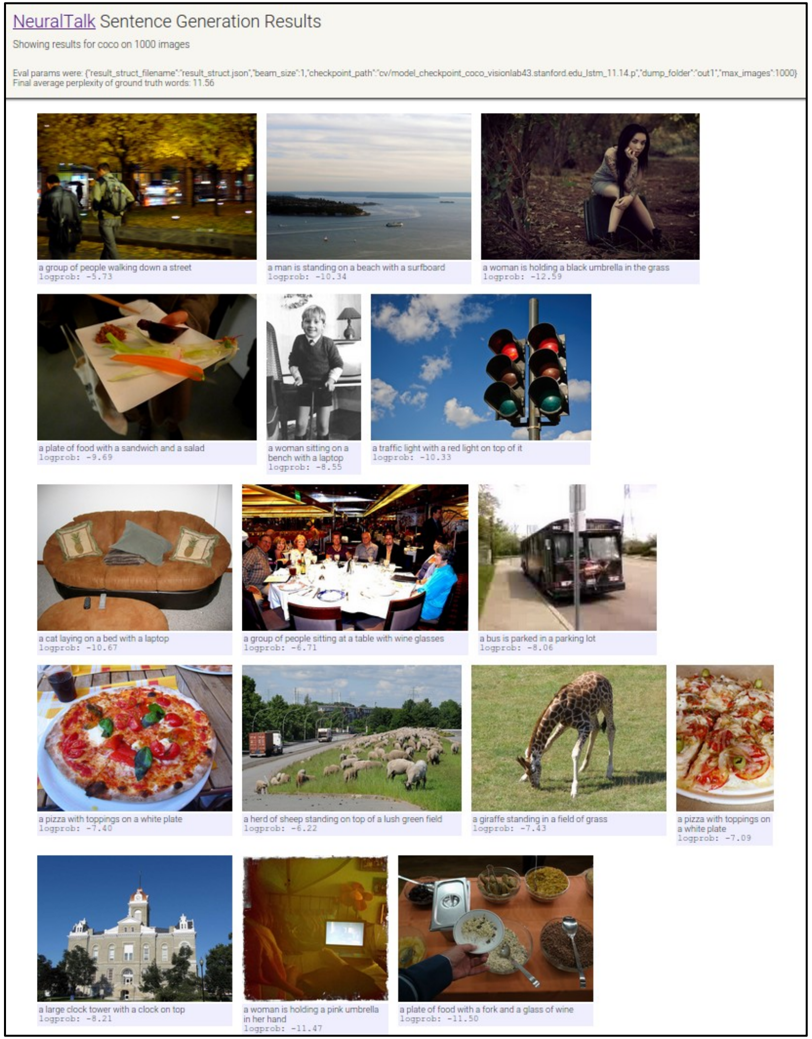
a toilet with a seat up in a bathroom
`logprob: -13.44`

a woman holding a teddy bear in front of a mirror
`logprob: -9.65`

a horse is standing in the middle of a road
`logprob: -10.34`

See predictions on 1000 COCO images:

http://bit.ly/neuraltalkdemo



What this approach Doesn't do:

- There is no *reasoning*

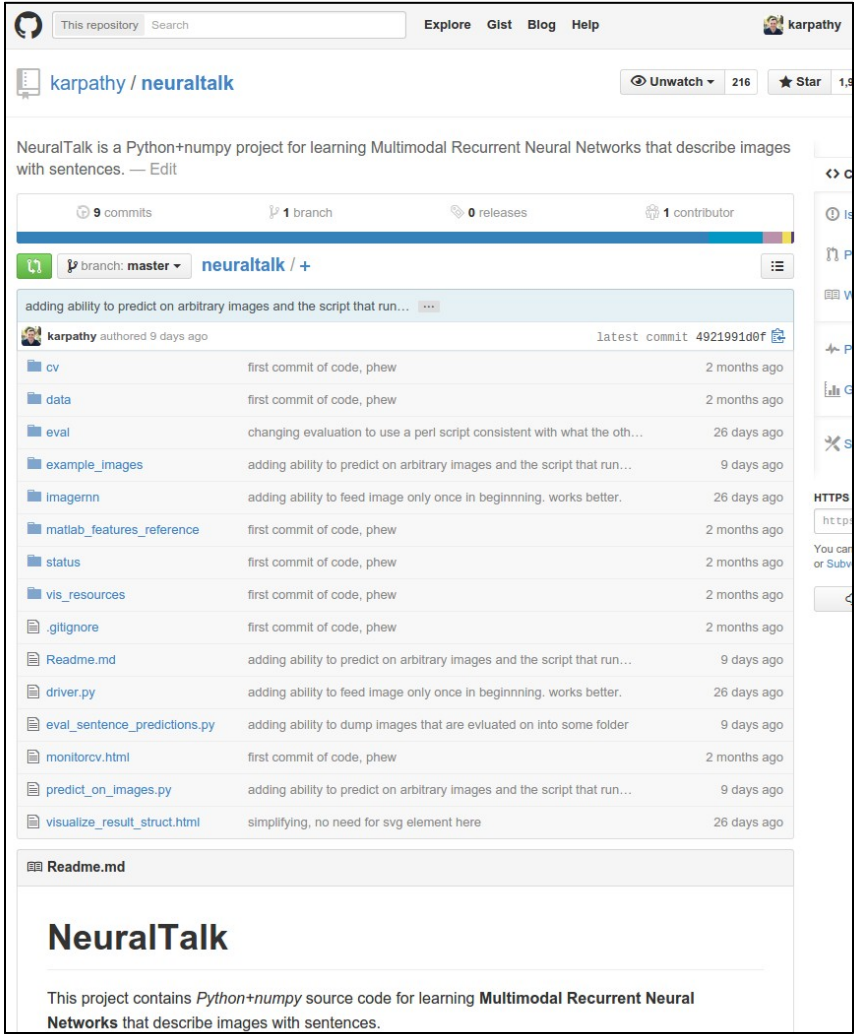
- A single glance is taken at the image, no objects are detected, etc.

- We can't just describe any image

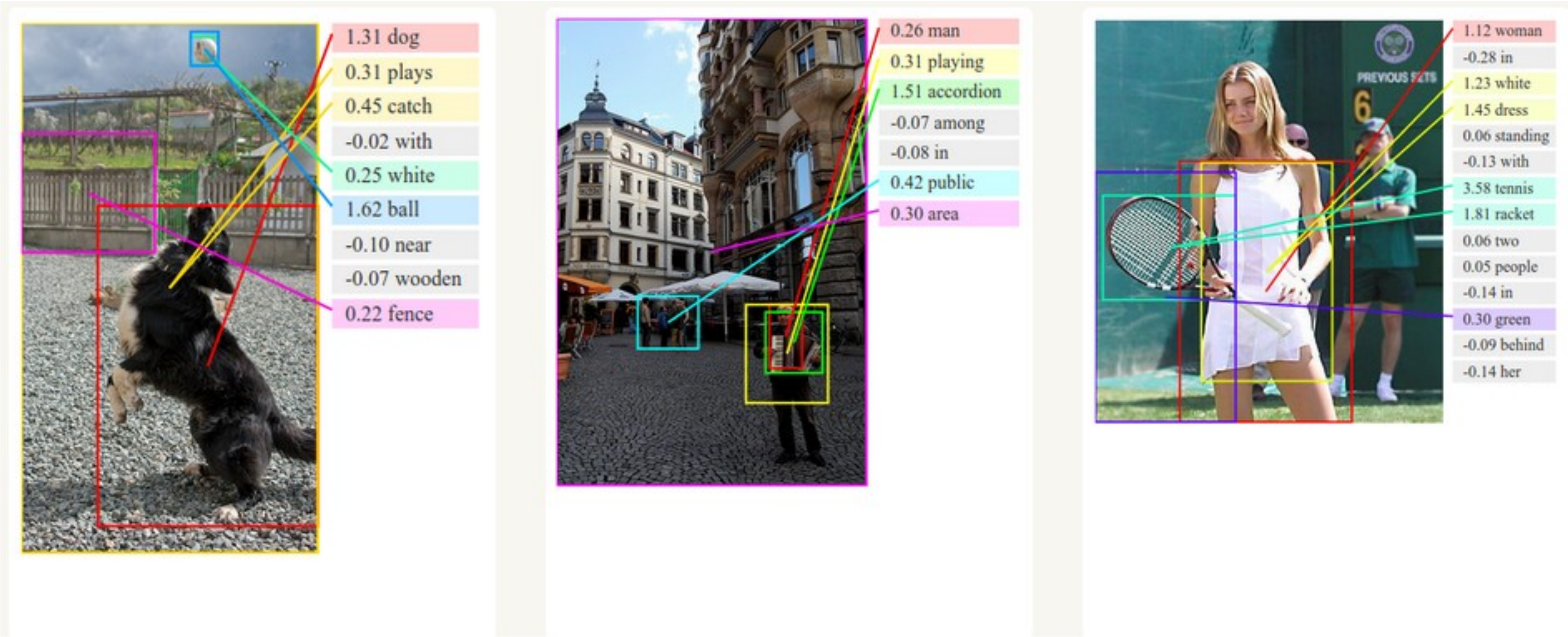
NeuralTalk

- Code on Github
- Both RNN/LSTM

- Python+numpy (CPU)
- Matlab+Caffe if you want to run on new images (for now)

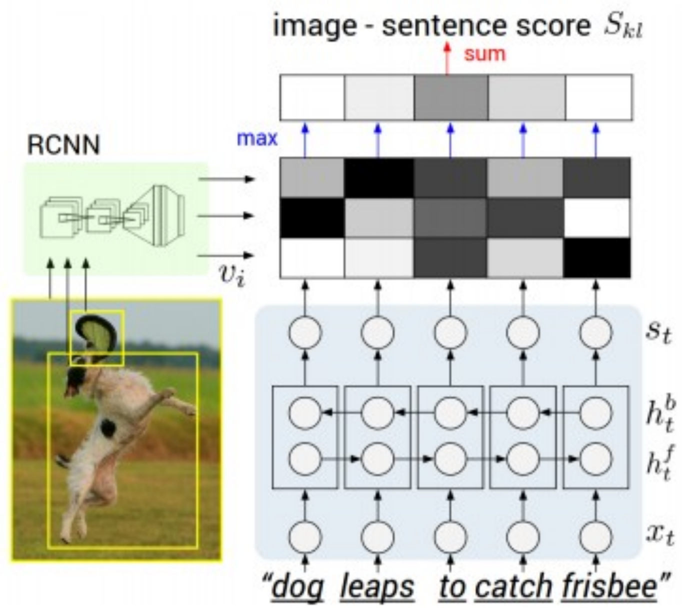


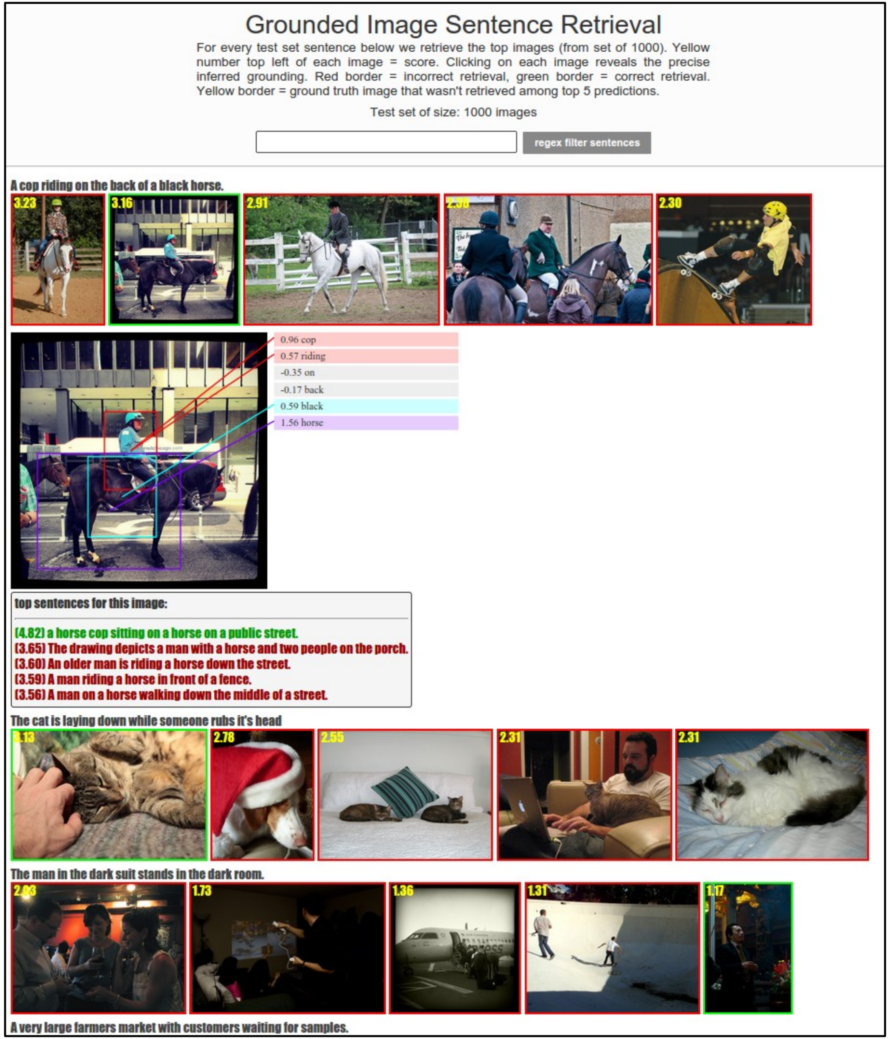
Ranking model



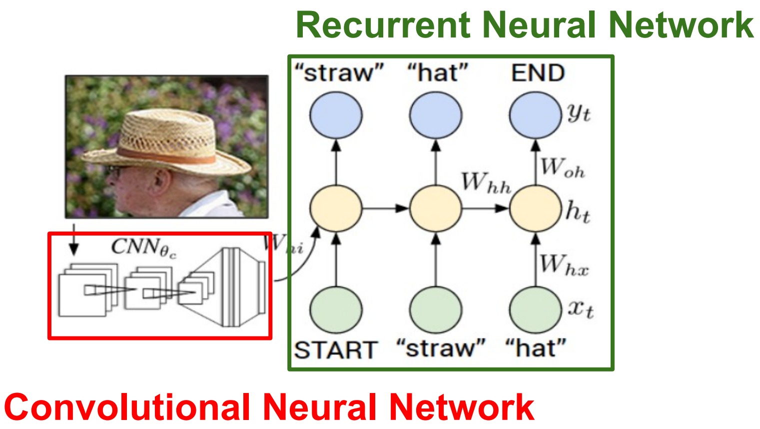
Ranking model web demo:

http://bit.ly/rankingdemo





Summary



Neural Networks:

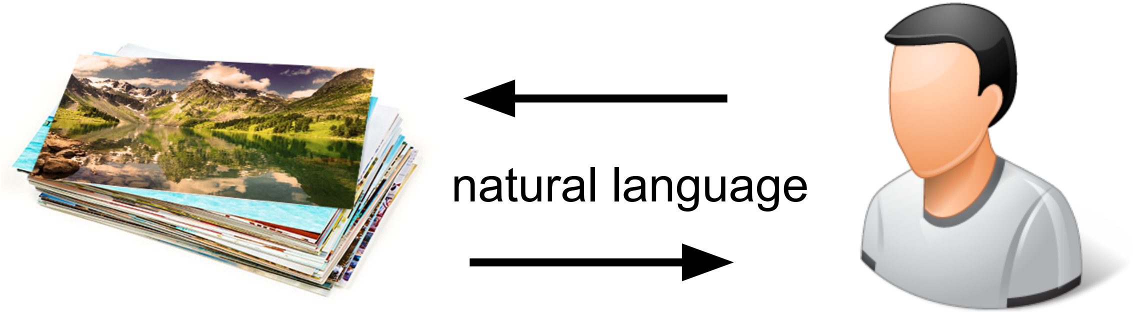
- input->output **end-to-end** optimization
- stackable / composable like Lego
- easily support Transfer Learning
- work very well.

Summary

1. image -> sentence
2. sentence -> image

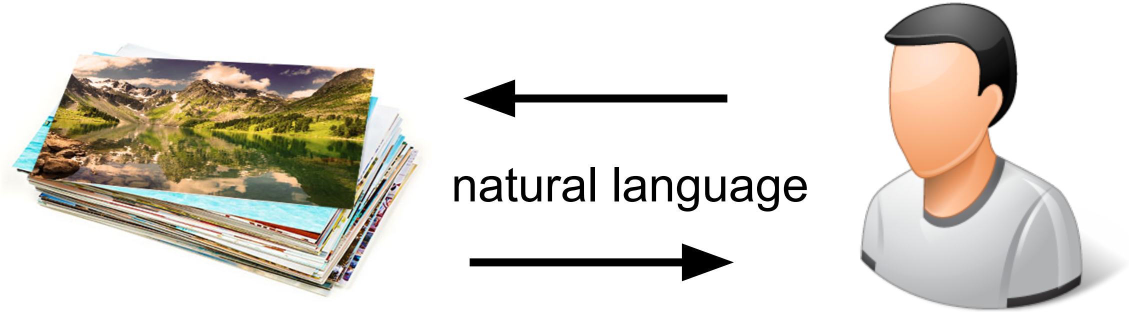
Summary

1. image -> sentence
2. sentence -> image



Summary

1. image -> sentence
2. sentence -> image



Thank you!