Lecture 11:

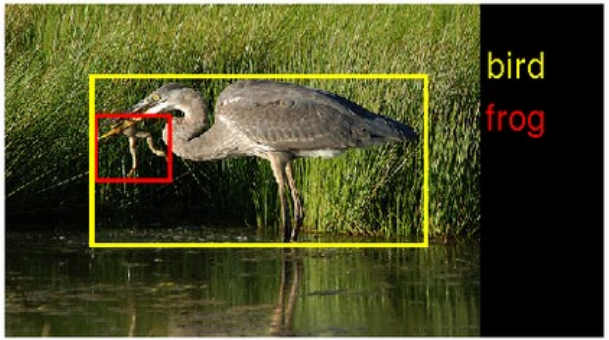
Beyond Image Classification

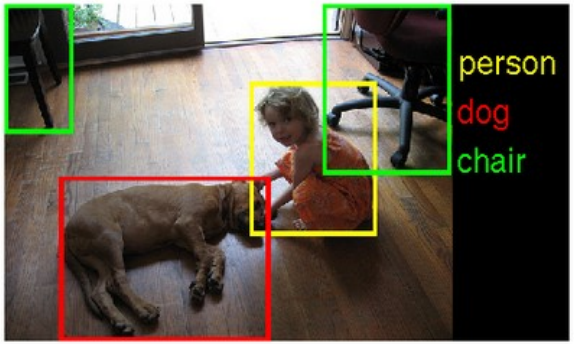
Localization

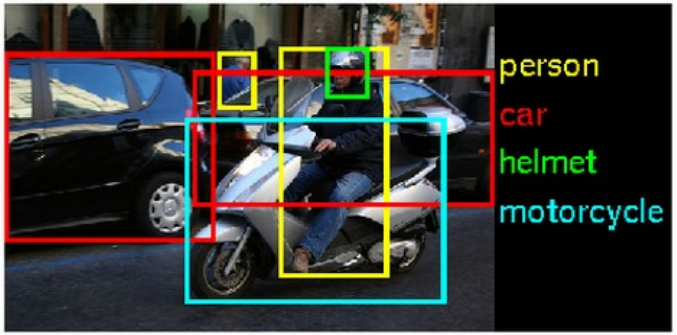
Model must output:

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

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]

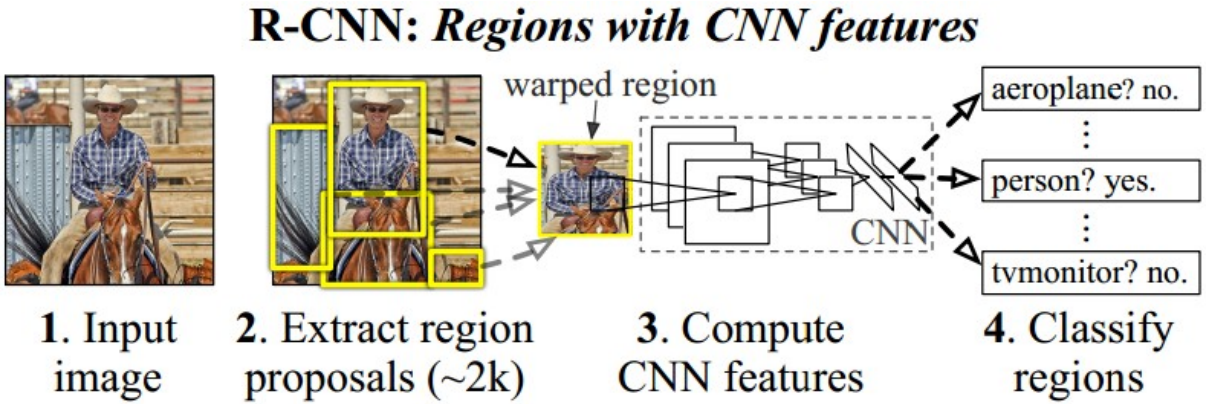
***Idea**: Turn a Detection Problem into an Image Classification problem (but over image regions).*

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**.

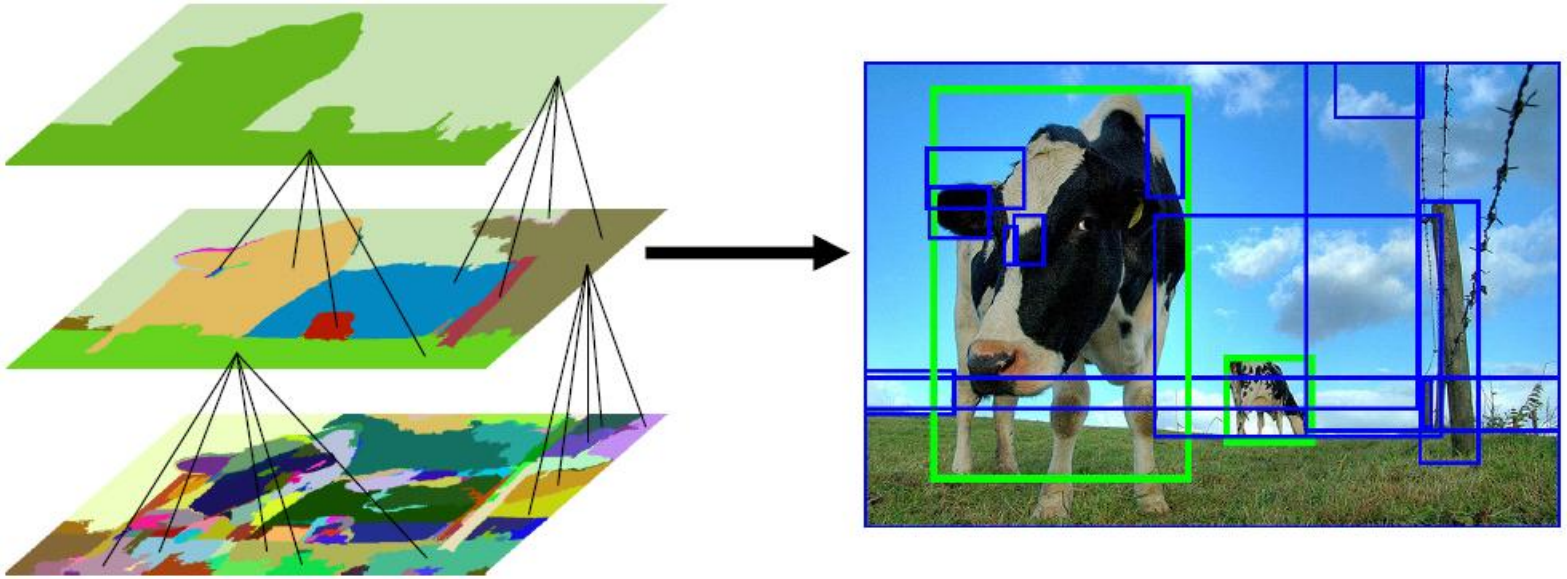
Rich feature hierarchies for accurate object detection and semantic segmentation
[Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik]

***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]

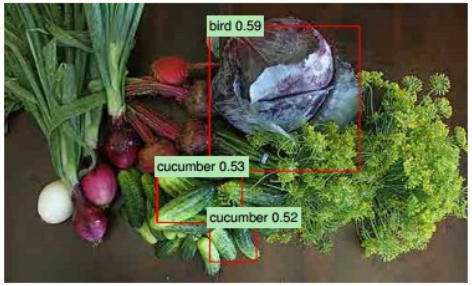


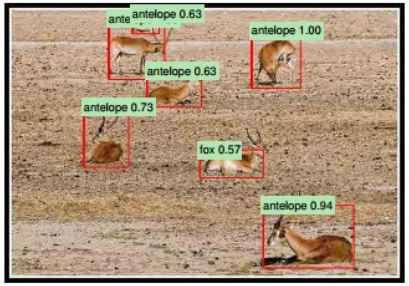
Gives on average ~2,000 candidate region proposals per image.

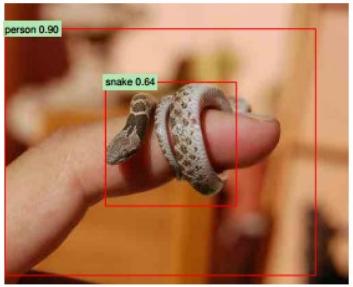
(This paradigm currently outperform the "sliding window" approach)

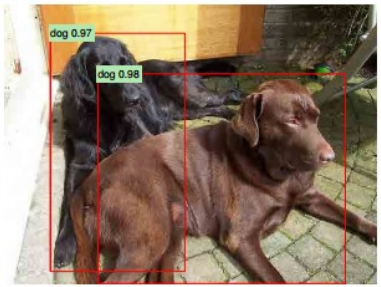
Rich feature hierarchies for accurate object detection and semantic segmentation

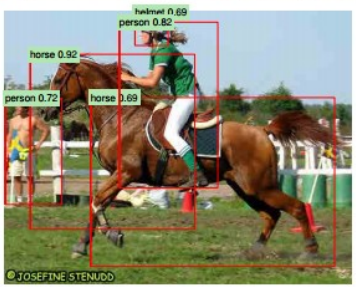
[Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik]

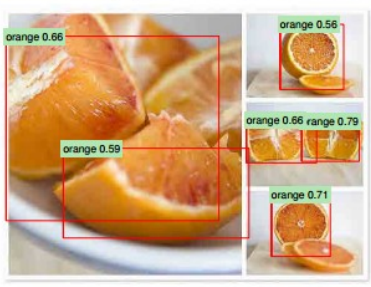


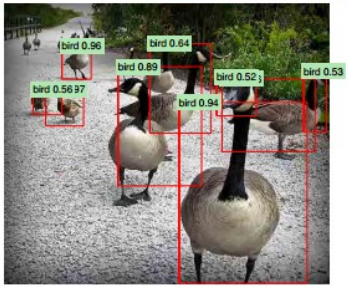


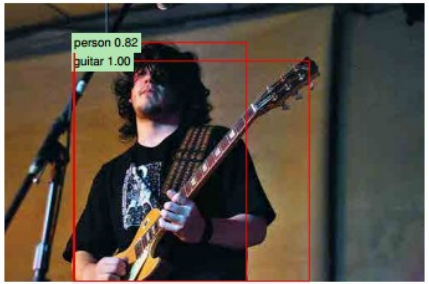


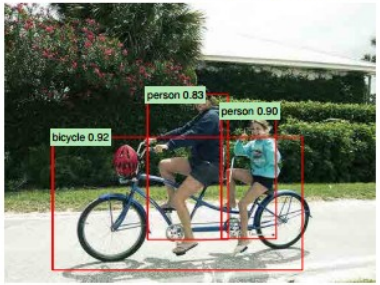


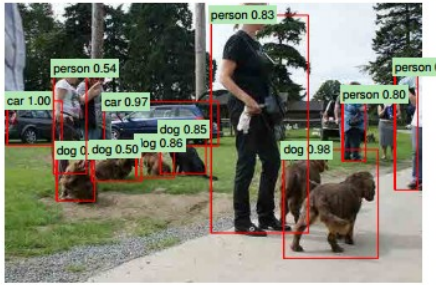




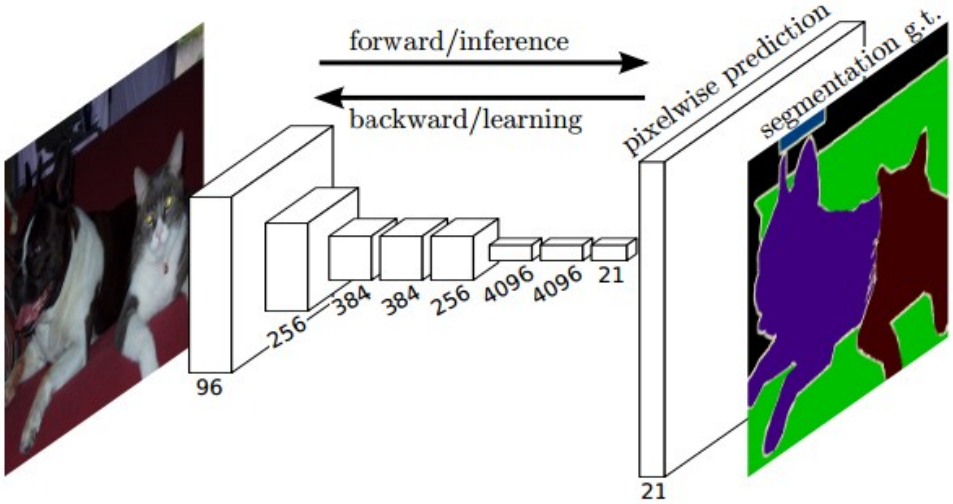


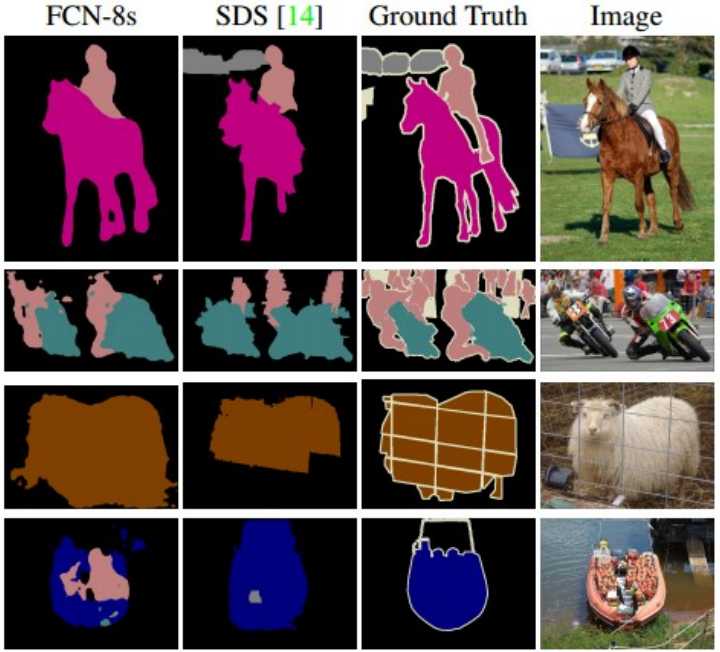






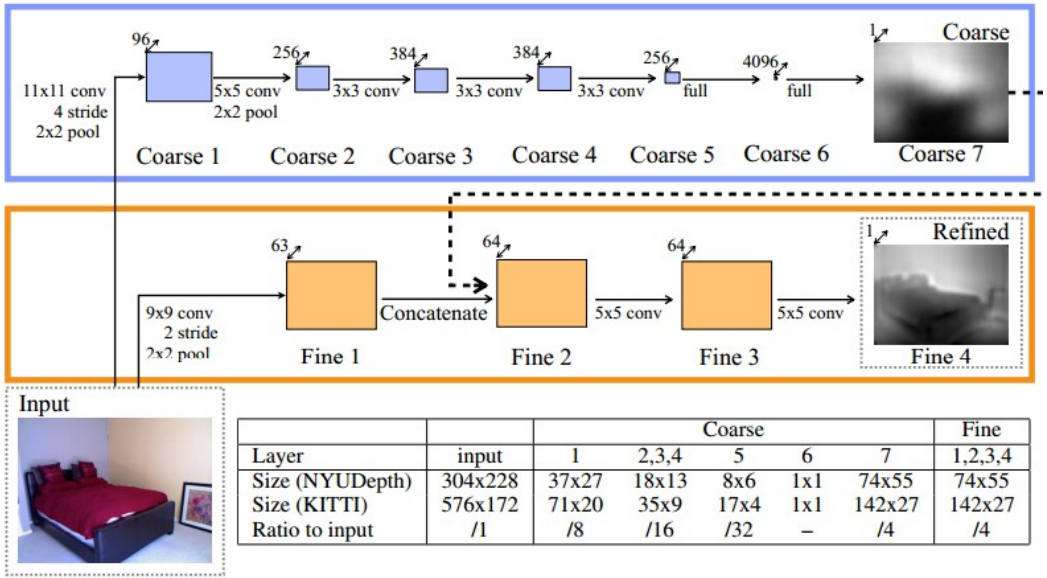
Segmentation



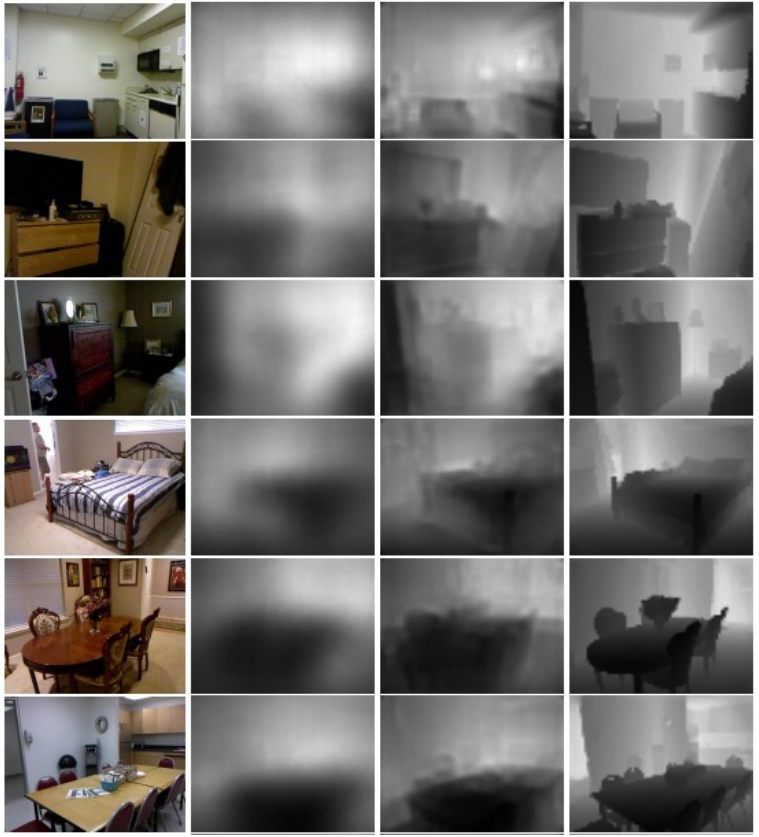


Fully Convolutional Networks for Semantic Segmentation
Long, Shelhamer, Darrell

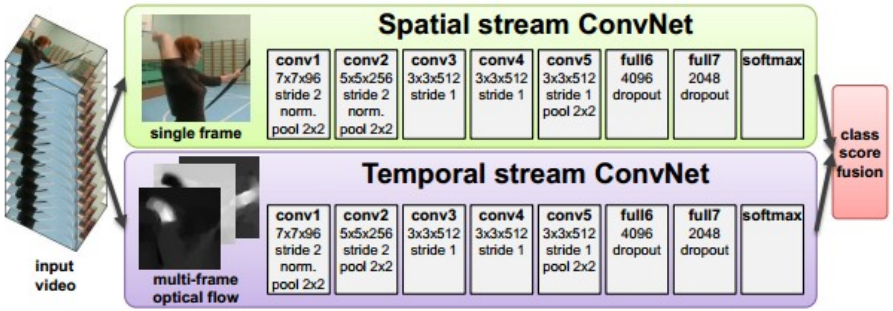
Depth Map Prediction from a Single Image using a Multi-Scale Deep Network
[Eigen et al.], 2014



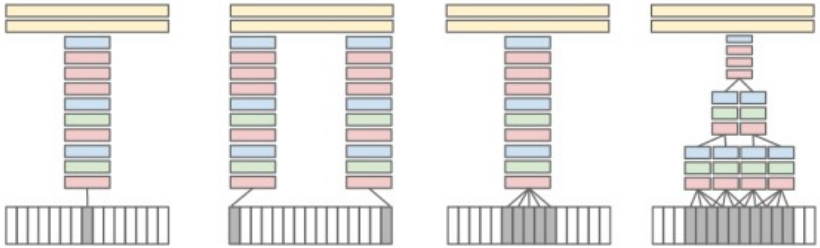
		Coarse					Fine
Layer	input	1	2,3,4	5	6	7	1,2,3,4
Size (NYUDepth)	304x228	37x27	18x13	8x6	1x1	74x55	74x55
Size (KITTI)	576x172	71x20	35x9	17x4	1x1	142x27	142x27
Ratio to input	/1	/8	/16	/32	–	/4	/4



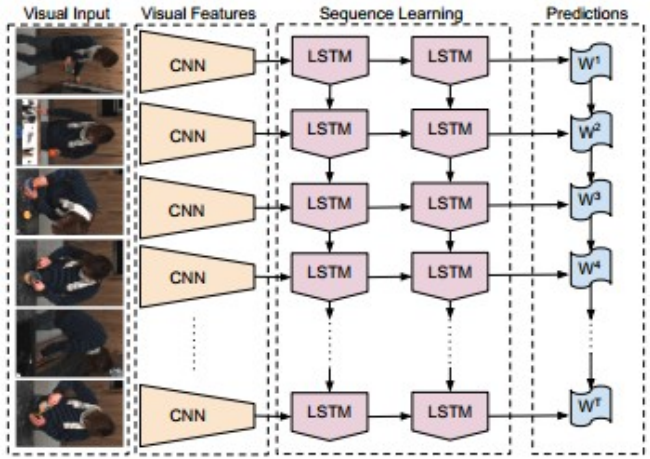
Video Classification



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



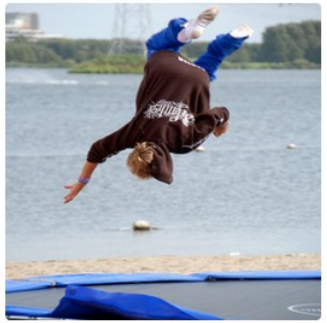
Long-term Recurrent Convolutional Networks for Visual Recognition and Description
[Donahue et al.], 2014

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."

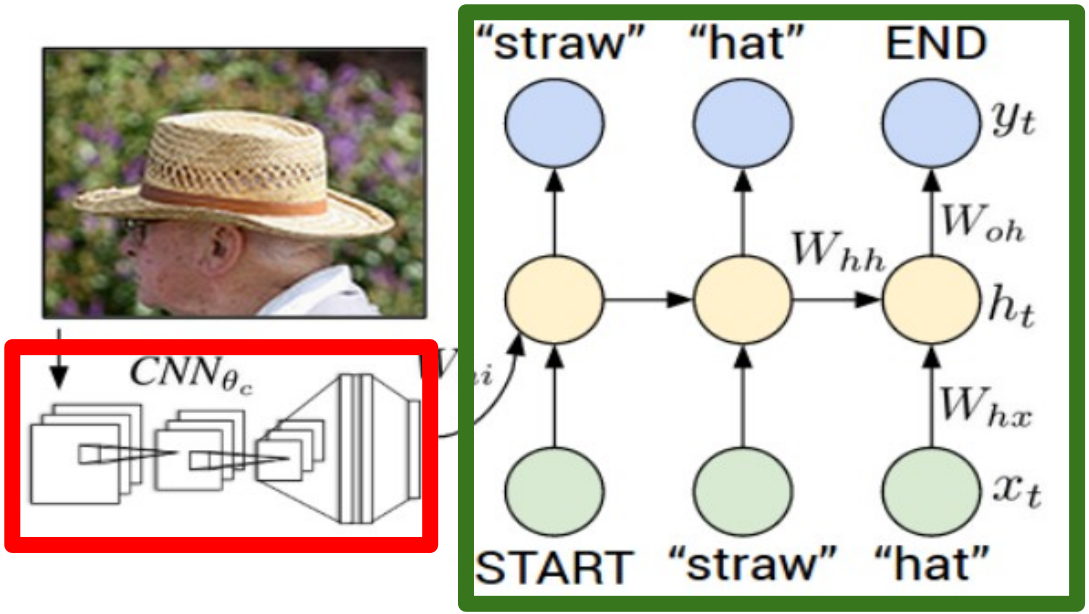
"girl in pink dress is jumping in air."

"black and white dog jumps over bar."

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

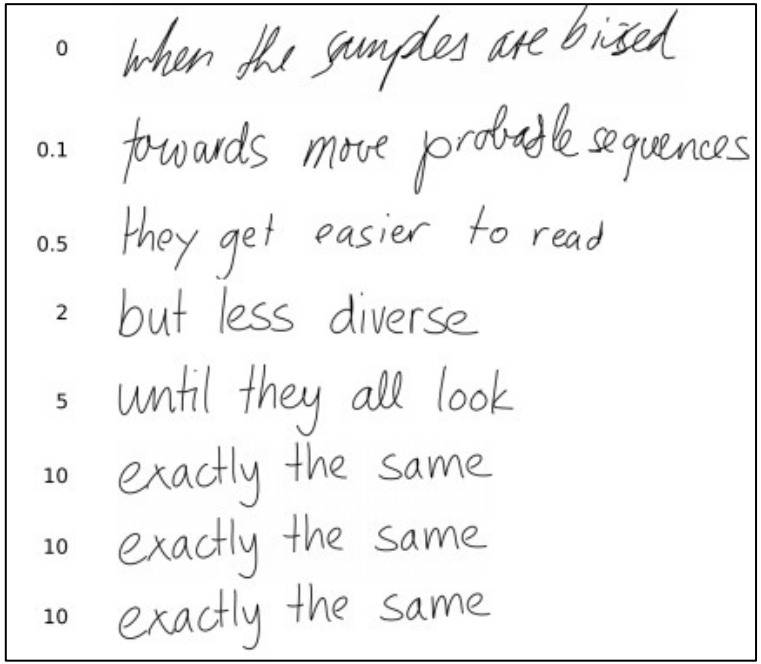
"man in blue wetsuit is surfing on wave."

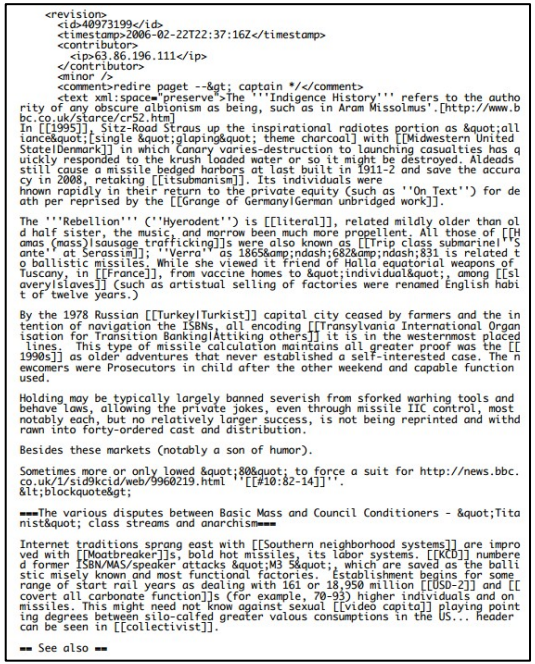
Recurrent Neural Network



Convolutional Neural Network

Recurrent Networks are good at modeling sequences...

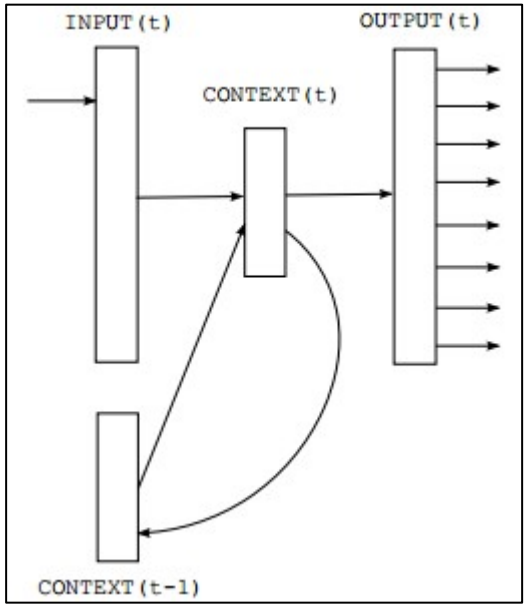




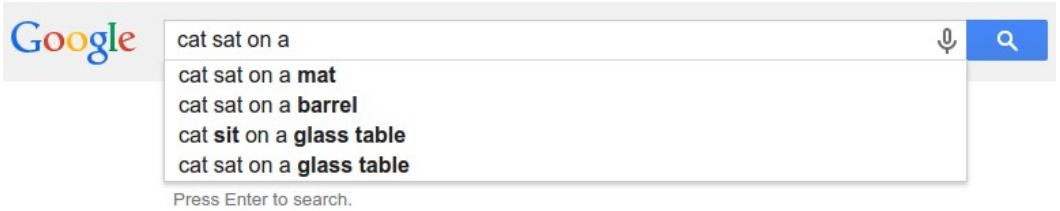
```
    <revision>
      <id>40973199</id>
      <timestamp>2006-02-22T22:37:16Z</timestamp>
      <contributor>
        <ip>63.86.196.111</ip>
      </contributor>
      <minor />
      <comment>redire paget --&gt; captain */</comment>
      <text xml:space="preserve">The '''Indigence History''' refers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



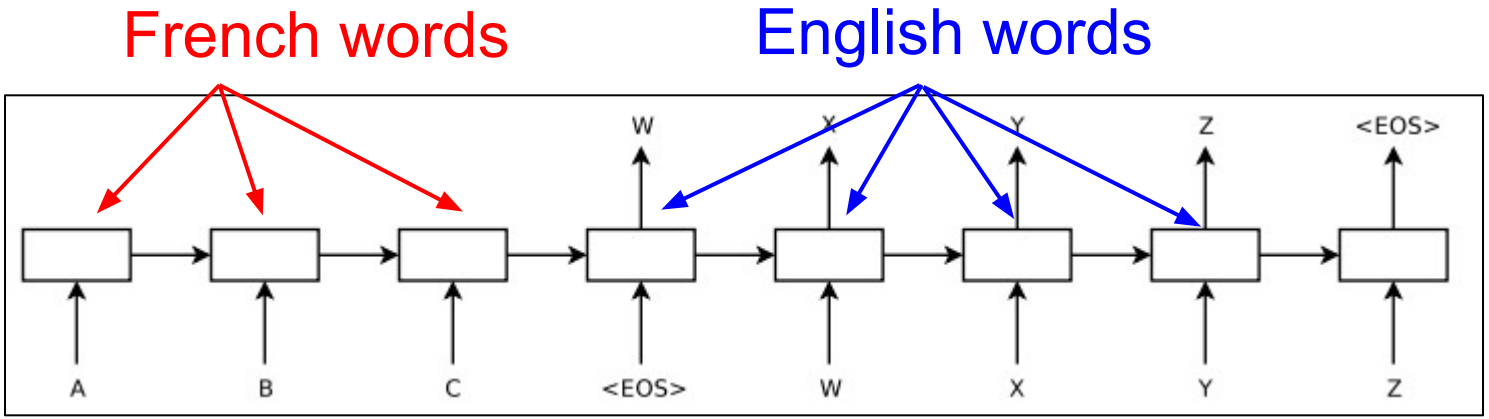
Word-level language model. Similar to:



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]

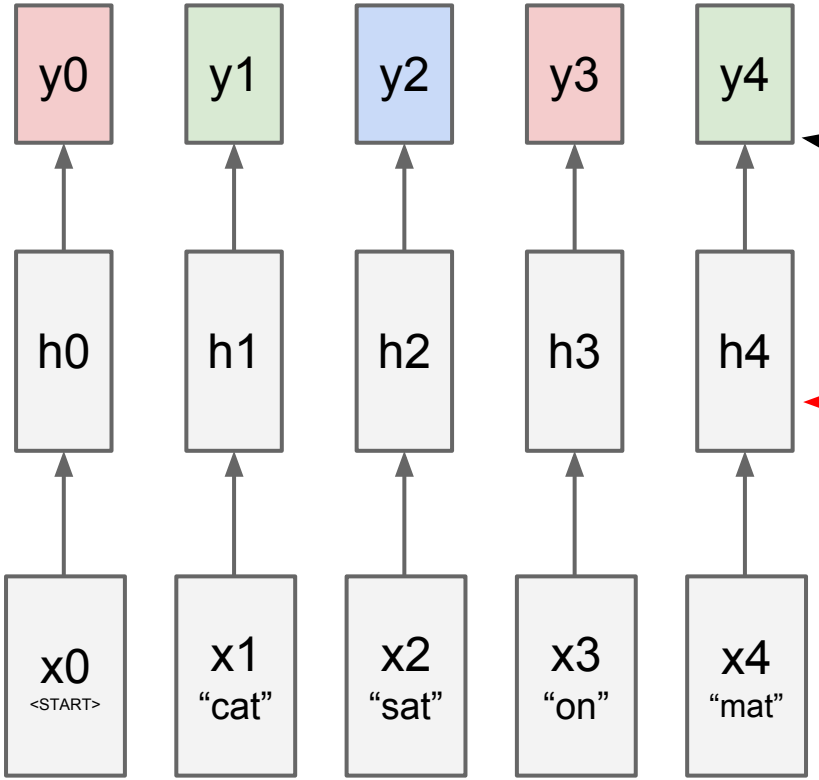
Suppose we had the training sentence “cat sat on mat”

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])

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

"cat sat on mat"



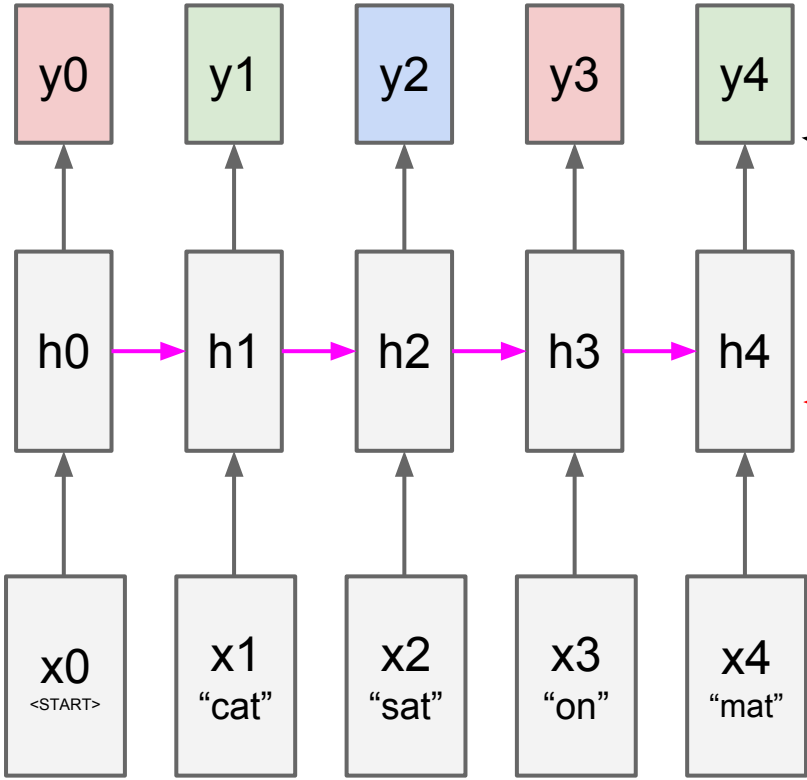
10,001-D class scores:
Softmax over 10,000 words and a special <END> token.
y4 = Why * h4

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)

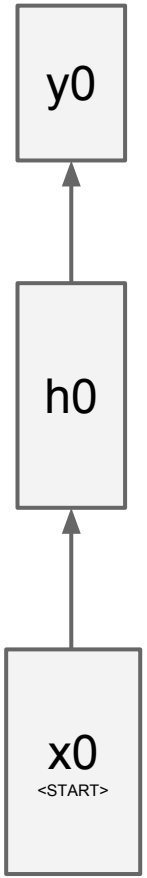
300 (learnable) numbers
associated with each word
in vocabulary

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

P(next word | previous words)

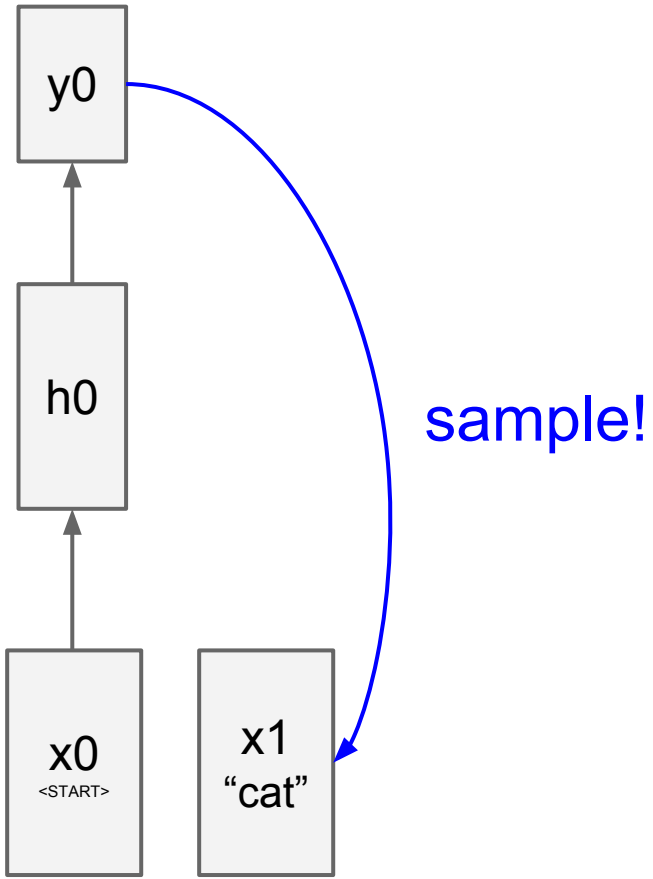
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



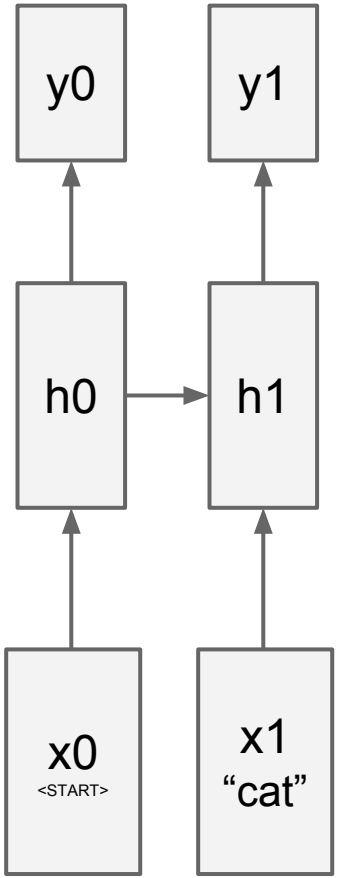
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



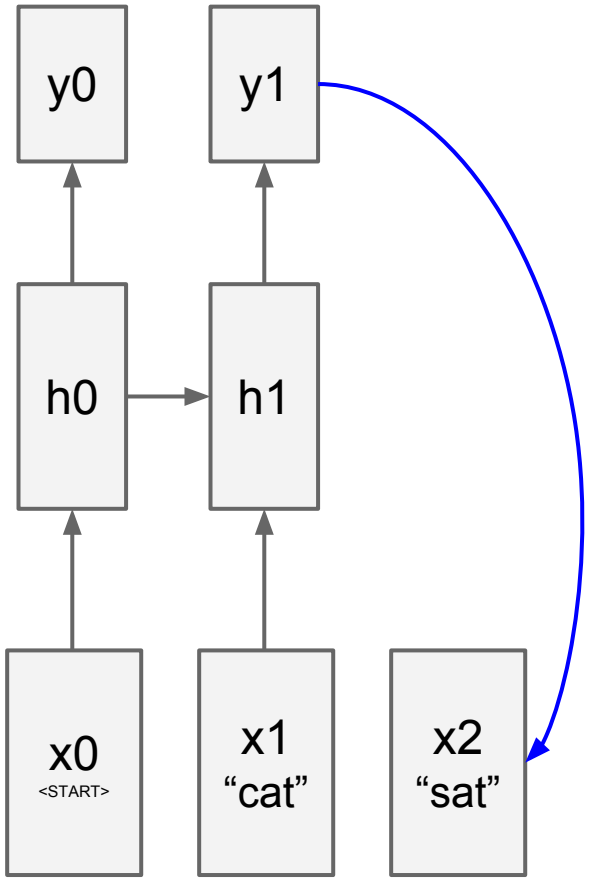
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



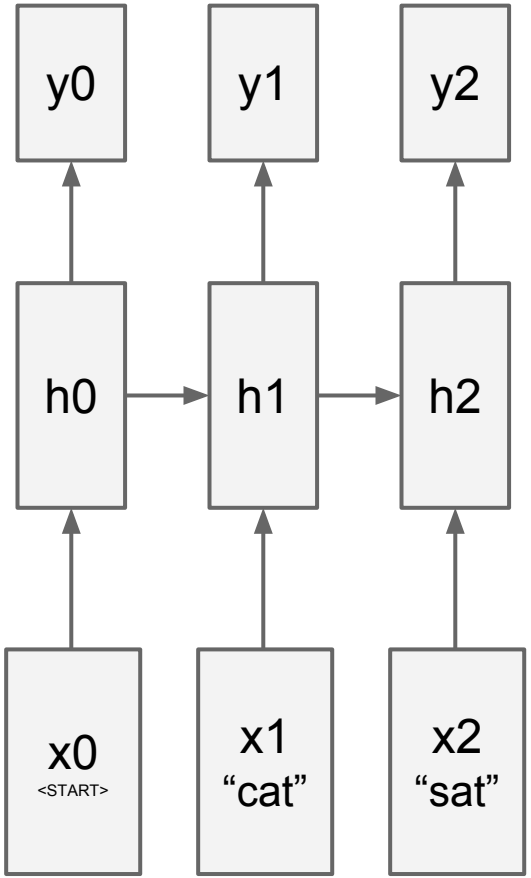
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



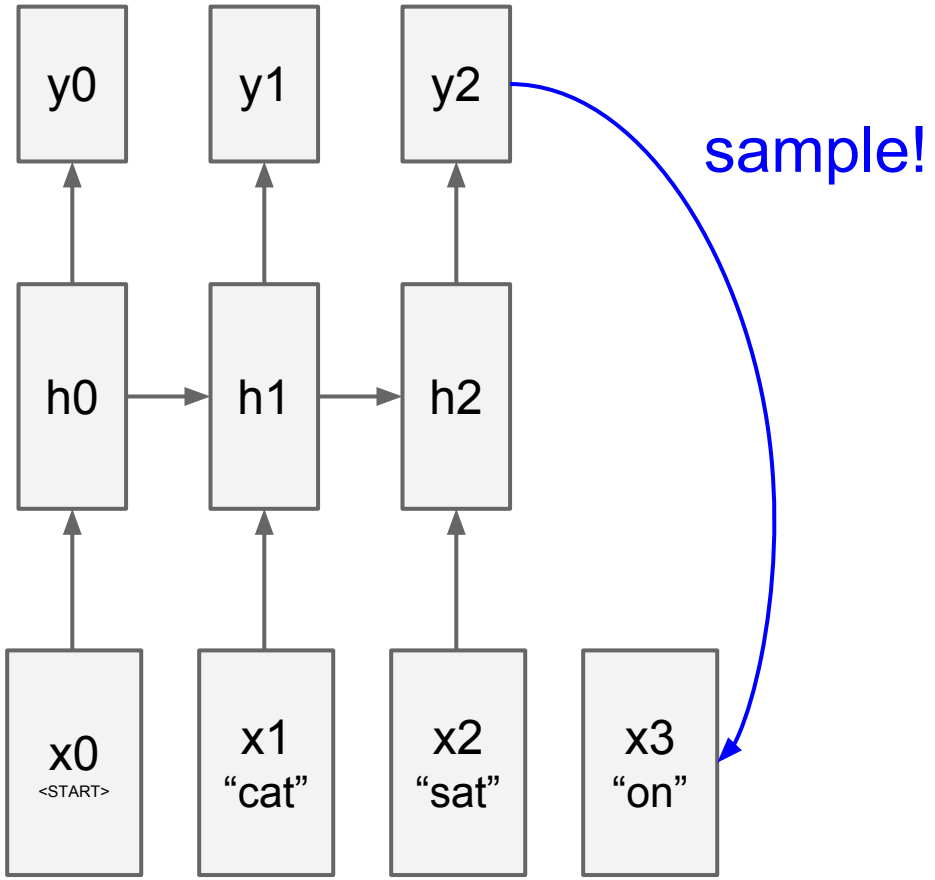
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



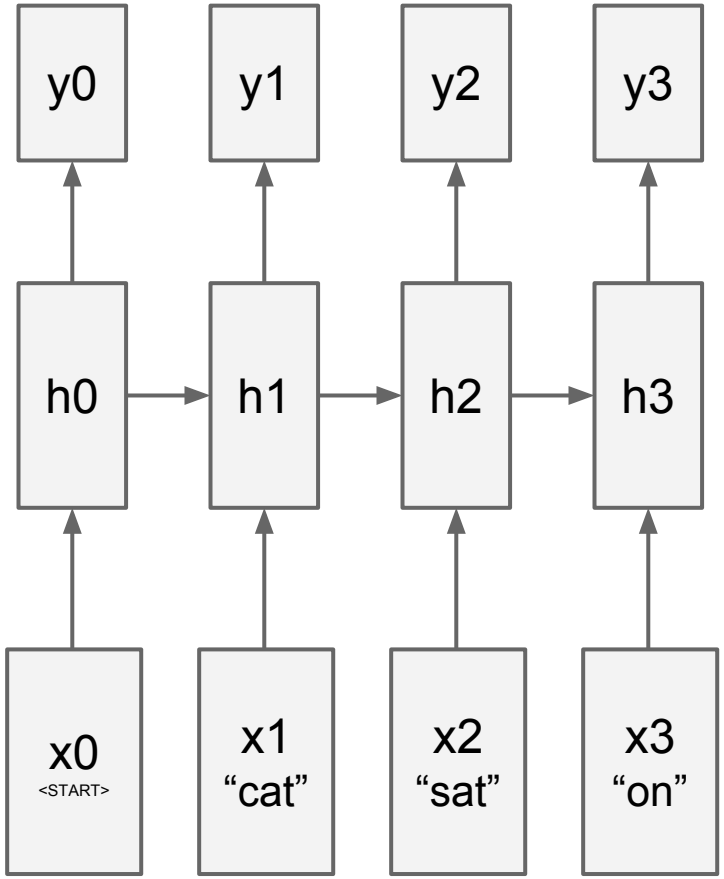
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



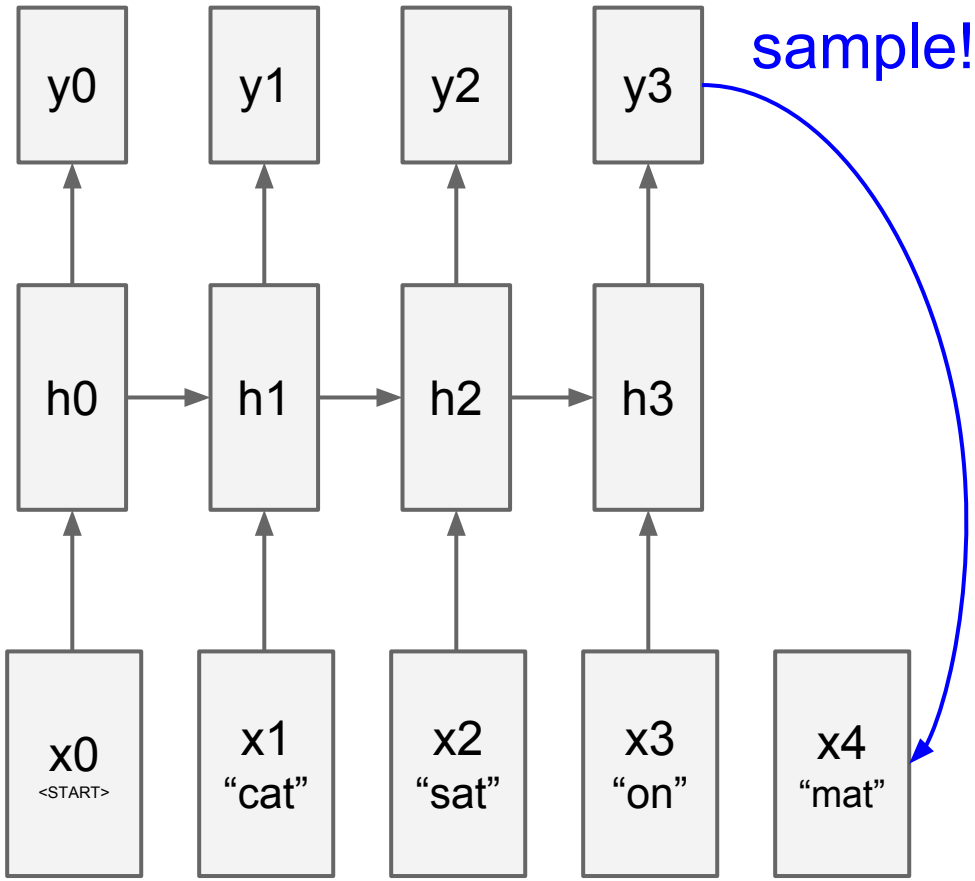
Training this on a lot of sentences would give us a language model. A way to predict

P(next word | previous words)



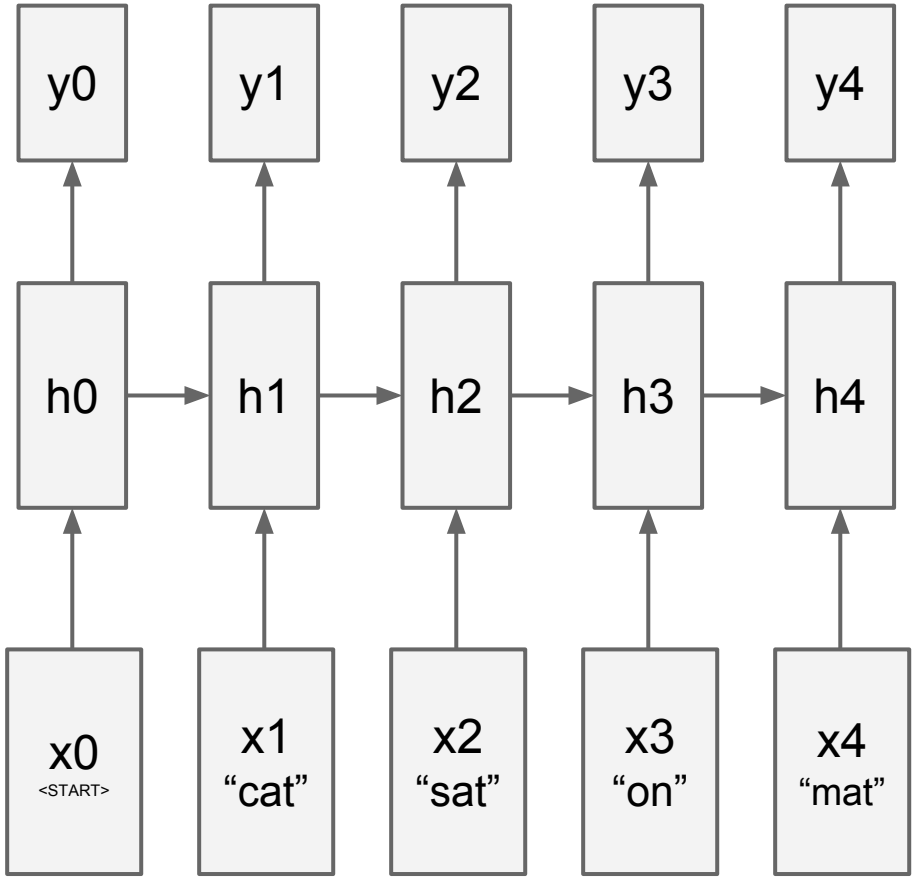
Training this on a lot of sentences would give us a language model. A way to predict

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)

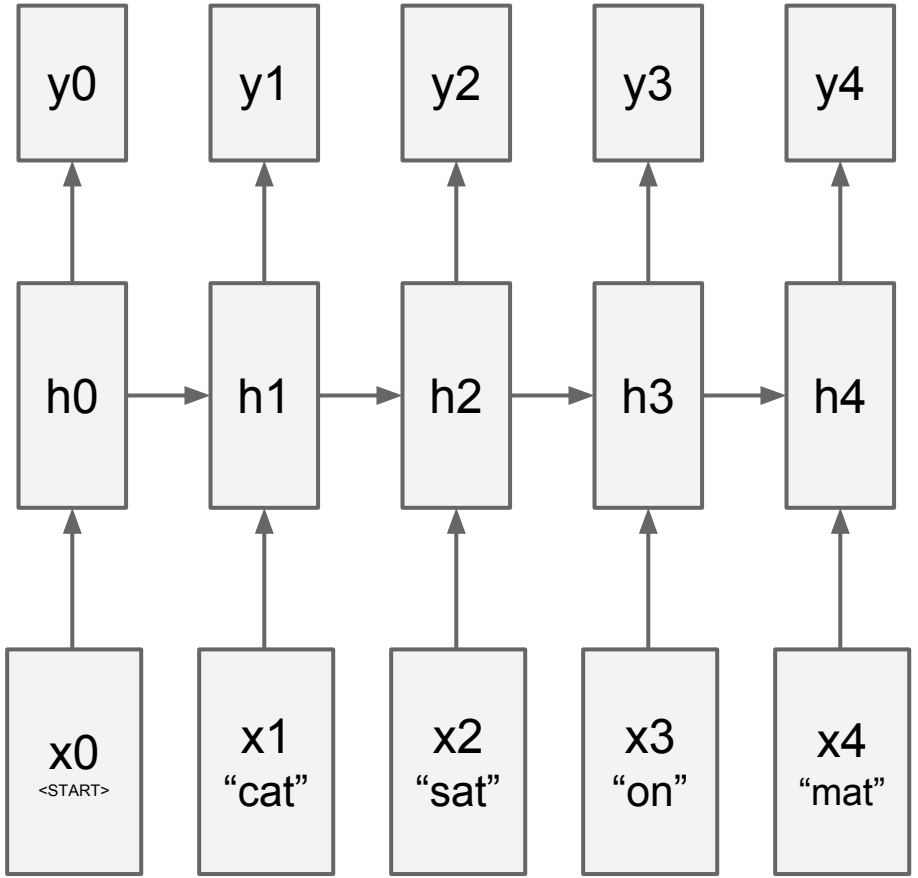


Image Sentence Datasets

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

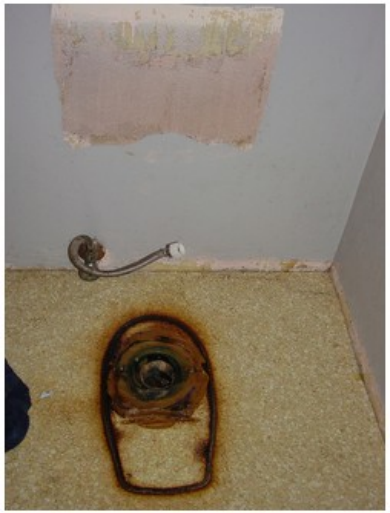
Well, I can kind of see it

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

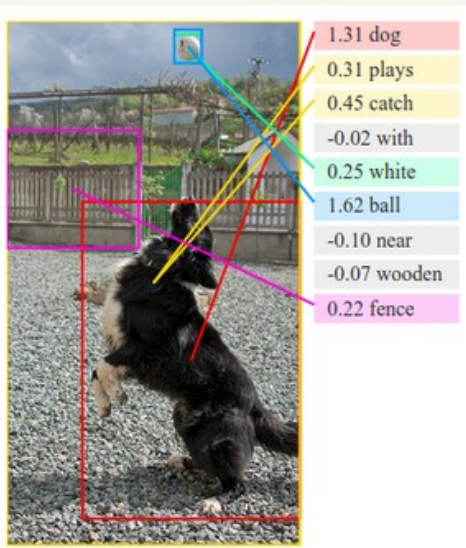
a woman holding a teddy bear in front of a mirror
logprob: -9.65

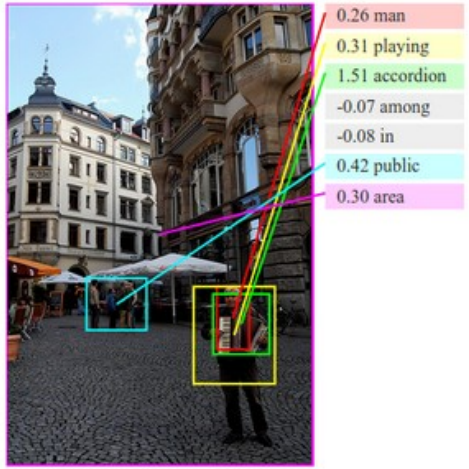
a horse is standing in the middle of a road
logprob: -10.34

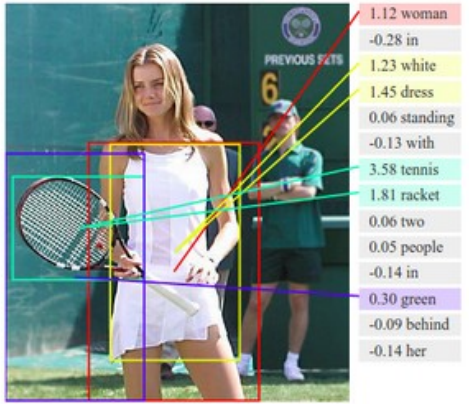
More examples in Web demo: http://bit.ly/neuraltalkdemo

Ranking and Retrieval

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



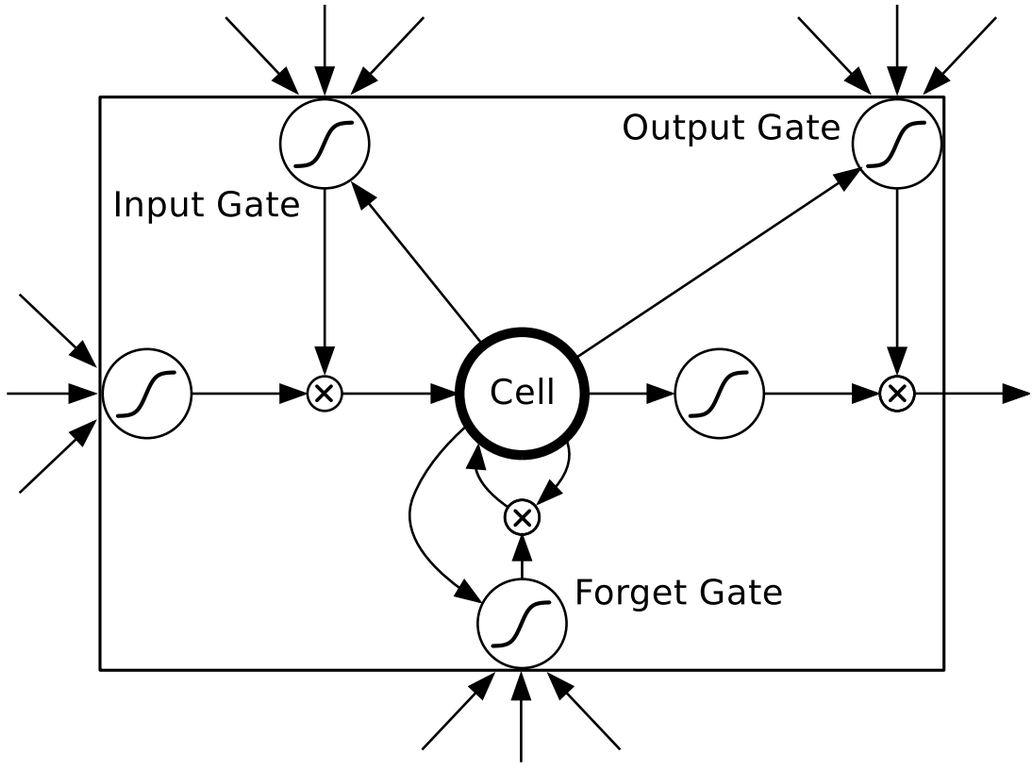




More examples in Web demo: http://bit.ly/rankingdemo

Long Short-Term Memory

- LSTM is an RNN architecture designed to have a longer memory. It uses linear memory cells surrounded by multiplicative gate units to store information



Input gate: scales input to cell (write)

Output gate: scales output from cell (read)

Forget gate: scales old cell value (reset)

- S. Hochreiter and J. Schmidhuber, “Long Short-term Memory” Neural Computation 1997

Neural Turing Machines

Alex Graves gravesa@google.com
Greg Wayne gregwayne@google.com
Ivo Danihelka danihelka@google.com

Google DeepMind, London, UK

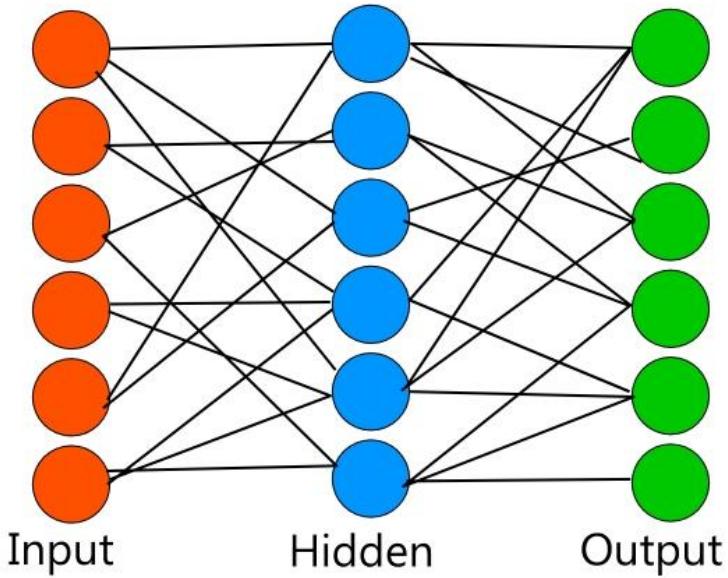
Goal: “Solve intelligence”

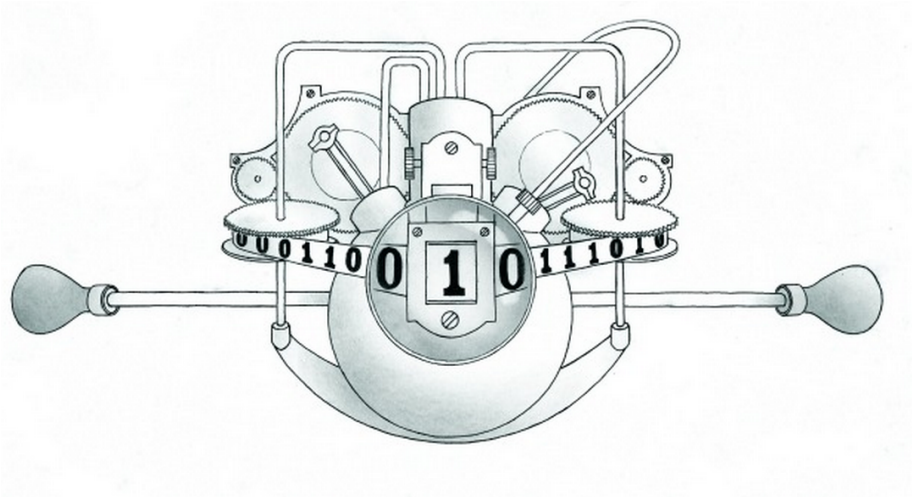
Price tag: *$400 million*

Abstract

We extend the capabilities of neural networks by coupling them to external memory resources, which they can interact with by attentional processes. The combined system is analogous to a Turing Machine or Von Neumann architecture but is differentiable end-to-end, allowing it to be efficiently trained with gradient descent. Preliminary results demonstrate that *Neural Turing Machines* can infer simple algorithms such as copying, sorting, and associative recall from input and output examples.

Building a Learning Machine

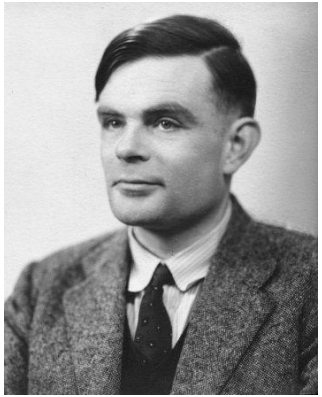




"Learning"
Input-Output mapping ~ rule

Formal model of solving a computational problem
rules + memory

Turing machine

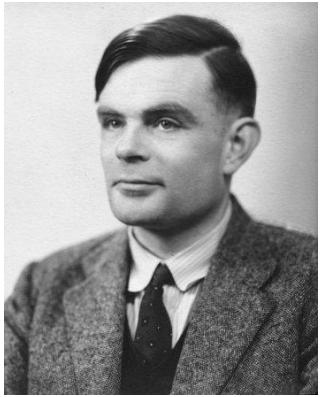


- What can be computed?

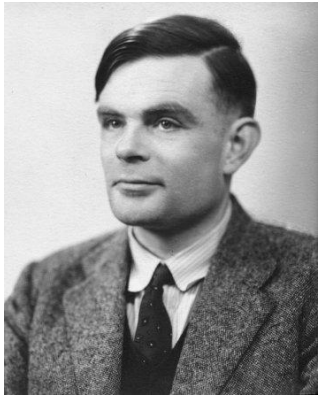
- Computability = instructions that lead to completion of task

http://plato.stanford.edu/entries/turing-machine

Turing machine



Turing machine

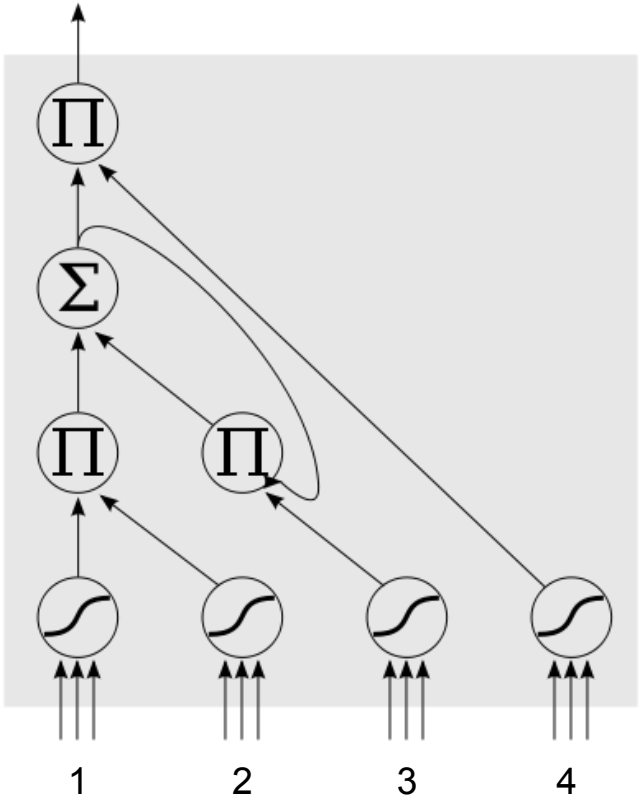


1. Tape (“memory”)
2. Read and write device (“head”)
3. Keeps track of current state (“state register”)
4. Instructions
    a. “If machine in $state_{current}$ and tape value is 0, go to $state_{next}$ and move left 1 space”

Recurrent neural networks

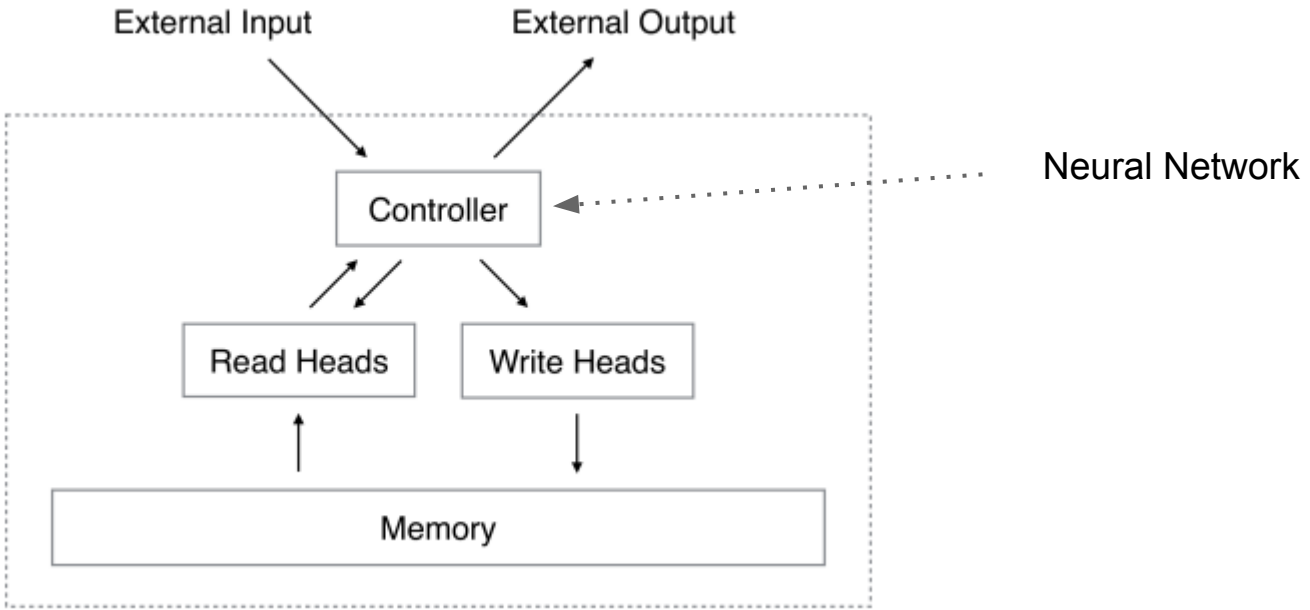
Long Short-Term Memory

1. Input
2. Input gate
3. “Remember” gate
4. Output gate



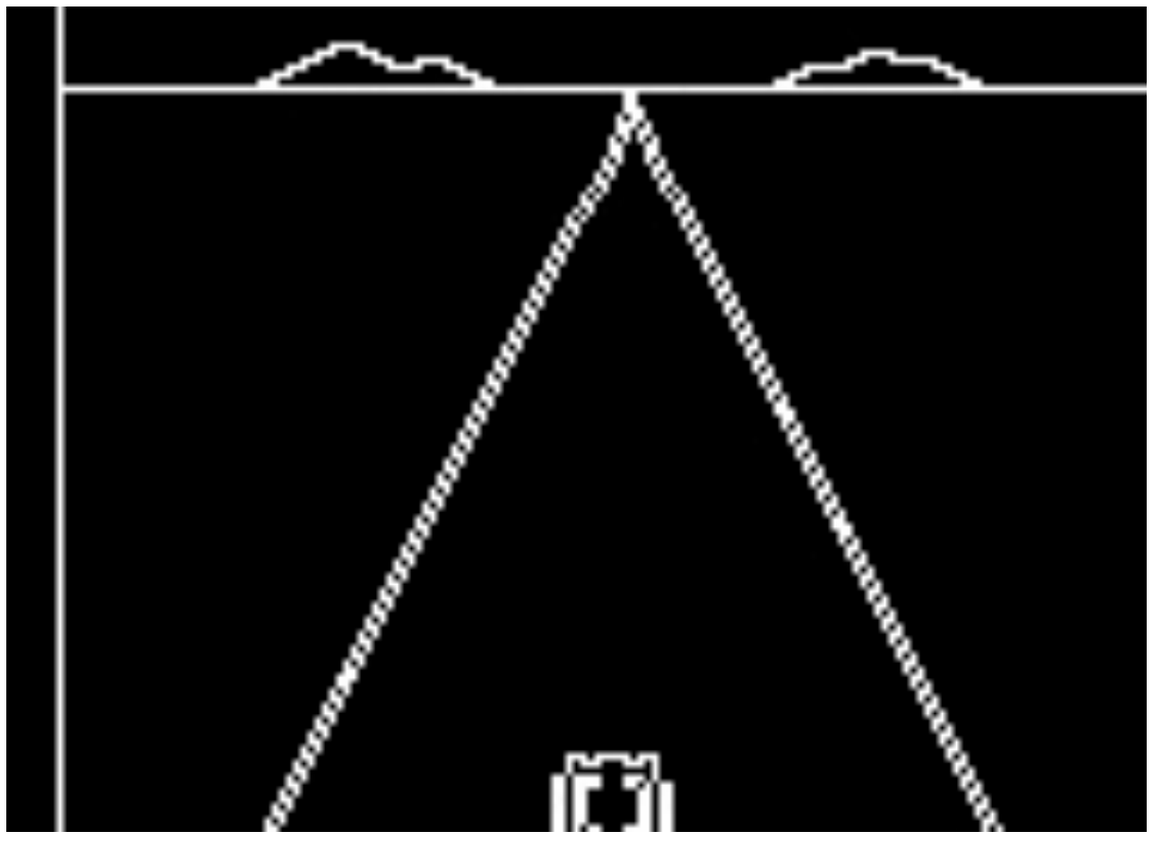
Somewhat complicated, lots of parameters

Neural Turing Machines

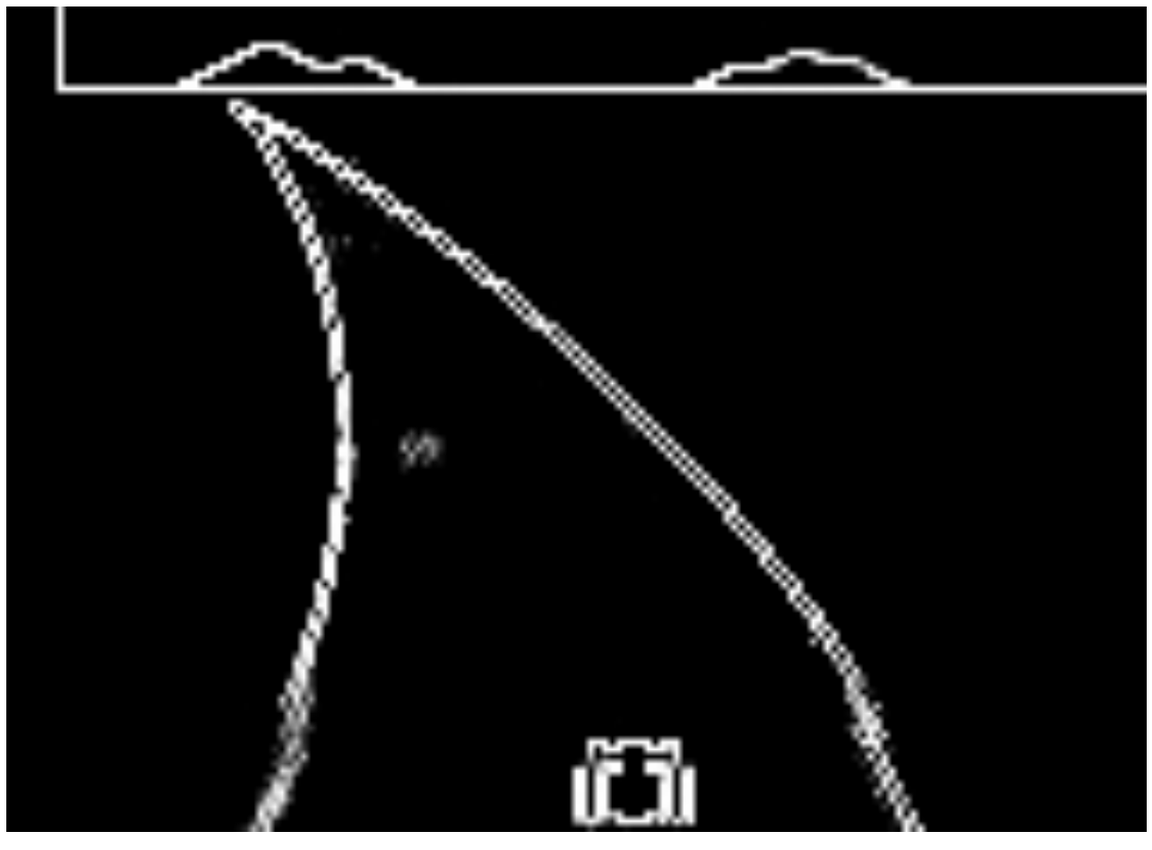


Atari Experiments

Real



Generated



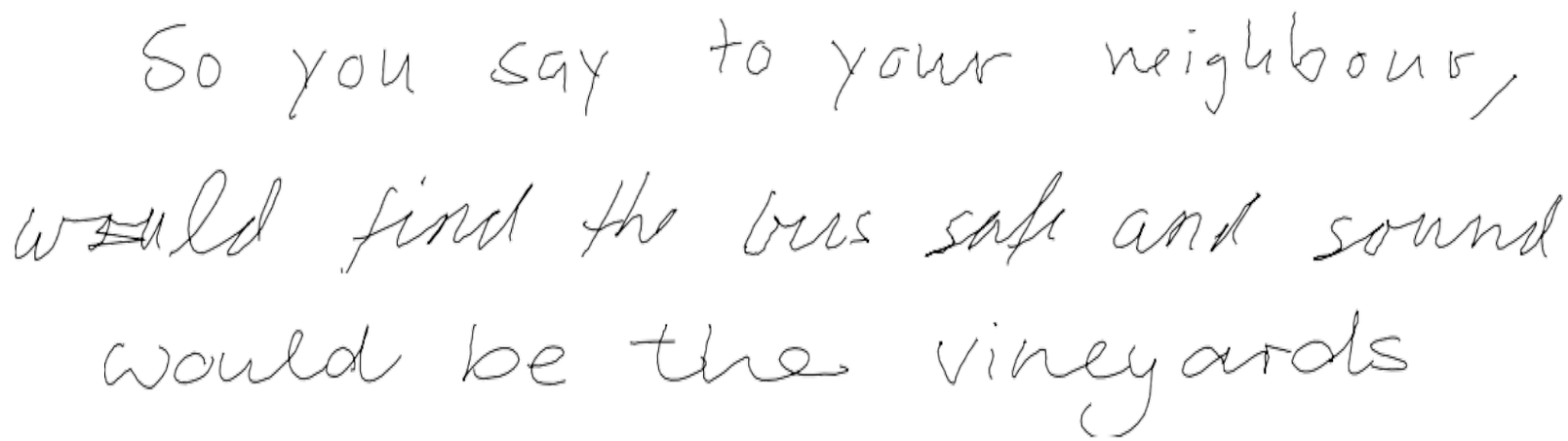
Karol Gregor, Ivo Danihelka, Andriy Mnih, Daan Wierstra…

Real

Generated

Handwriting Experiments

- Task: generate pen trajectories by predicting one (x,y) point at a time
- Data: IAM online handwriting, 10K training sequences, many writers, unconstrained style, captured from a whiteboard



- First problem: what to use for the density model?

Which is Real?

that a doctor should be

that a doctor should be

thaet a docaus should ae

that a doctor should be

that a doctor should be

that a doctor should be

Which is Real?

of presentlee alty inremmbering

of preseng reality in remmbering

of present reality in remembering

of present reality in remembering

of present reality in remembering

Which is Real?

from his travels it might have been

from his travels it might have been

from his travels 'it might have been

from his travels it mightmare been

from his travels it might have been

from his travels - it might have been

Biased Sampling

when the samples are biased

towards more probable sequences

they get easier to read

but less interesting to look at.

Primed Sampling

when the sample starts with real data

(prison welfare Officer complement)

if continues in the same style

(He dismissed the idea)

Primed *and* Biased

Take the breath away where they are

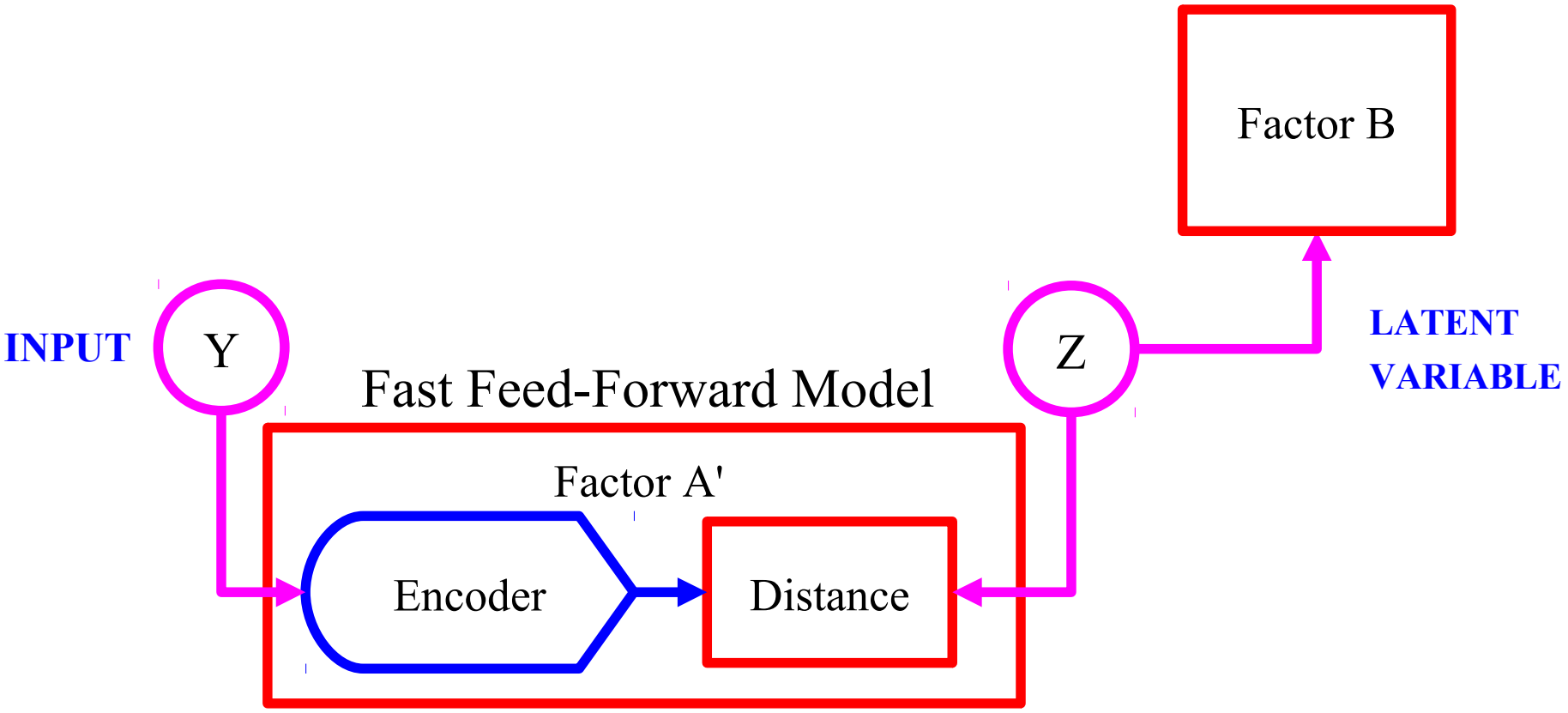
when the network is primed
and biased, it writes
in a cleaned up version
of the original style

Demo

http://www.cs.toronto.edu/~graves/handwriting.html

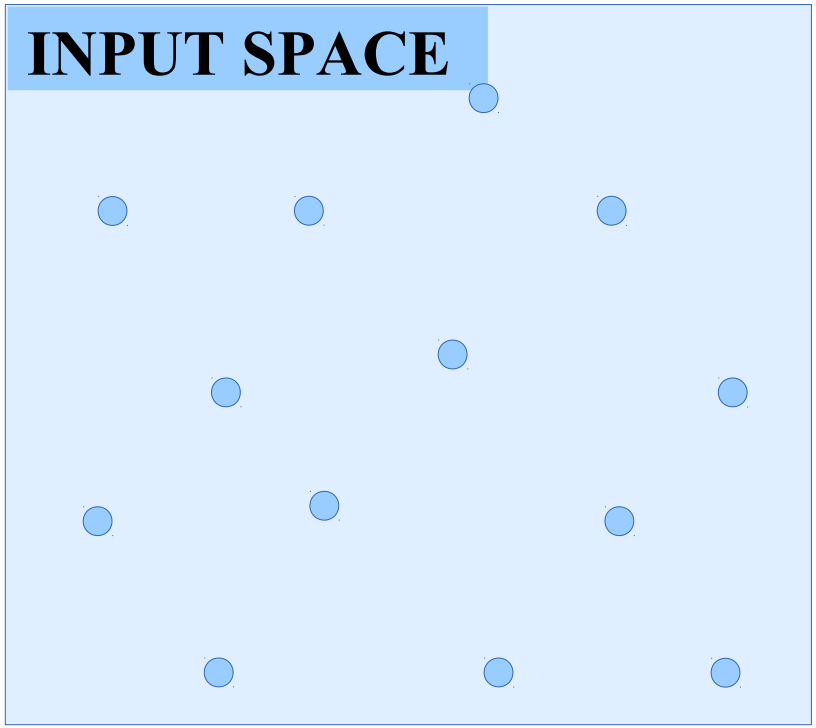
Encoder Architecture

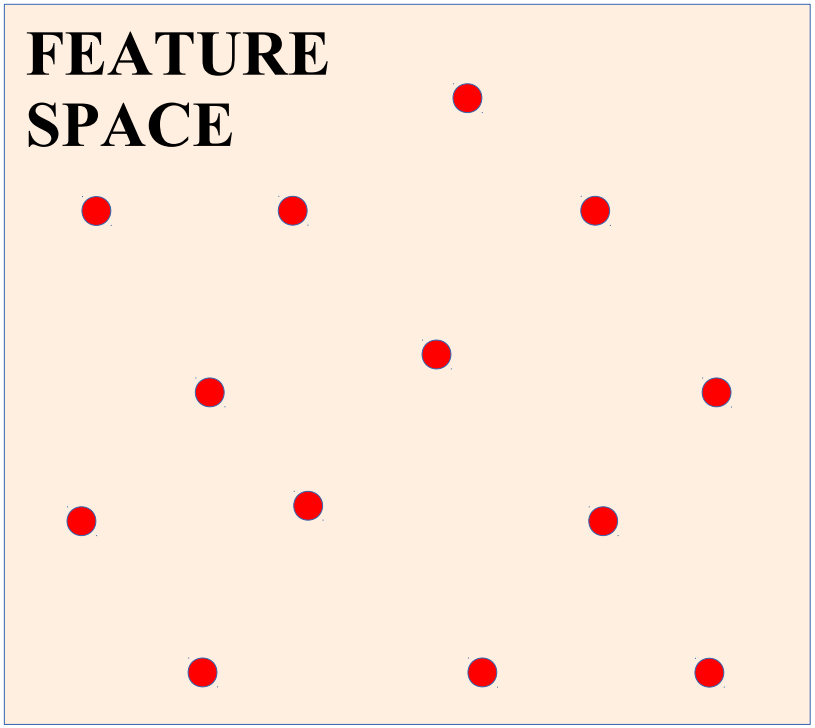
Examples: most ICA models, Product of Experts



Why Limit the Information Content of the Code?

- Training sample
- Input vector which is NOT a training sample
- Feature vector

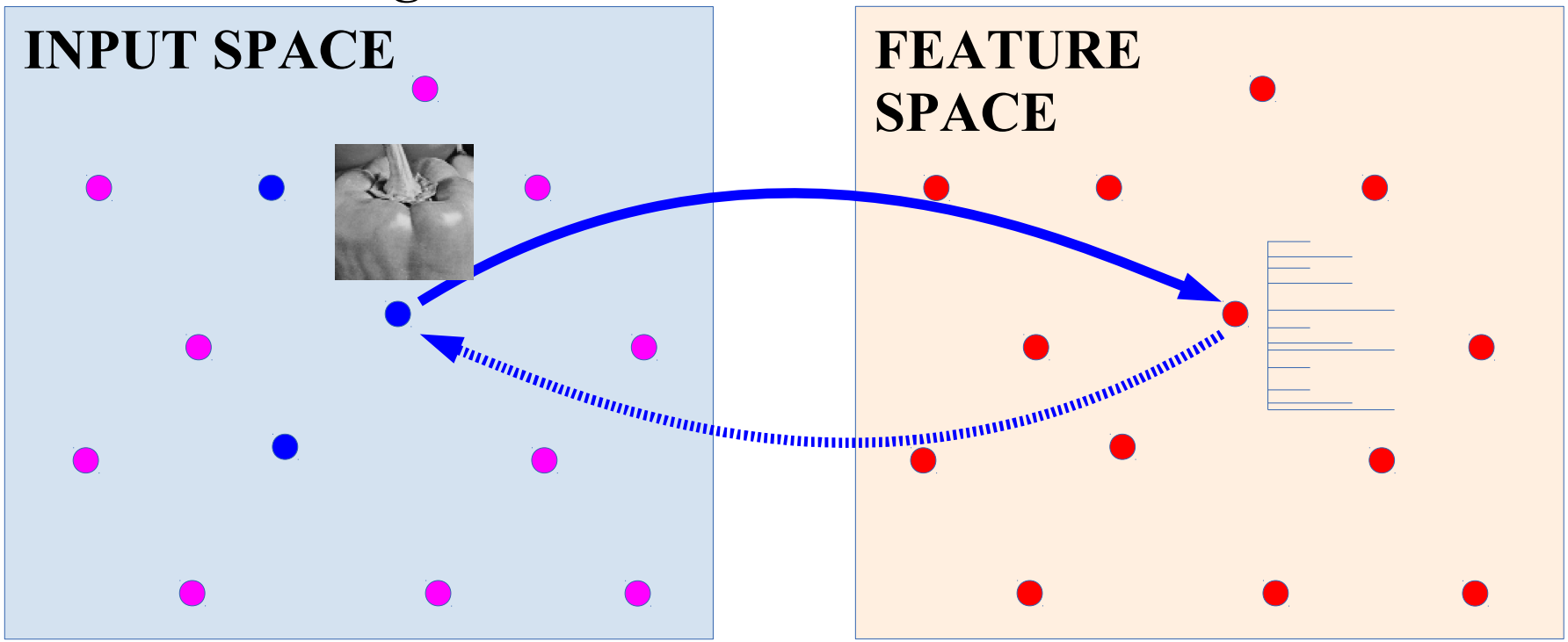




Why Limit the Information Content of the Code?

- **Training sample**
- **Input vector which is NOT a training sample**
- **Feature vector**

Training based on minimizing the reconstruction error over the training set

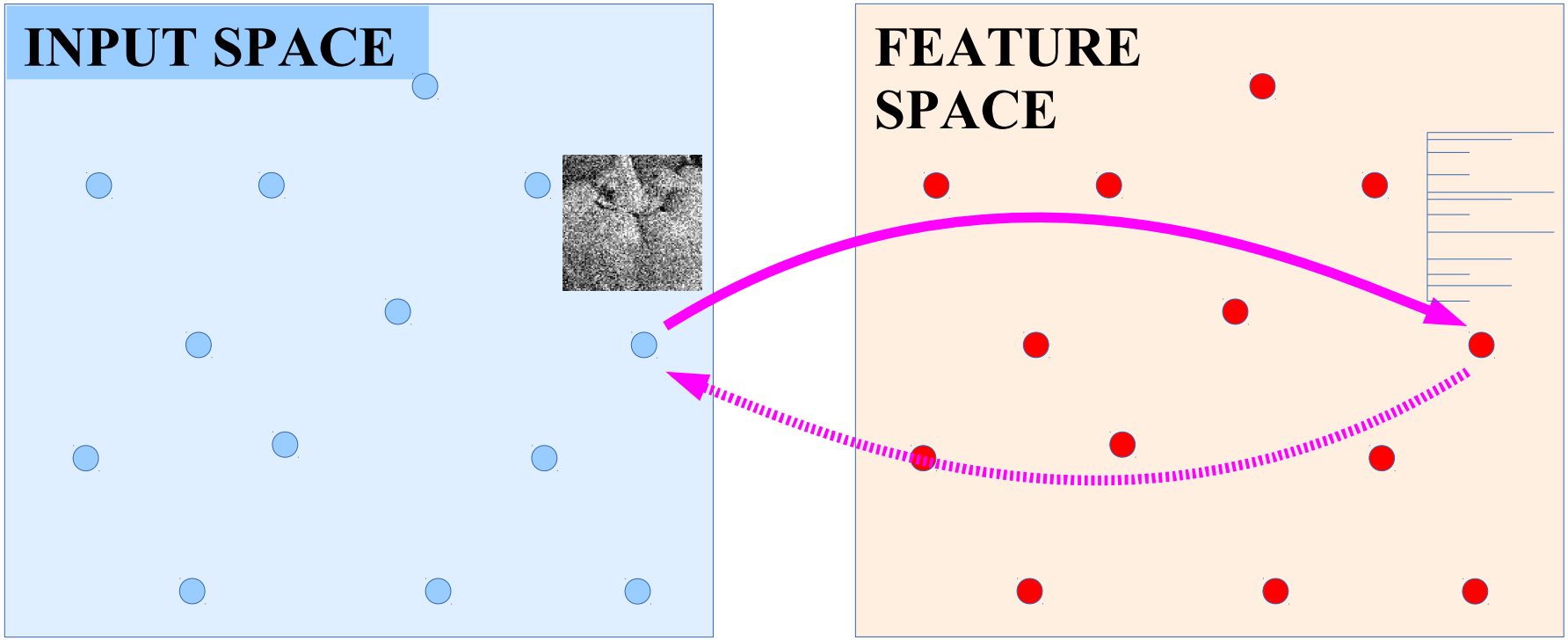


Why Limit the Information Content of the Code?

- **Training sample**
- **Input vector which is NOT a training sample**
- **Feature vector**

BAD: machine does not learn structure from training data!!

It just copies the data.

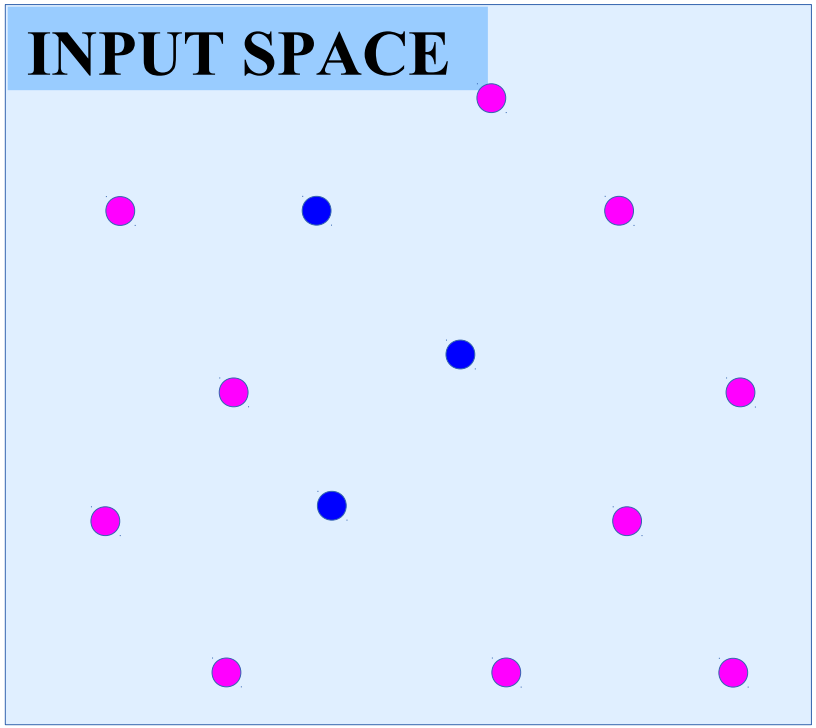


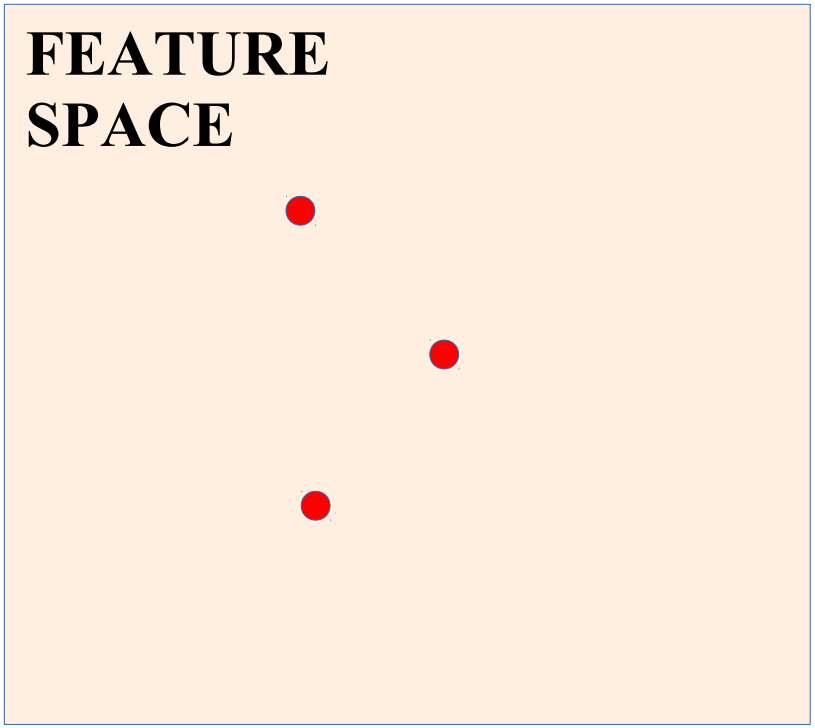
Why Limit the Information Content of the Code?

Y LeCun
MA Ranzato

- **Training sample**
- **Input vector which is NOT a training sample**
- **Feature vector**

IDEA: reduce number of available codes.

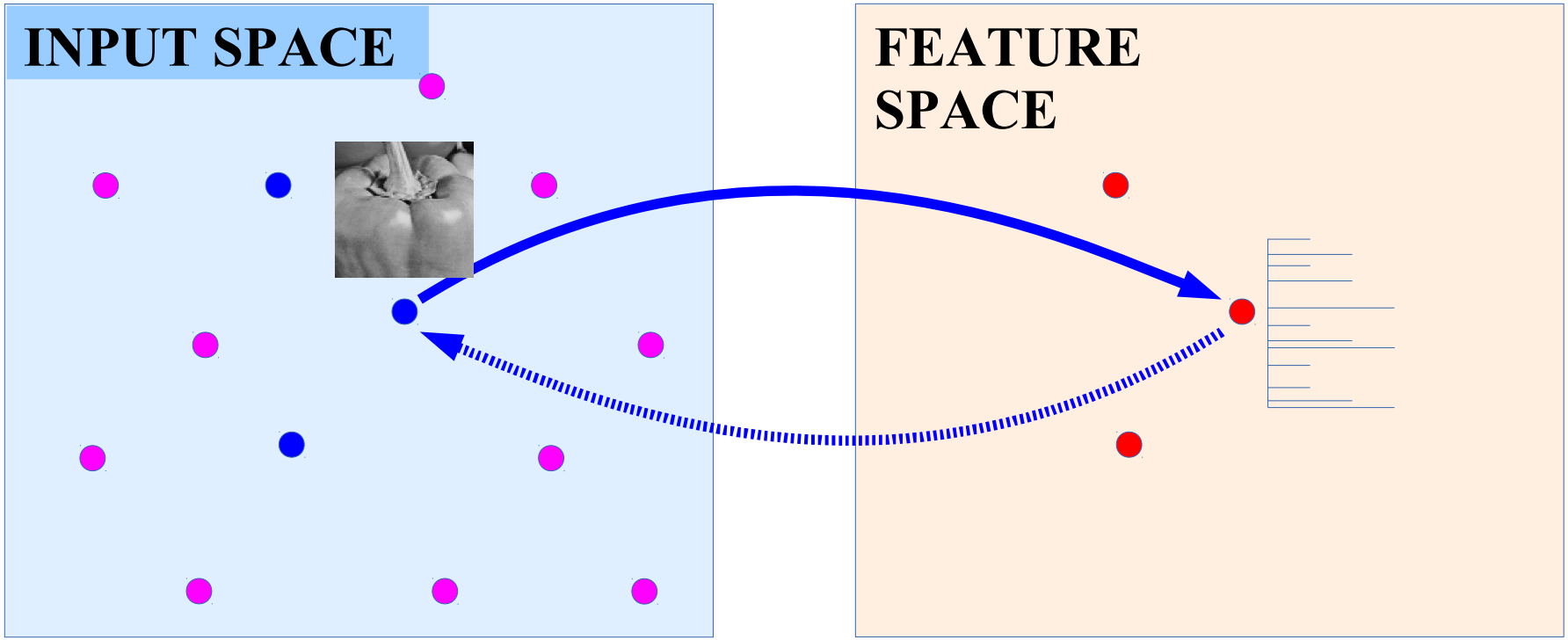




Why Limit the Information Content of the Code?

- **Training sample**
- **Input vector which is NOT a training sample**
- **Feature vector**

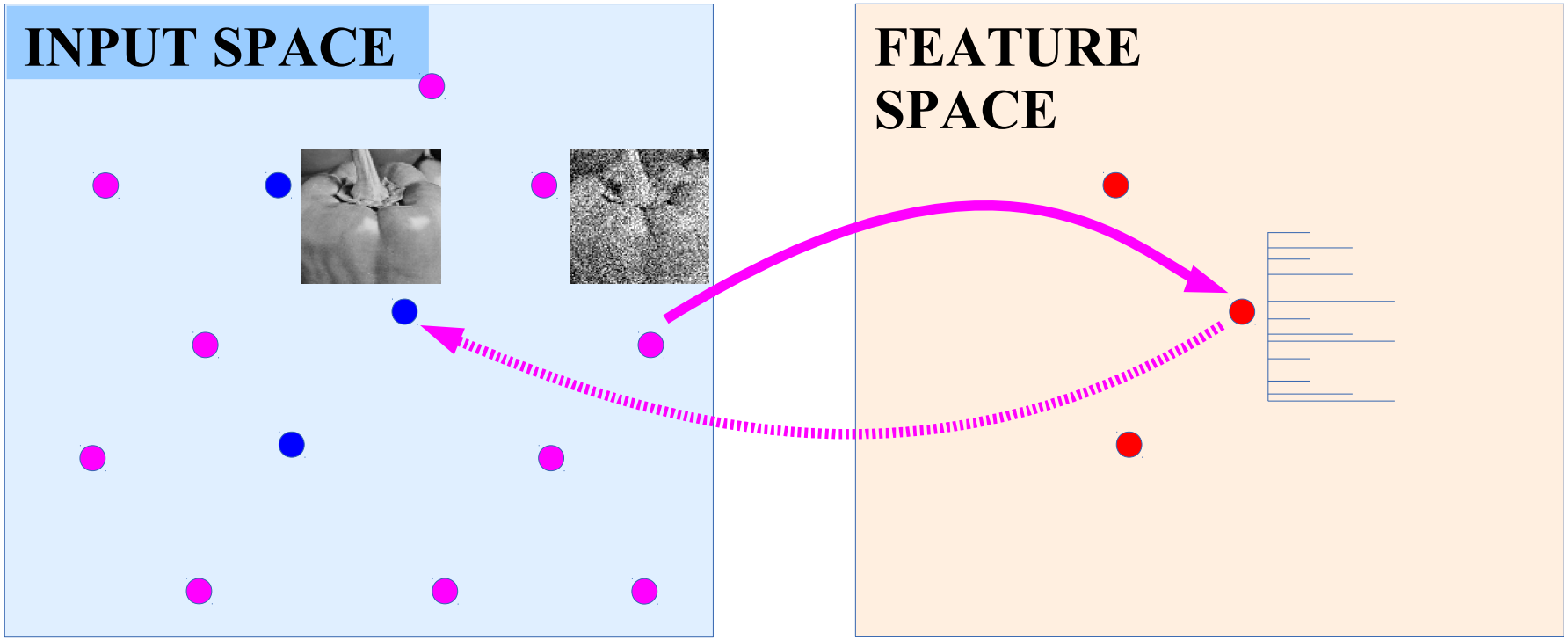
IDEA: reduce number of available codes.



Why Limit the Information Content of the Code?

- **Training sample**
- **Input vector which is NOT a training sample**
- **Feature vector**

IDEA: reduce number of available codes.



Adversarial Examples

- What is adversarial example? We can let the network to misclassify an image by adding a imperceptible (for human) perturbation.

- Why do adversarial examples exist? Deep Neural Networks learn input-output mappings that are discontinuous to a significant extent.

- Interesting observation: the adversarial examples generated for network A can also make network B fail.

Generate Adversarial Examples

Input image: $x \in \mathbb{R}^m$

Classifier: $f : \mathbb{R}^m \longrightarrow \{1 \ldots k\}$

Target label: $l \in \{1 \ldots k\}$

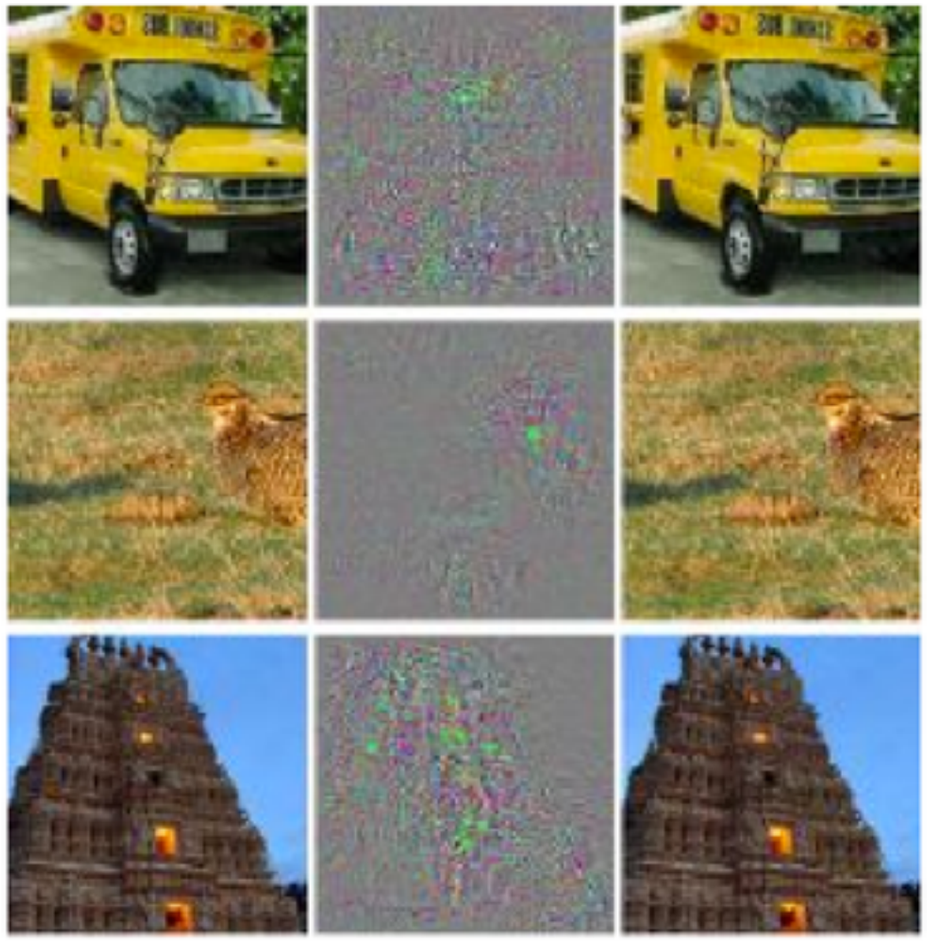
Minimize $\|r\|_2$ **subject to:**

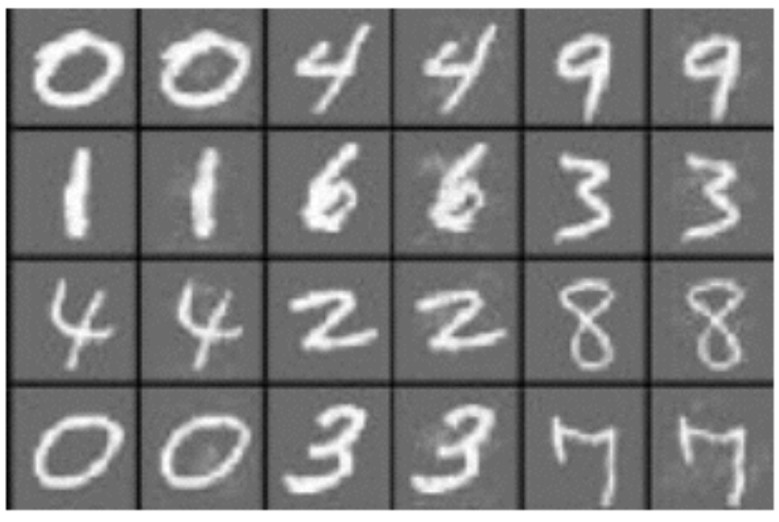
1. $f(x + r) = l$
2. $x + r \in [0, 1]^m$

x+r is the closest image to ***x*** classified as ***l*** by ***f***.

When $f(x) \neq l$:

Minimize $c|r| + \text{loss}_f(x + r, l)$ **subject to** $x + r \in [0, 1]^m$





Intriguing properties

- Properties:
  - Visually hard to distinguish the generated adversarial examples.
  - Cross model generalization. (different hyper-parameters)
  - Cross training-set generalization. (different training set)
- Observation:
  - adversarial examples are universal.
  - back-feeding adversarial examples to training might improve generalization of the model.

Experiment

Cross-model generalization of adversarial examples.

	softmax1	softmax2	softmax3	N100-100-10	N200-200-10	AE400-10	Av. distortion
softmax with $\lambda = 10^{-4}$	100%	11.7%	22.7%	2%	3.9%	2.7%	0.062
softmax with $\lambda = 10^{-2}$	87.1%	100%	35.2%	35.9%	27.3%	9.8%	0.1
softmax with $\lambda = 1$	71.9%	76.2%	100%	48.1%	47%	34.4%	0.14
N100-100-10	28.9%	13.7%	21.1%	100%	6.6%	2%	0.058
N200-200-10	38.2%	14%	23.8%	20.3%	100%	2.7%	0.065
AE400-10	23.4%	16%	24.8%	9.4%	6.6%	100%	0.086
Gaussian noise, stddev=0.1	5.0%	10.1%	18.3%	0%	0%	0.8%	0.1
Gaussian noise, stddev=0.3	15.6%	11.3%	22.7%	5%	4.3%	3.1%	0.3

Experiment

Cross training-set generalization - baseline (no distortion)

Model	Error on P_1	Error on P_2	Error on Test	Min Av. Distortion
M_1: 100-100-10 trained on P_1	0%	2.4%	2%	0.062
M_1': 123-456-10 trained on P_1	0%	2.5%	2.1%	0.059
M_2: 100-100-10 trained on P_2	2.3%	0%	2.1%	0.058

Cross training-set generalization error rate

	M_1	M_1'	M_2
Distorted for M_1 (av. stddev=0.062)	100%	26.2%	5.9%
Distorted for M_1' (av. stddev=0.059)	6.25%	100%	5.1%
Distorted for M_2 (av. stddev=0.058)	8.2%	8.2%	100%
Gaussian noise with stddev=0.06	2.2%	2.6%	2.4%
Distorted for M_1 amplified to stddev=0.1	100%	98%	43%
Distorted for M_1' amplified to stddev=0.1	96%	100%	22%
Distorted for M_2 amplified to stddev=0.1	27%	50%	100%
Gaussian noise with stddev=0.1	2.6%	2.8%	2.7%

magnify distortion

The Opposite Direction

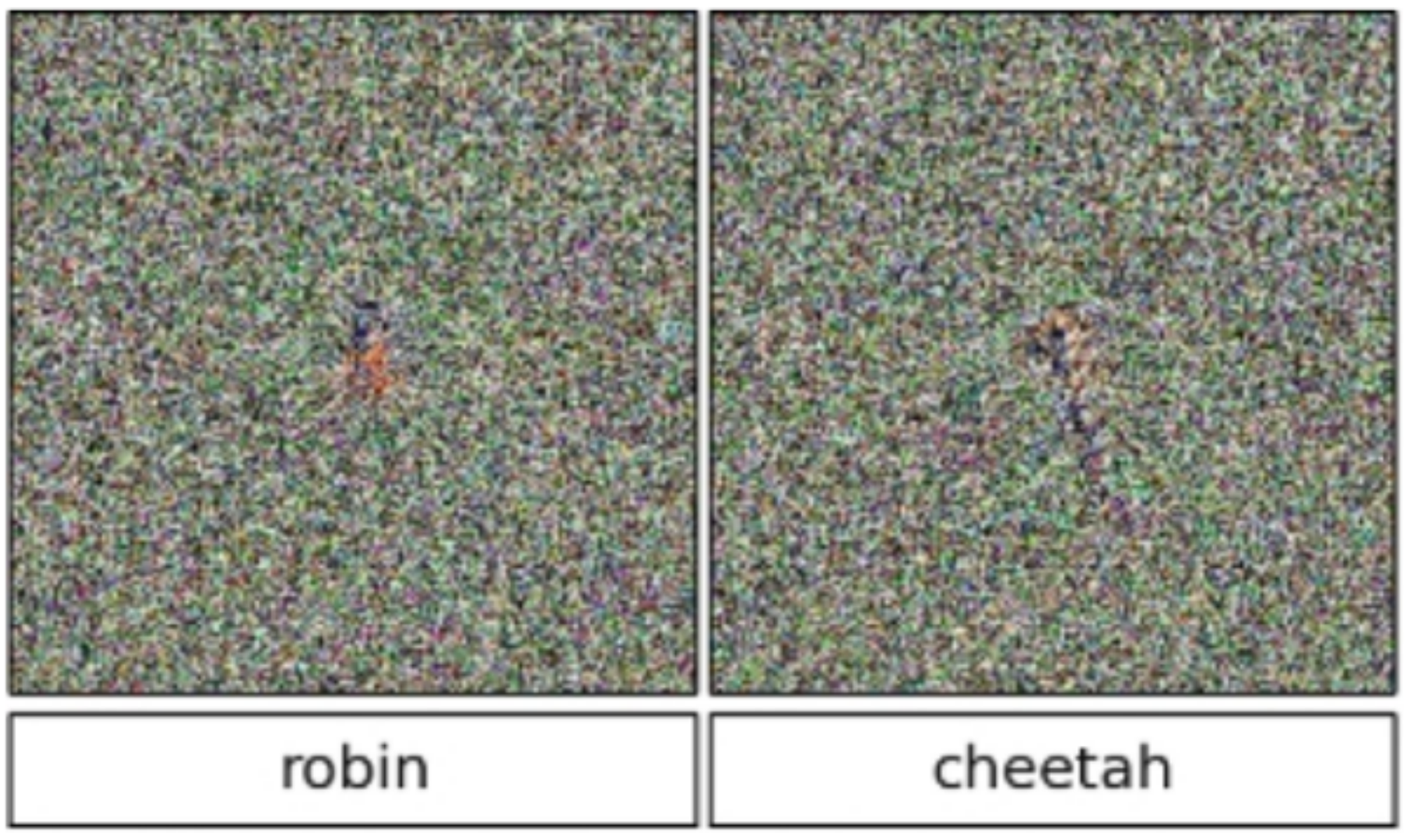
Imperceptible adversarial examples that cause misclassification.

Unrecognizable images that make DNN believe

Nguyen, Anh, Jason Yosinski, and Jeff Clune. "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images." arXiv preprint arXiv:1412.1897 (2014).

Fooling Examples

Problem statement: producing images that are completely unrecognizable to humans, but that state-of-the-art Deep Neural Networks believe to be recognizable objects with high confidence (99%).



DNN Models

- ImageNet: AlexNet. (Caffe version)
  - 42.6% error rate. Original error rate is 40.7%.
- MNIST: LeNet (Caffe version)
  - 0.94% error rate. Original error rate is 0.8%.

Generating Images with Evolution (one class)

- Evolutionary Algorithms (EAs) are inspired by Darwinian evolution.

- Contains a population of organisms (images).

- Organisms will be randomly perturbed and selected based on *fitness function*.

- Fitness function: in our case, is the highest prediction value a DNN believes that the image belongs to a class.

Generating Images with Evolution (multi-class)

- Algorithm: Multi-dimensional archive of phenotypic elites MAP-Elites.

- Procedures:
  - Randomly choose an organism, mutate it randomly.
  - Show the mutated organism to the DNN. If the prediction score is higher than the current highest score of ***ANY*** class, make the organism as the champion of that class.

Slide References

- Stanford CS 231n
- Univ of Notre Dame: CSE 60647, Spring 2014
- Bayesian Behavior Lab, UNW
- Princeton, COS598 Spring 2015: The Unreasonable Effectiveness of Big Visual Data
- Slides from Deep Learning tutorials
- Slides from DeepMind