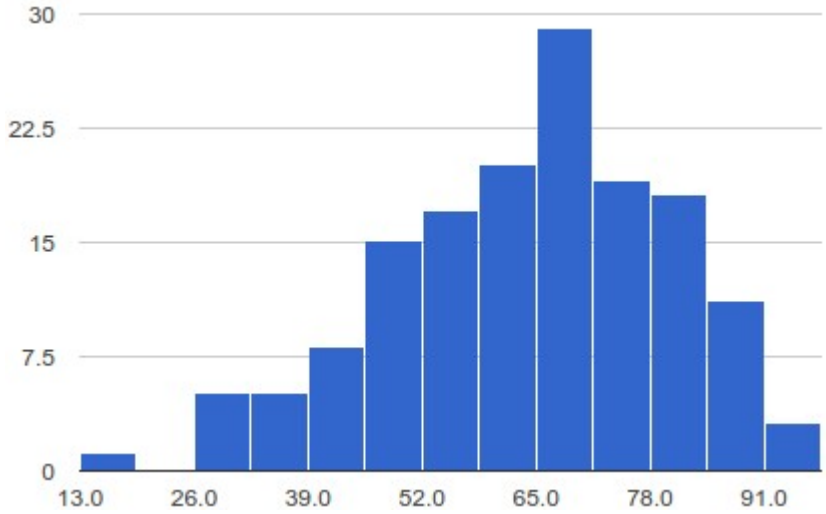
Administrative

- The midterm has been graded!
- Office hours: Andrej instead of Fei-Fei, 4pm, Fei-Fei office

Midterm statistics



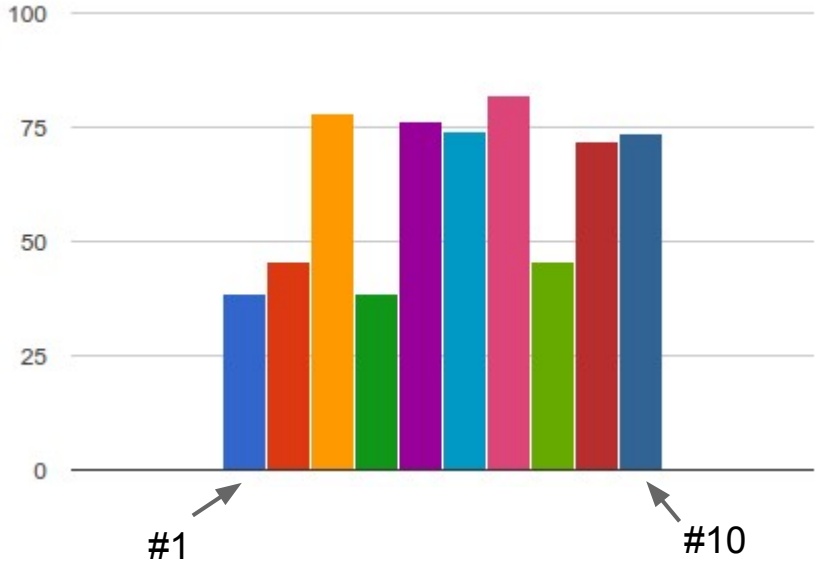
Mean: 64
Median: 65
Max grade: 94

top 10%: 84+
top 20%: 78+
top 30%: 73+
top 40%: 69+
top 50%: 65+

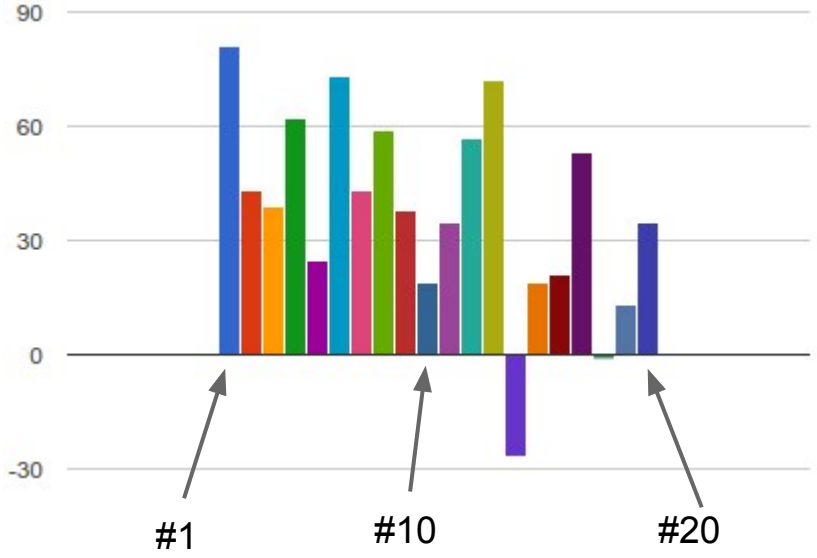
top scores descending:
94, 93, 92, 90, 90, 90, 89, 89, 86, 86, 86, 86, 85, 85, 84, 83, 83 ...

Midterm statistics

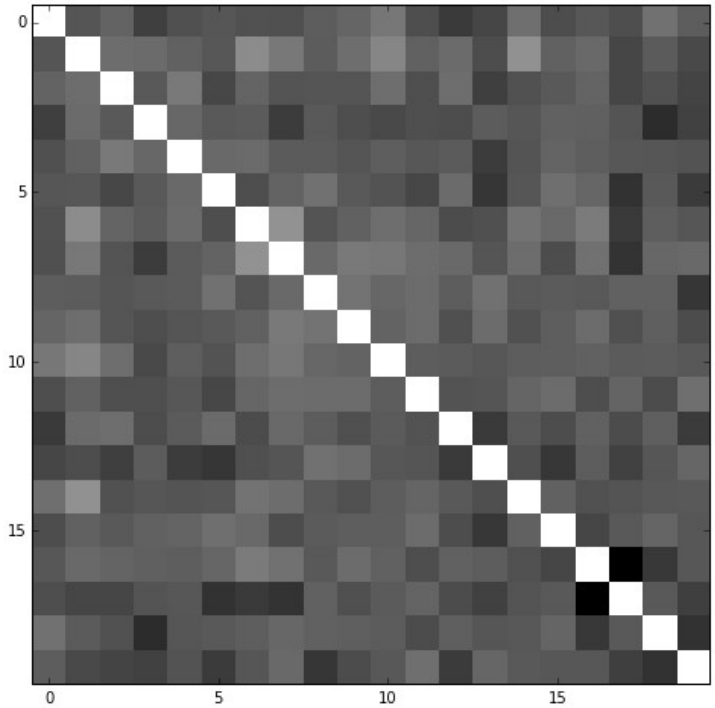
10 Multiple Choice questions:
average grade per question



20 True False questions:
average grade per question



Midterm statistics



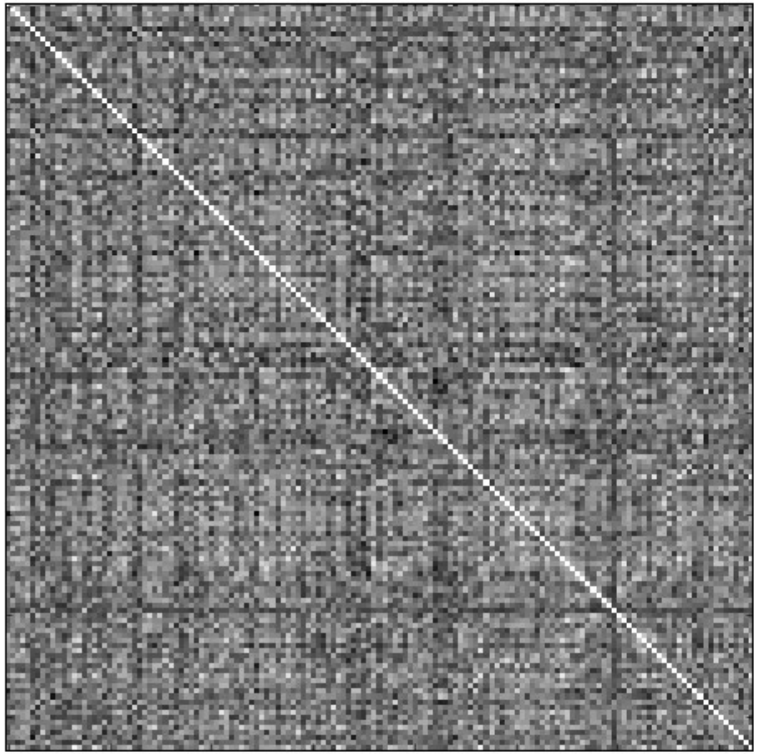
True False questions correlation matrix

Correlated T/F questions:

pair, correlation:
(7,8), 0.37
(2,15), 0.37
(2,7), 0.34
(2,11), 0.3
(7,17), 0.24

(17,18), -0.47
(4,19), -0.21

Midterm statistics



Can also do the same thing but swap axes:

Student Correlation Matrix

lowest correlation: -0.7
highest correlation: 0.95

fun notes:

- noone gave exact same TF answers
- noone got all TF questions correct

Midterm statistics

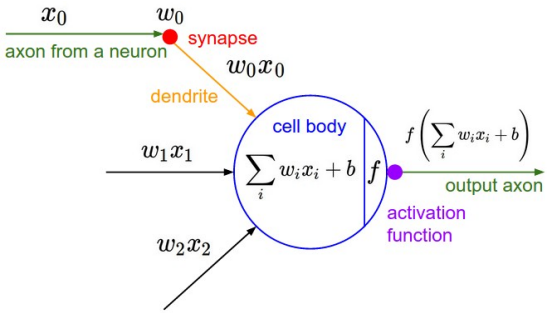
BONUS:
6 people got the Bonus Question
9 people gave admirable effort (>0 & < 3 points awarded)

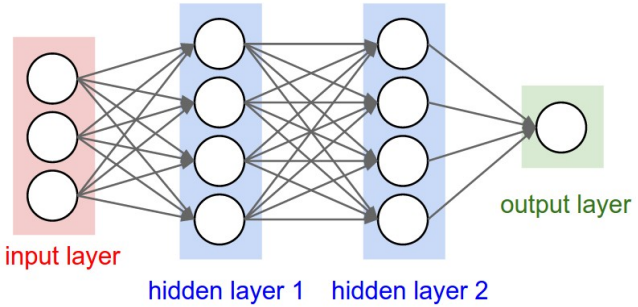
Lecture 11:

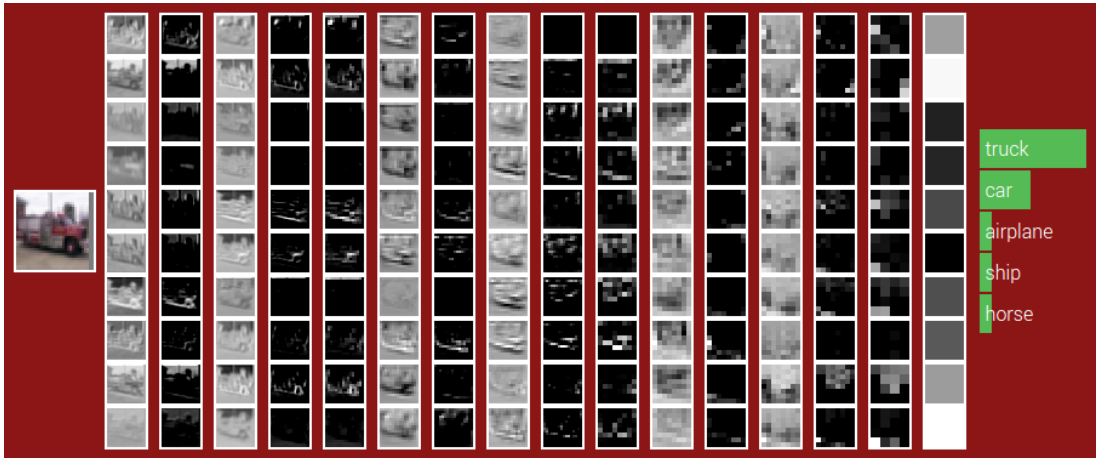
Beyond Image Classification

Where we are...

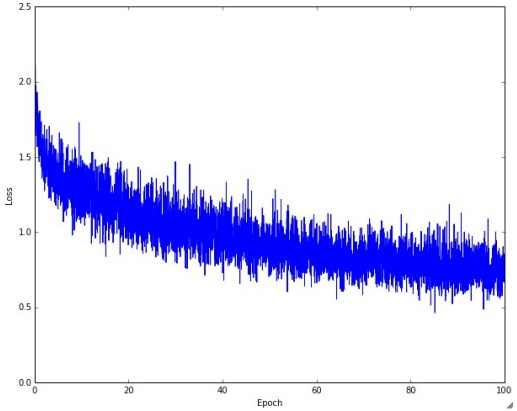
We've introduced (Convolutional) Neural Nets

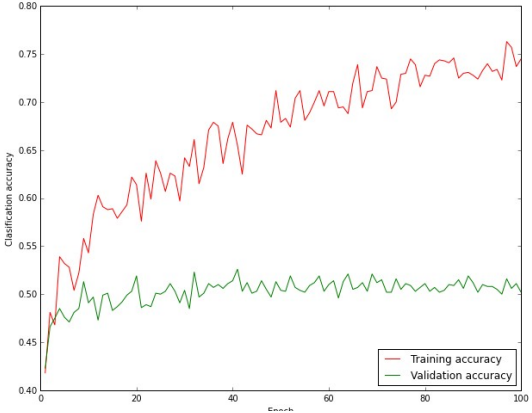


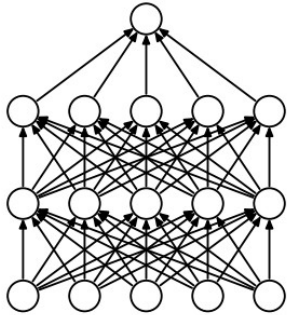




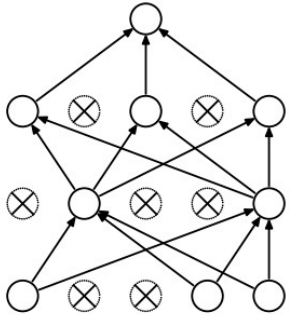
We've seen how to train them



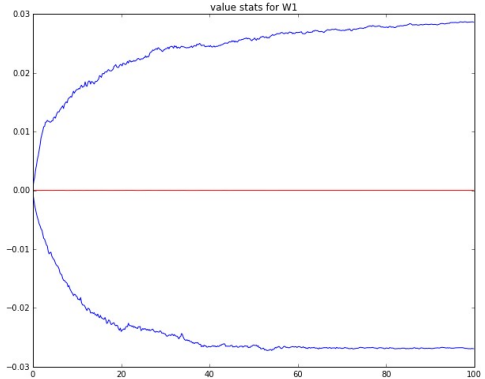


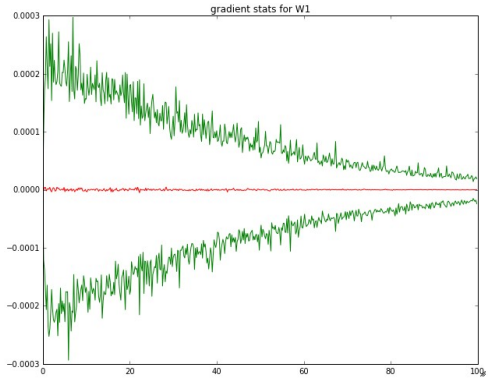


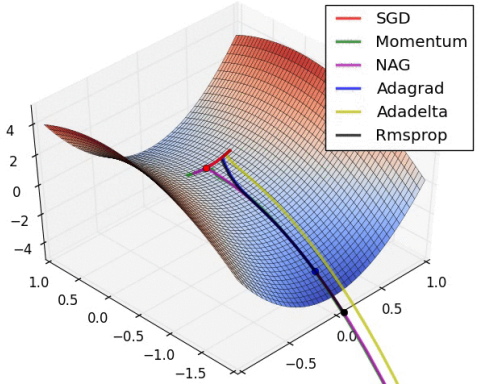
(a) Standard Neural Net



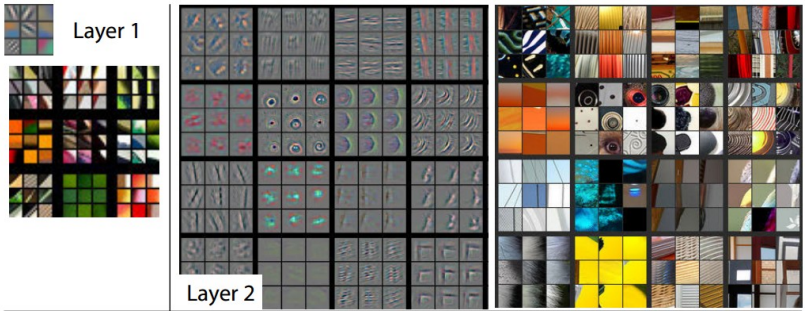
(b) After applying dropout.

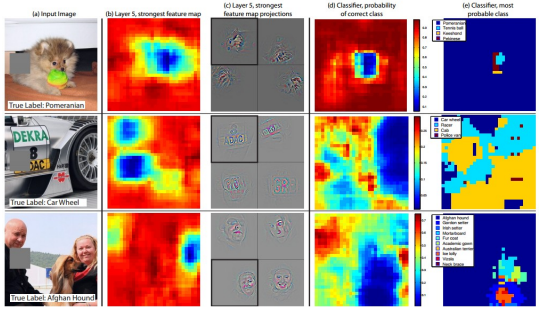


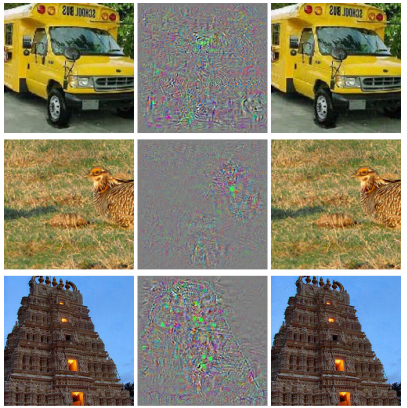


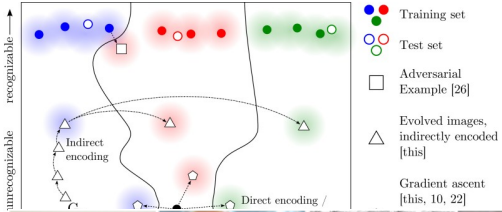


We've looked at how they work



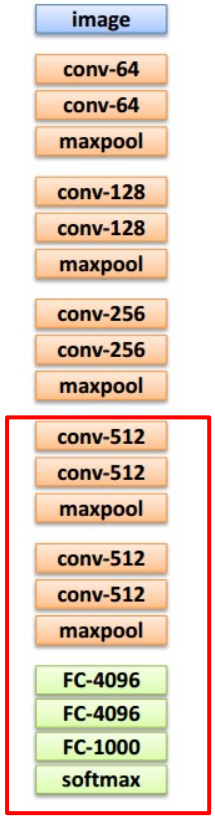


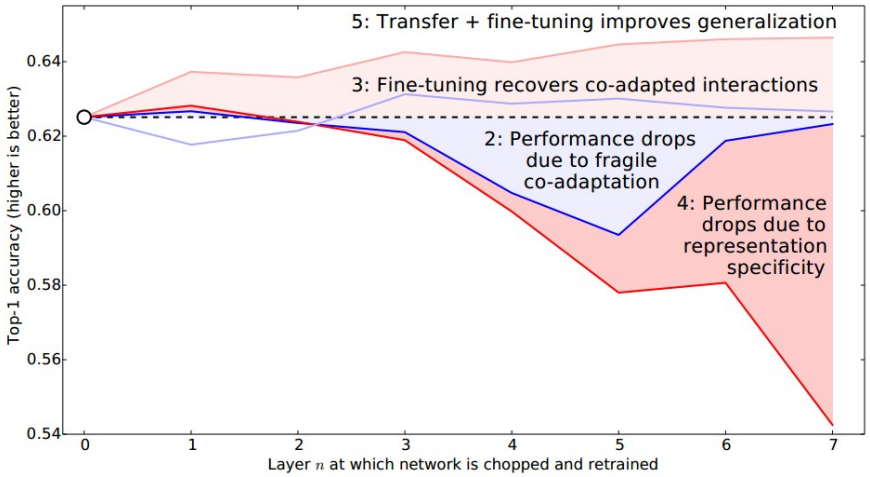


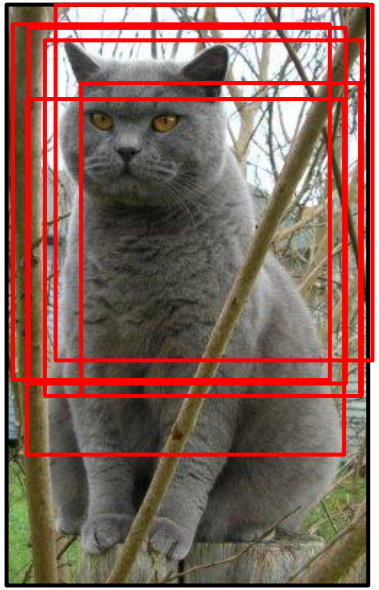


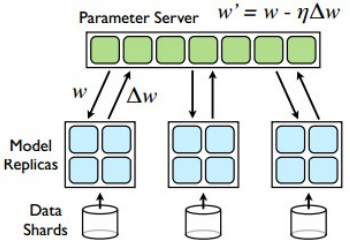
rule, ruler	king crab, Alaska crab	sidewinder	saltshaker, salt shaker	reel	hatchet	schipperke
pencil box, pencil case	pizza, pizza pie	maze, labyrinth	pill bottle	stethoscope	vase	schipperke
rubber eraser, rubber	strawberry	gar, garfish	water bottle	whistle	pitcher, ewer	groenendael
ballpoint, ballpoint pen	orange	valley, vale	lotion	ice lolly, lolly	coffeepot	doormat, welcome mat
pencil sharpener	fig	hammerhead	hair spray	hair spray	mask	teddy, teddy bear
carpenter's kit, tool kit	ice cream, icecream	sea snake	beer bottle	maypole	cup	jigsaw puzzle

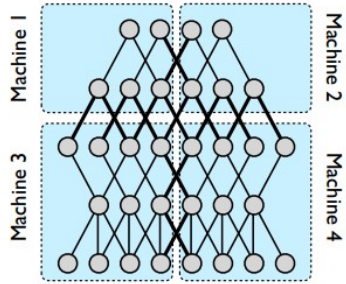
And how they are applied in practice

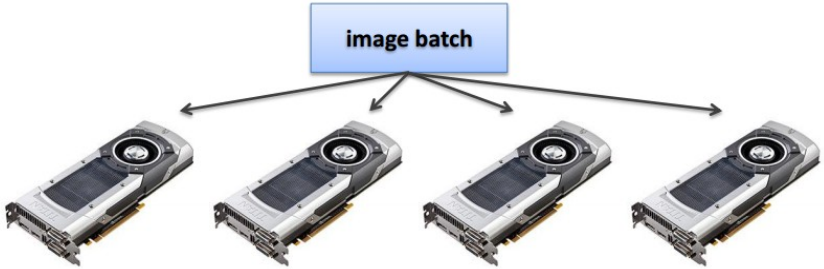












But one thing has remained the same...

(assume given set of discrete labels)
{dog, cat, truck, plane, ...}

cat

Lecture 11:

Beyond Image Classification

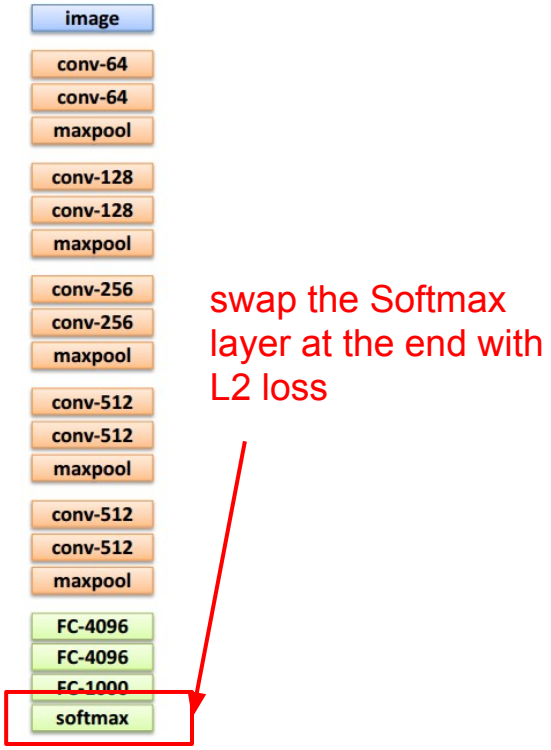
Localization

Model must output:

- class (integer)
- x1,y1,x2,y2 bounding box coordinates

Very Deep Convolutional Networks for Large-Scale Image Recognition,
Simonyan et al., 2014
OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks,
Sermanet et al., 2014

Idea: train a Localization net
Take out Softmax loss, swap in L2 (regression) loss, **fine-tune** the classification network.



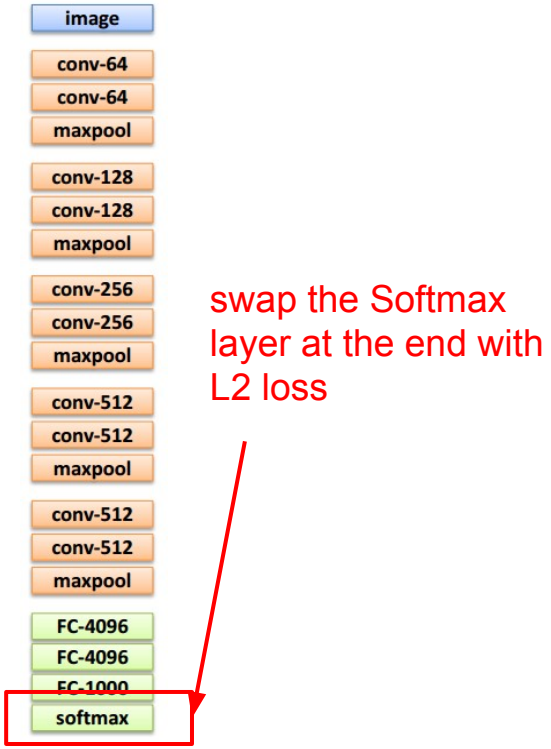
Very Deep Convolutional Networks for Large-Scale Image Recognition,
Simonyan et al., 2014
OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks,
Sermanet et al., 2014

Idea: train a Localization net
Take out Softmax loss, swap in L2 (regression) loss, **fine-tune** the classification network.

predictions: instead of class scores, now interpreted as the 4 bounding box coords **(also 4D vector from net)**

targets: true bounding box **4D vector of [x1,y1,x2,y2]**

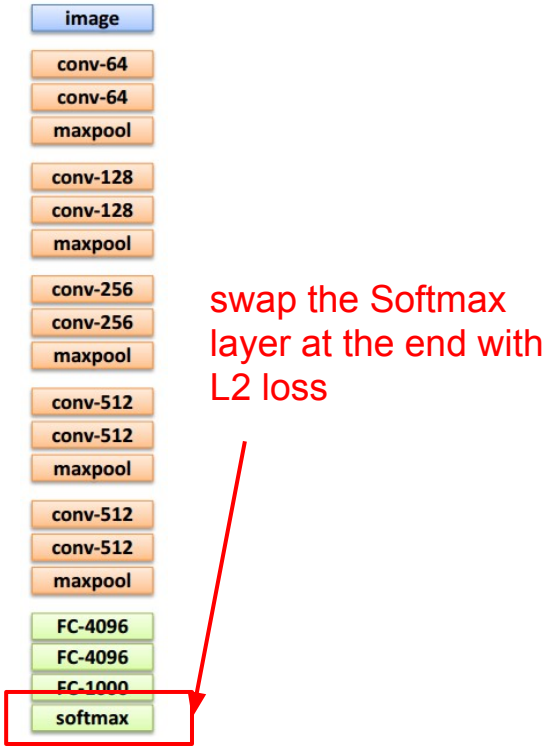
$$L_i = \|f - y_i\|_2^2$$

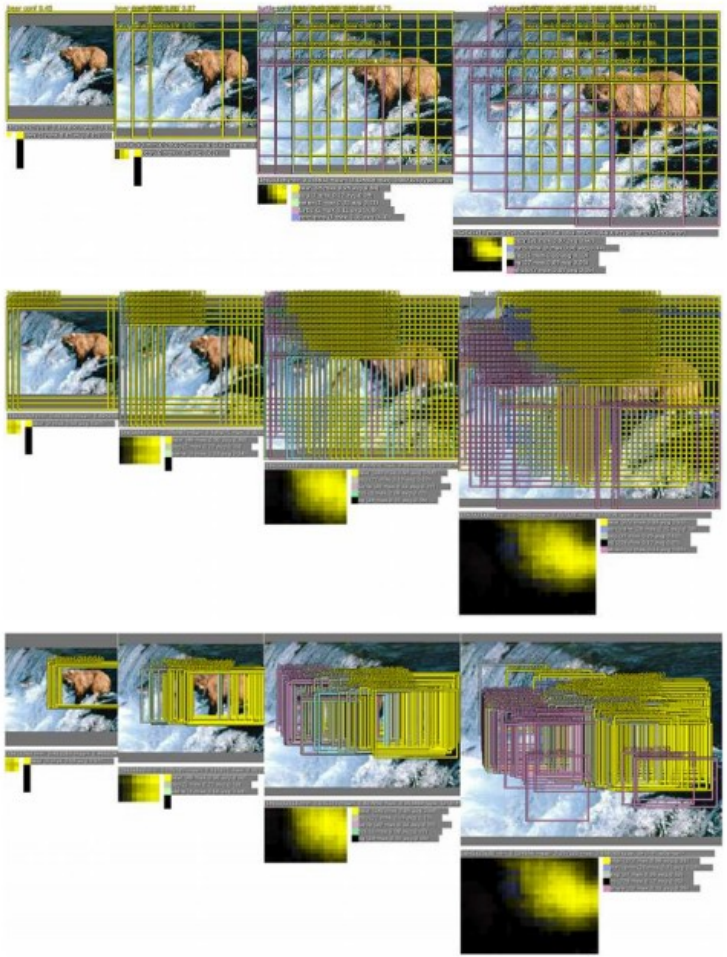


Very Deep Convolutional Networks for Large-Scale Image Recognition,
Simonyan et al., 2014
OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks,
Sermanet et al., 2014

In practice:

- It works better to predict a **4D vector for every class** (e.g. 4000D vector for 1000 ImageNet classes). During training only backprop the loss for the correct class
- apply at **multiple locations and scales**





OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks,
Sermanet et al., 2014

greedy merging procedure

Tip: Note that we can apply ConvNets over all positions in an image very efficiently

Normal ConvNet:

[224x224x3] -> … -> [7x7x512] -> [1x1x4096] -> … -> [1x1x1000]

last volume first FC layer class scores

Tip: Note that we can apply ConvNets over all positions in an image very efficiently

Normal ConvNet:

[224x224x3] -> … -> [7x7x512] -> [1x1x4096] -> … -> [1x1x1000]

last volume first FC layer class scores

Convert the first FC layer into a CONV layer:

Note: This FC layer is equivalent to CONV layer with:

receptive field size of 7x7, pad 0, stride 1, and 4096 neurons

Convert later FC layers: receptive field sizes 1x1, pad 0, stride 1

Tip: Note that we can apply ConvNets over all positions in an image very efficiently

Normal ConvNet:

[224x224x3] -> … -> [7x7x512] -> [1x1x4096] -> … -> [1x1x1000]

last volume first FC layer class scores

Convert the first FC layer into a CONV layer:

Note: This FC layer is equivalent to CONV layer with:

receptive field size of 7x7, pad 0, stride 1, and 4096 neurons

Convert later FC layers: receptive field sizes 1x1, pad 0, stride 1

Modified ConvNet:

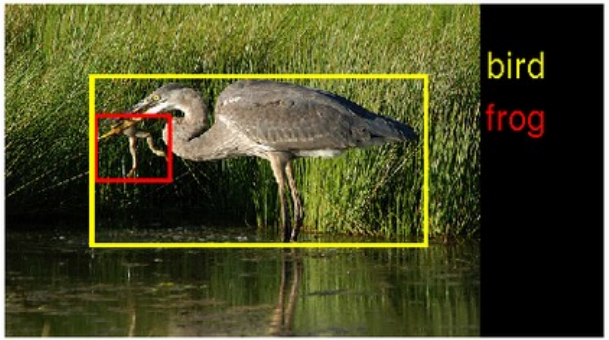
[384x384x3] -> … -> [**12**x**12**x512] -> [**6**x**6**x4096] -> … -> [**6**x**6**x1000]

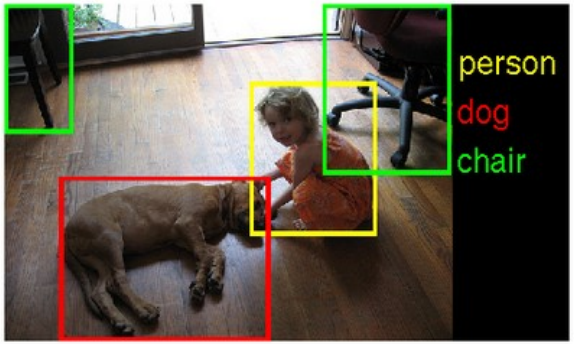
last volume 7x7 CONV 1x1 CONV

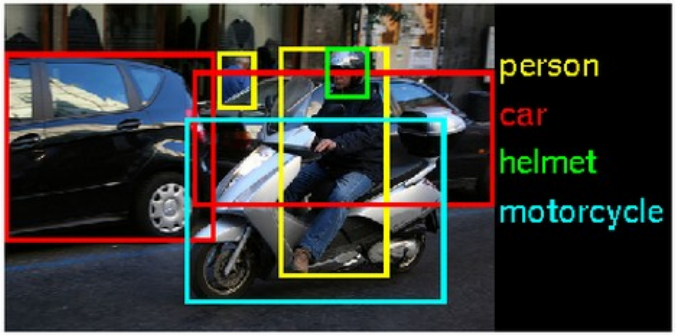
class scores volume!

This is very efficient, can be seen as evaluating the FC layer in parallel on many locations in the image. This is a common trick. Same trick used in localization.

Detection







Model must output:

A set of detections

Each detection has:

- confidence
- class (integer)
- x1,y1,x2,y2 bounding box coordinates

Rich feature hierarchies for accurate object detection and semantic segmentation
[Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik]

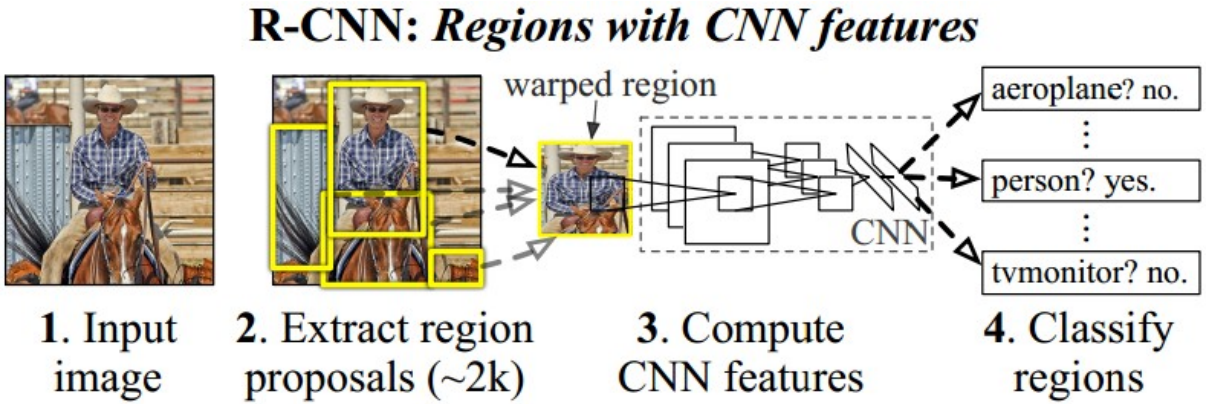
***Idea**: Turn a Detection Problem into an Image Classification problem (but over image regions).*

Content of every labeled bounding box for is a positive example for a class.

Every other bounding box in the image is a special **negative class**.

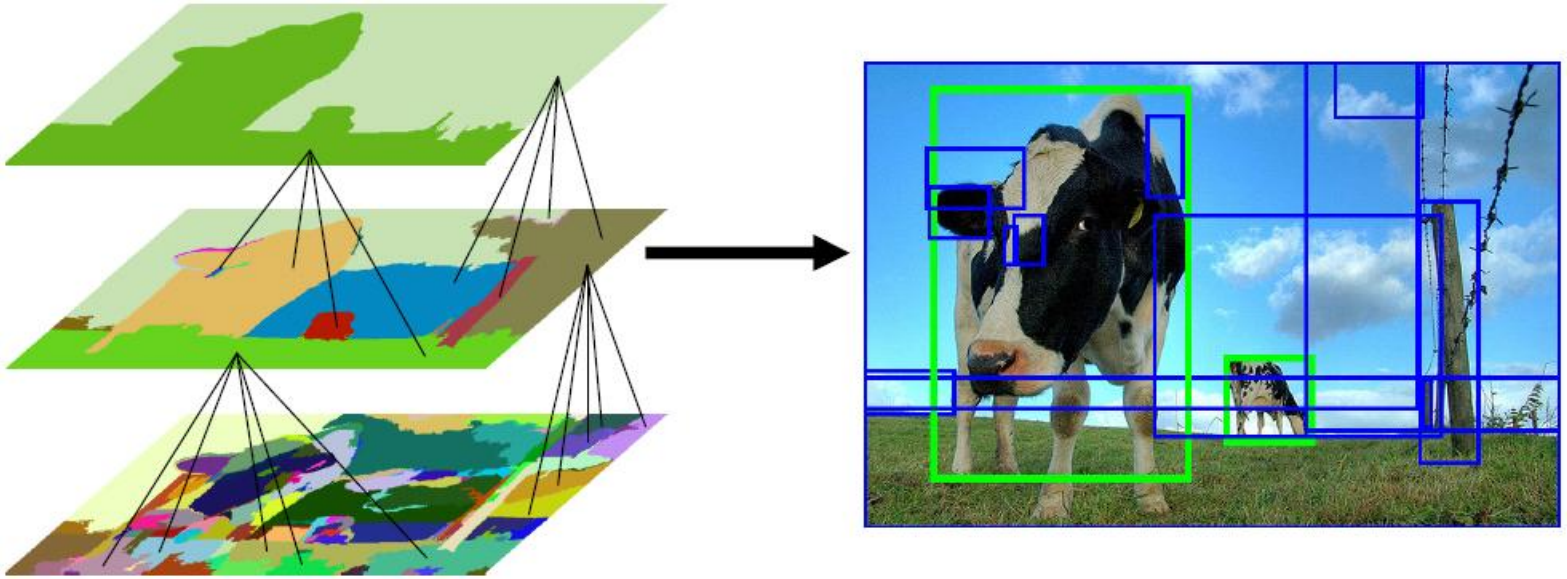
Rich feature hierarchies for accurate object detection and semantic segmentation
[Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik]

***Idea**: Turn a Detection Problem into an Image Classification problem (but over image regions).*



Selective Search for Object Recognition

[J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, A. W. M. Smeulders]

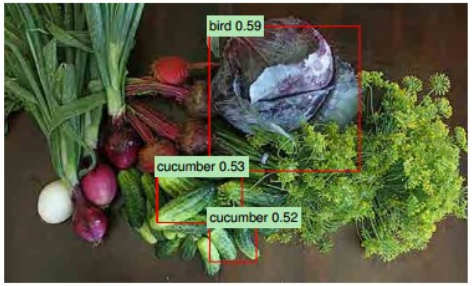


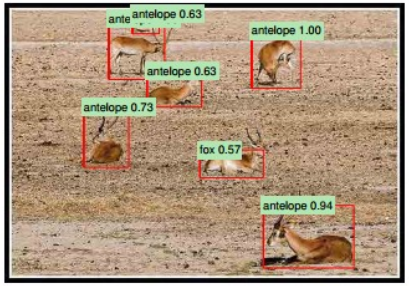
Gives on average ~2,000 candidate region proposals per image.

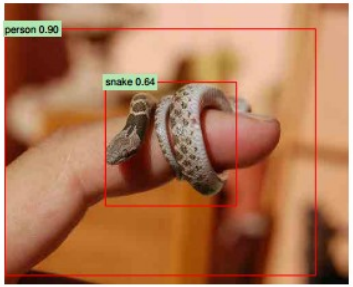
(This paradigm currently outperform the "sliding window" approach)

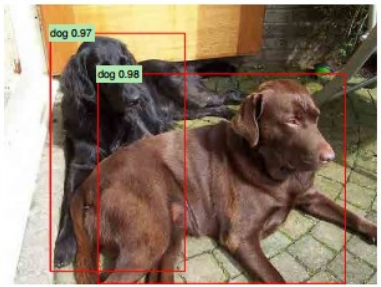
Rich feature hierarchies for accurate object detection and semantic segmentation

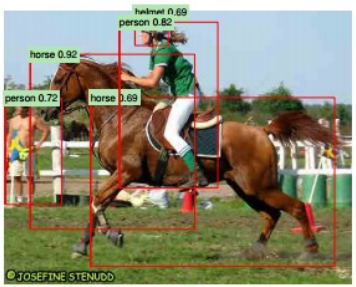
[Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik]

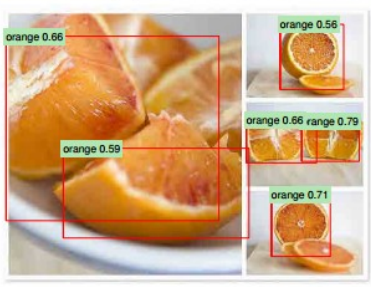


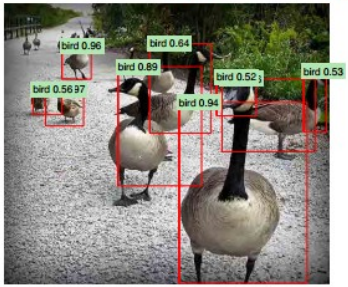


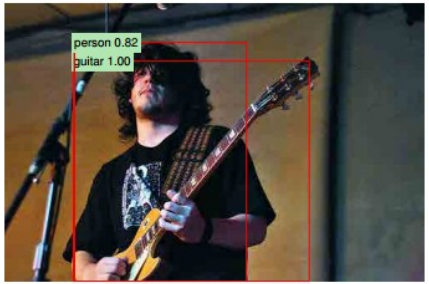


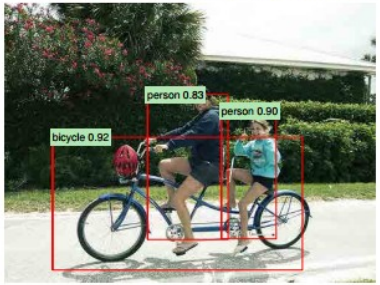


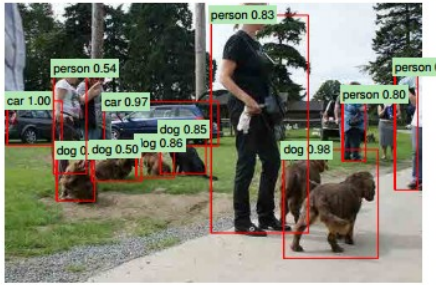




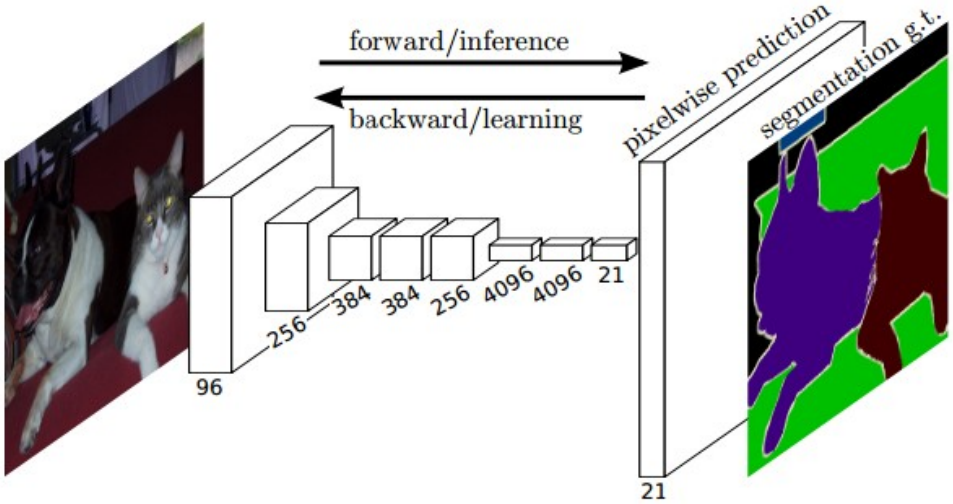


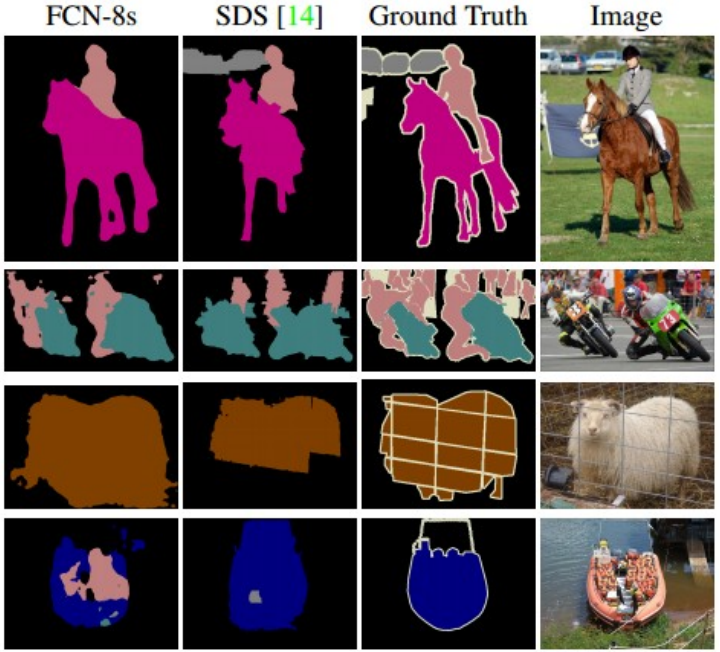






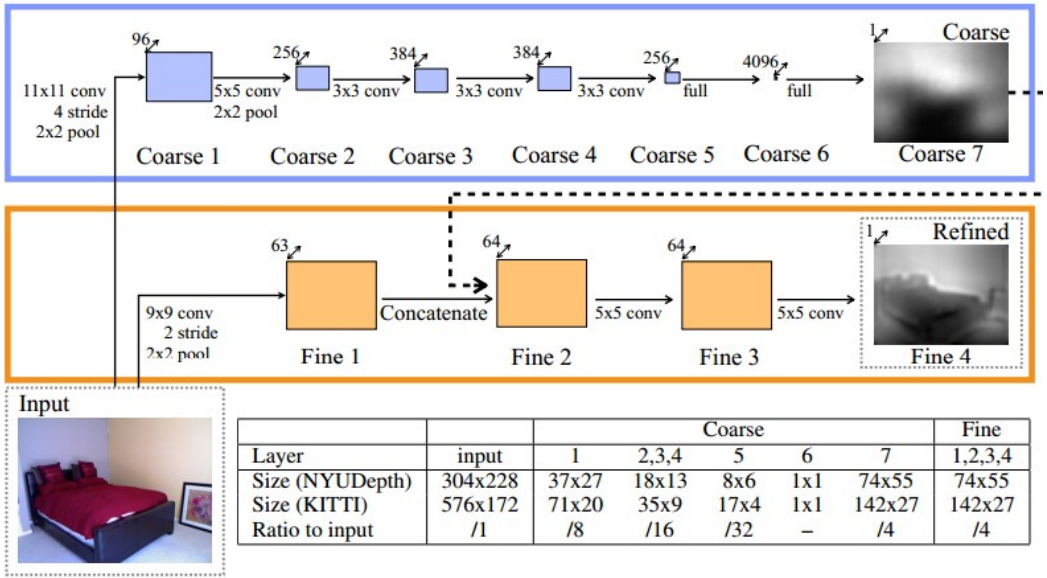
Segmentation



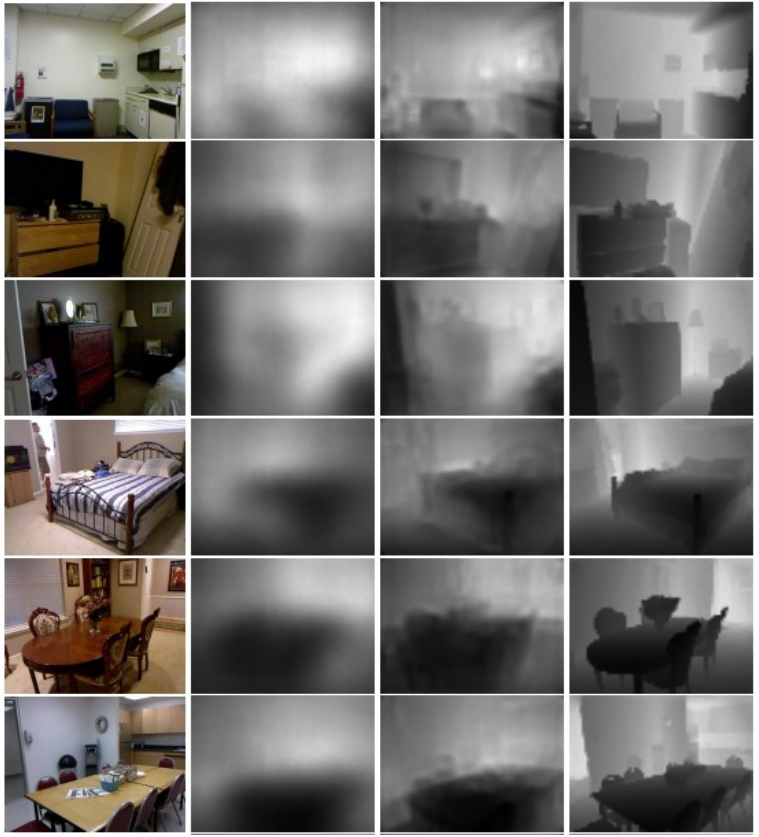


Fully Convolutional Networks for Semantic Segmentation
Long, Shelhamer, Darrell

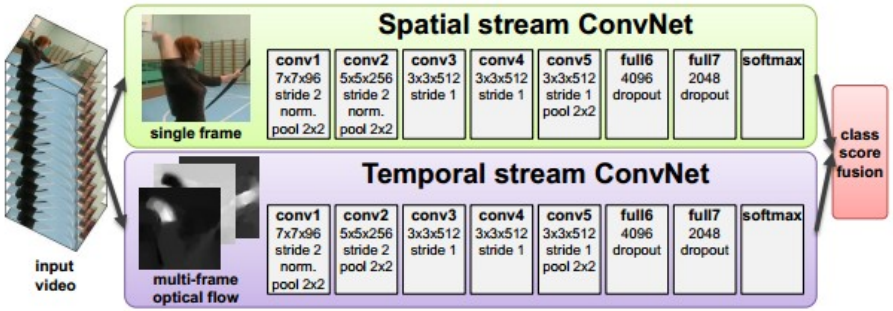
Depth Map Prediction from a Single Image using a Multi-Scale Deep Network
[Eigen et al.], 2014



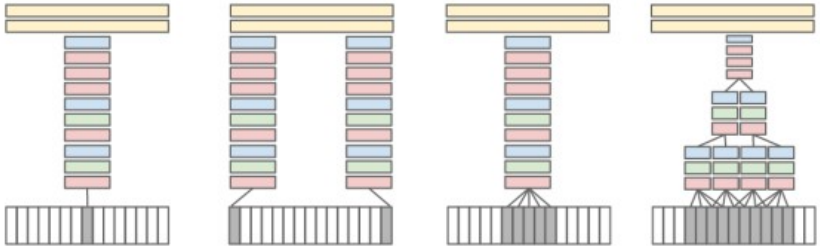
		Coarse					Fine
Layer	input	1	2,3,4	5	6	7	1,2,3,4
Size (NYUDepth)	304x228	37x27	18x13	8x6	1x1	74x55	74x55
Size (KITTI)	576x172	71x20	35x9	17x4	1x1	142x27	142x27
Ratio to input	/1	/8	/16	/32	–	/4	/4



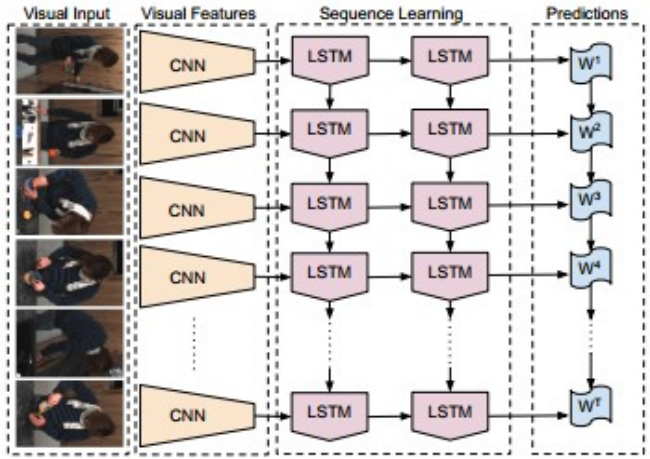
Video Classification



Two-Stream Convolutional Networks for Action Recognition in Videos *[Simonyan et al.], 2014*



Large-scale Video Classification with Convolutional Neural Networks
[Karpathy et al.], 2014



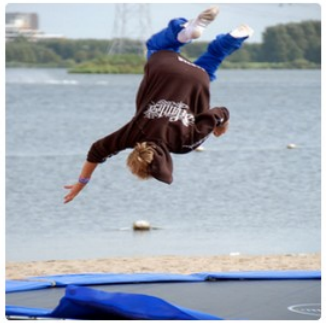
Long-term Recurrent Convolutional Networks for Visual Recognition and Description
[Donahue et al.], 2014

Image Captioning

"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."

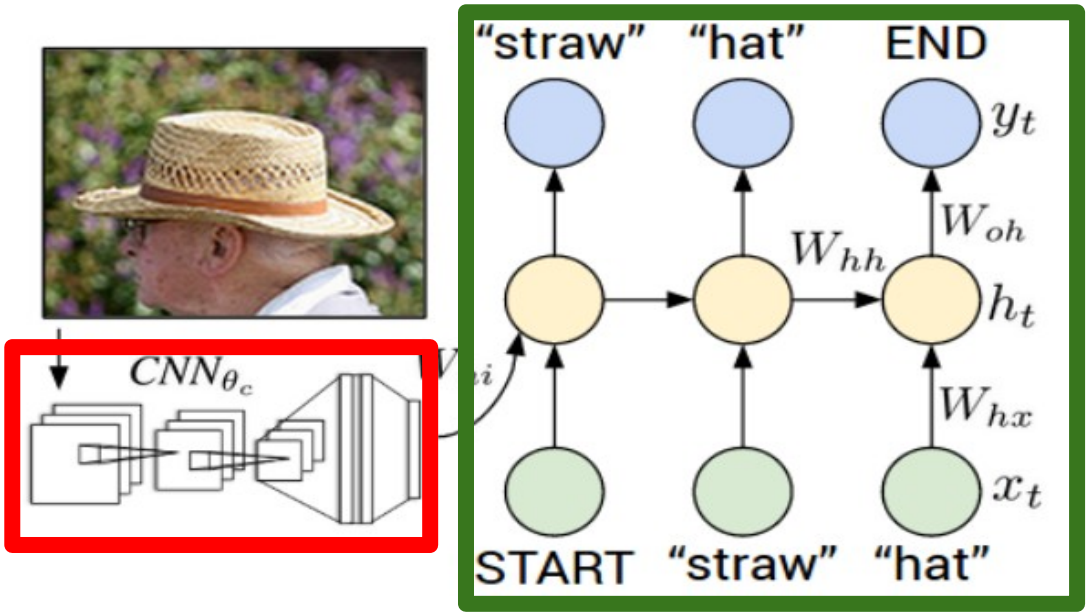
"black and white dog jumps over bar."

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

"man in blue wetsuit is surfing on wave."

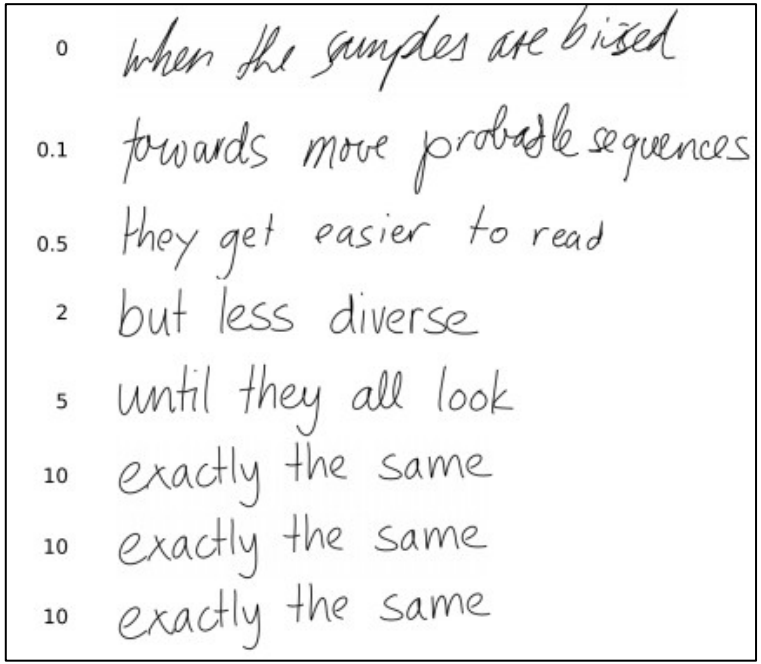
Neural Networks practitioner

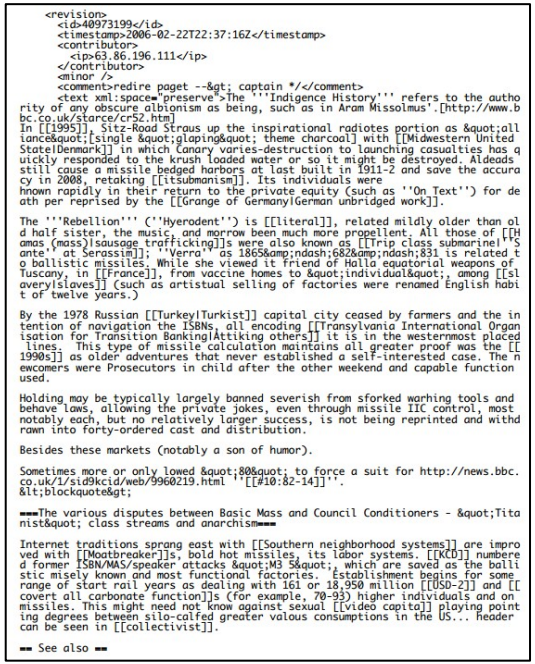
Recurrent Neural Network



Convolutional Neural Network

Recurrent Networks are good at modeling sequences...

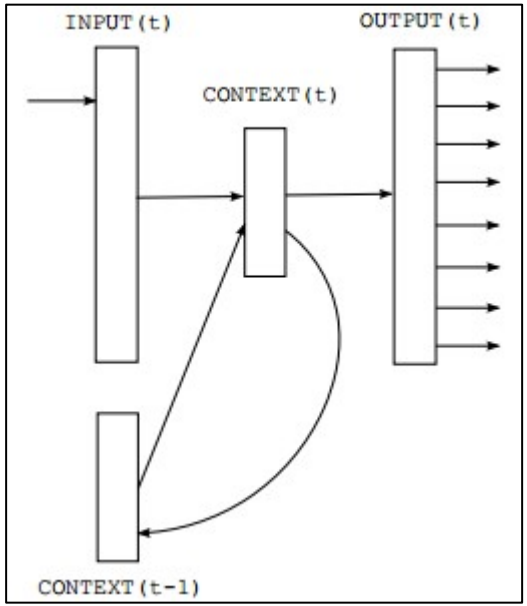




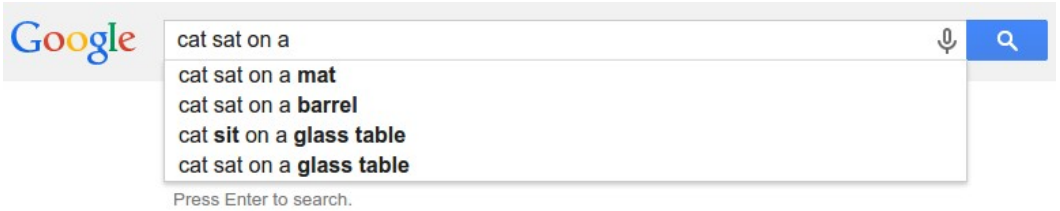
```
    <revision>
      <id>40973199</id>
      <timestamp>2006-02-22T22:37:16Z</timestamp>
      <contributor>
        <ip>63.86.196.111</ip>
      </contributor>
      <minor />
      <comment>redire paget --&gt; captain */</comment>
      <text xml:space="preserve">The '''Indigence History''' refers 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 S&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...



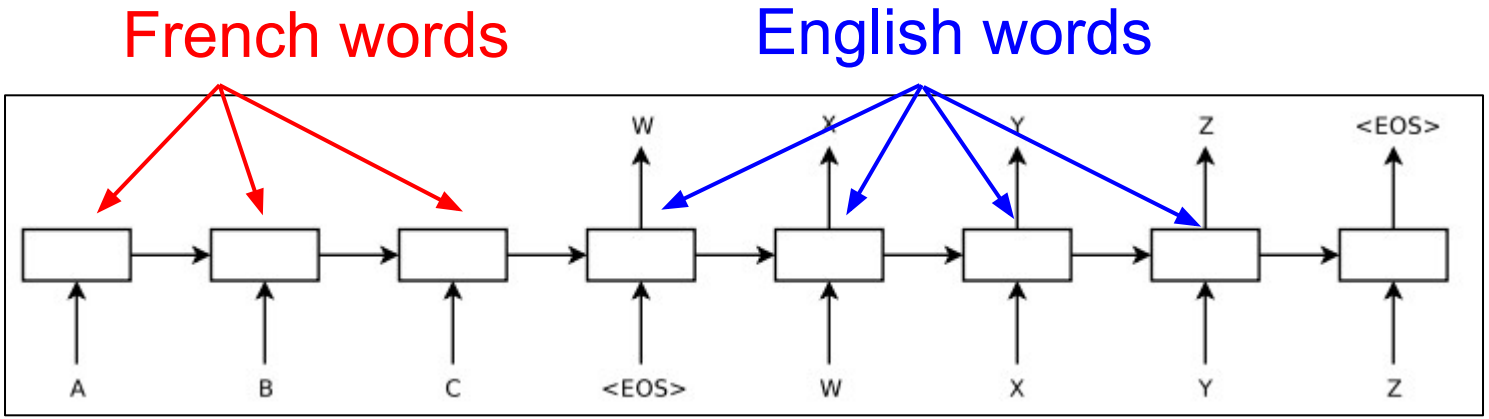
Word-level language model. Similar to:



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]

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])
P(<E>| [<S>, cat, sat, on, mat])

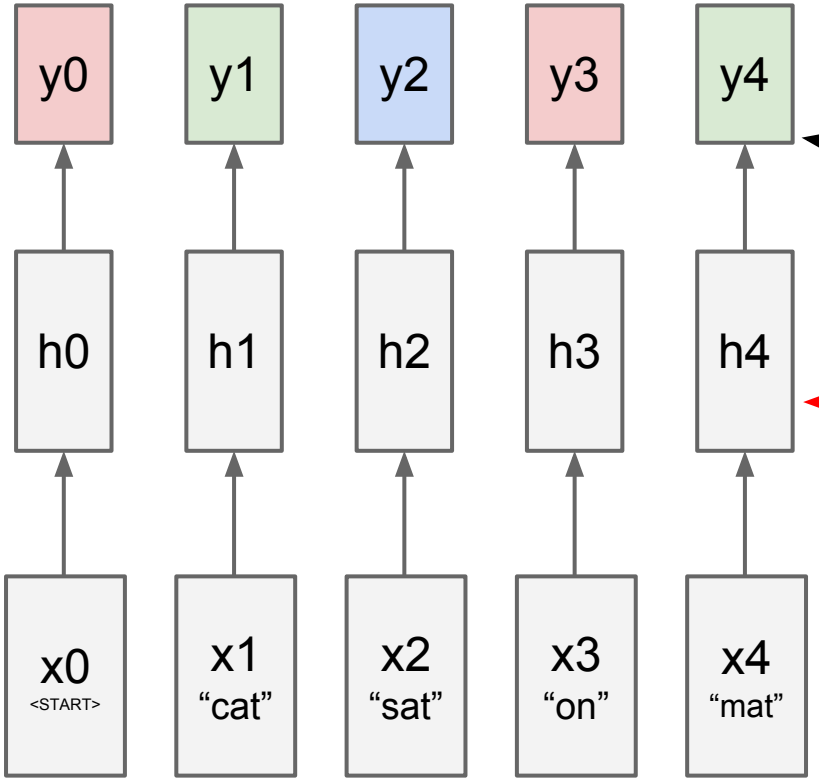
Suppose we had the training sentence “cat sat on mat”

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

First, suppose we had only a finite, 1-word history:
i.e. want these to be high:
P(cat | <S>)
P(sat | cat)
P(on | sat)
P(mat | on)
P(<E>| mat)

Vanilla 2-layer classification net for each word given previous word:

"cat sat on mat"



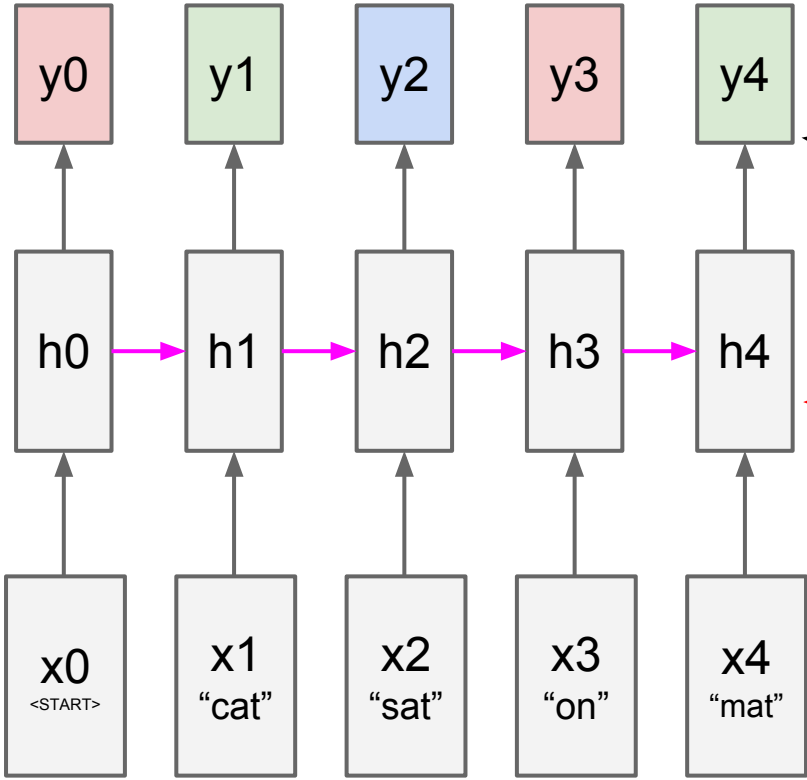
10,001-D class scores:
Softmax over 10,000 words and a special <END> token.
y4 = Why * h4

hidden layer (e.g. 500-D vectors)
h4 = max(0, Wxh * x4)

300 (learnable) numbers associated with each word in vocabulary

Turn it into RNN: (#anticlimatic)

"cat sat on mat"



10,001-D class scores:
Softmax over 10,000 words and
a special <END> token.
y4 = Why * h4

hidden layer (e.g. 500-D vectors)
h4 = max(0, Wxh * x4 + Whh * h3)

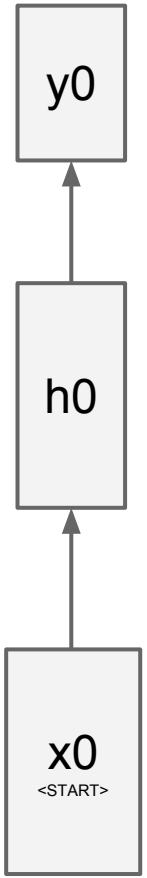
300 (learnable) numbers
associated with each word
in vocabulary

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

P(next word | previous words)

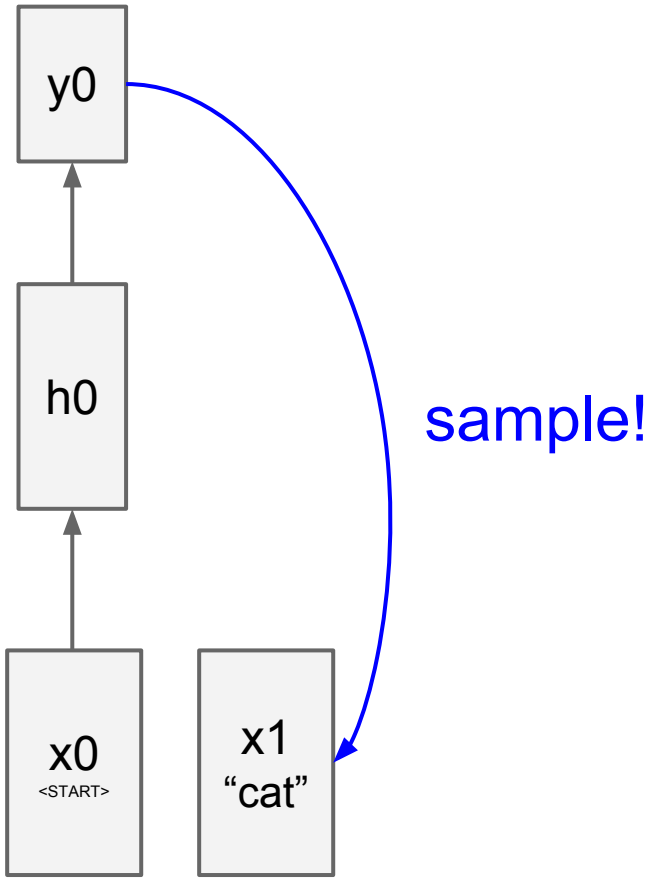
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



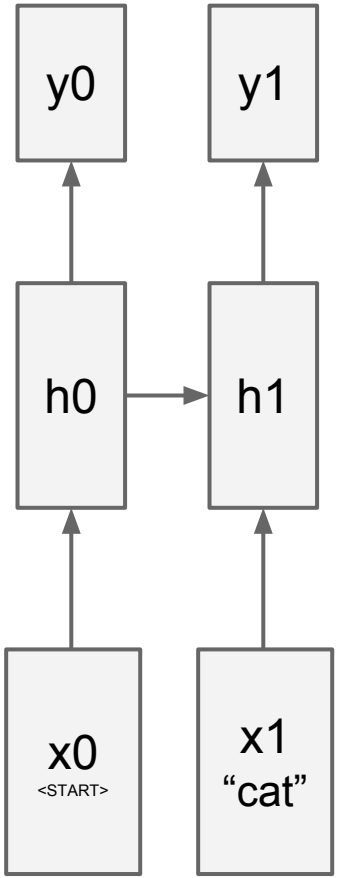
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



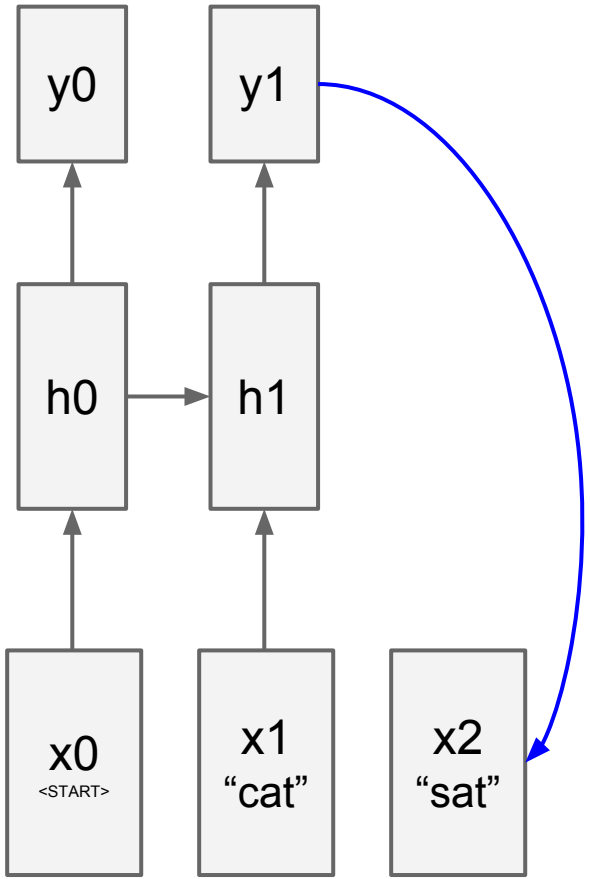
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



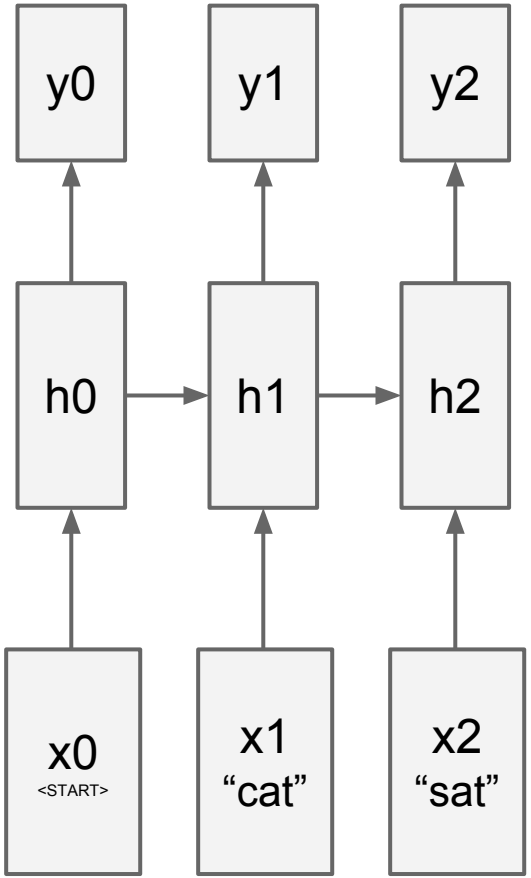
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



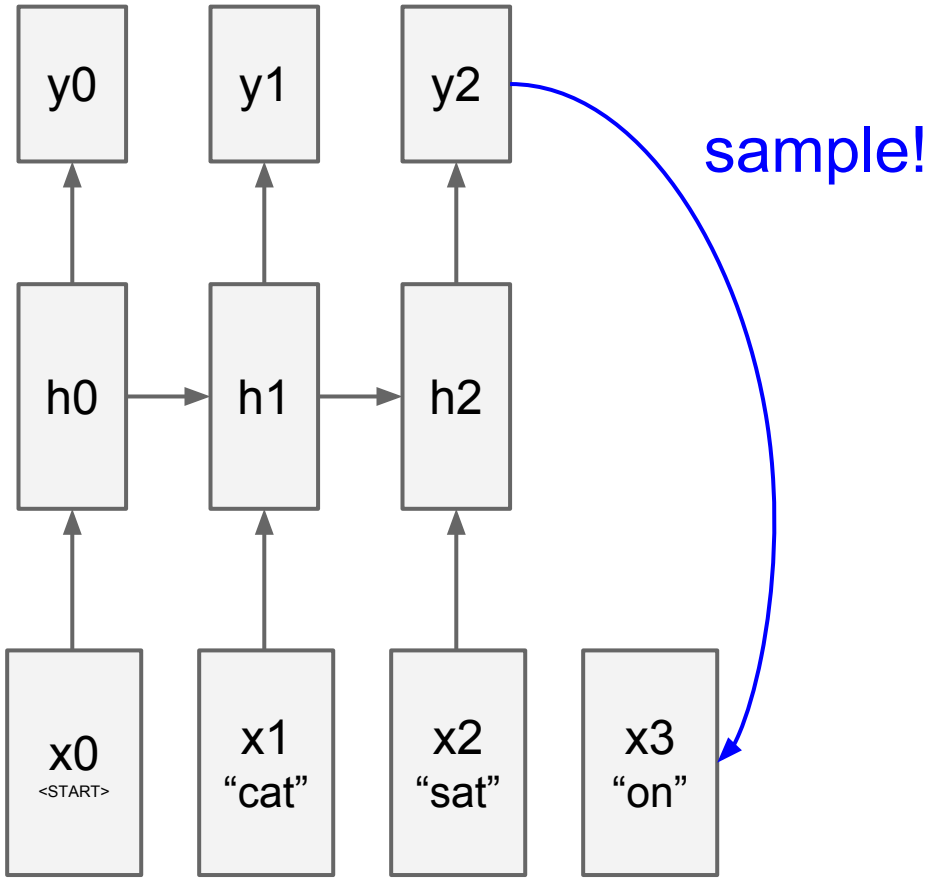
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



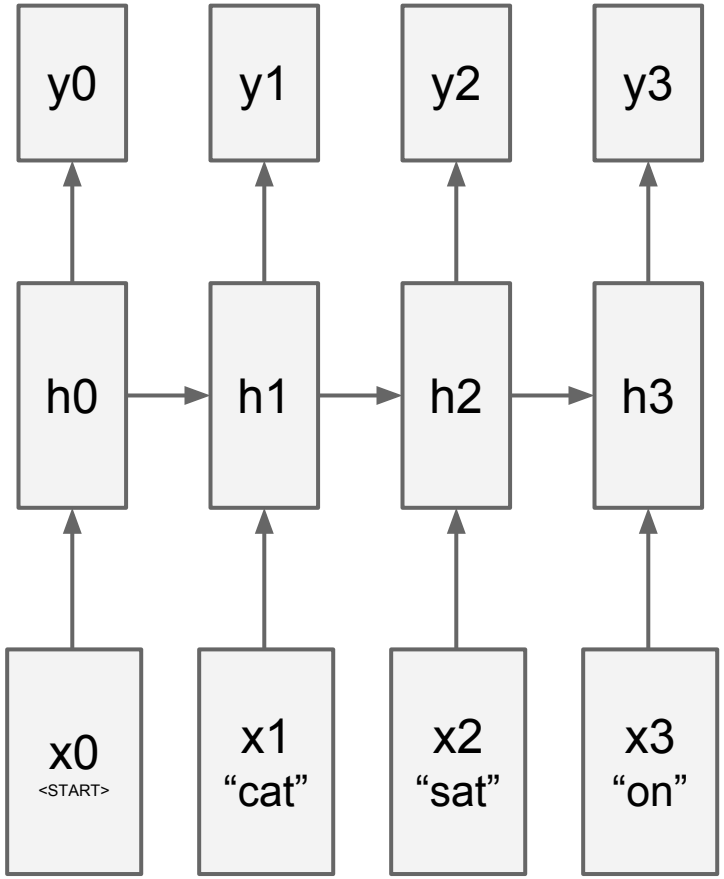
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



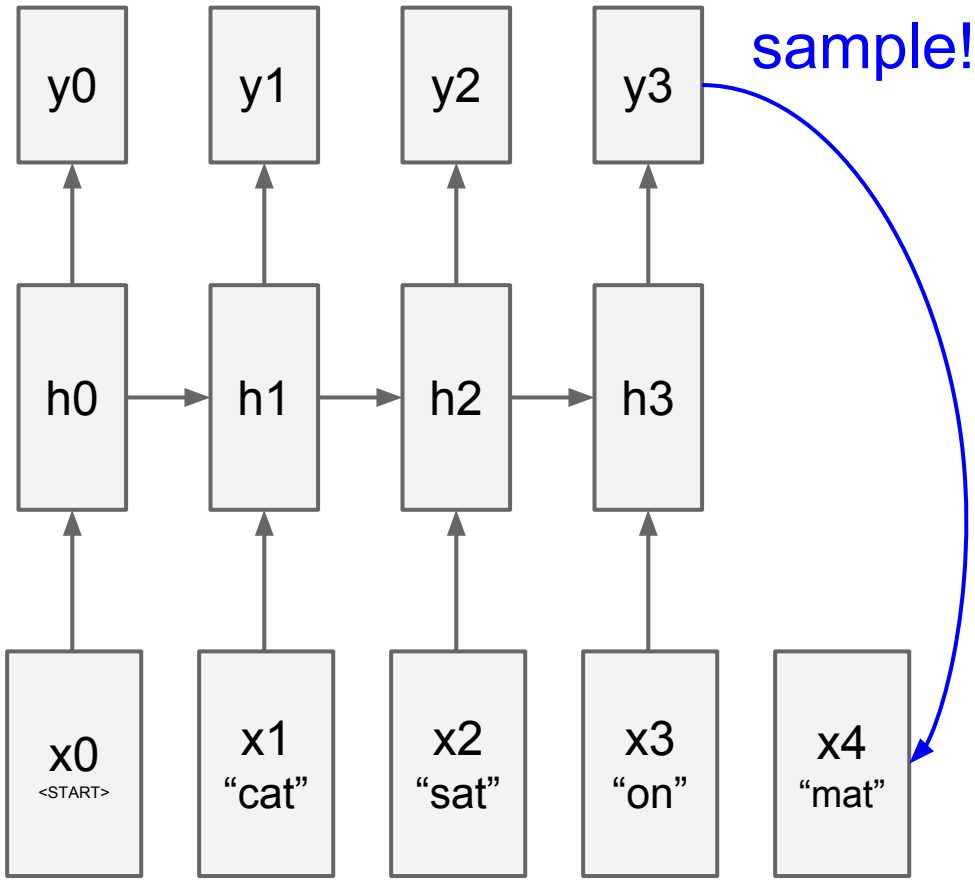
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



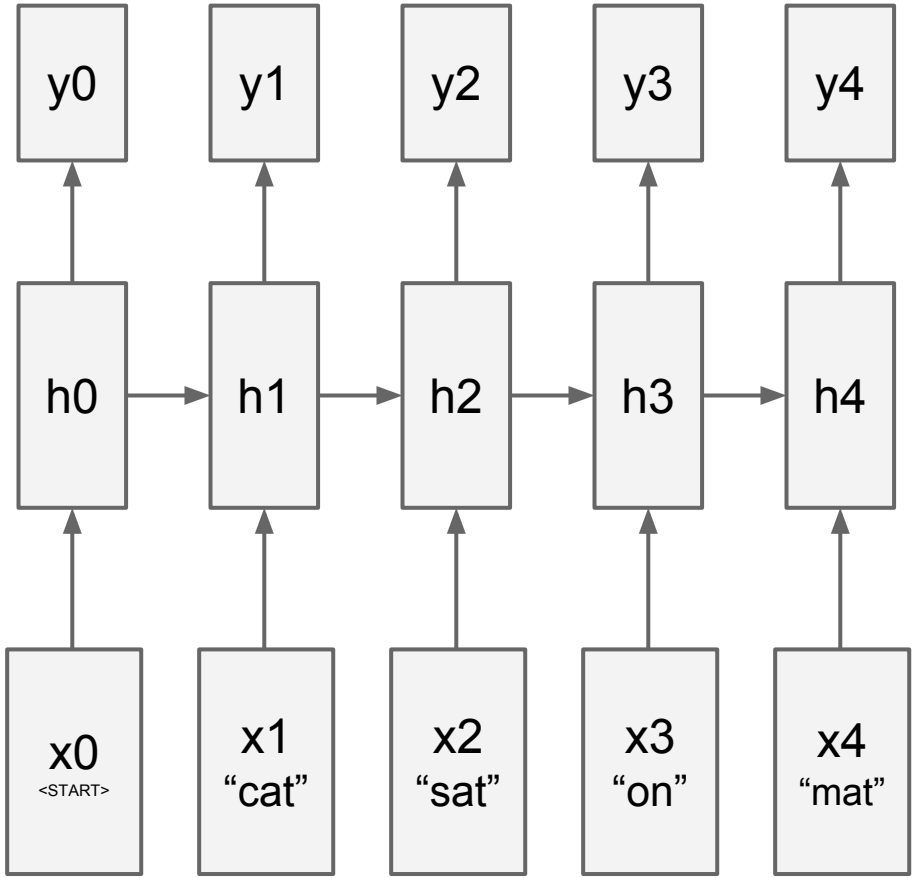
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



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

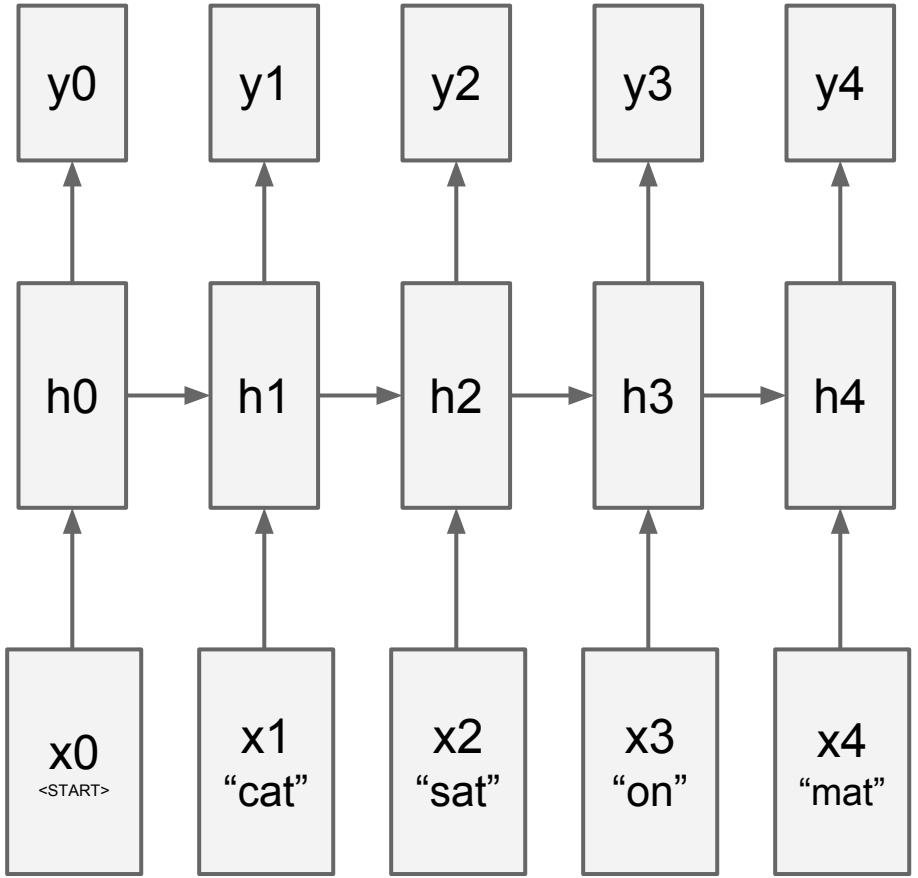
P(next word | previous words)



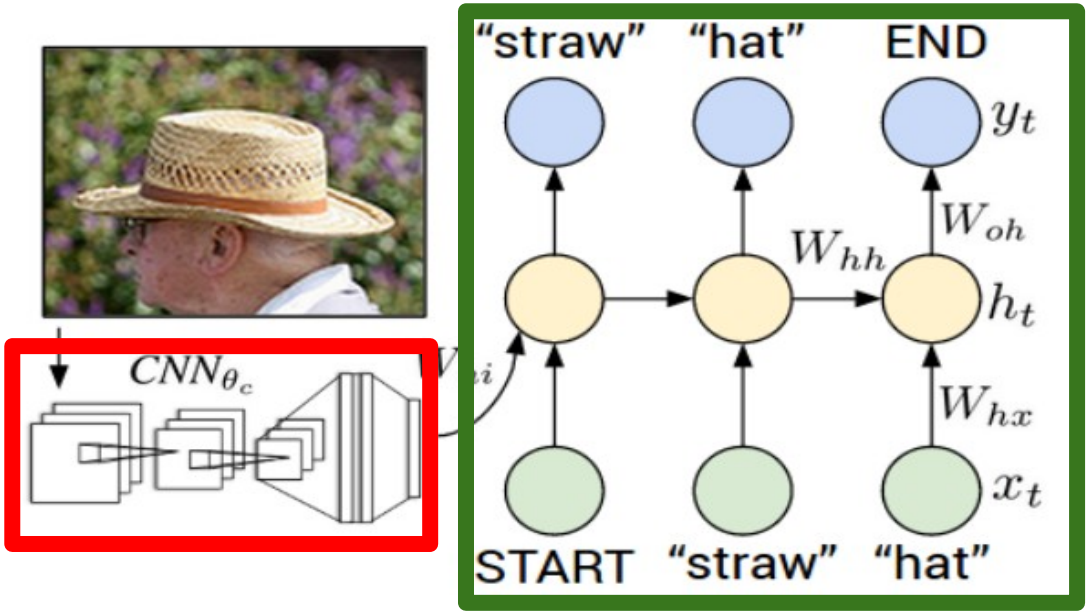
samples <END>? done.

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

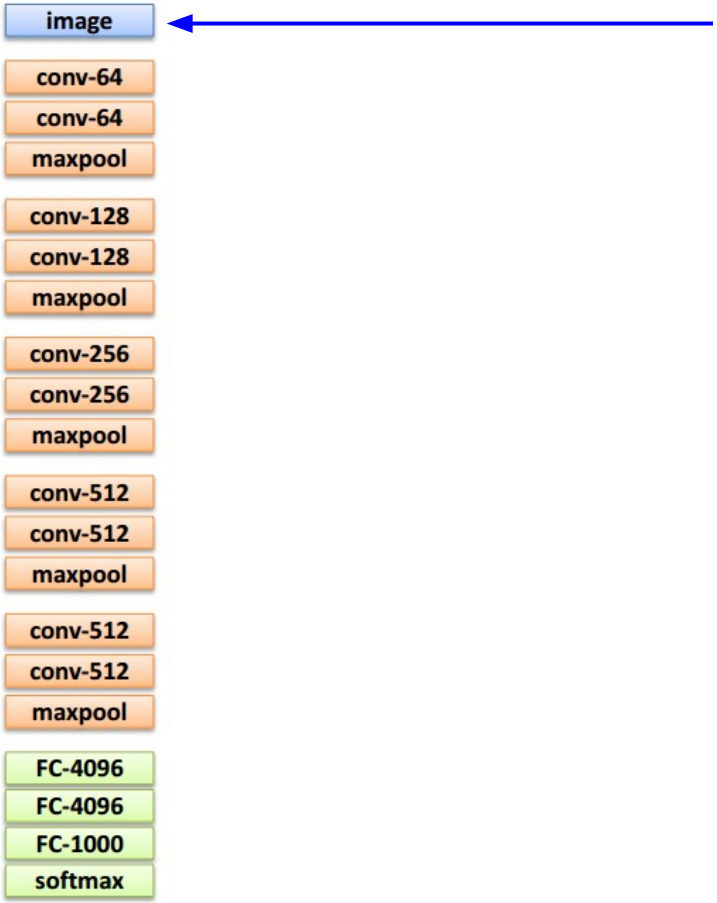
P(next word | previous words)



Recurrent Neural Network



Convolutional Neural Network

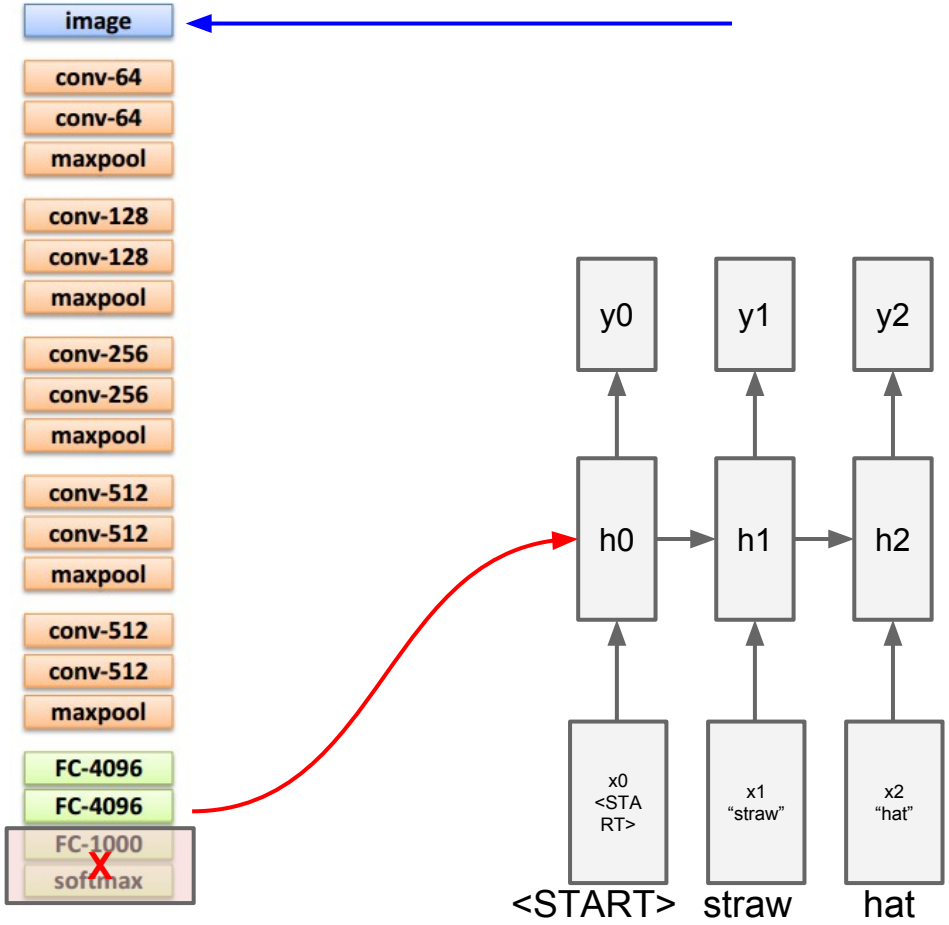


“straw hat”

training example

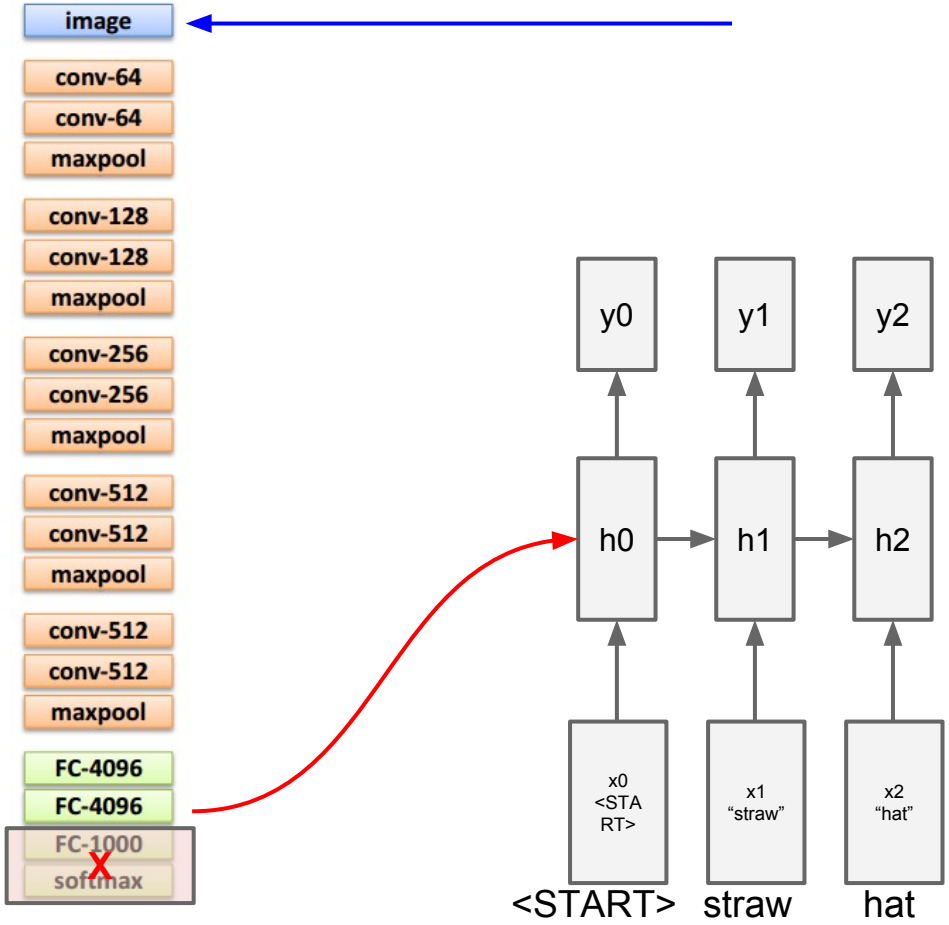
"straw hat"

training example



“straw hat”

training example



“straw hat”

training example

before:

h0 = max(0, Wxh * x0)

now:

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

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
<S>

<START>

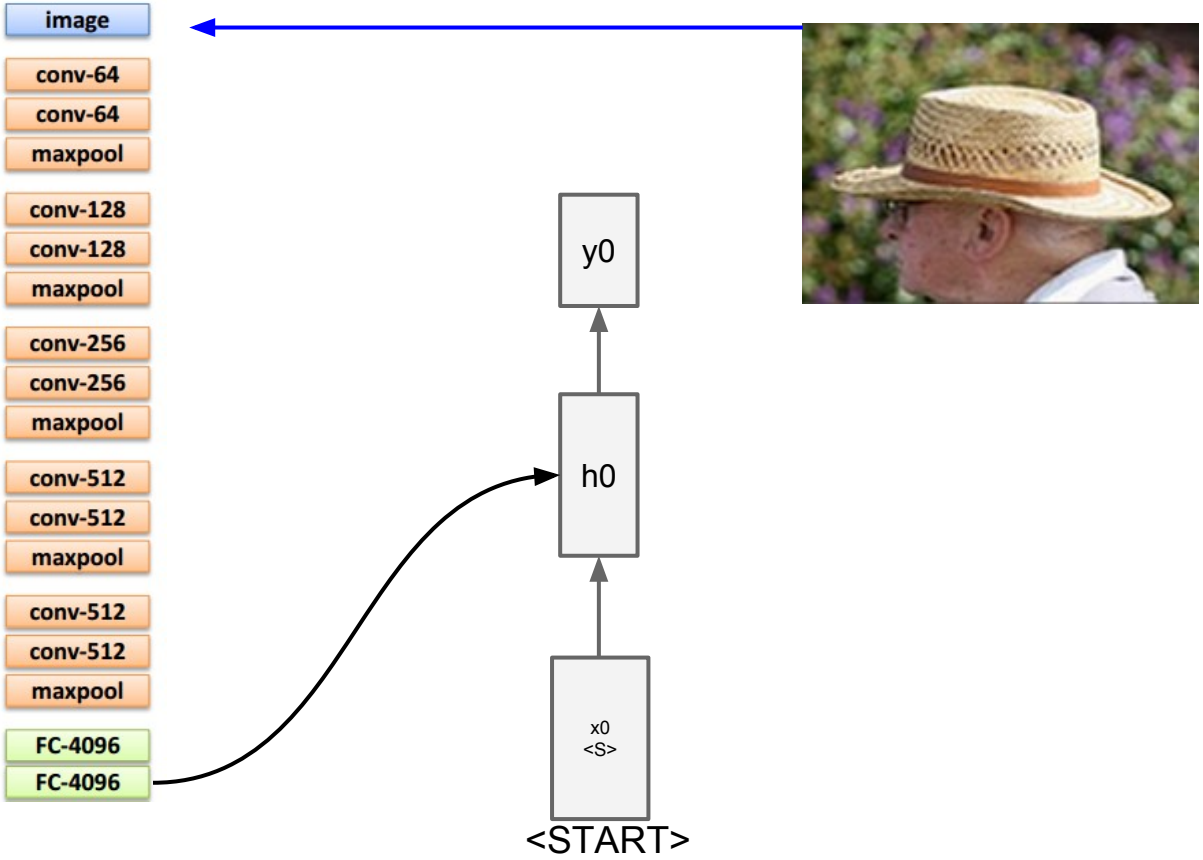


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
y0
h0
x0
<S>
<START>
test image

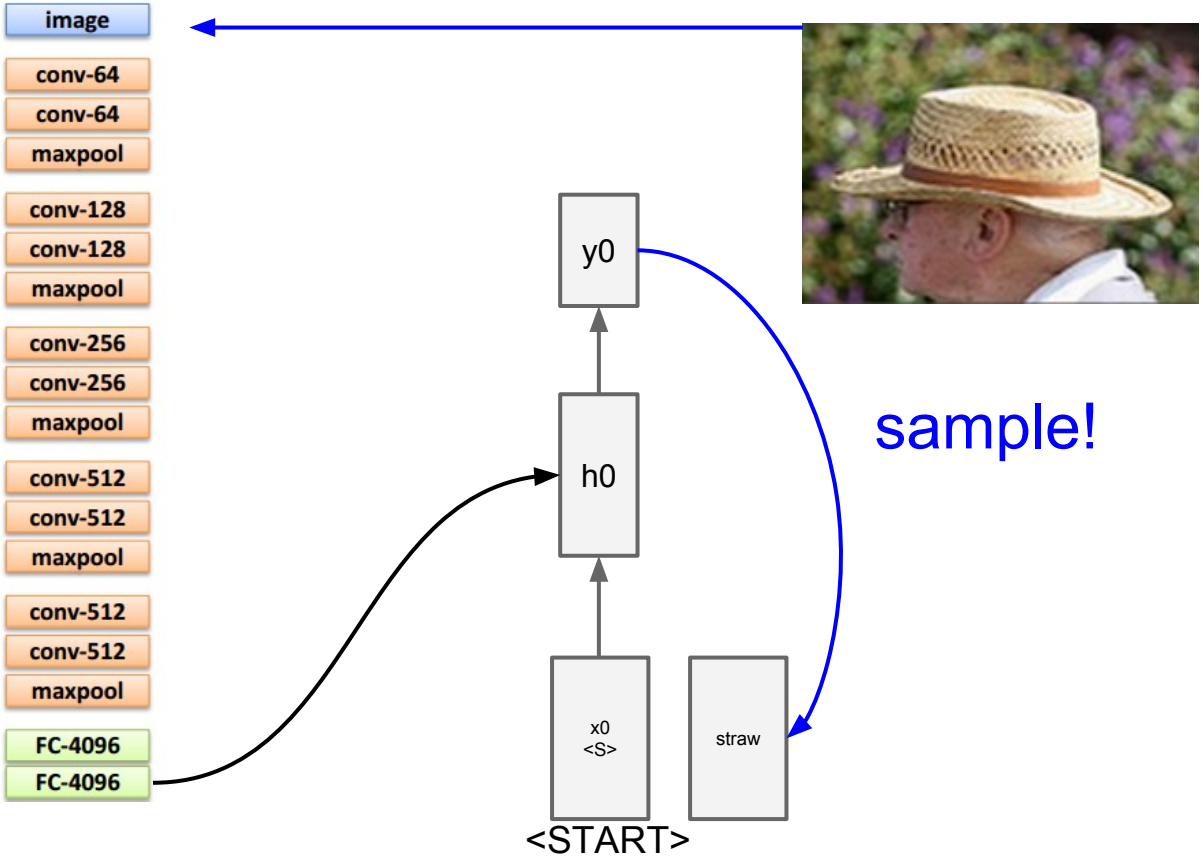


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
y0
h0
x0
<S>
straw
<START>
test image
sample!

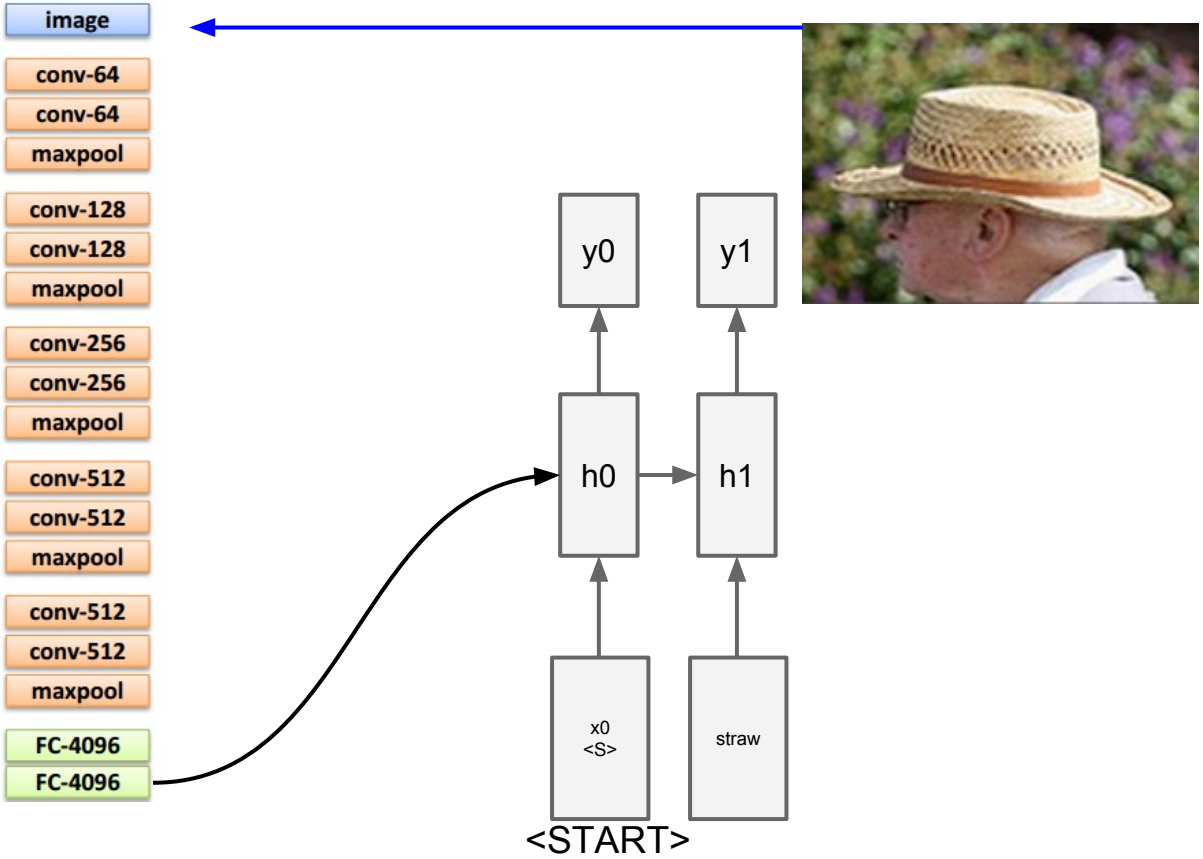


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
y0
y1
h0
h1
x0
<S>
straw
<START>
test image

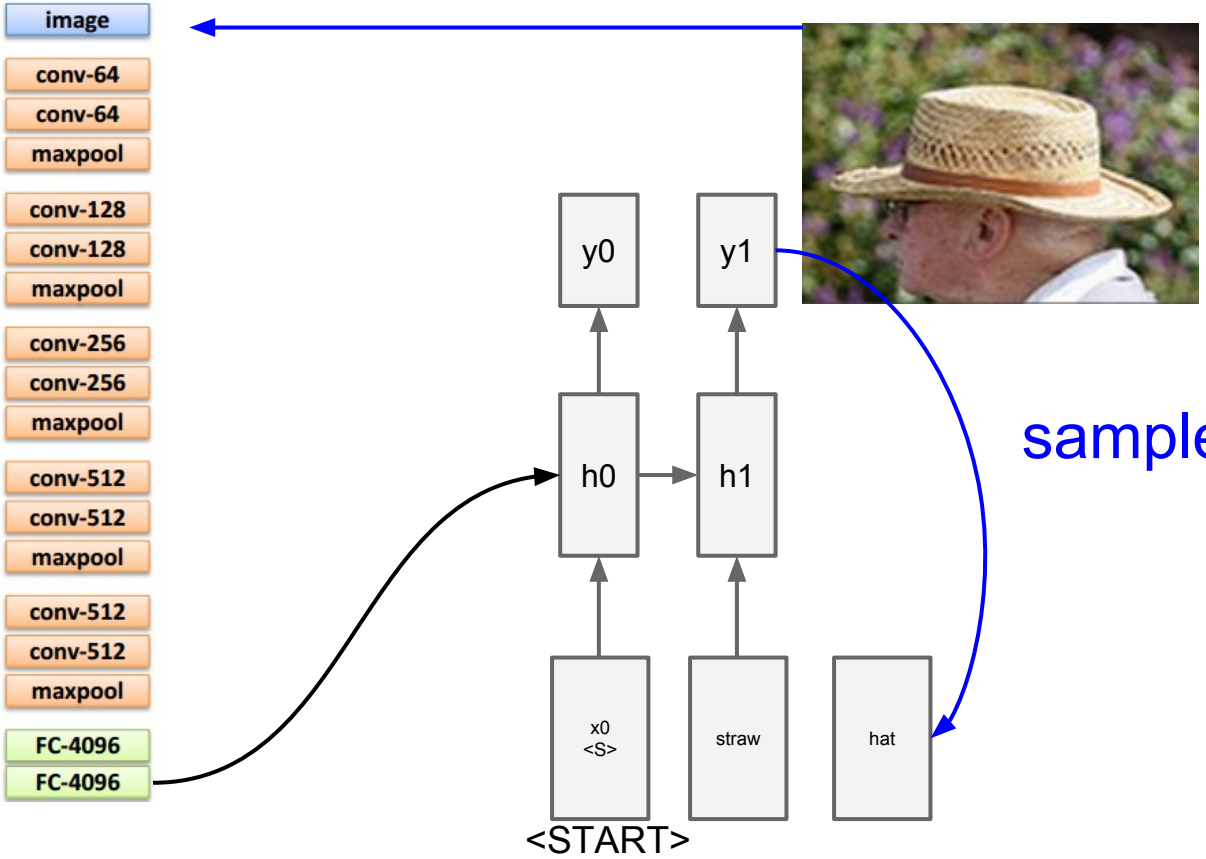


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
test image
y0
y1
h0
h1
x0
<S>
straw
hat
<START>
sample!

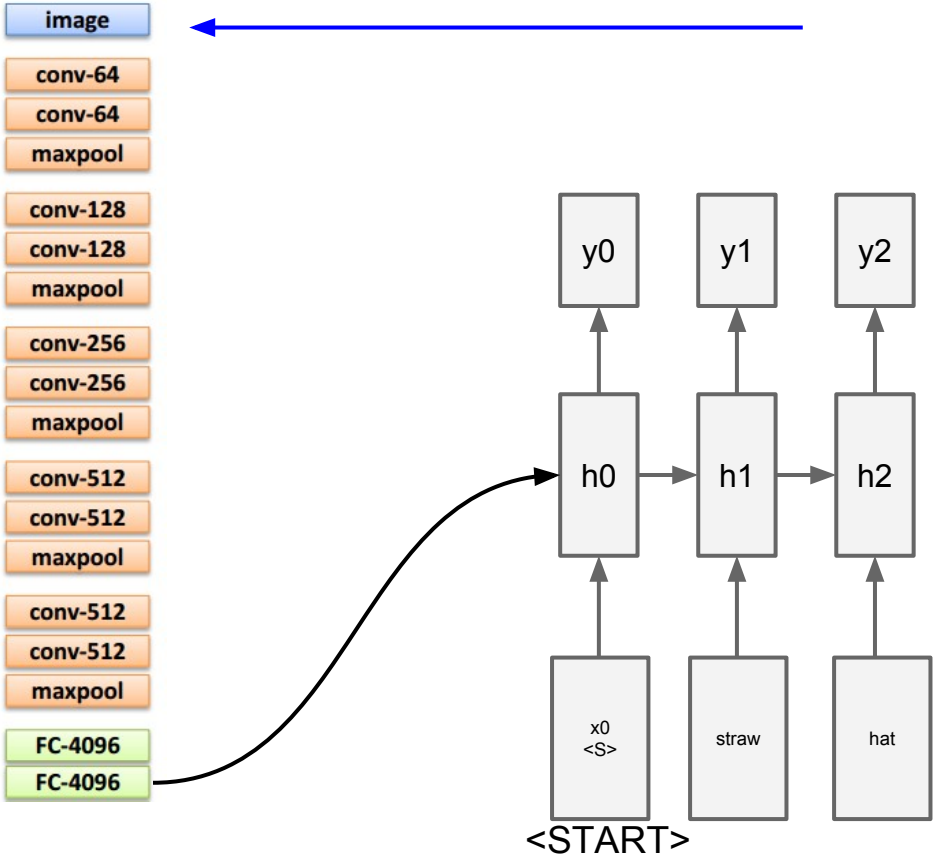
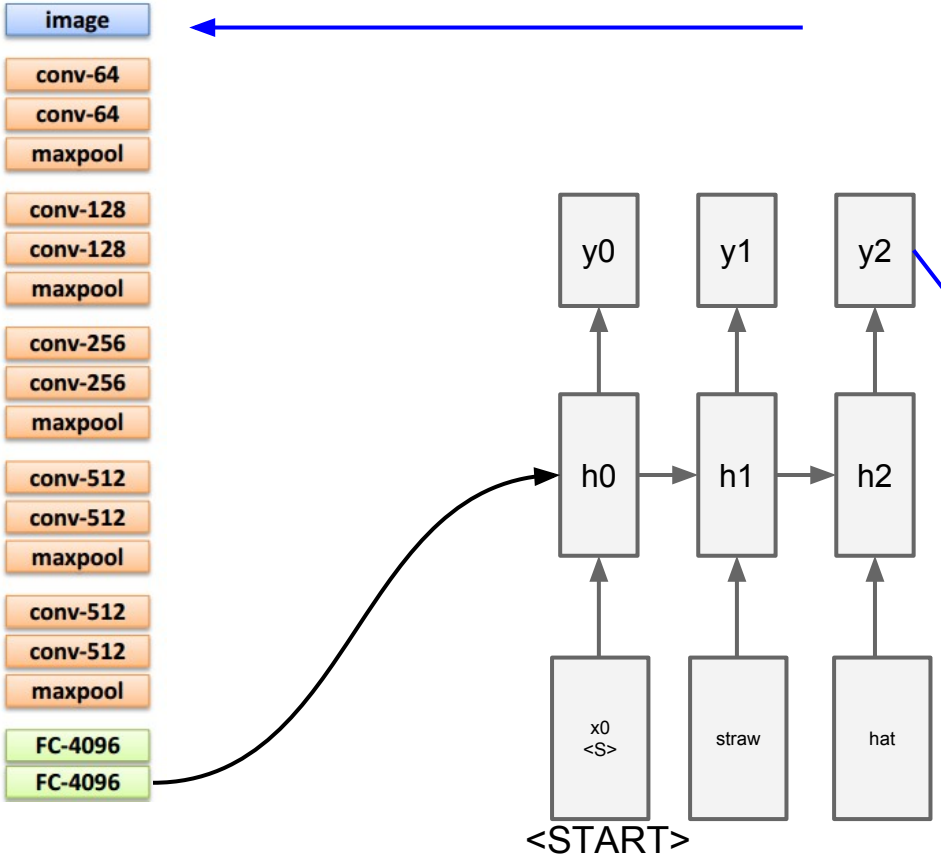


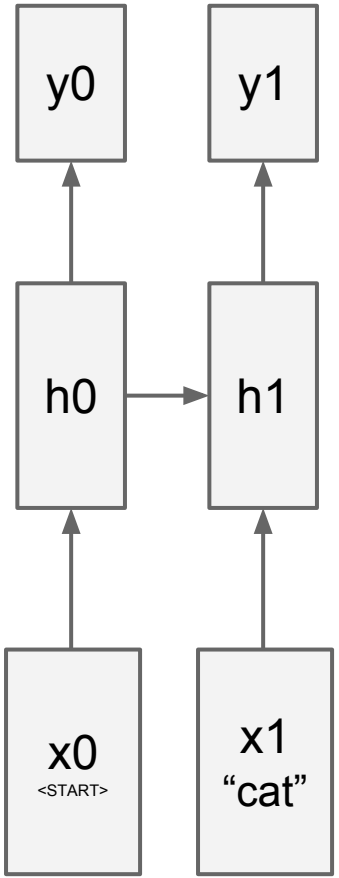
image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
y0
y1
y2
h0
h1
h2
x0
<S>
straw
hat
<START>

test image



test image

sample!
<END> token
=> finish.



RNN vs. LSTM

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

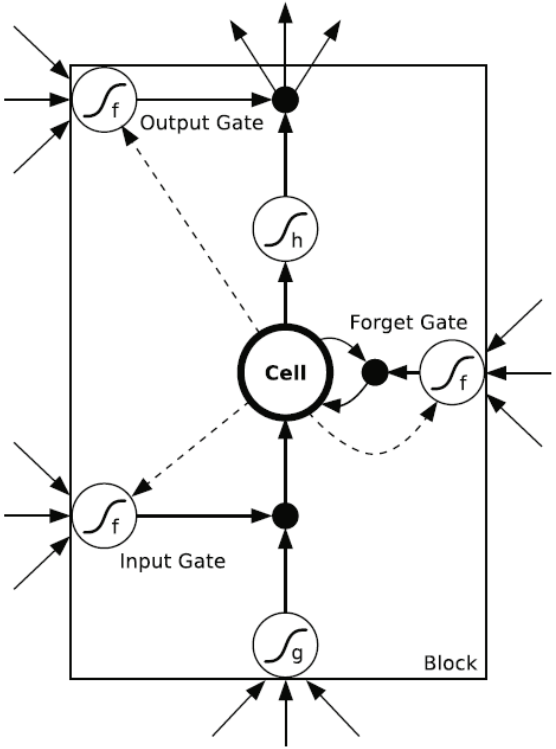
LSTM just changes the form of the equation for **h** such that:

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

RNN vs. LSTM

- Use LSTM over RNN
- Do not be intimidated by pictures that try to draw an LSTM, e.g.:

(it's just a particular funny form of forward function, backpropagation as usual)



$$\text{LSTM} : h_t^{l-1}, h_{t-1}^l, c_{t-1}^l \rightarrow h_t^l, c_t^l$$

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} T_{2n,4n} \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}$$

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

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

Recurrent Neural Network Regularization
[Zaremba, Sutskever, Vinyals]

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

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

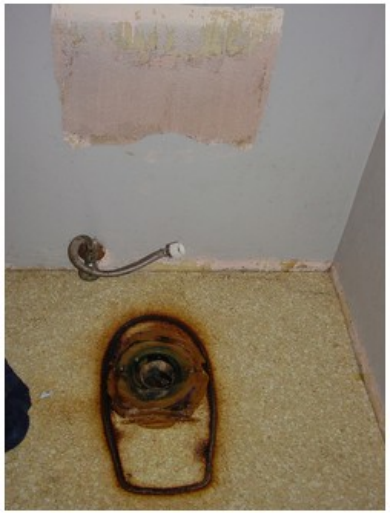
Well, I can kind of see it

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

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

Not sure what happened there...



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

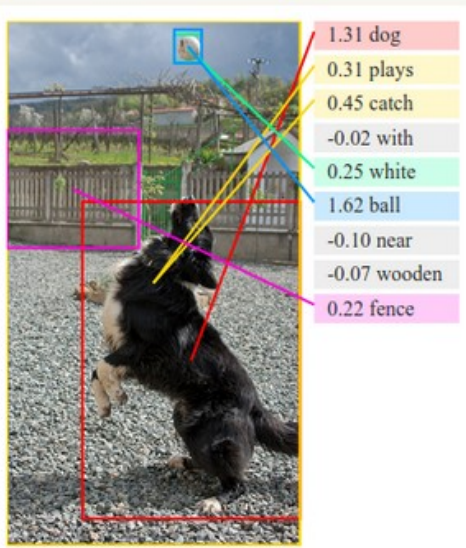
a woman holding a teddy bear in front of a mirror
logprob: -9.65

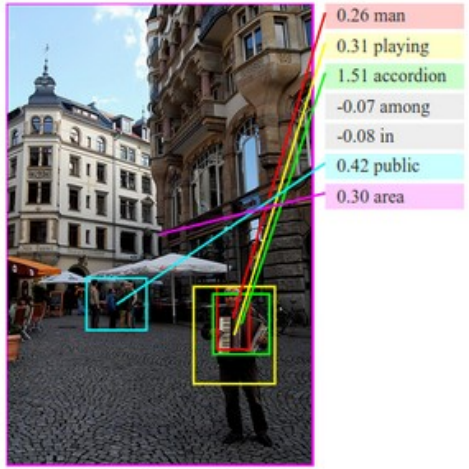
a horse is standing in the middle of a road
logprob: -10.34

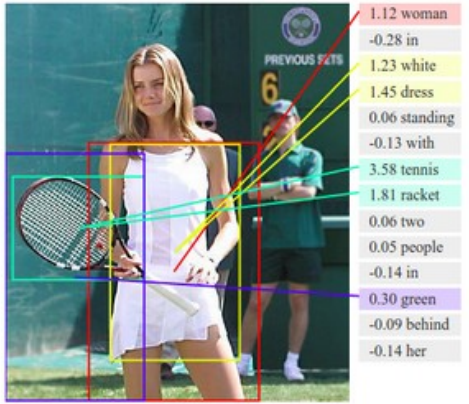
More examples in Web demo: http://bit.ly/neuraltalkdemo

Ranking and Retrieval

Each example is a query test sentence, the most likely retrieved image & the grounding:

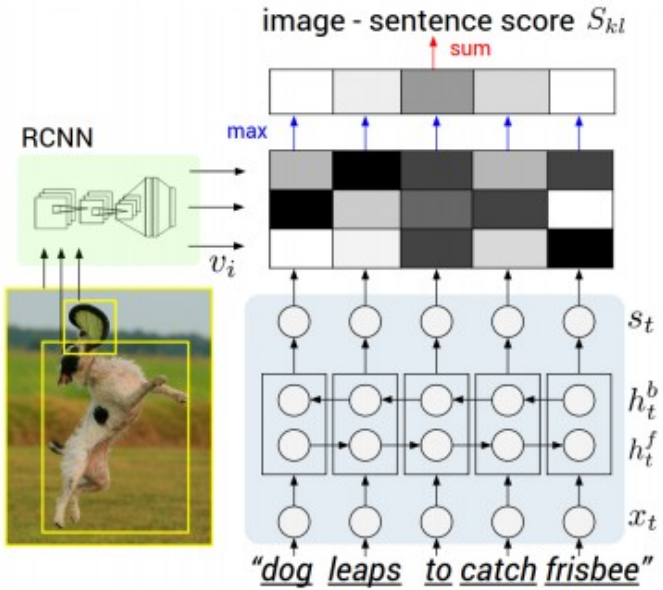




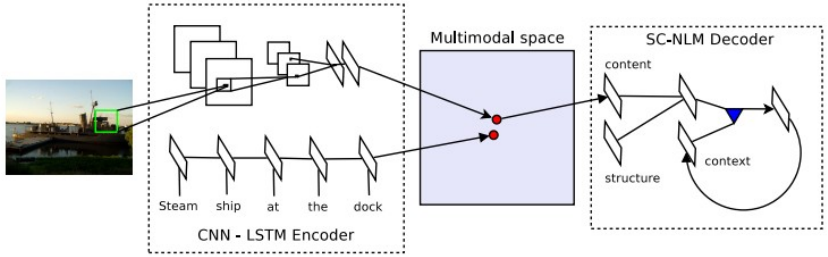


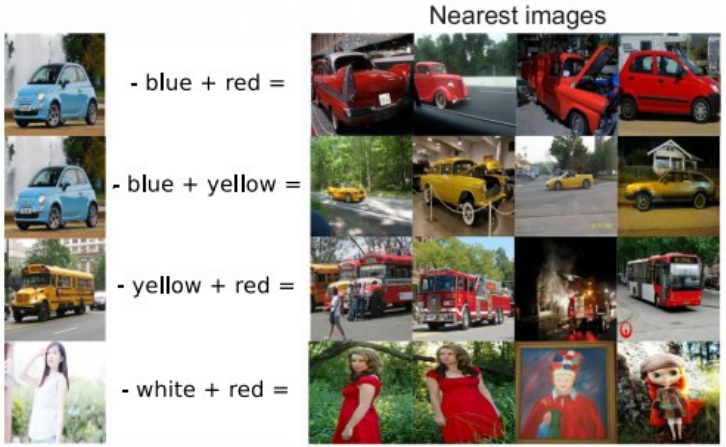
More examples in Web demo: http://bit.ly/rankingdemo

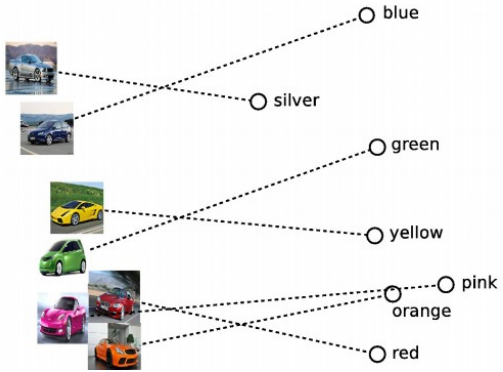
Ranking and Retrieval



Deep Visual-Semantic Alignments for Generating Image Descriptions
[Karpathy and Fei-Fei, 2015]

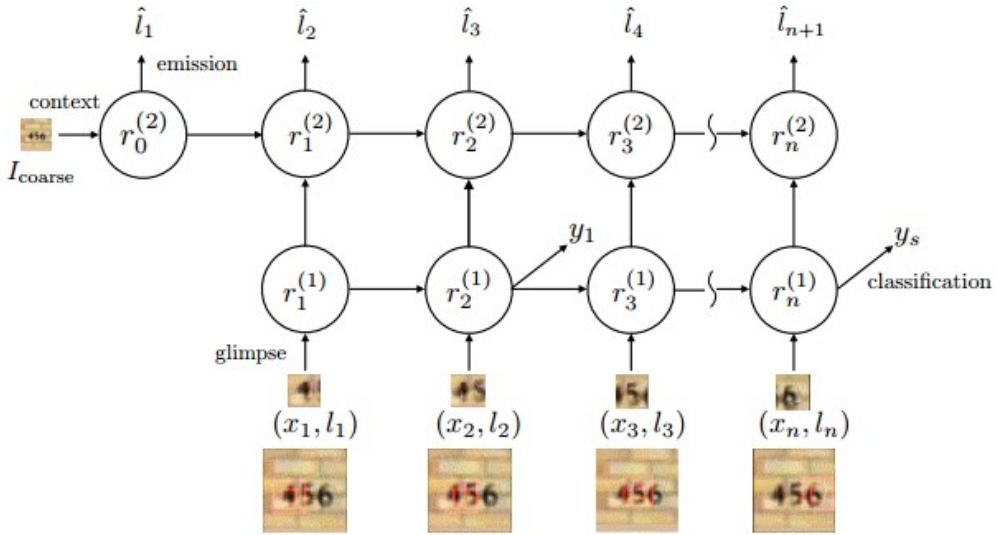






Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models
[Kiros, Salakhutdinov, Zemel, 2014]

Recurrent Attention Models



Multiple Object Recognition with Visual Attention
[Jimmy Lei Ba, Volodymyr Mnih, Koray Kavukcuoglu], 2014

web demo
http://www.psi.toronto.edu/~jimmy/dram/
also DRAW: https://www.youtube.com/watch?v=Zt7MI9eKEo

Summary:

- We looked at many Computer Vision tasks beyond Image Classification and how they are addressed with Convolutional Neural Networks

Next: Guest Lecture:

Evan Shelhamer

Caffe