



III
p

max planck institut
informatik

SIC Saarland Informatics
Campus

High Level Computer Vision

Data Preprocessing & Recurrent Neural Network (RNNs)

@ May 29, 2023

Bernt Schiele

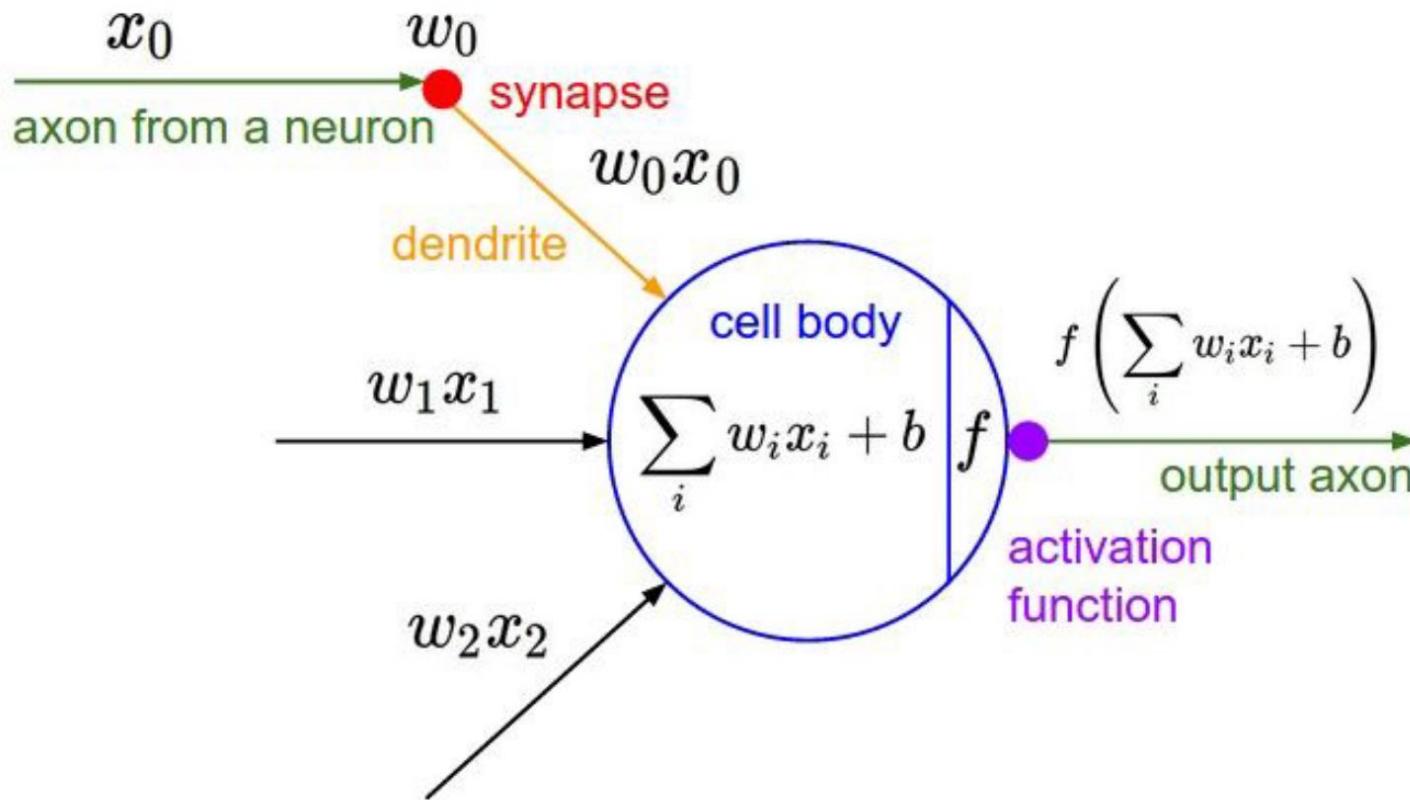
cms.sic.saarland/hlcvss24/

**Max Planck Institute for Informatics & Saarland University,
Saarland Informatics Campus Saarbrücken**

Overview Today's Lecture

- Data Preprocessing
 - ▶ Activation functions
 - ▶ Batch normalization
- Recurrent Neural Networks (RNNs)
 - ▶ Motivation & flexibility of RNNs
 - ▶ Language modeling
 - including “unreasonable effectiveness of RNNs”
 - ▶ RNNs for image description / captioning
 - ▶ Standard RNN and a particularly successful RNN:
Long Short Term Memory (LSTM)
 - including “visualizations of RNN cells”

Activation Functions



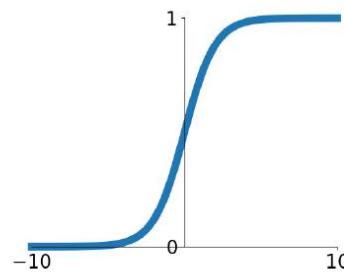
slide credit: Fei-Fei, Justin Johnson, Serena Yeung



Activation Functions

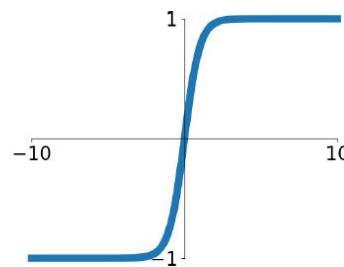
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



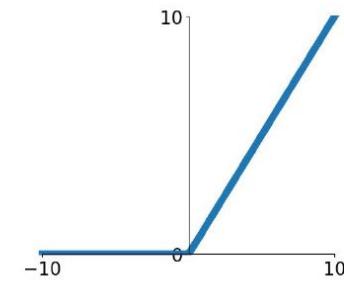
tanh

$$\tanh(x)$$

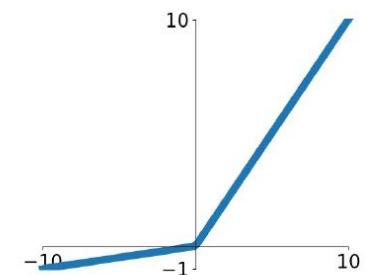


ReLU

$$\max(0, x)$$



Leaky ReLU

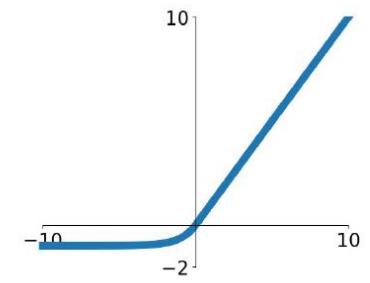
$$\max(0.1x, x)$$


Maxout

$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

ELU

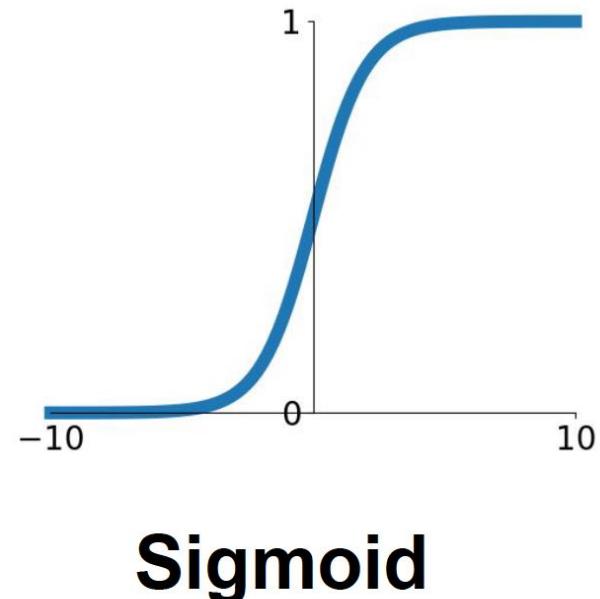
$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



slide credit: Fei-Fei, Justin Johnson, Serena Yeung



Activation Functions



$$\sigma(x) = 1/(1 + e^{-x})$$

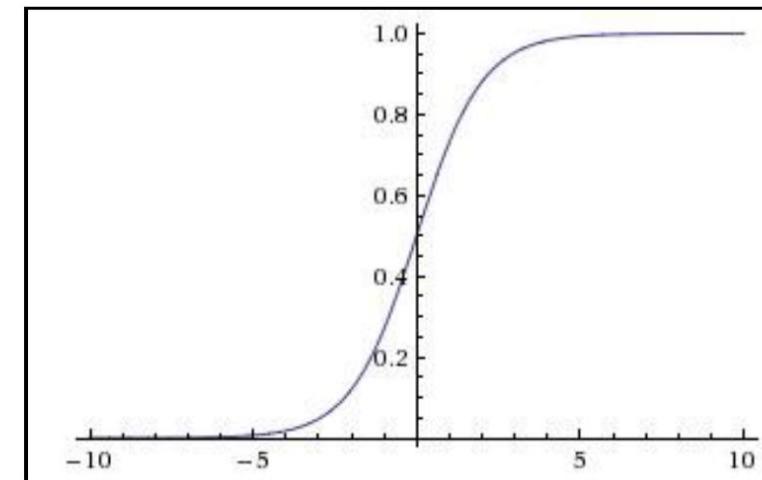
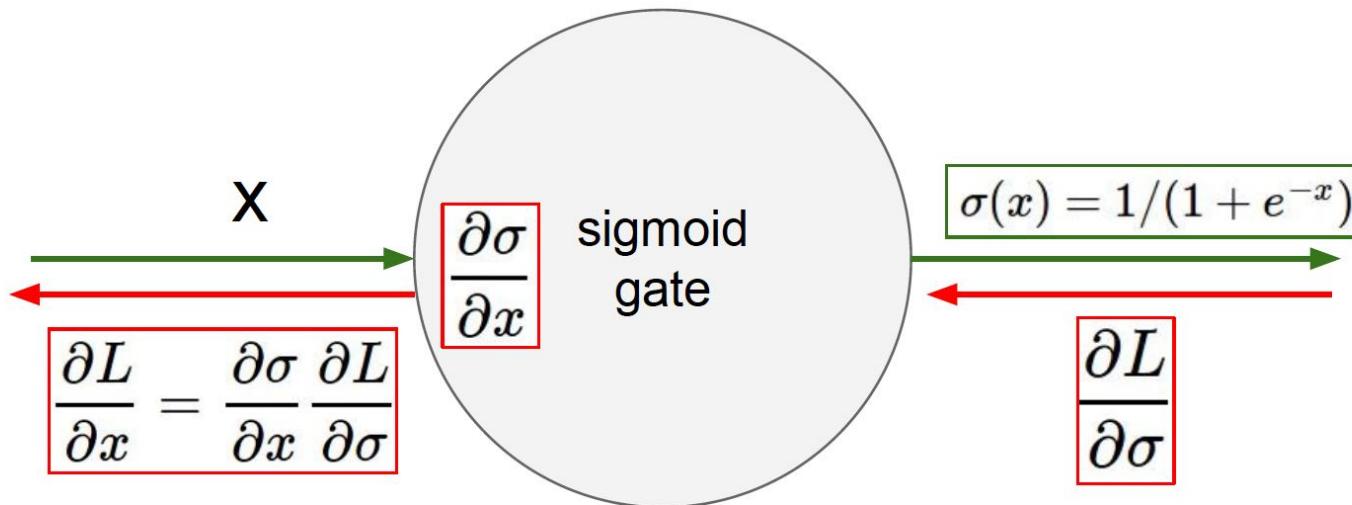
- Squashes numbers to range [0,1]
- Historically popular since they have nice interpretation as a saturating “firing rate” of a neuron

3 problems:

1. **Saturated neurons “kill” the gradients**

slide credit: Fei-Fei, Justin Johnson, Serena Yeung





What happens when $x = -10$?

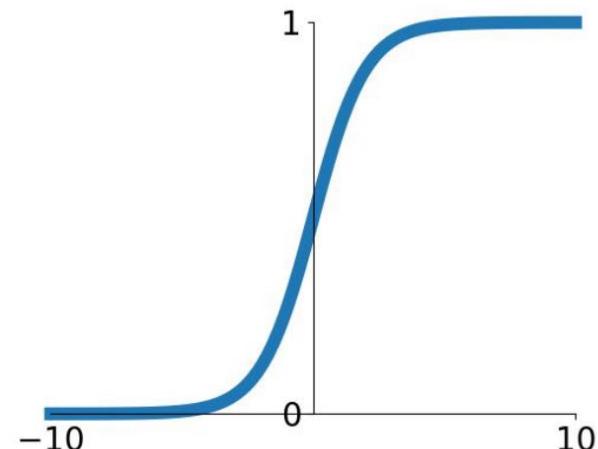
What happens when $x = 0$?

What happens when $x = 10$?

slide credit: Fei-Fei, Justin Johnson, Serena Yeung



Activation Functions



Sigmoid

$$\sigma(x) = 1/(1 + e^{-x})$$

- Squashes numbers to range [0,1]
- Historically popular since they have nice interpretation as a saturating “firing rate” of a neuron

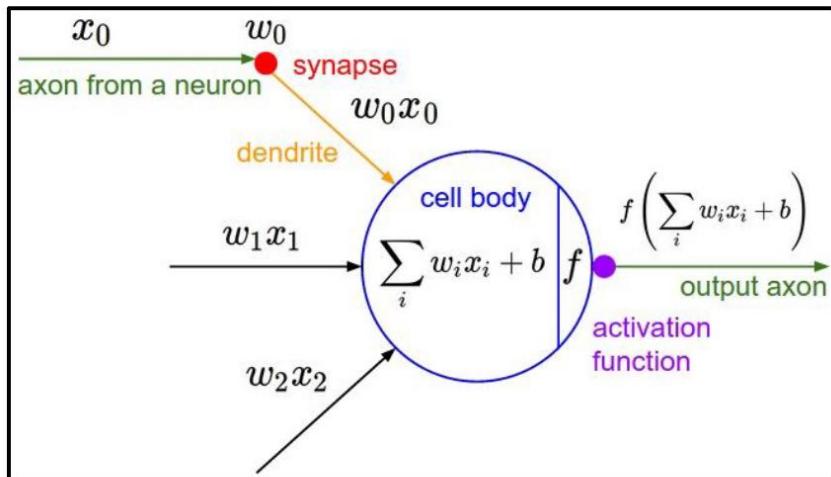
3 problems:

1. Saturated neurons “kill” the gradients
2. Sigmoid outputs are not zero-centered

slide credit: Fei-Fei, Justin Johnson, Serena Yeung



Consider what happens when the input to a neuron (x) is always positive:



$$f \left(\sum_i w_i x_i + b \right)$$

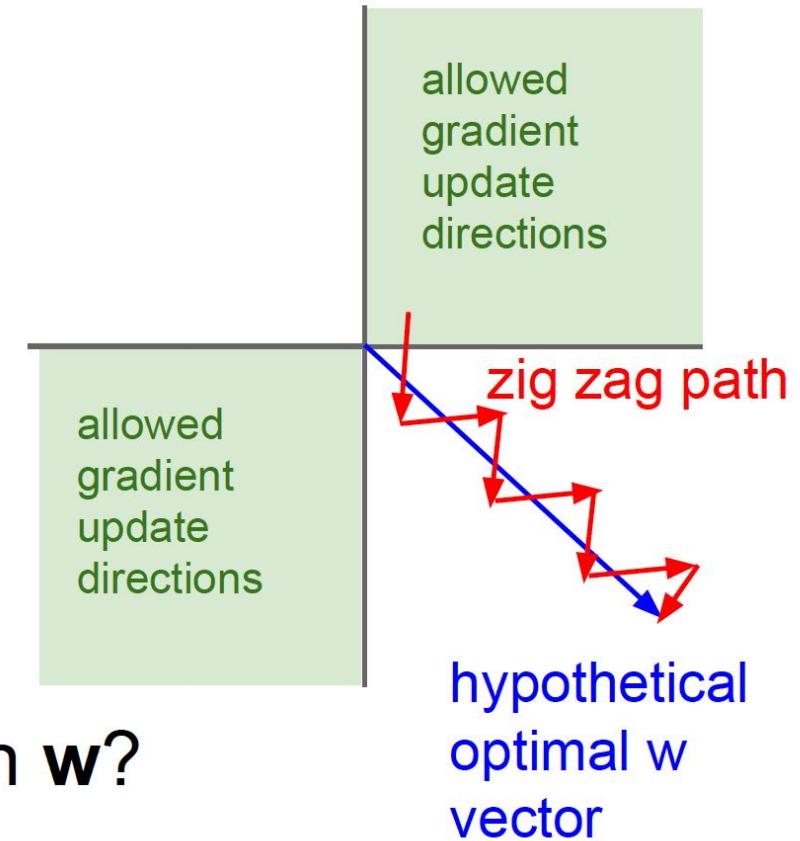
What can we say about the gradients on w ?

slide credit: Fei-Fei, Justin Johnson, Serena Yeung



Consider what happens when the input to a neuron is always positive...

$$f \left(\sum_i w_i x_i + b \right)$$

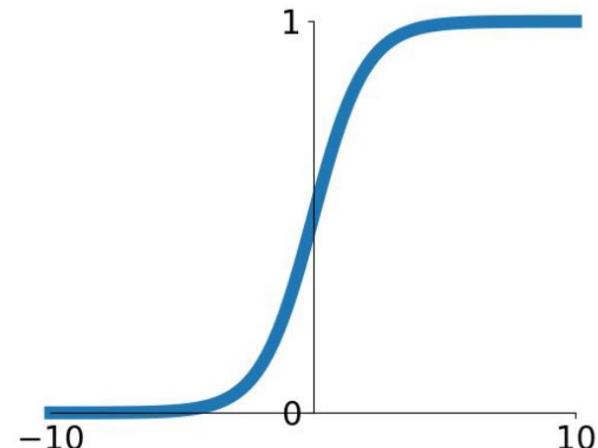


What can we say about the gradients on w ?
Always all positive or all negative :(
(this is also why you want zero-mean data!)

slide credit: Fei-Fei, Justin Johnson, Serena Yeung



Activation Functions



Sigmoid

$$\sigma(x) = 1/(1 + e^{-x})$$

- Squashes numbers to range [0,1]
- Historically popular since they have nice interpretation as a saturating “firing rate” of a neuron

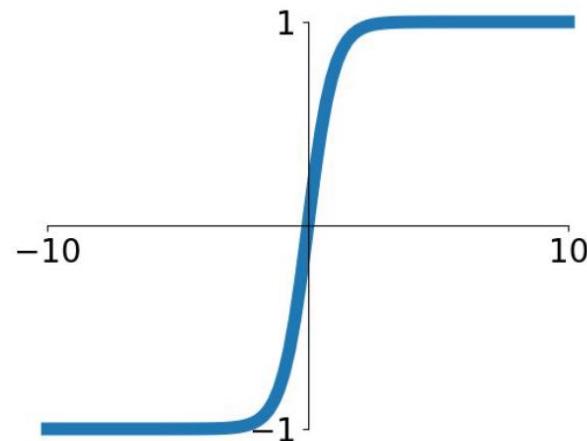
3 problems:

1. Saturated neurons “kill” the gradients
2. Sigmoid outputs are not zero-centered
3. $\exp()$ is a bit compute expensive

slide credit: Fei-Fei, Justin Johnson, Serena Yeung



Activation Functions



tanh(x)

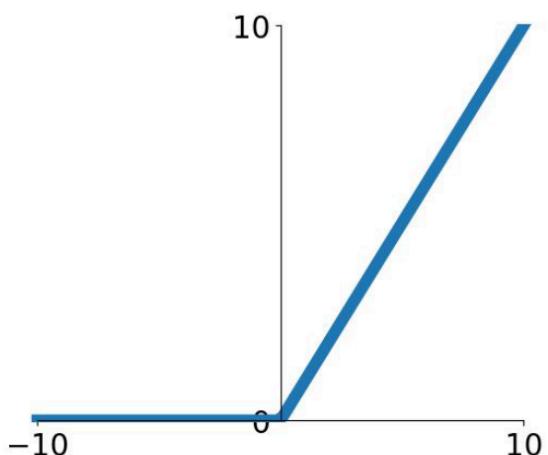
- Squashes numbers to range [-1,1]
- zero centered (nice)
- still kills gradients when saturated :(

[LeCun et al., 1991]

slide credit: Fei-Fei, Justin Johnson, Serena Yeung



Activation Functions



- Computes $f(x) = \max(0, x)$
- Does not saturate (in +region)
- Very computationally efficient
- Converges much faster than sigmoid/tanh in practice (e.g. 6x)
- Actually more biologically plausible than sigmoid

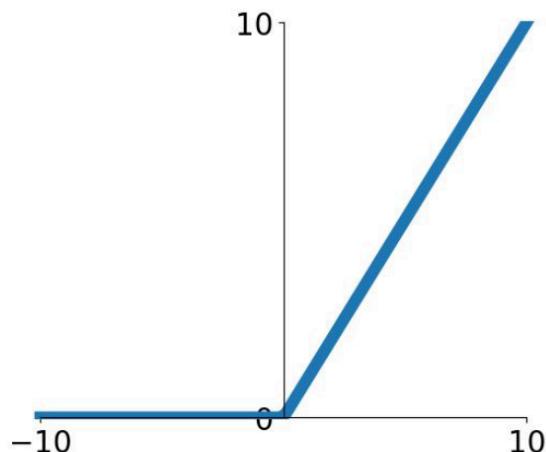
ReLU
(Rectified Linear Unit)

[Krizhevsky et al., 2012]

slide credit: Fei-Fei, Justin Johnson, Serena Yeung



Activation Functions



ReLU
(Rectified Linear Unit)

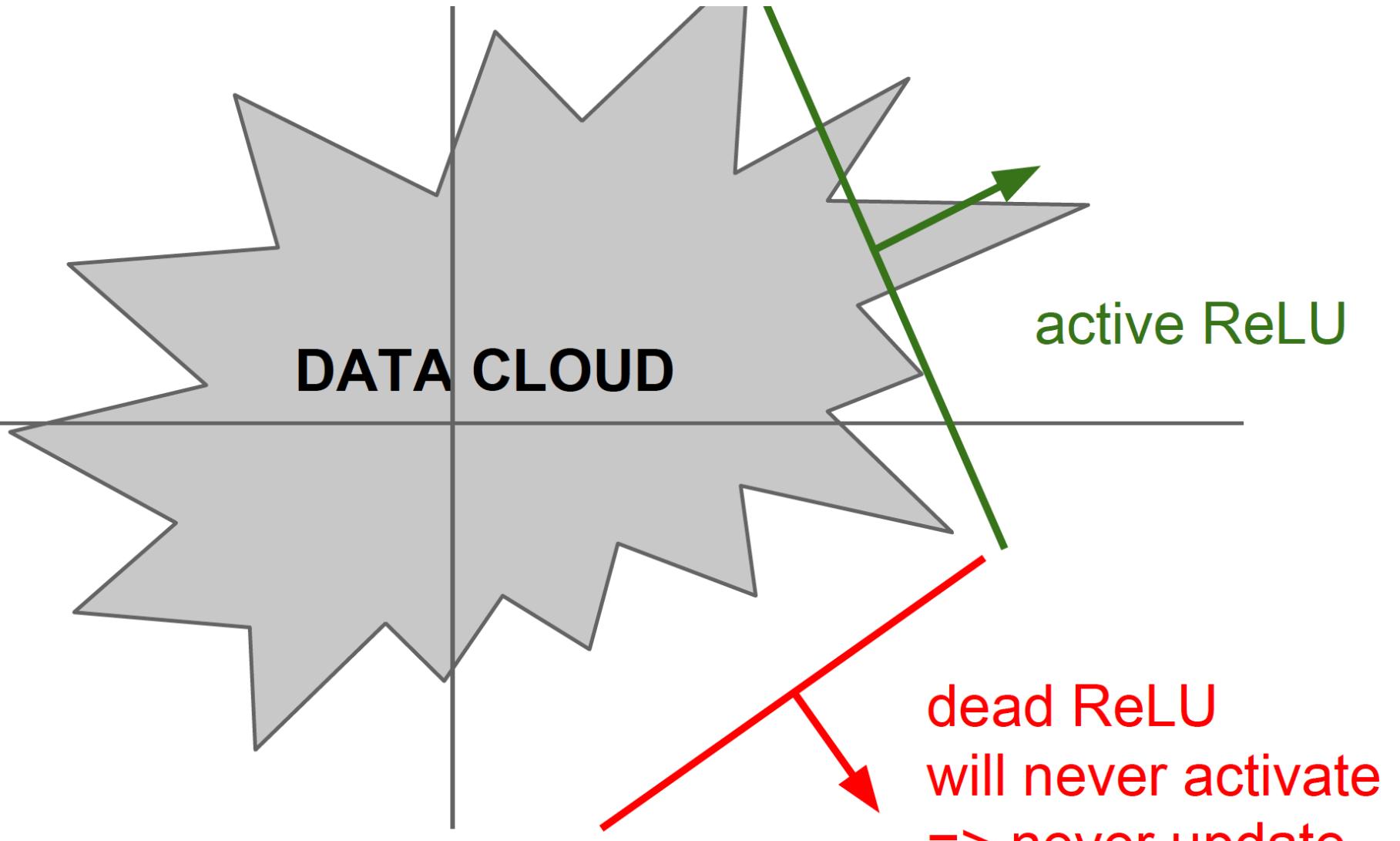
- Computes $f(x) = \max(0, x)$
- Does not saturate (in +region)
- Very computationally efficient
- Converges much faster than sigmoid/tanh in practice (e.g. 6x)
- Actually more biologically plausible than sigmoid

- Not zero-centered output
- An annoyance:

hint: what is the gradient when $x < 0$?

slide credit: Fei-Fei, Justin Johnson, Serena Yeung

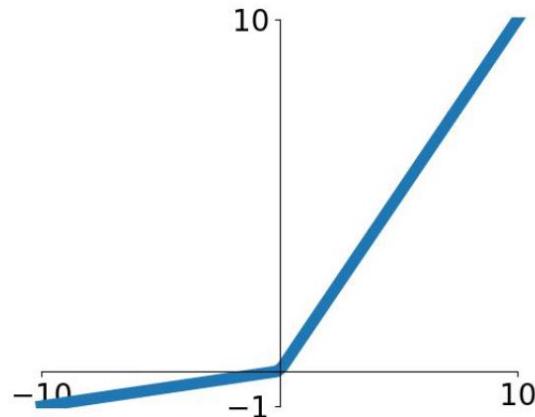




slide credit: Fei-Fei, Justin Johnson, Serena Yeung

Activation Functions

[Mass et al., 2013]
[He et al., 2015]



Leaky ReLU

$$f(x) = \max(0.01x, x)$$

- Does not saturate
- Computationally efficient
- Converges much faster than sigmoid/tanh in practice! (e.g. 6x)
- will not “die”.

Parametric Rectifier (PReLU)

$$f(x) = \max(\alpha x, x)$$

backprop into α
(parameter)

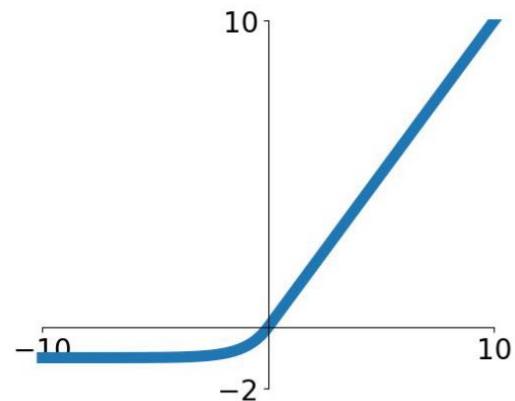
slide credit: Fei-Fei, Justin Johnson, Serena Yeung



Activation Functions

[Clevert et al., 2015]

Exponential Linear Units (ELU)



$$f(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha (\exp(x) - 1) & \text{if } x \leq 0 \end{cases}$$

- All benefits of ReLU
- Closer to zero mean outputs
- Negative saturation regime compared with Leaky ReLU adds some robustness to noise

- Computation requires $\exp()$

slide credit: Fei-Fei, Justin Johnson, Serena Yeung



Maxout “Neuron”

[Goodfellow et al., 2013]

- Does not have the basic form of dot product -> nonlinearity
- Generalizes ReLU and Leaky ReLU
- Linear Regime! Does not saturate! Does not die!

$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

generalized version of leaky relu

Problem: doubles the number of parameters/neuron :(

slide credit: Fei-Fei, Justin Johnson, Serena Yeung



TLDR: In practice:

- Use ReLU. Be careful with your learning rates
- Try out Leaky ReLU / Maxout / ELU
- Try out tanh but don't expect much
- Don't use sigmoid

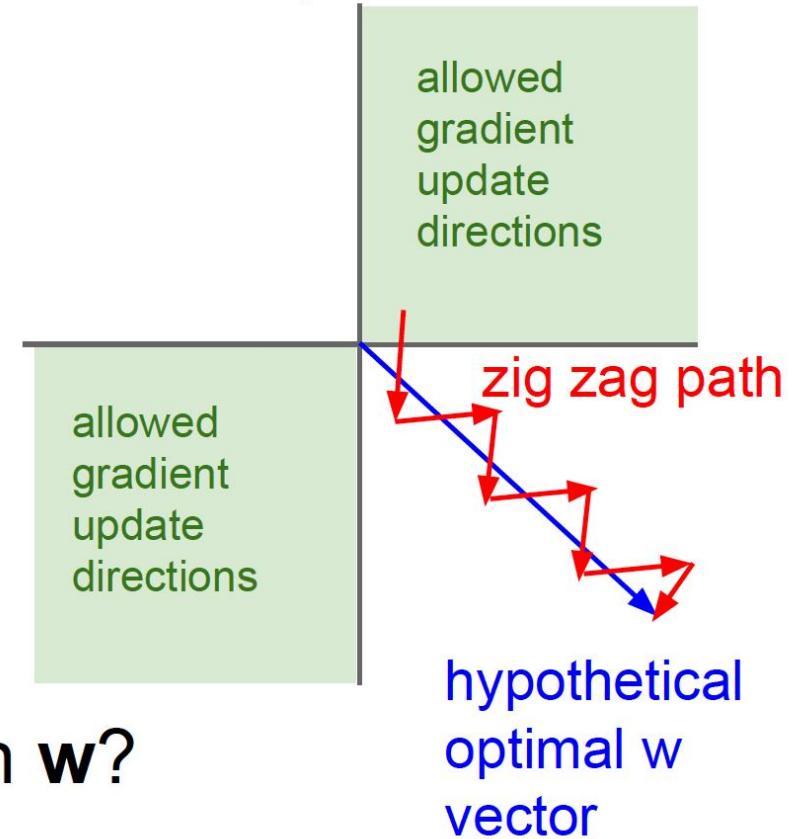
slide credit: Fei-Fei, Justin Johnson, Serena Yeung



Data Preprocessing

Remember: Consider what happens when the input to a neuron is always positive...

$$f \left(\sum_i w_i x_i + b \right)$$

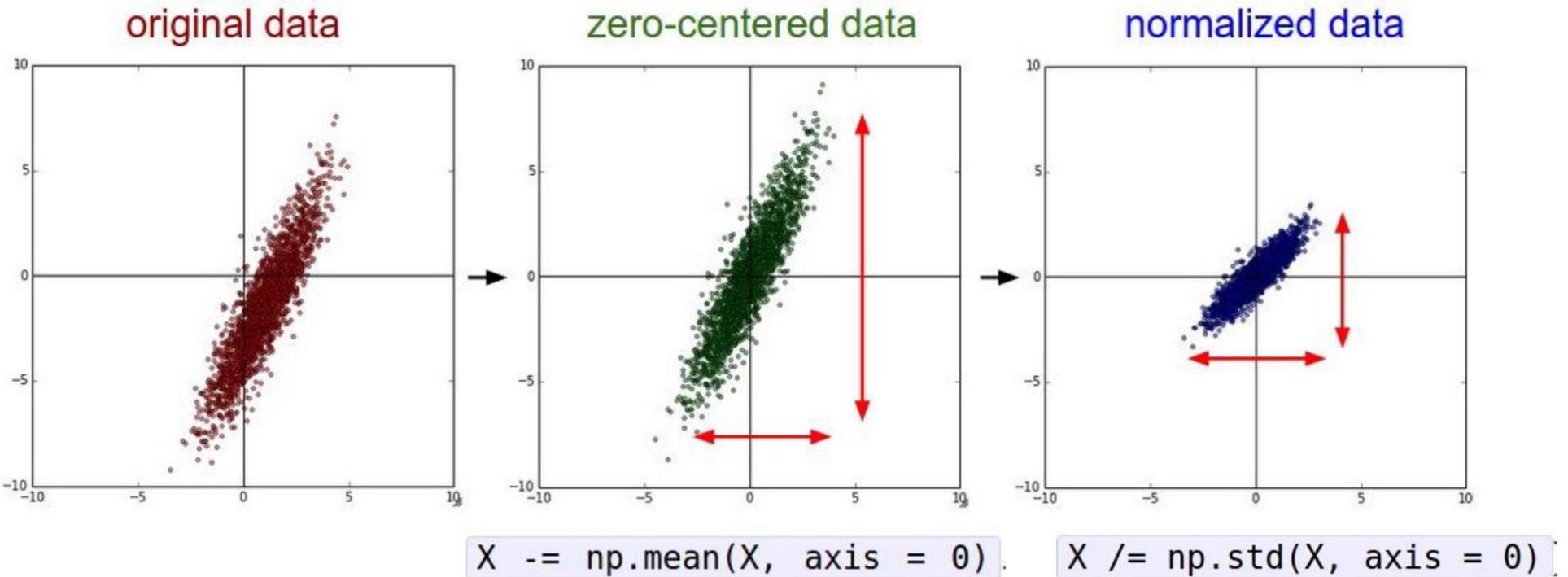


What can we say about the gradients on w ?
Always all positive or all negative :(
(this is also why you want zero-mean data!)

slide credit: Fei-Fei, Justin Johnson, Serena Yeung



Step 1: Preprocess the data



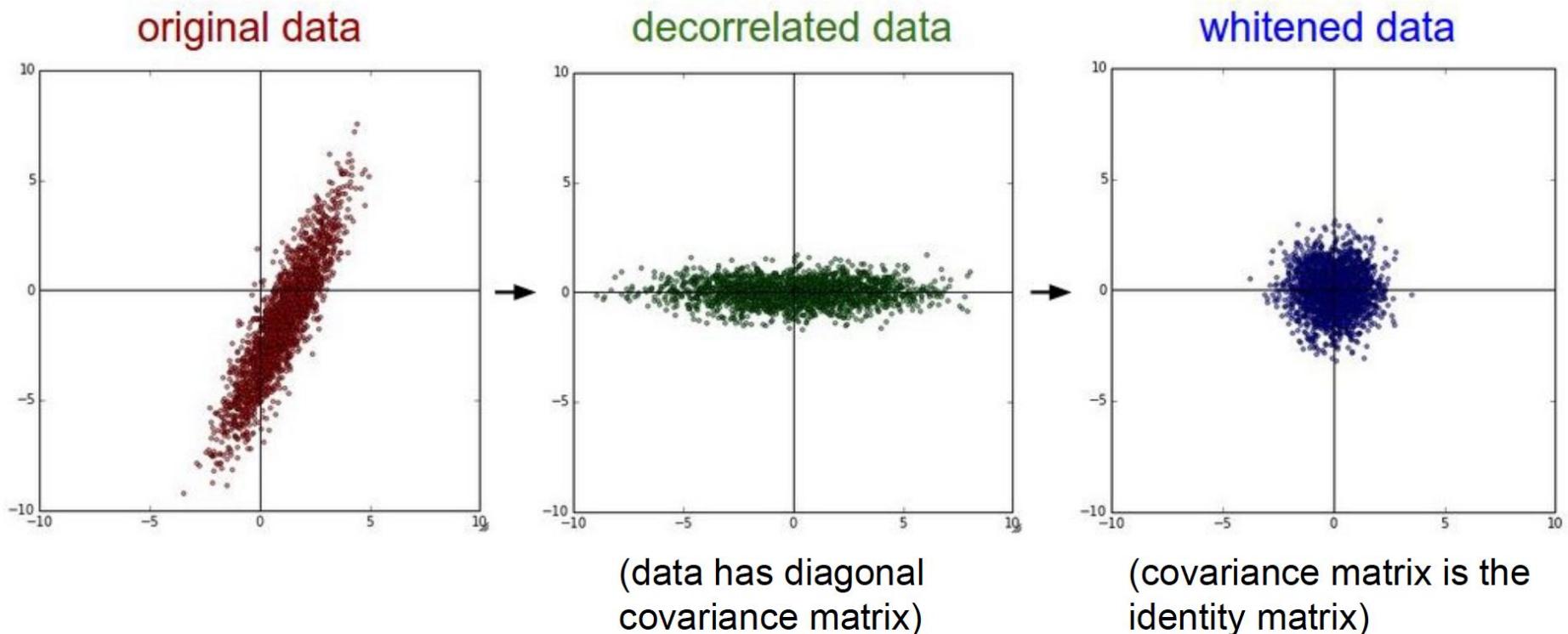
(Assume X [NxD] is data matrix,
each example in a row)

slide credit: Fei-Fei, Justin Johnson, Serena Yeung



Step 1: Preprocess the data

In practice, you may also see **PCA** and **Whitening** of the data



slide credit: Fei-Fei, Justin Johnson, Serena Yeung

TLDR: In practice for Images: center only

e.g. consider CIFAR-10 example with [32,32,3] images

- Subtract the mean image (e.g. AlexNet)
(mean image = [32,32,3] array)
- Subtract per-channel mean (e.g. VGGNet)
(mean along each channel = 3 numbers)

Not common to normalize variance, to do PCA or whitening

slide credit: Fei-Fei, Justin Johnson, Serena Yeung



Batch Normalization

Batch Normalization

[Ioffe and Szegedy, 2015]

“you want zero-mean unit-variance activations? just make them so.”

consider a batch of activations at some layer. To make each dimension zero-mean unit-variance, apply:

$$\hat{x}^{(k)} = \frac{x^{(k)} - \text{E}[x^{(k)}]}{\sqrt{\text{Var}[x^{(k)}]}}$$

this is a vanilla
differentiable function...

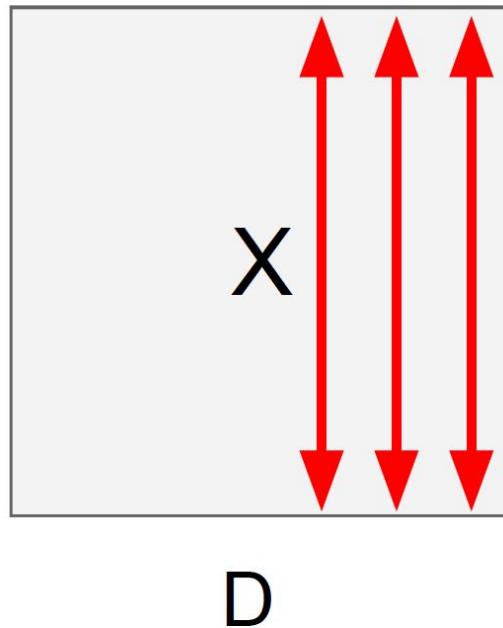
slide credit: Fei-Fei, Justin Johnson, Serena Yeung



Batch Normalization

[Ioffe and Szegedy, 2015]

“you want zero-mean unit-variance activations? just make them so.”



1. compute the empirical mean and variance independently for each dimension.

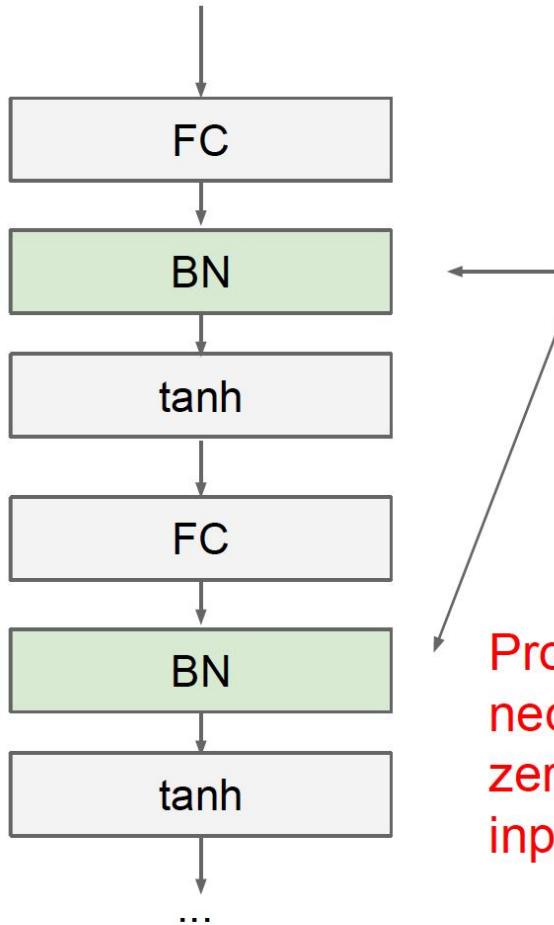
2. Normalize

$$\hat{x}^{(k)} = \frac{x^{(k)} - \text{E}[x^{(k)}]}{\sqrt{\text{Var}[x^{(k)}]}}$$

slide credit: Fei-Fei, Justin Johnson, Serena Yeung

Batch Normalization

[Ioffe and Szegedy, 2015]



Usually inserted after Fully Connected or Convolutional layers, and before nonlinearity.

Problem: do we necessarily want a zero-mean unit-variance input?

$$\hat{x}^{(k)} = \frac{x^{(k)} - \mathbb{E}[x^{(k)}]}{\sqrt{\text{Var}[x^{(k)}]}}$$

slide credit: Fei-Fei, Justin Johnson, Serena Yeung



Batch Normalization

[Ioffe and Szegedy, 2015]

for each channel, keep it and update for each batch

for test time: either we keep the running avg for test time, or we run the whole training set once after training to get these paras

Normalize:

$$\hat{x}^{(k)} = \frac{x^{(k)} - \text{E}[x^{(k)}]}{\sqrt{\text{Var}[x^{(k)}]}}$$

And then allow the network to squash the range if it wants to:

$$y^{(k)} = \gamma^{(k)} \hat{x}^{(k)} + \beta^{(k)}$$

Note, the network can learn:

$$\gamma^{(k)} = \sqrt{\text{Var}[x^{(k)}]}$$

$$\beta^{(k)} = \text{E}[x^{(k)}]$$

to recover the identity mapping.



Batch Normalization

[Ioffe and Szegedy, 2015]

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1\dots m}\}$;
Parameters to be learned: γ, β

Output: $\{y_i = \text{BN}_{\gamma, \beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m} \sum_{i=1}^m x_i \quad // \text{mini-batch mean}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m} \sum_{i=1}^m (x_i - \mu_{\mathcal{B}})^2 \quad // \text{mini-batch variance}$$

$$\hat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \quad // \text{normalize}$$

$$y_i \leftarrow \gamma \hat{x}_i + \beta \equiv \text{BN}_{\gamma, \beta}(x_i) \quad // \text{scale and shift}$$

- Improves gradient flow through the network
- Allows higher learning rates
- Reduces the strong dependence on initialization
- Acts as a form of regularization in a funny way, and slightly reduces the need for dropout, maybe



Batch Normalization

[Ioffe and Szegedy, 2015]

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1\dots m}\}$;

Parameters to be learned: γ, β

Output: $\{y_i = \text{BN}_{\gamma, \beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m} \sum_{i=1}^m x_i \quad // \text{mini-batch mean}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m} \sum_{i=1}^m (x_i - \mu_{\mathcal{B}})^2 \quad // \text{mini-batch variance}$$

$$\hat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \quad // \text{normalize}$$

$$y_i \leftarrow \gamma \hat{x}_i + \beta \equiv \text{BN}_{\gamma, \beta}(x_i) \quad // \text{scale and shift}$$

Note: at test time BatchNorm layer functions differently:

The mean/std are not computed based on the batch. Instead, a single fixed empirical mean of activations during training is used.

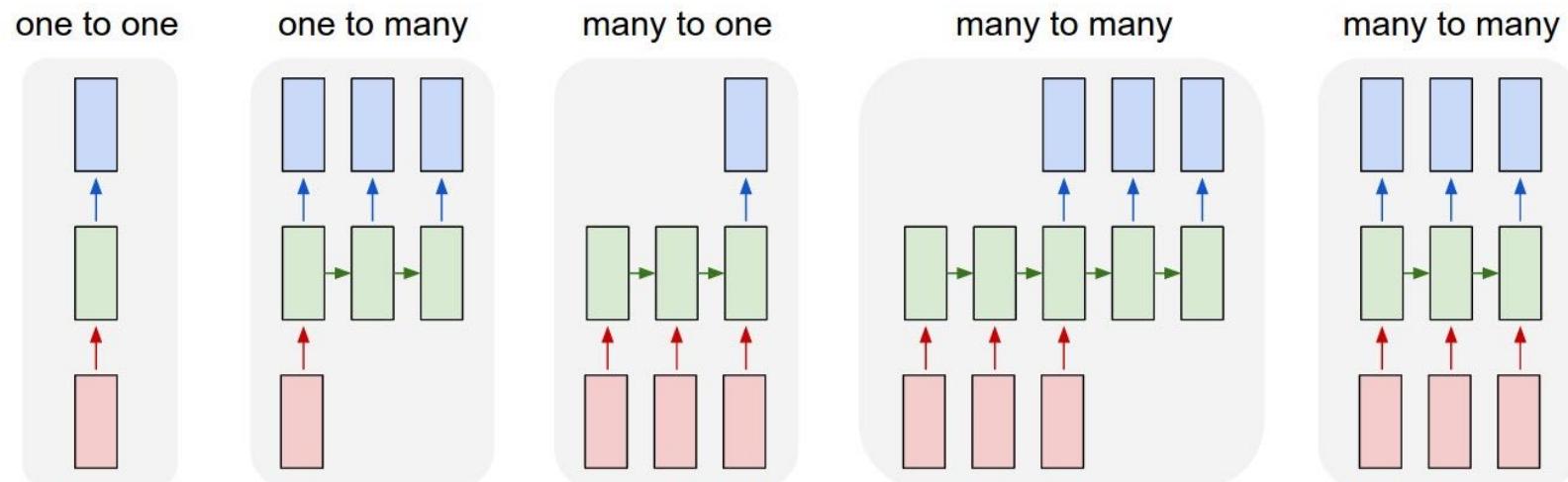
(e.g. can be estimated during training with running averages)



Overview Today's Lecture

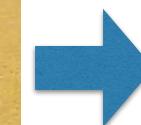
- Data Preprocessing
 - ▶ Activation functions
 - ▶ Batch normalization
- Recurrent Neural Networks (RNNs)
 - ▶ Motivation & flexibility of RNNs
 - ▶ Language modeling
 - including “unreasonable effectiveness of RNNs”
 - ▶ RNNs for image description / captioning
 - ▶ Standard RNN and a particularly successful RNN:
Long Short Term Memory (LSTM)
 - including “visualizations of RNN cells”

Recurrent Networks offer a lot of flexibility:

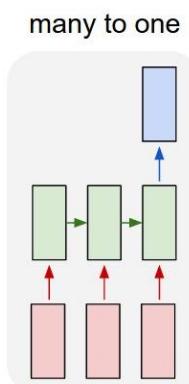


Sequences in Vision

Sequences in the input...
(many-to-one)

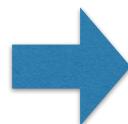


- Running
- Jumping
- Dancing
- Fighting
- Eating

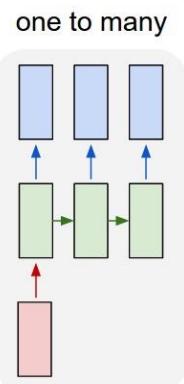


Sequences in Vision

Sequences in the output...
(one-to-many)

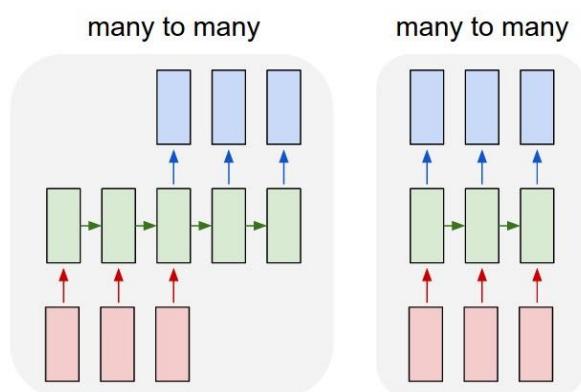


A happy brown dog.



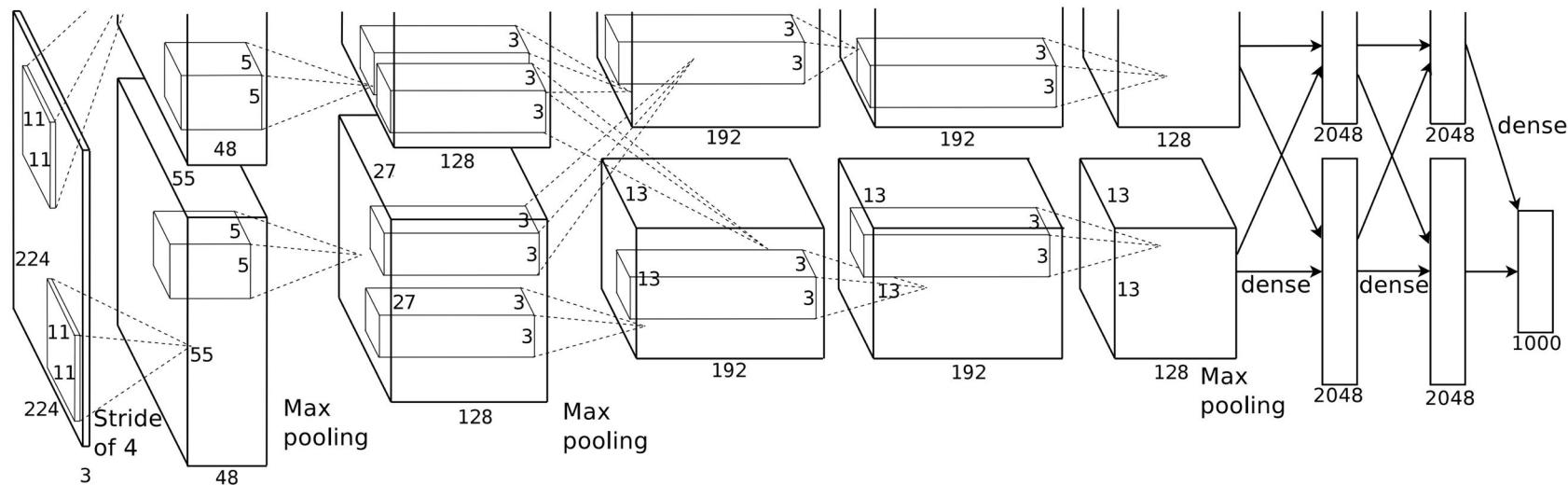
Sequences in Vision

Sequences everywhere!
(many-to-many)



A dog jumps over a hurdle.

ConvNets



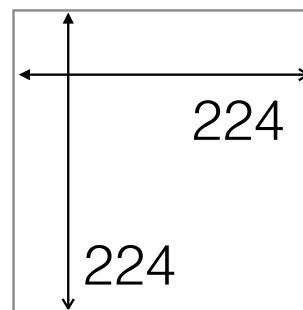
Krizhevsky et al., NIPS 2012

Problem #1



3

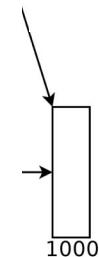
fixed-size, static input



Problem #2

output is a single choice from a fixed list of options

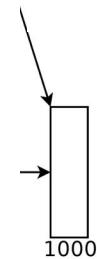
- cat
- dog
- horse
- fish
- snake



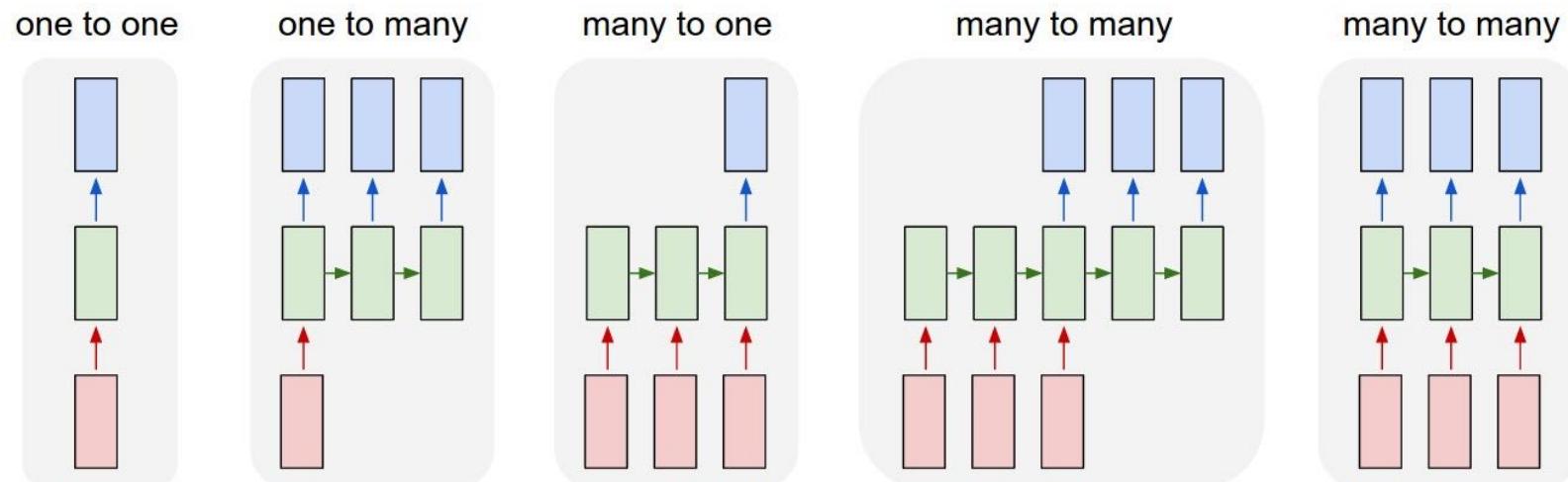
Problem #2

output is a single choice from a fixed list of options

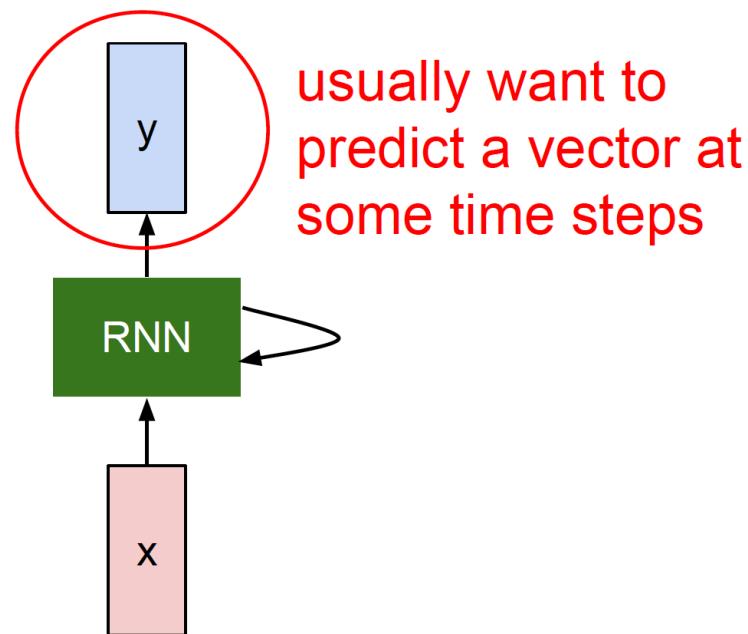
- a happy brown dog
- a big brown dog
- a happy red dog
- a big red dog
- ...



Recurrent Networks offer a lot of flexibility:



Recurrent Neural Network (RNN)



slide credit: Fei-Fei, Justin Johnson, Serena Yeung

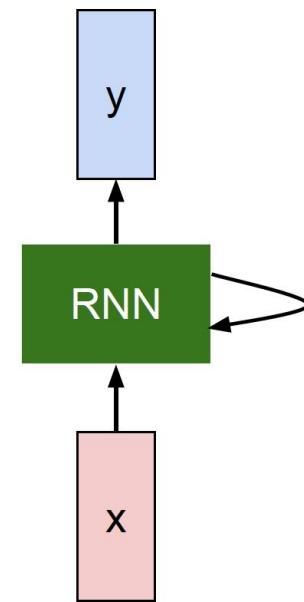


Recurrent Neural Network (RNN)

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

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

new state / old state input vector at
 \ some function some time step
 some function
 with parameters W



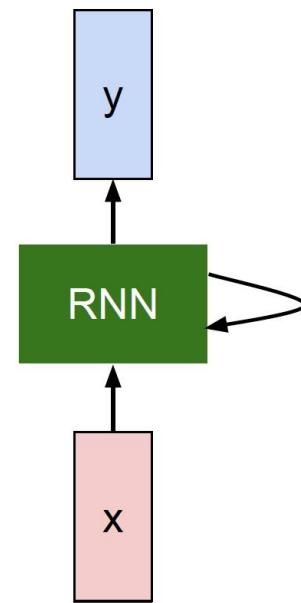
slide credit: Fei-Fei, Justin Johnson, Serena Yeung

Recurrent Neural Network (RNN)

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

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

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

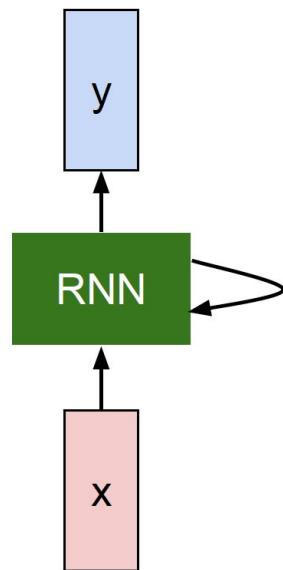


slide credit: Fei-Fei, Justin Johnson, Serena Yeung



(Simple) Recurrent Neural Network

The state consists of a single “hidden” vector \mathbf{h} :



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

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

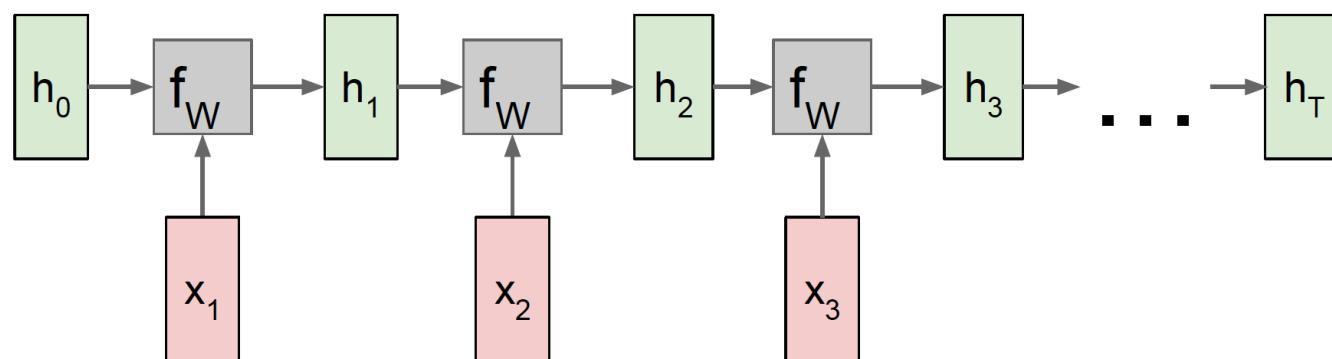
$$y_t = W_{hy}h_t$$

Sometimes called a “Vanilla RNN” or an “Elman RNN” after Prof. Jeffrey Elman

slide credit: Fei-Fei, Justin Johnson, Serena Yeung

RNN: Computational Graph

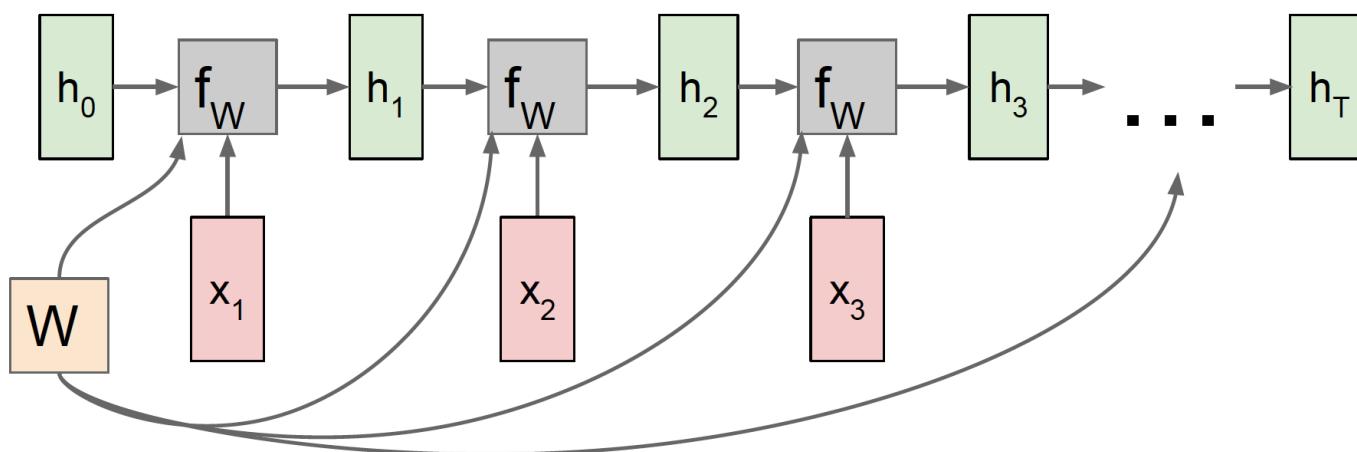
$$h_t = f_W(h_{t-1}, x_t)$$



slide credit: Fei-Fei, Justin Johnson, Serena Yeung

RNN: Computational Graph

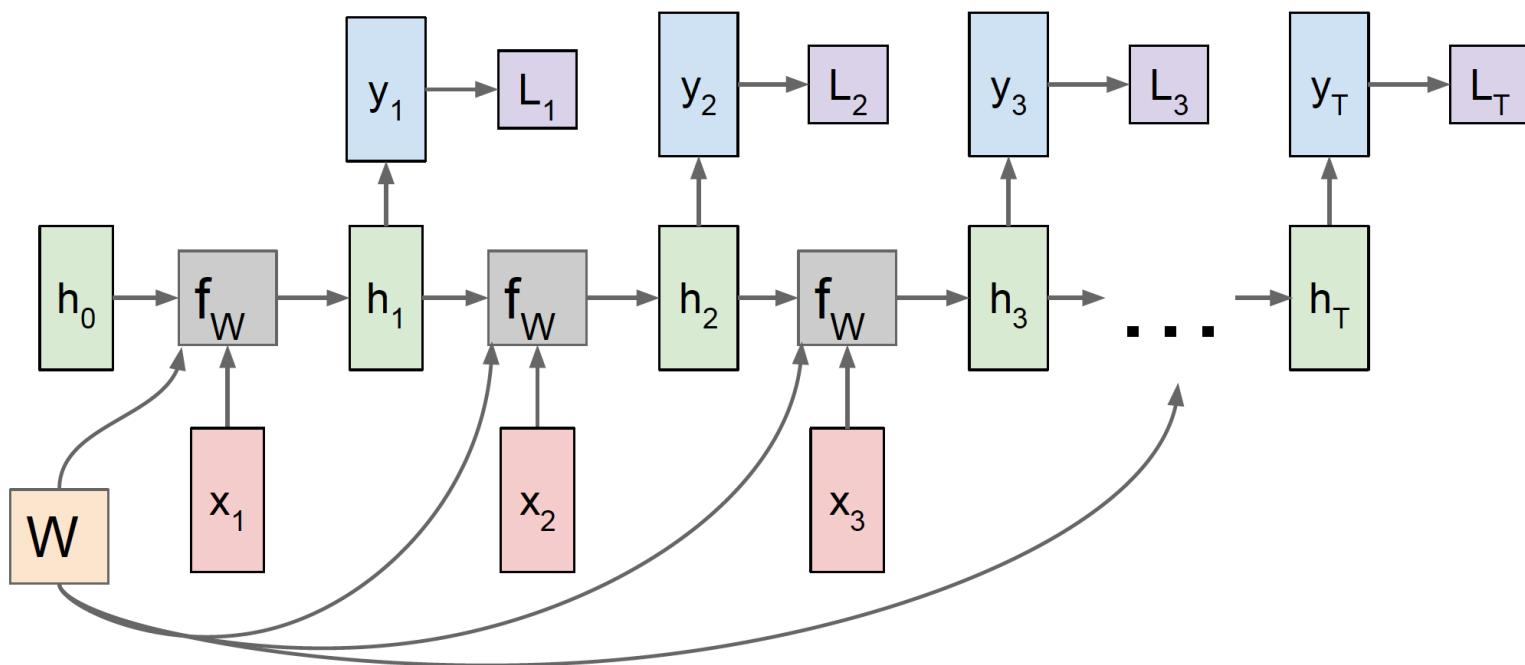
Re-use the same weight matrix at every time-step



slide credit: Fei-Fei, Justin Johnson, Serena Yeung



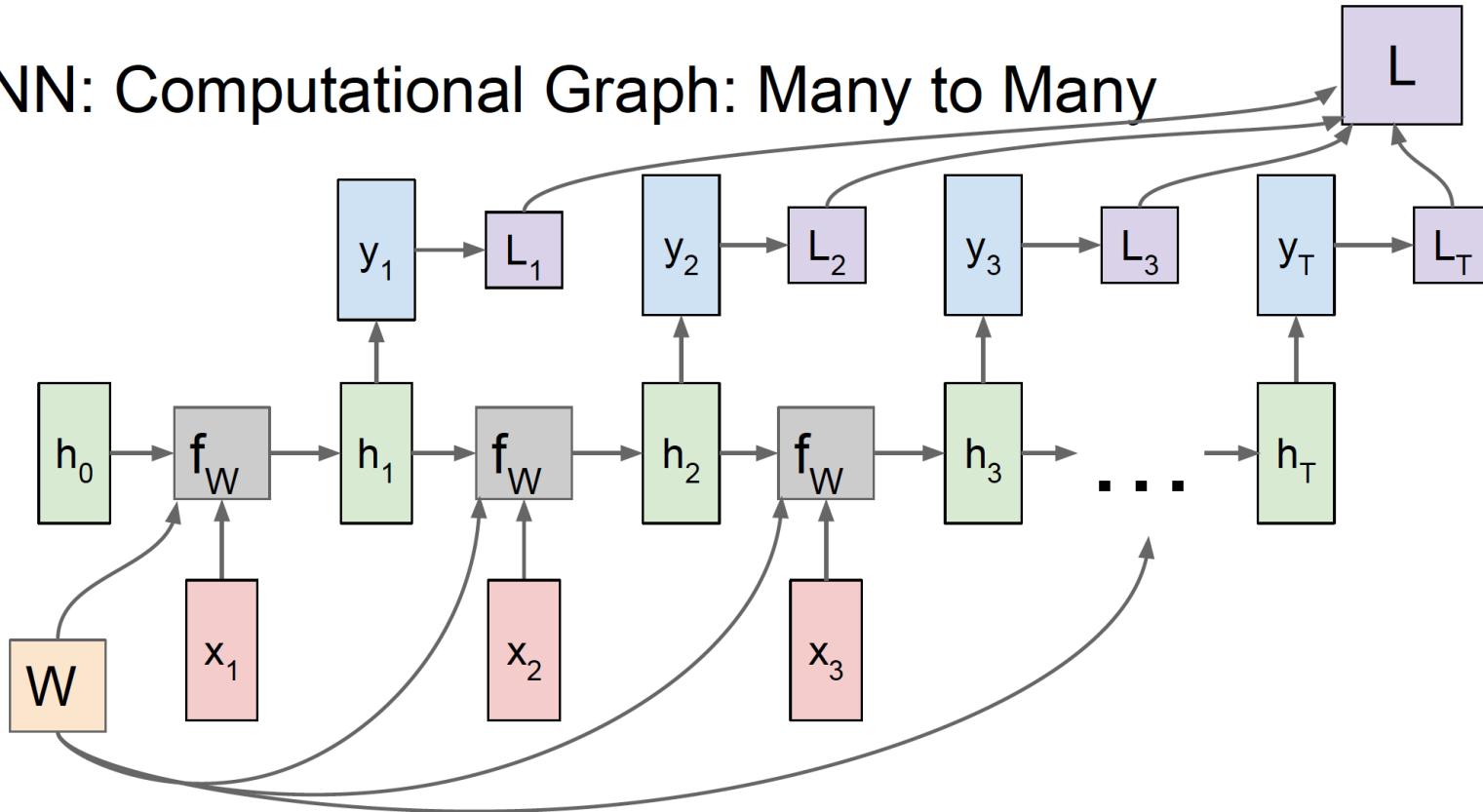
RNN: Computational Graph: Many to Many



slide credit: Fei-Fei, Justin Johnson, Serena Yeung

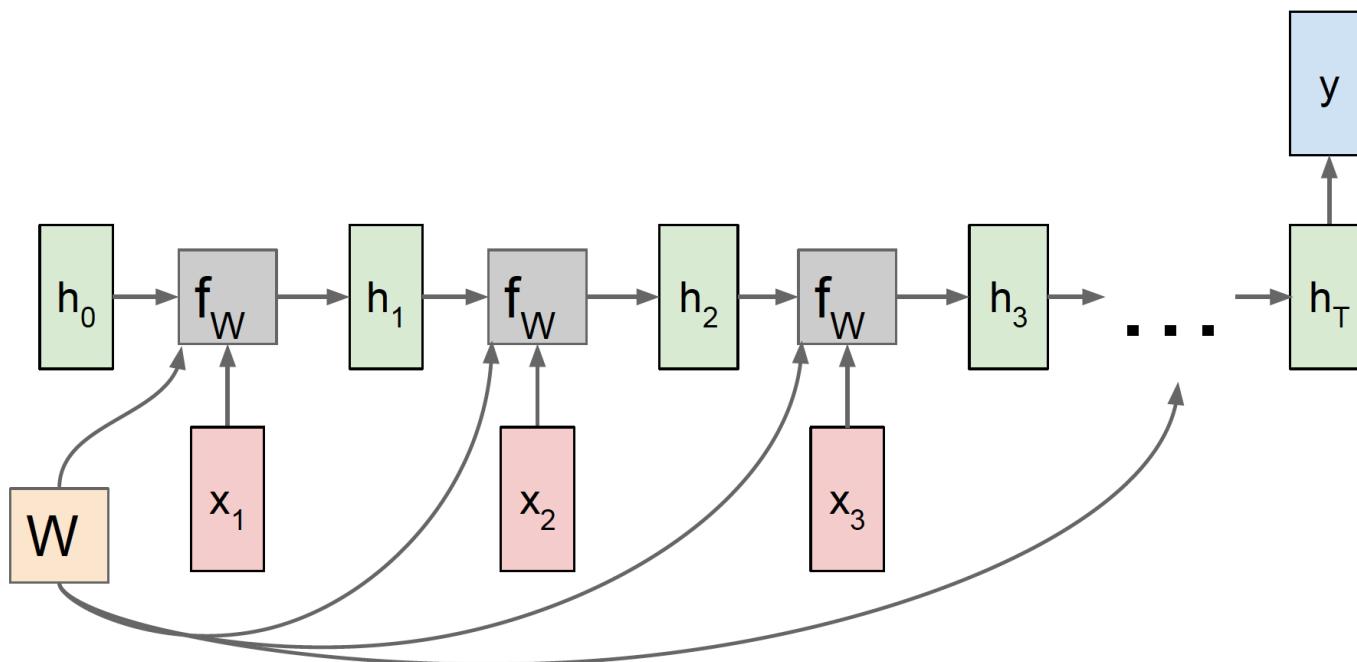


RNN: Computational Graph: Many to Many



slide credit: Fei-Fei, Justin Johnson, Serena Yeung

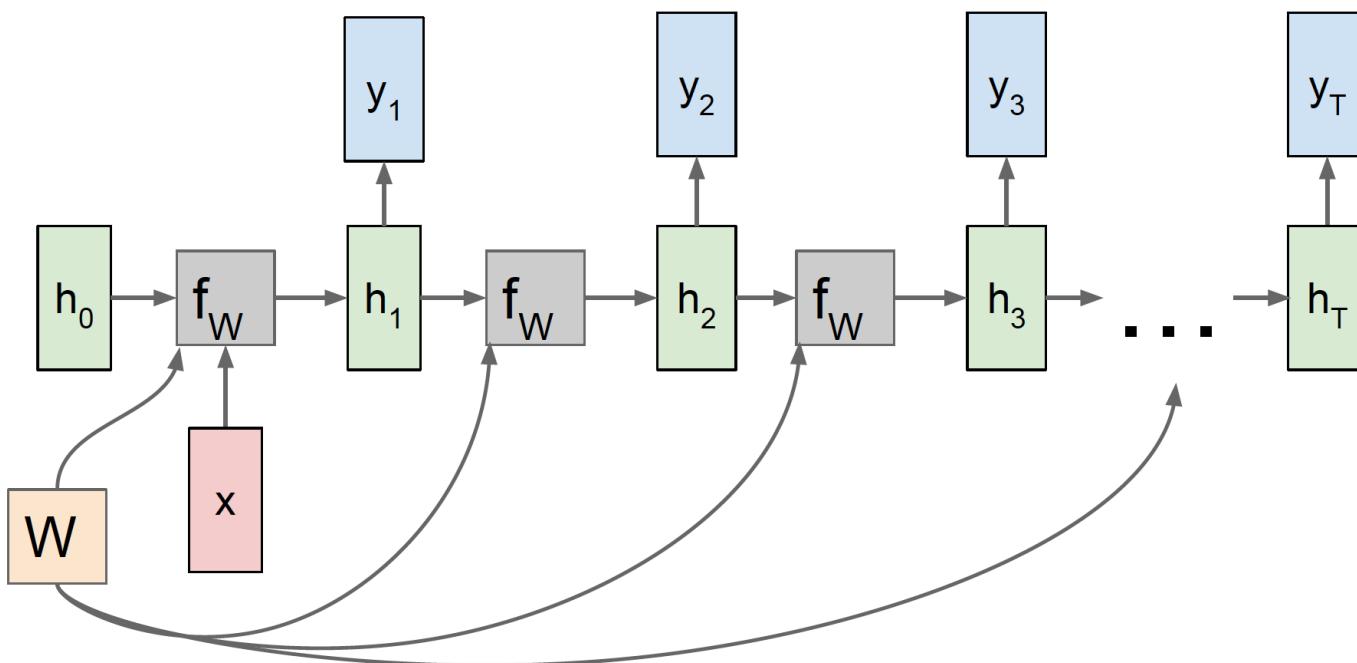
RNN: Computational Graph: Many to One



slide credit: Fei-Fei, Justin Johnson, Serena Yeung



RNN: Computational Graph: One to Many

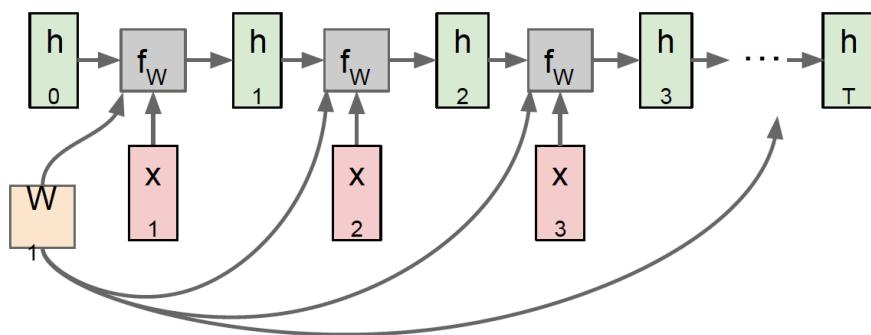


slide credit: Fei-Fei, Justin Johnson, Serena Yeung

Sequence to Sequence

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

Many to one: Encode input sequence in a single vector

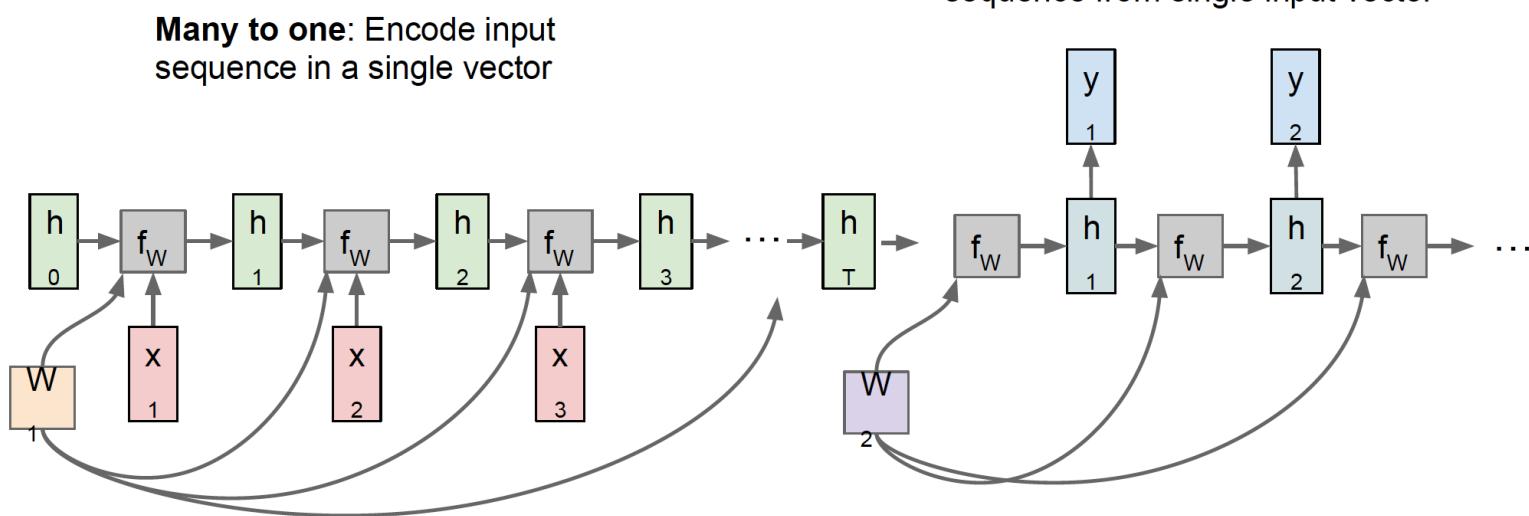


Sutskever et al, "Sequence to Sequence Learning with Neural Networks", NIPS 2014

slide credit: Fei-Fei, Justin Johnson, Serena Yeung

Sequence to Sequence

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

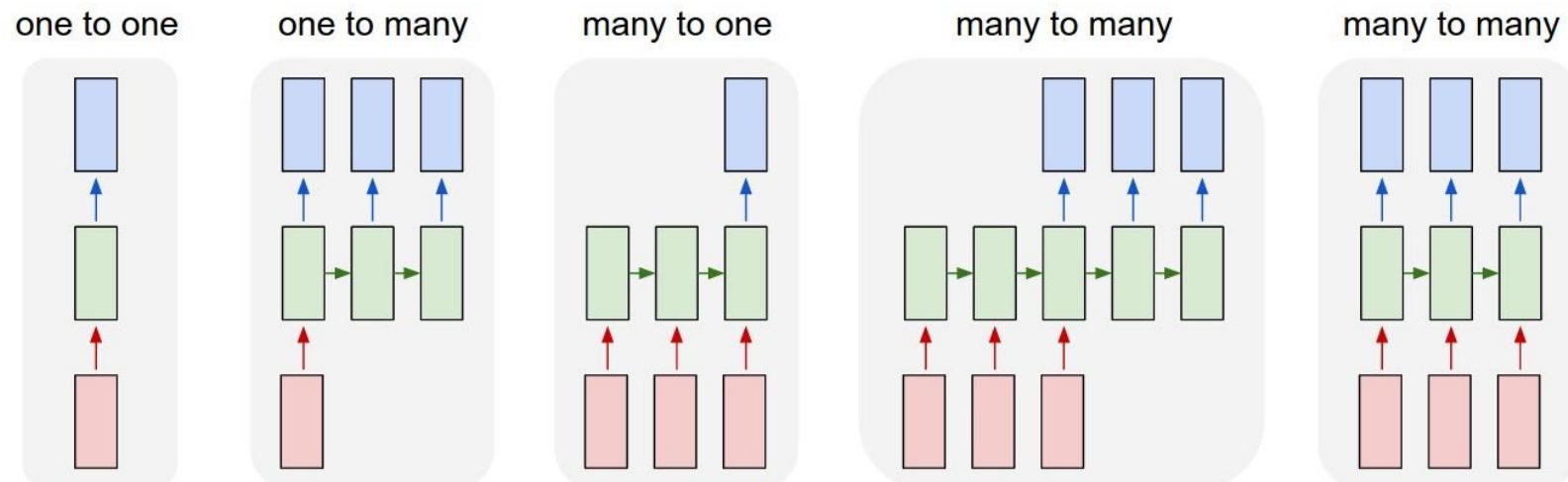


Sutskever et al, "Sequence to Sequence Learning with Neural Networks", NIPS 2014

slide credit: Fei-Fei, Justin Johnson, Serena Yeung

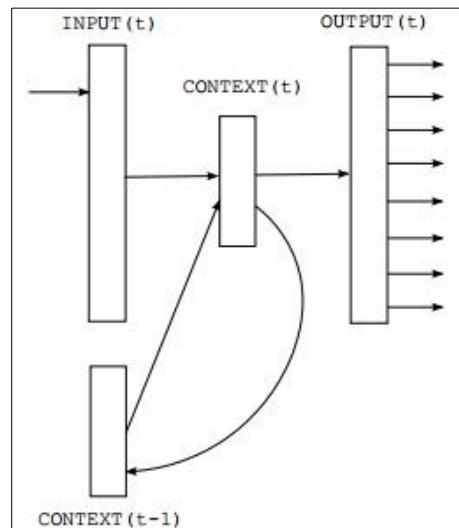


Recurrent Networks offer a lot of flexibility:

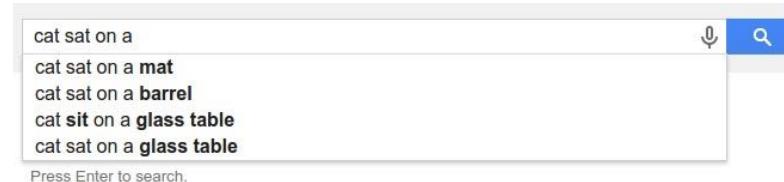


slide credit: Andrej Karpathy

Language Models



Word-level language model. Similar to:



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

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

We want to train a **language model**:

$P(\text{next word} \mid \text{previous words})$

i.e. want these to be high:

$P(\text{cat} \mid [\text{<S>}])$

$P(\text{sat} \mid [\text{<S>}, \text{cat}])$

$P(\text{on} \mid [\text{<S>}, \text{cat}, \text{sat}])$

$P(\text{mat} \mid [\text{<S>}, \text{cat}, \text{sat}, \text{on}])$

$P(\text{<E>} \mid [\text{<S>}, \text{cat}, \text{sat}, \text{on}, \text{mat}])$

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

We want to train a **language model**:

$P(\text{next word} \mid \text{previous words})$

First, suppose we had only a finite, 1-word history:

i.e. want these to be high:

$P(\text{cat} \mid \langle S \rangle)$

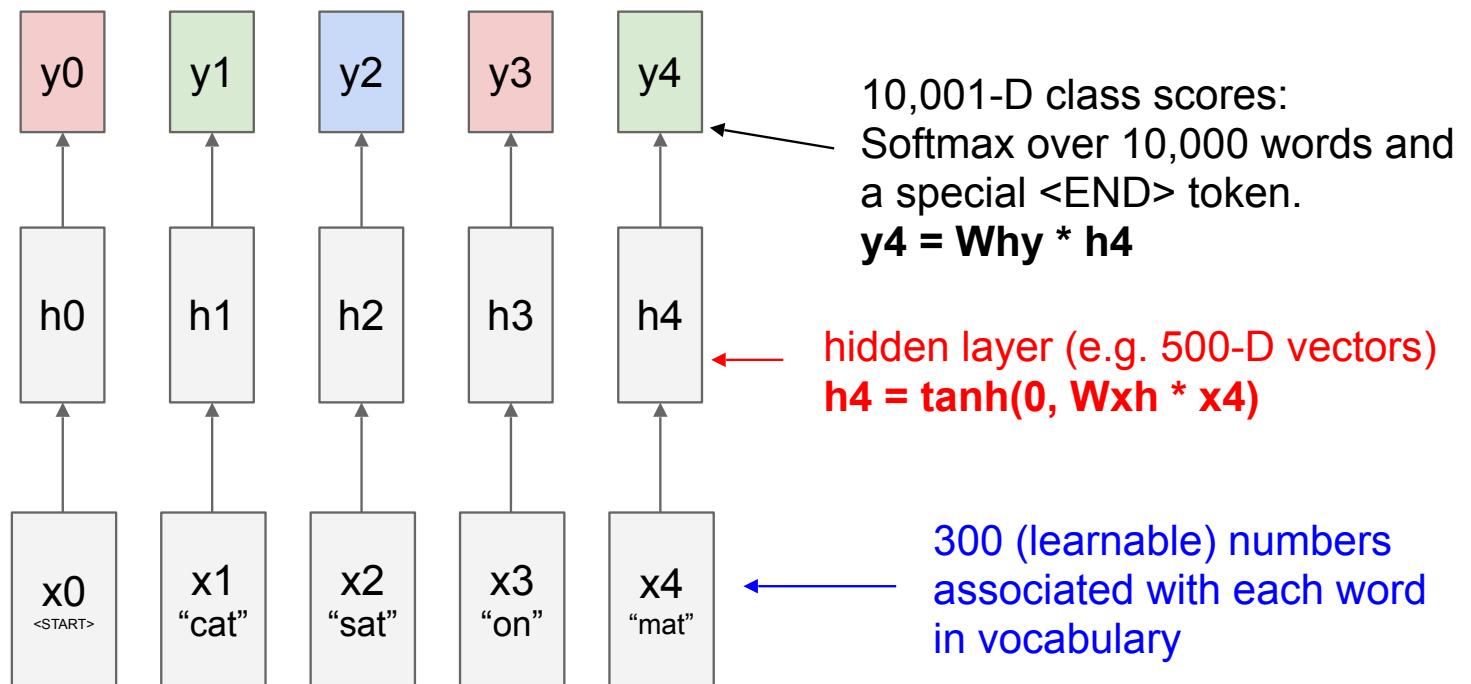
$P(\text{sat} \mid \text{cat})$

$P(\text{on} \mid \text{sat})$

$P(\text{mat} \mid \text{on})$

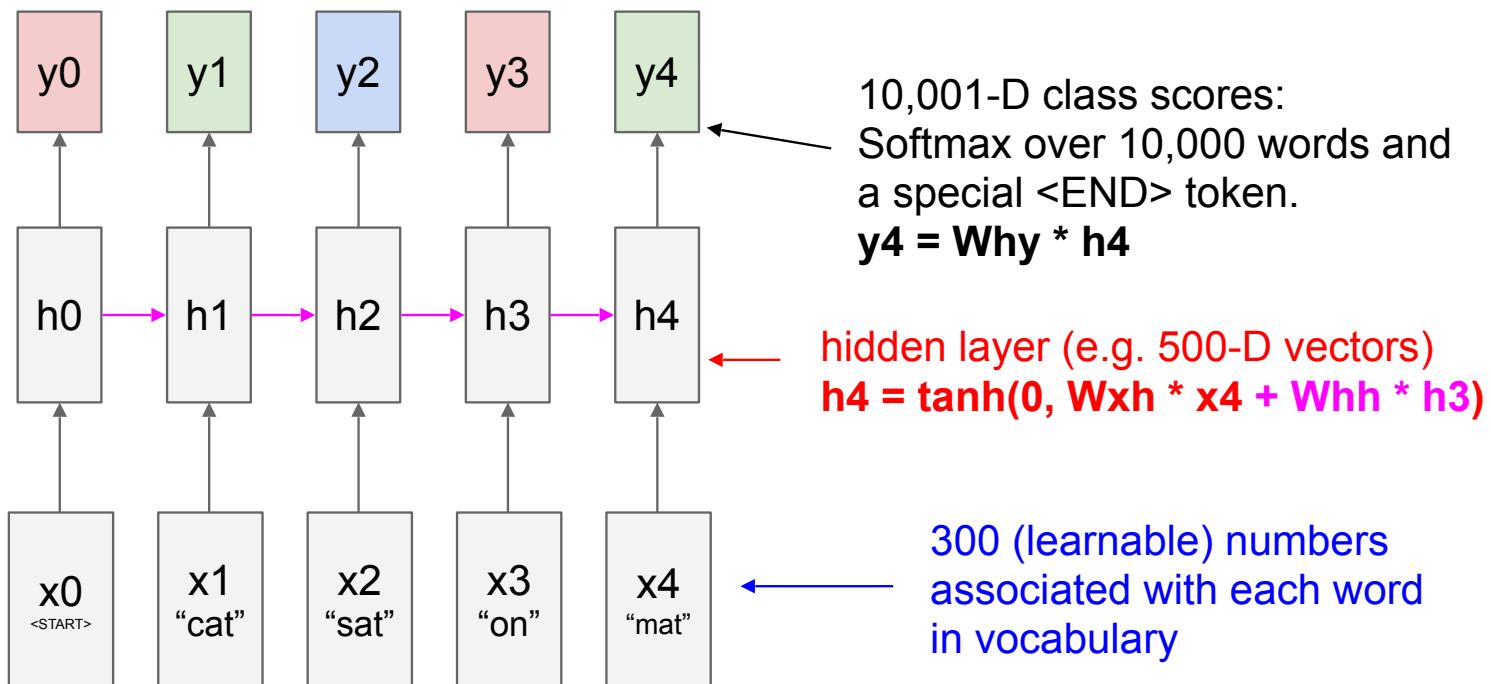
$P(\langle E \rangle \mid \text{mat})$

“cat sat on mat”



Recurrent Neural Network:

“cat sat on mat”



slide credit: Andrej Karpathy

Generating Sentences...

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

$P(\text{next word} \mid \text{previous words})$

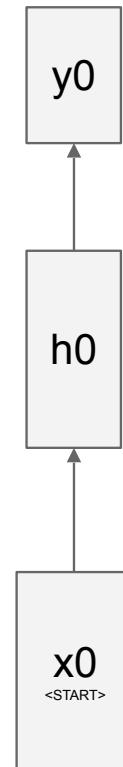


slide credit: Andrej Karpathy

Generating Sentences...

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

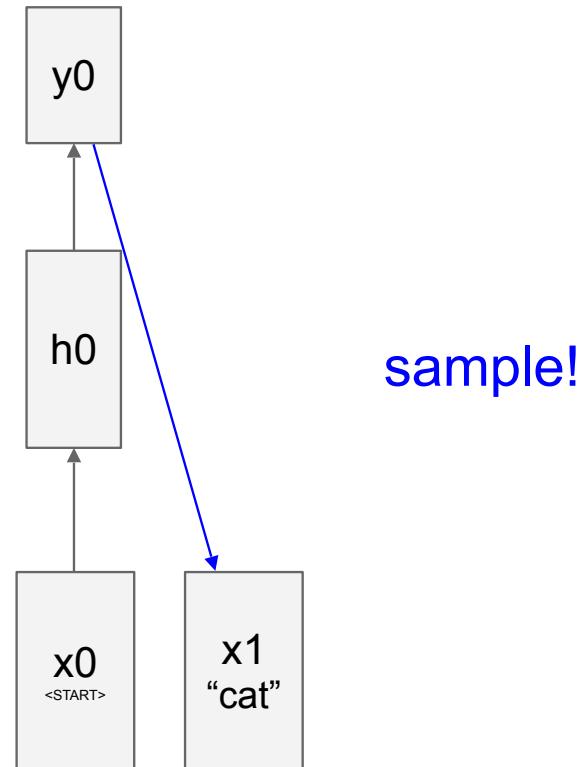
$P(\text{next word} \mid \text{previous words})$



Generating Sentences...

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

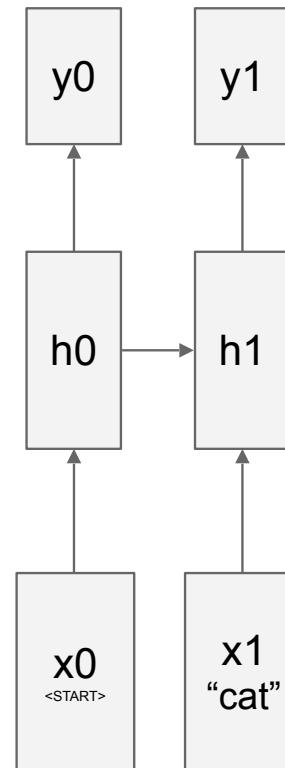
$P(\text{next word} \mid \text{previous words})$



Generating Sentences...

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

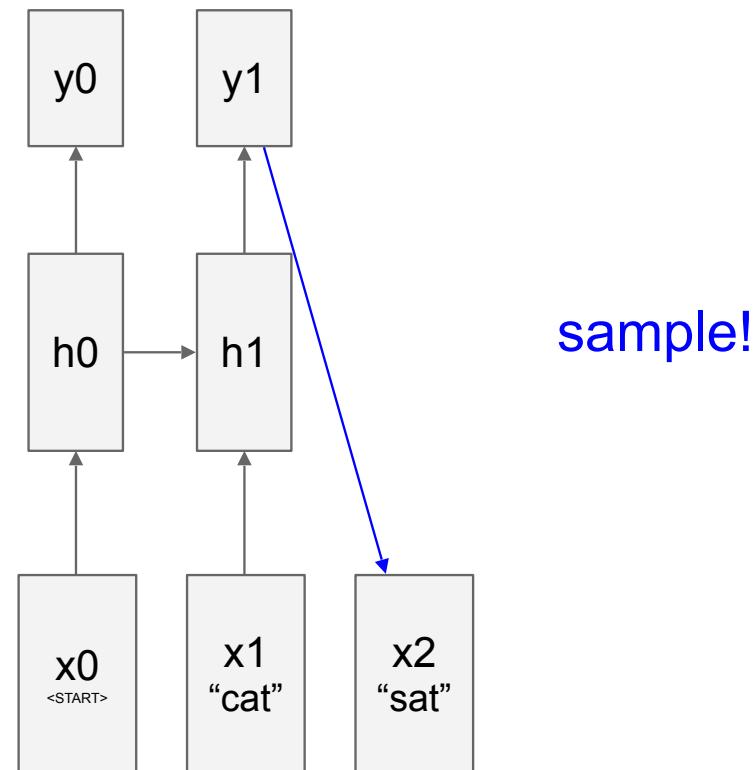
$P(\text{next word} \mid \text{previous words})$



Generating Sentences...

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

$P(\text{next word} \mid \text{previous words})$

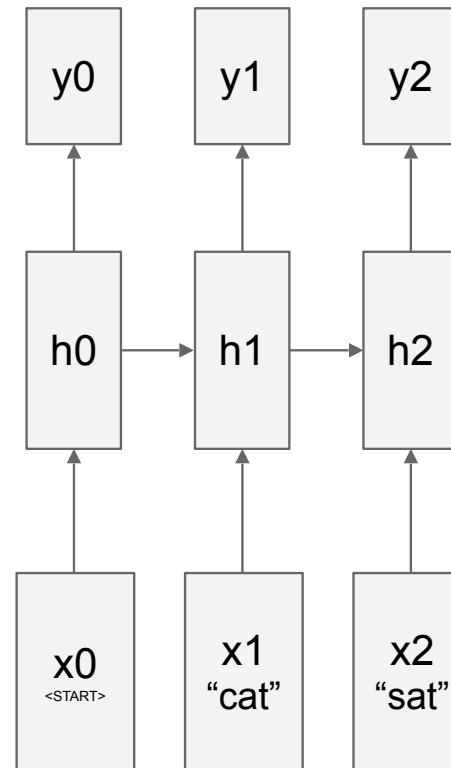


sample!

Generating Sentences...

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

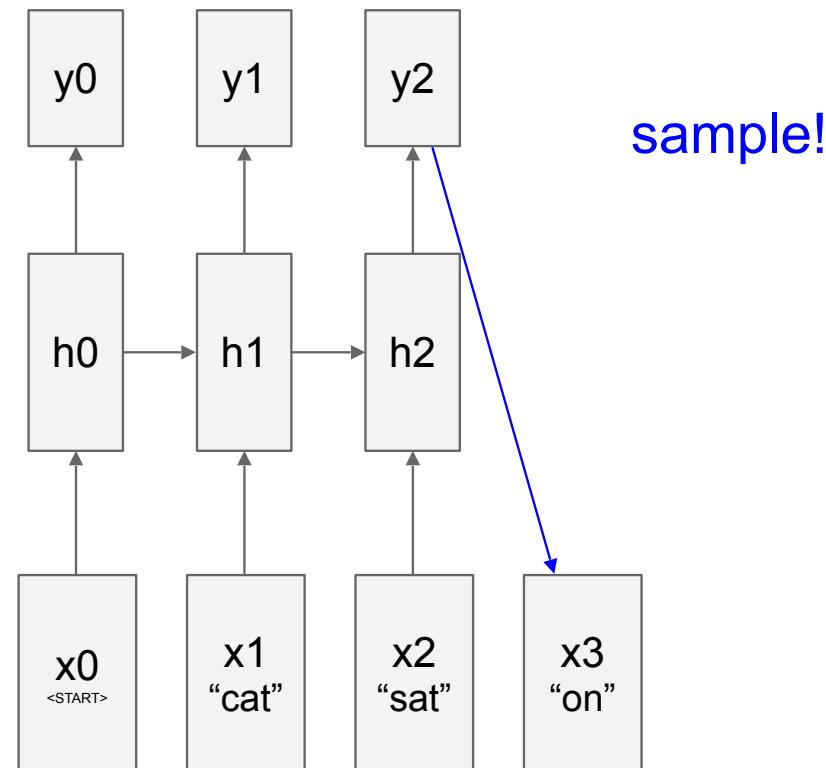
$P(\text{next word} \mid \text{previous words})$



Generating Sentences...

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

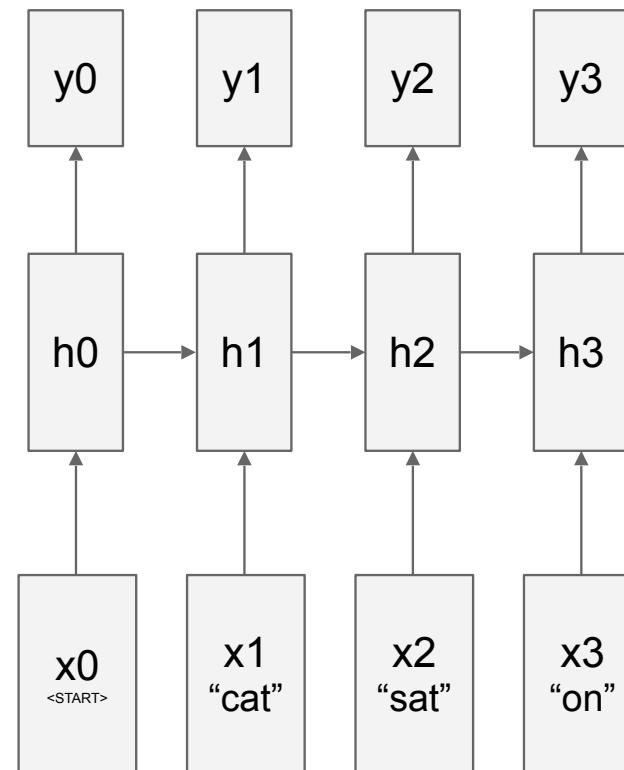
$P(\text{next word} \mid \text{previous words})$



Generating Sentences...

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

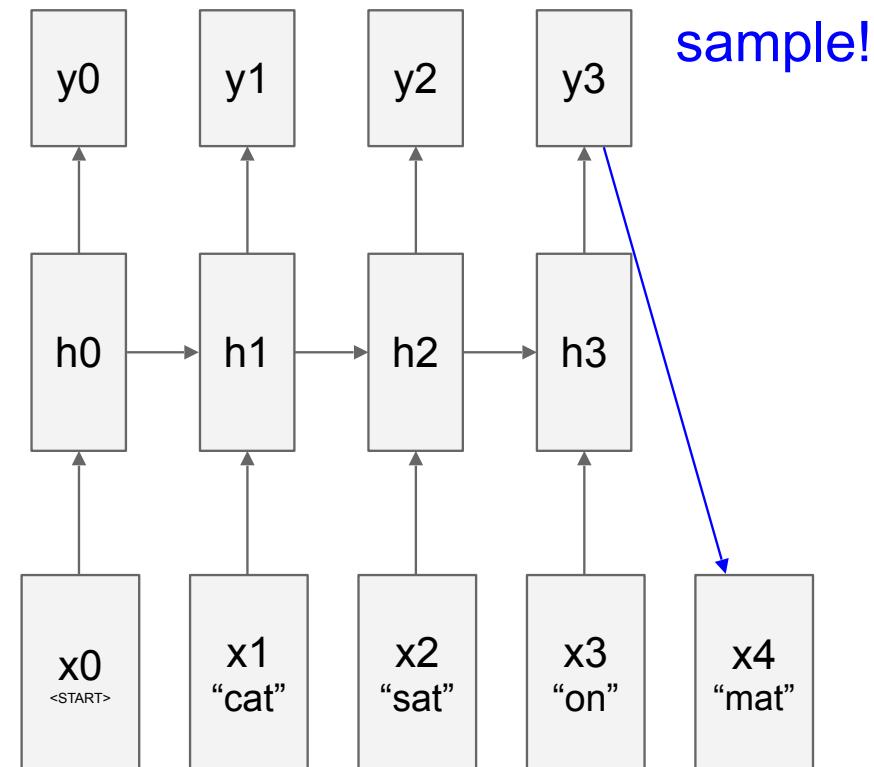
$P(\text{next word} \mid \text{previous words})$



Generating Sentences...

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

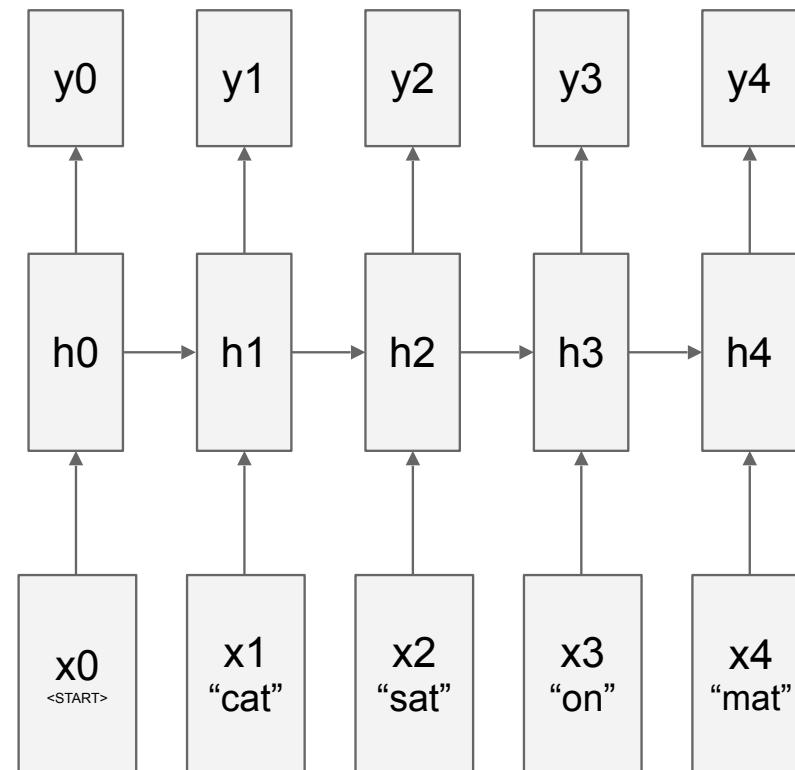
$P(\text{next word} \mid \text{previous words})$



Generating Sentences...

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

$P(\text{next word} \mid \text{previous words})$

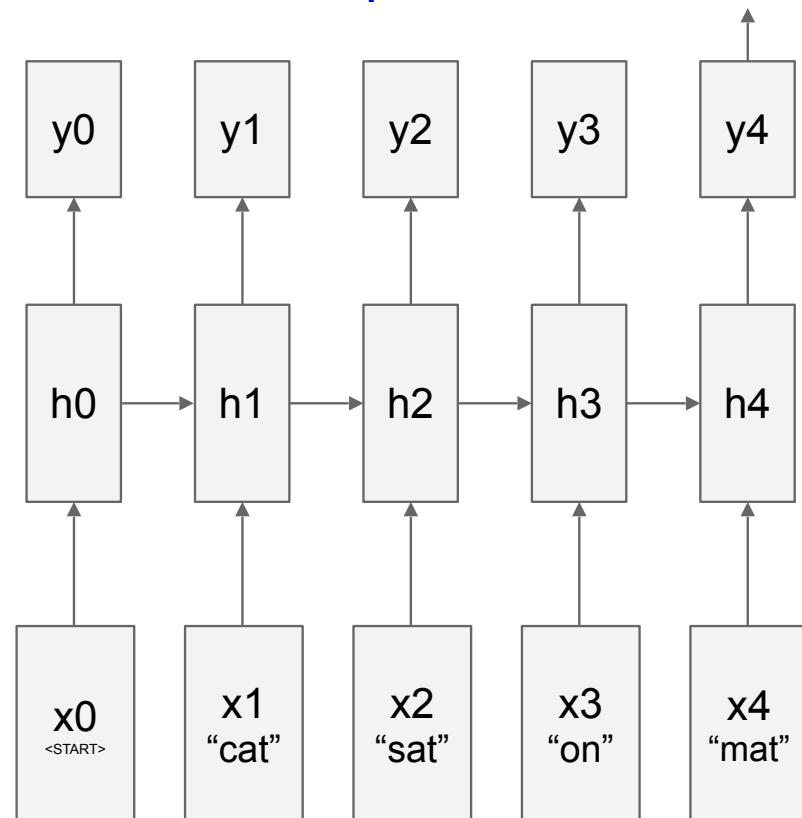


Generating Sentences...

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

$P(\text{next word} \mid \text{previous words})$

samples `<END>`? done.

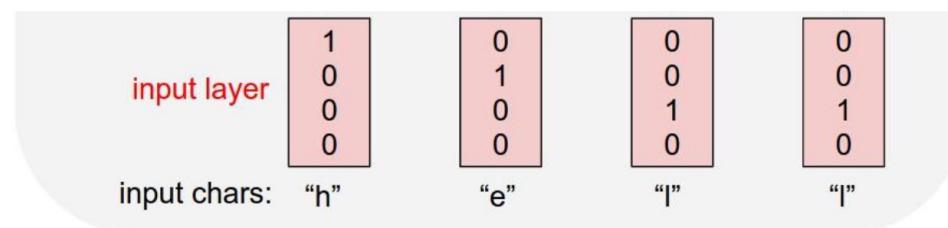


Example...

Example: Character-level Language Model

Vocabulary:
[h,e,l,o]

Example training
sequence:
“hello”



slide credit: Fei-Fei, Justin Johnson, Serena Yeung



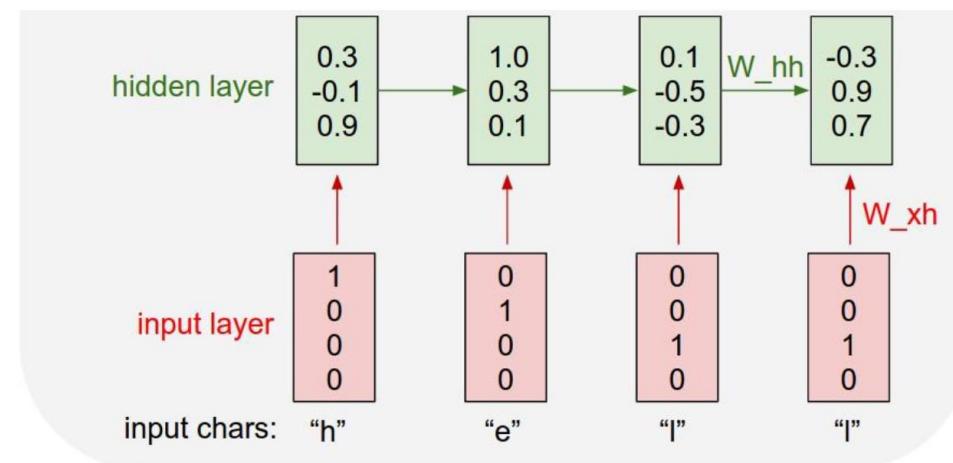
Example...

Example: Character-level Language Model

Vocabulary:
[h,e,l,o]

Example training
sequence:
“hello”

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



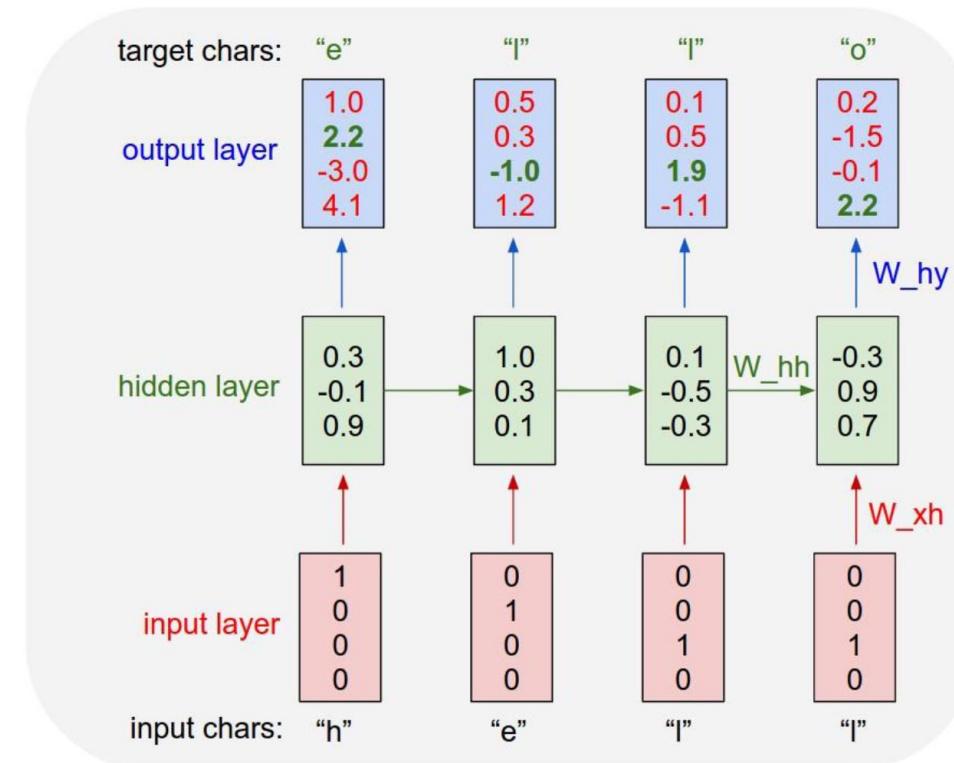
slide credit: Fei-Fei, Justin Johnson, Serena Yeung

Example...

Example: Character-level Language Model

Vocabulary:
[h,e,l,o]

Example training
sequence:
“hello”



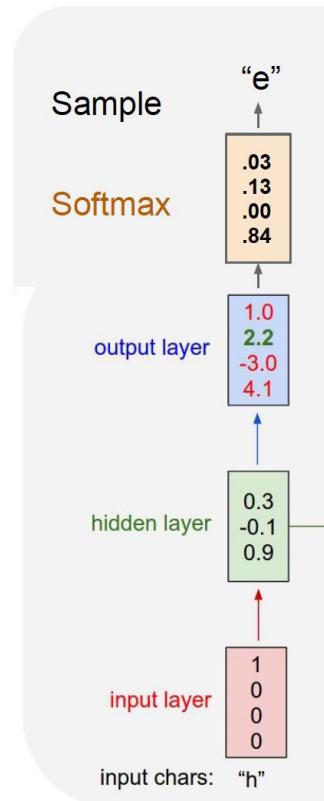
slide credit: Fei-Fei, Justin Johnson, Serena Yeung

Example...

Example: Character-level Language Model Sampling

Vocabulary:
[h,e,l,o]

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



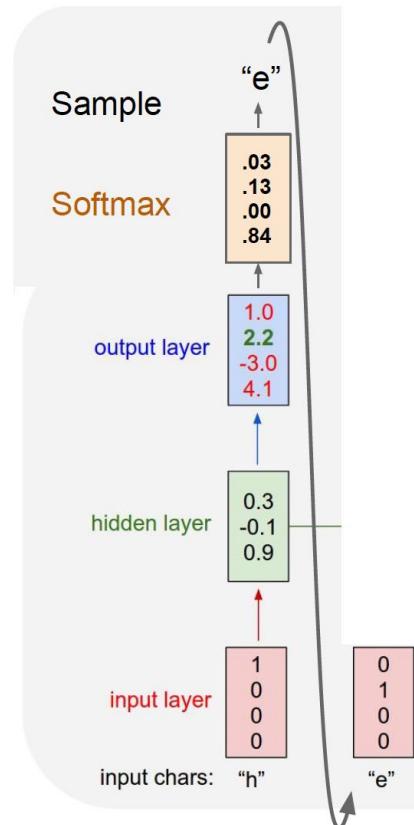
slide credit: Fei-Fei, Justin Johnson, Serena Yeung

Example...

Example: Character-level Language Model Sampling

Vocabulary:
[h,e,l,o]

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



slide credit: Fei-Fei, Justin Johnson, Serena Yeung

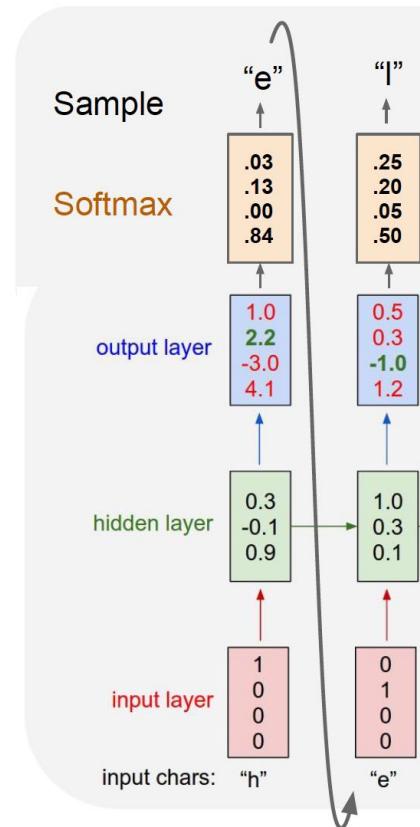


Example...

Example: Character-level Language Model Sampling

Vocabulary:
[h,e,l,o]

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



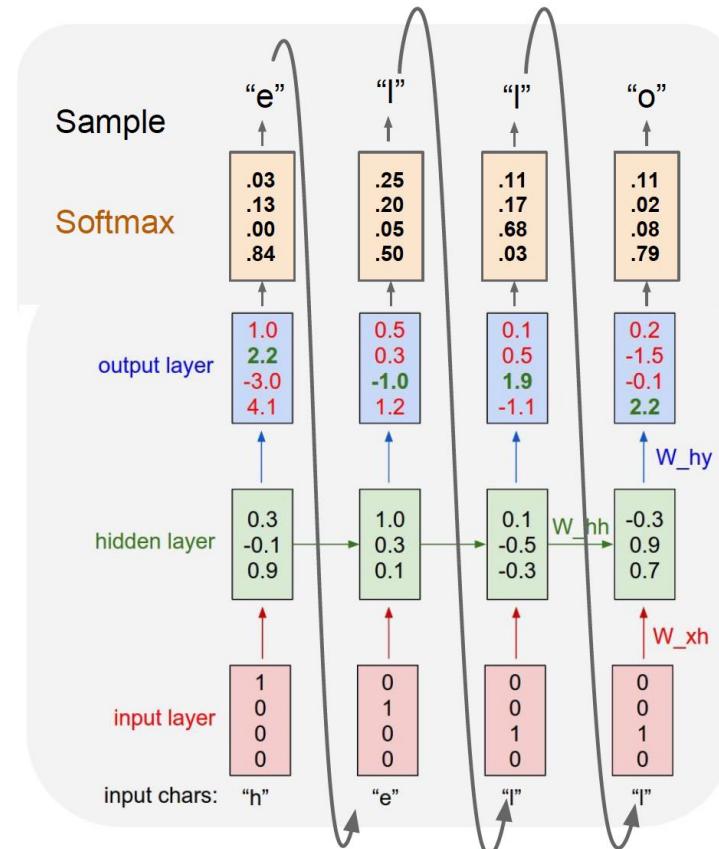
slide credit: Fei-Fei, Justin Johnson, Serena Yeung

Example...

Example: Character-level Language Model Sampling

Vocabulary:
[h,e,l,o]

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

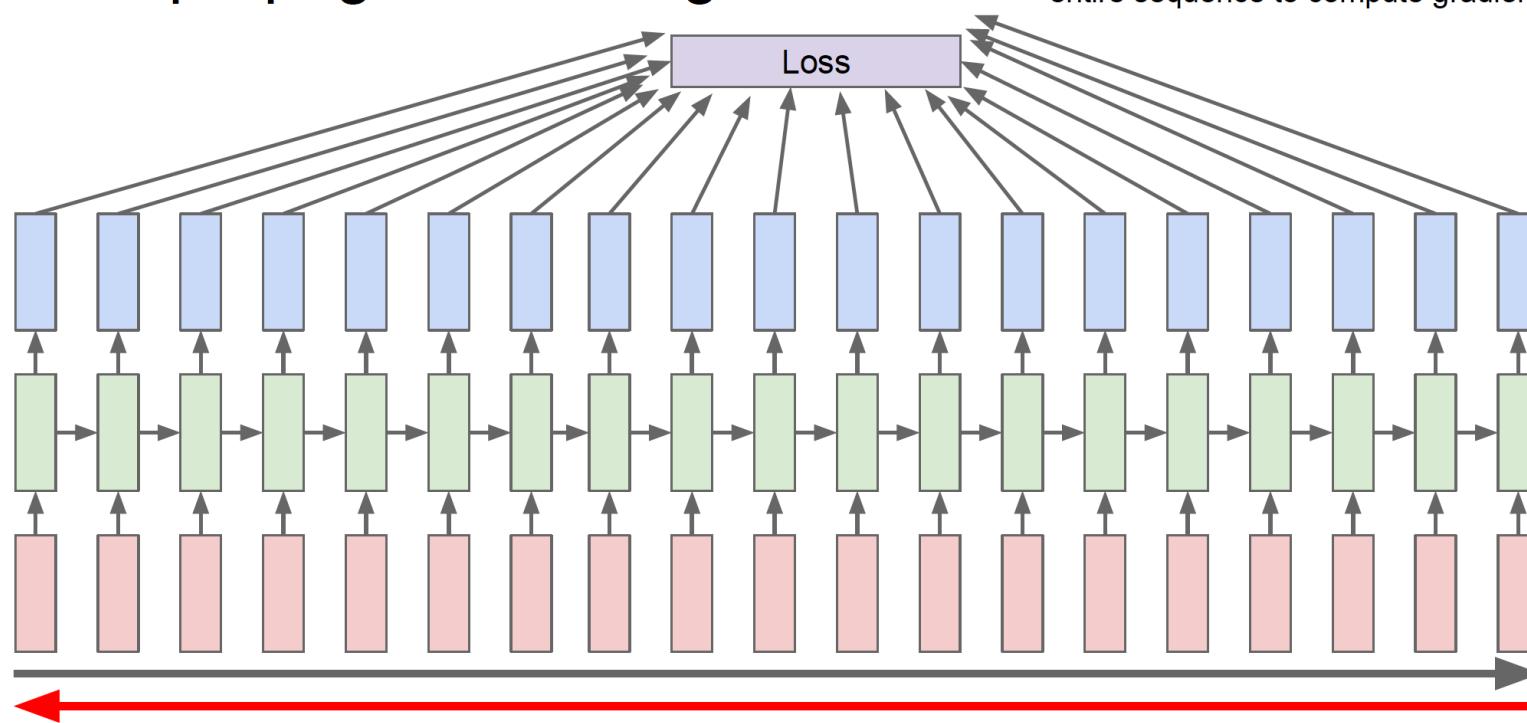


slide credit: Fei-Fei, Justin Johnson, Serena Yeung

Learning via Backpropagation...

Backpropagation through time

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

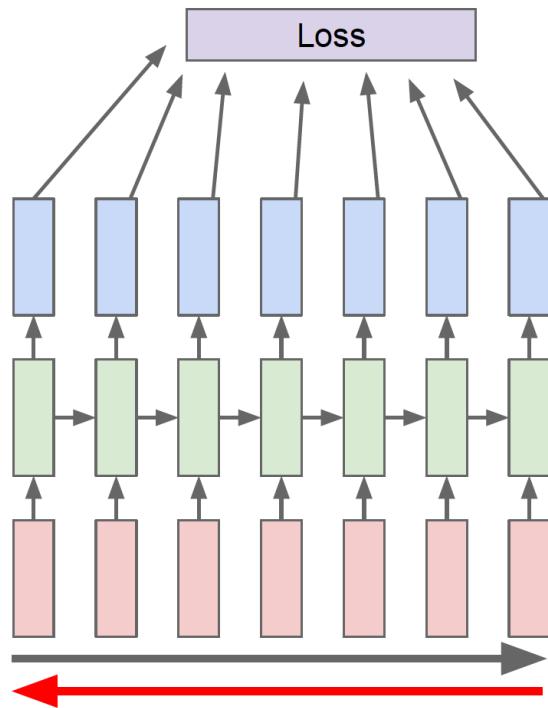


slide credit: Fei-Fei, Justin Johnson, Serena Yeung



Learning via Backpropagation...

Truncated Backpropagation through time

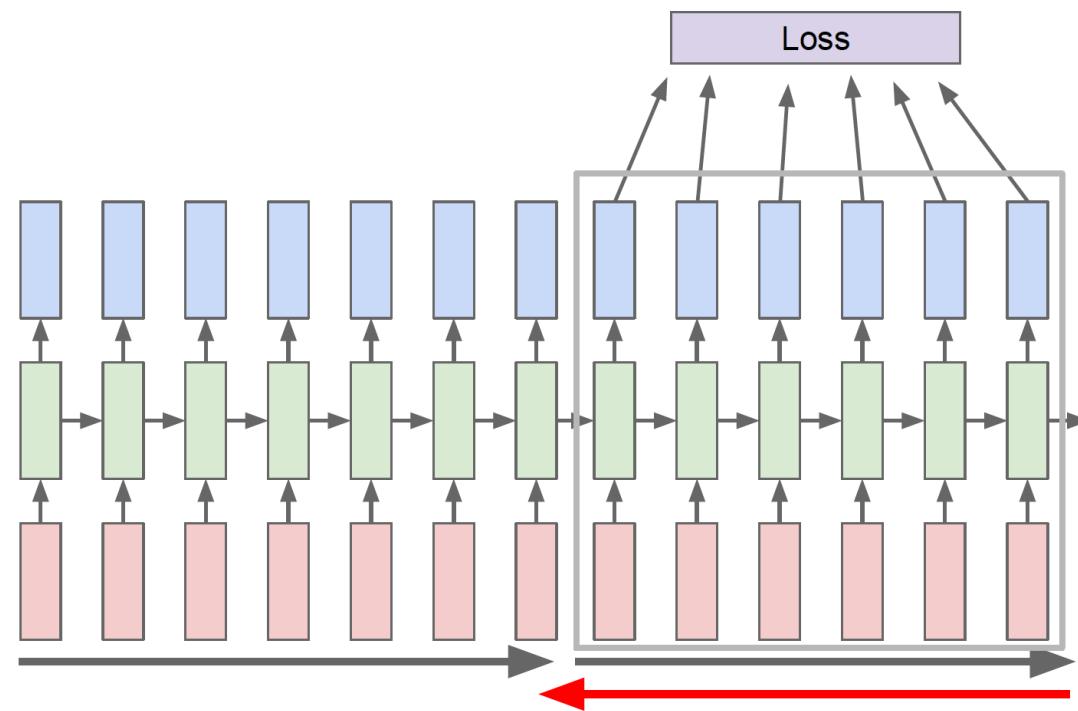


Run forward and backward
through chunks of the
sequence instead of whole
sequence

slide credit: Fei-Fei, Justin Johnson, Serena Yeung

Learning via Backpropagation...

Truncated Backpropagation through time

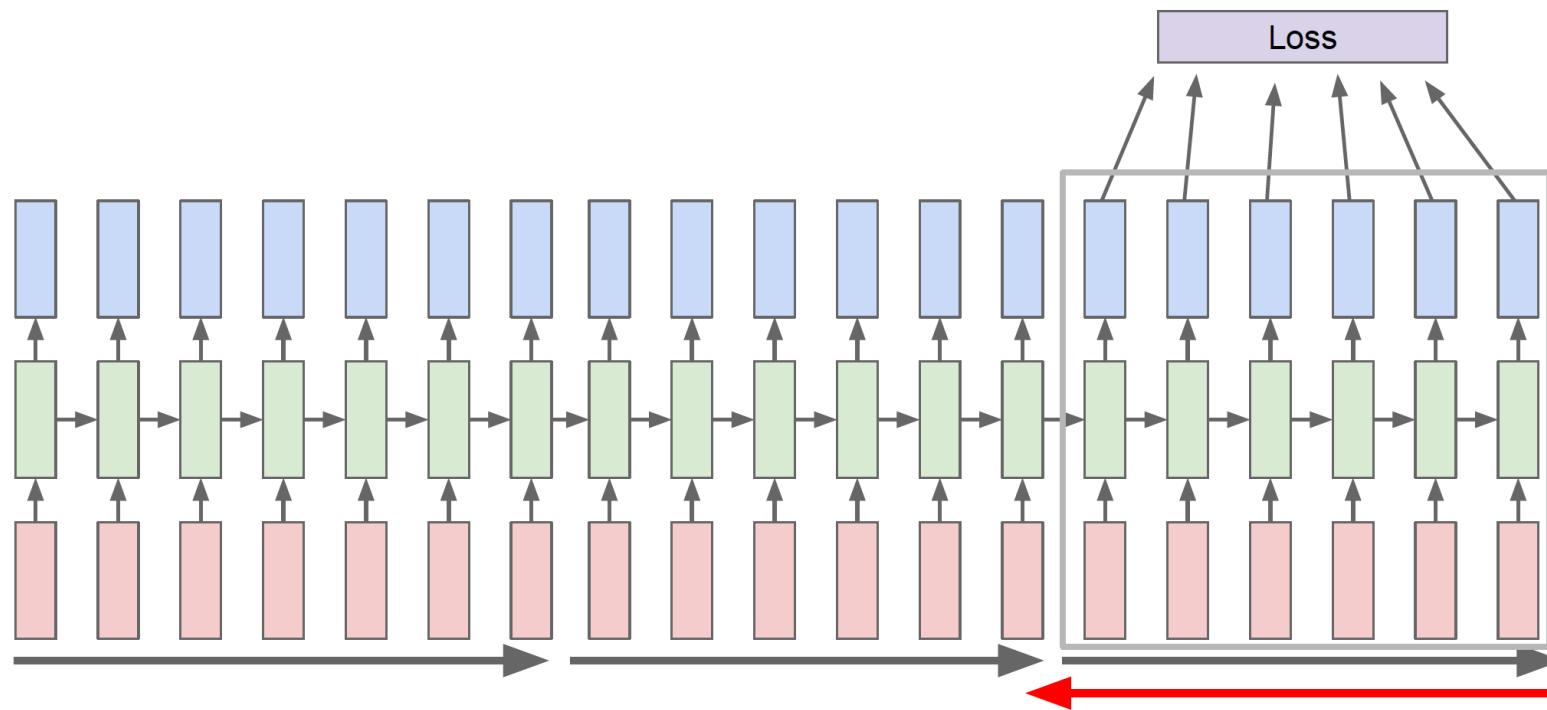


Carry hidden states
forward in time forever,
but only backpropagate
for some smaller
number of steps

slide credit: Fei-Fei, Justin Johnson, Serena Yeung

Learning via Backpropagation...

Truncated Backpropagation through time



slide credit: Fei-Fei, Justin Johnson, Serena Yeung

slide credit: Andrej Karpathy

“The Unreasonable Effectiveness of Recurrent Neural Networks”

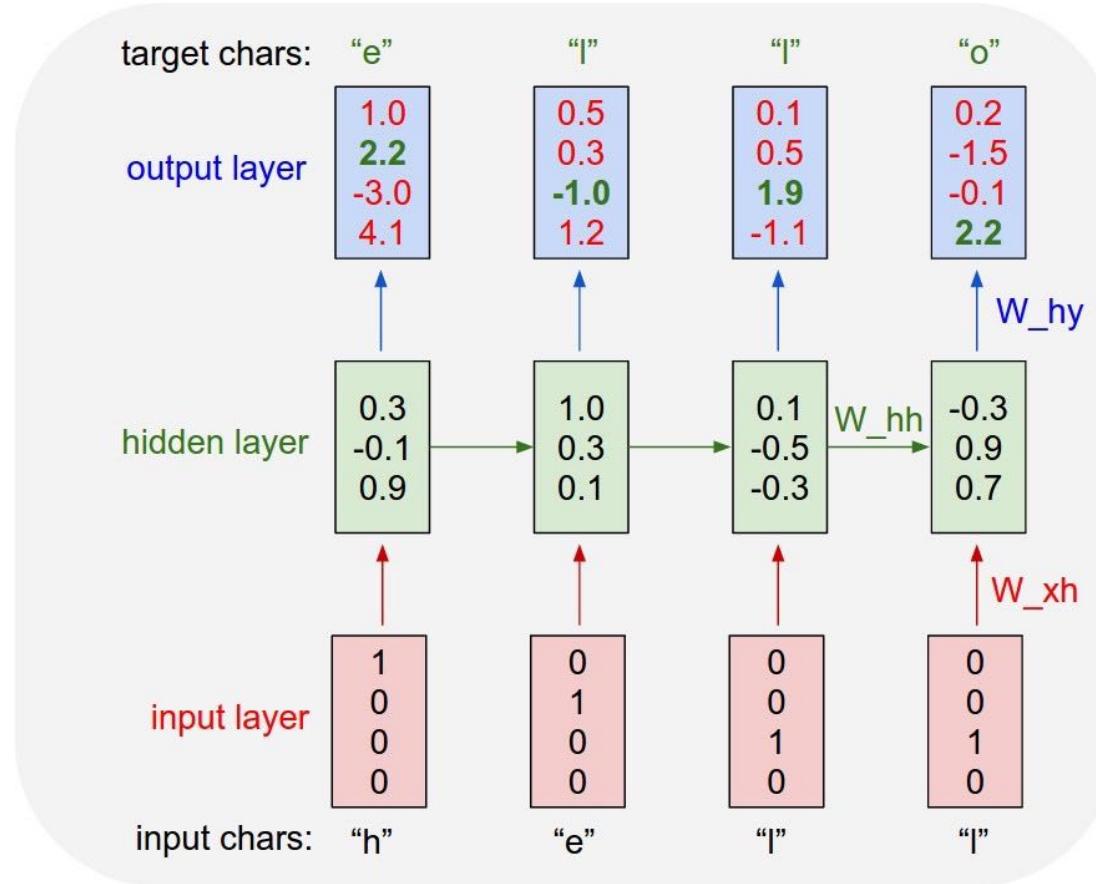
karpathy.github.io

Character-level language model example

Vocabulary:
[h,e,l,o]

Example training
sequence:
“hello”

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



Sonnet 116 – Let me not ...

by William Shakespeare

Let me not to the marriage of true minds
Admit impediments. Love is not love
Which alters when it alteration finds,
Or bends with the remover to remove:
O no! it is an ever-fixed mark
That looks on tempests and is never shaken;
It is the star to every wandering bark,
Whose worth's unknown, although his height be taken.
Love's not Time's fool, though rosy lips and cheeks
Within his bending sickle's compass come:
Love alters not with his brief hours and weeks,
But bears it out even to the edge of doom.
If this be error and upon me proved,
I never writ, nor no man ever loved.

slide credit: Andrej Karpathy

The Stacks Project

[home](#) [about](#) [tags explained](#) [tag lookup](#) [browse](#) [search](#) [bibliography](#) [recent comments](#) [blog](#) [add slogans](#)

Browse chapters

Part	Chapter	online	TeX source	view pdf
Preliminaries	1. Introduction	online	tex	pdf
	2. Conventions	online	tex	pdf
	3. Set Theory	online	tex	pdf
	4. Categories	online	tex	pdf
	5. Topology	online	tex	pdf
	6. Sheaves on Spaces	online	tex	pdf
	7. Sites and Sheaves	online	tex	pdf
	8. Stacks	online	tex	pdf
	9. Fields	online	tex	pdf
	10. Commutative Algebra	online	tex	pdf

Parts

- [Preliminaries](#)
- [Schemes](#)
- [Topics in Scheme Theory](#)
- [Algebraic Spaces](#)
- [Topics in Geometry](#)
- [Deformation Theory](#)
- [Algebraic Stacks](#)
- [Miscellany](#)

Statistics

The Stacks project now consists of

- o 455910 lines of code
- o 14221 tags (56 inactive tags)
- o 2366 sections

For $\bigoplus_{n=1,\dots,m} \mathcal{L}_{m_n} = 0$, hence we can find a closed subset \mathcal{H} in \mathcal{H} and any sets \mathcal{F} on X , U is a closed immersion of S , then $U \rightarrow T$ is a separated algebraic space.

Proof. Proof of (1). It also start we get

$$S = \text{Spec}(R) = U \times_X U \times_X U$$

and the comparicoly in the fibre product covering we have to prove the lemma generated by $\coprod Z \times_U U \rightarrow V$. Consider the maps M along the set of points Sch_{fppf} and $U \rightarrow U$ is the fibre category of S in U in Section, ?? and the fact that any U affine, see Morphisms, Lemma ???. Hence we obtain a scheme S and any open subset $W \subset U$ in $\text{Sh}(G)$ such that $\text{Spec}(R') \rightarrow S$ is smooth or an

$$U = \bigcup U_i \times_{S_i} U_i$$

which has a nonzero morphism we may assume that f_i is of finite presentation over S . We claim that $\mathcal{O}_{X,x}$ is a scheme where $x, x', s'' \in S'$ such that $\mathcal{O}_{X,x'} \rightarrow \mathcal{O}'_{X',x'}$ is separated. By Algebra, Lemma ?? we can define a map of complexes $\text{GL}_{S'}(x'/S'')$ and we win. \square

To prove study we see that $\mathcal{F}|_U$ is a covering of X' , and \mathcal{T}_i is an object of $\mathcal{F}_{X/S}$ for $i > 0$ and \mathcal{F}_p exists and let \mathcal{F}_i be a presheaf of \mathcal{O}_X -modules on \mathcal{C} as a \mathcal{F} -module. In particular $\mathcal{F} = U/\mathcal{F}$ we have to show that

$$\widetilde{M}^\bullet = \mathcal{I}^\bullet \otimes_{\text{Spec}(k)} \mathcal{O}_{S,s} - i_X^{-1} \mathcal{F}$$

is a unique morphism of algebraic stacks. Note that

$$\text{Arrows} = (\text{Sch}/S)^{\text{opp}}_{fppf}, (\text{Sch}/S)_{fppf}$$

and

$$V = \Gamma(S, \mathcal{O}) \longrightarrow (U, \text{Spec}(A))$$

is an open subset of X . Thus U is affine. This is a continuous map of X is the inverse, the groupoid scheme S .

Proof. See discussion of sheaves of sets. \square

The result for prove any open covering follows from the less of Example ???. It may replace S by $X_{\text{spaces},\text{étale}}$ which gives an open subspace of X and T equal to S_{Zar} , see Descent, Lemma ???. Namely, by Lemma ?? we see that R is geometrically regular over S .

Lemma 0.1. Assume (3) and (3) by the construction in the description.

Suppose $X = \lim |X|$ (by the formal open covering X and a single map $\underline{\text{Proj}}_X(\mathcal{A}) = \text{Spec}(B)$ over U compatible with the complex

$$\text{Set}(\mathcal{A}) = \Gamma(X, \mathcal{O}_{X,\mathcal{O}_X}).$$

When in this case of to show that $\mathcal{Q} \rightarrow \mathcal{C}_{Z/X}$ is stable under the following result in the second conditions of (1), and (3). This finishes the proof. By Definition ?? (without element is when the closed subschemes are catenary. If T is surjective we may assume that T is connected with residue fields of S . Moreover there exists a closed subspace $Z \subset X$ of X where U in X' is proper (some defining as a closed subset of the uniqueness it suffices to check the fact that the following theorem

(1) f is locally of finite type. Since $S = \text{Spec}(R)$ and $Y = \text{Spec}(R)$.

Proof. This is form all sheaves of sheaves on X . But given a scheme U and a surjective étale morphism $U \rightarrow X$. Let $U \cap U = \coprod_{i=1,\dots,n} U_i$ be the scheme X over S at the schemes $X_i \rightarrow X$ and $U = \lim_i X_i$. \square

The following lemma surjective restrocomposes of this implies that $\mathcal{F}_{x_0} = \mathcal{F}_{x_0} = \mathcal{F}_{X,\dots,0}$

Lemma 0.2. Let X be a locally Noetherian scheme over S , $E = \mathcal{F}_{X/S}$. Set $\mathcal{I} = \mathcal{J}_1 \subset \mathcal{I}'_n$. Since $\mathcal{I}^n \subset \mathcal{I}^n$ are nonzero over $i_0 \leq p$ is a subset of $\mathcal{J}_{n,0} \circ \overline{A}_2$ works.

Lemma 0.3. In Situation ???. Hence we may assume $q' = 0$.

Proof. We will use the property we see that p is the next functor (??). On the other hand, by Lemma ?? we see that

$$D(\mathcal{O}_{X'}) = \mathcal{O}_X(D)$$

where K is an F -algebra where δ_{n+1} is a scheme over S . \square

Proof. Omitted. \square

Lemma 0.1. Let \mathcal{C} be a set of the construction.

Let \mathcal{C} be a gerber covering. Let \mathcal{F} be a quasi-coherent sheaves of \mathcal{O} -modules. We have to show that

$$\mathcal{O}_{\mathcal{O}_X} = \mathcal{O}_X(\mathcal{L})$$

Proof. This is an algebraic space with the composition of sheaves \mathcal{F} on X_{etale} we have

$$\mathcal{O}_X(\mathcal{F}) = \{morph_1 \times_{\mathcal{O}_X} (\mathcal{G}, \mathcal{F})\}$$

where \mathcal{G} defines an isomorphism $\mathcal{F} \rightarrow \mathcal{F}$ of \mathcal{O} -modules. \square

Lemma 0.2. This is an integer \mathcal{Z} is injective.

Proof. See Spaces, Lemma ??.

Lemma 0.3. Let S be a scheme. Let X be a scheme and X is an affine open covering. Let $\mathcal{U} \subset \mathcal{X}$ be a canonical and locally of finite type. Let X be a scheme. Let X be a scheme which is equal to the formal complex.

The following to the construction of the lemma follows.

Let X be a scheme. Let X be a scheme covering. Let

$$b : X \rightarrow Y' \rightarrow Y \rightarrow Y' \times_X Y \rightarrow X.$$

be a morphism of algebraic spaces over S and Y .

Proof. Let X be a nonzero scheme of X . Let X be an algebraic space. Let \mathcal{F} be a quasi-coherent sheaf of \mathcal{O}_X -modules. The following are equivalent

- (1) \mathcal{F} is an algebraic space over S .
- (2) If X is an affine open covering.

Consider a common structure on X and X the functor $\mathcal{O}_X(U)$ which is locally of finite type. \square

This since $\mathcal{F} \in \mathcal{F}$ and $x \in \mathcal{G}$ the diagram

$$\begin{array}{ccccc}
 S & \xrightarrow{\quad} & & & \\
 \downarrow & & & & \\
 \xi & \longrightarrow & \mathcal{O}_{X'} & & \\
 \text{gor}_* & & \uparrow & \searrow & \\
 & & = \alpha' & \longrightarrow & \\
 & & \downarrow & & \\
 & & = \alpha' & \longrightarrow & \alpha \\
 & & & & \\
 \text{Spec}(K_\psi) & & \text{Mor}_{Sets} & & X \\
 & & & & \downarrow \\
 & & & & d(\mathcal{O}_{X_{f/k}}, \mathcal{G})
 \end{array}$$

is a limit. Then \mathcal{G} is a finite type and assume S is a flat and \mathcal{F} and \mathcal{G} is a finite type f_* . This is of finite type diagrams, and

- the composition of \mathcal{G} is a regular sequence,
- $\mathcal{O}_{X'}$ is a sheaf of rings.

Proof. We have see that $X = \text{Spec}(R)$ and \mathcal{F} is a finite type representable by algebraic space. The property \mathcal{F} is a finite morphism of algebraic stacks. Then the cohomology of X is an open neighbourhood of U . \square

Proof. This is clear that \mathcal{G} is a finite presentation, see Lemmas ??.
A reduced above we conclude that U is an open covering of \mathcal{C} . The functor \mathcal{F} is a “field”

$$\mathcal{O}_{X,x} \rightarrow \mathcal{F}_{\overline{x}} \dashv^{-1} (\mathcal{O}_{X_{etale}}) \rightarrow \mathcal{O}_{X_\ell}^{-1} \mathcal{O}_{X_\lambda} (\mathcal{O}_{X_\eta}^\pi)$$

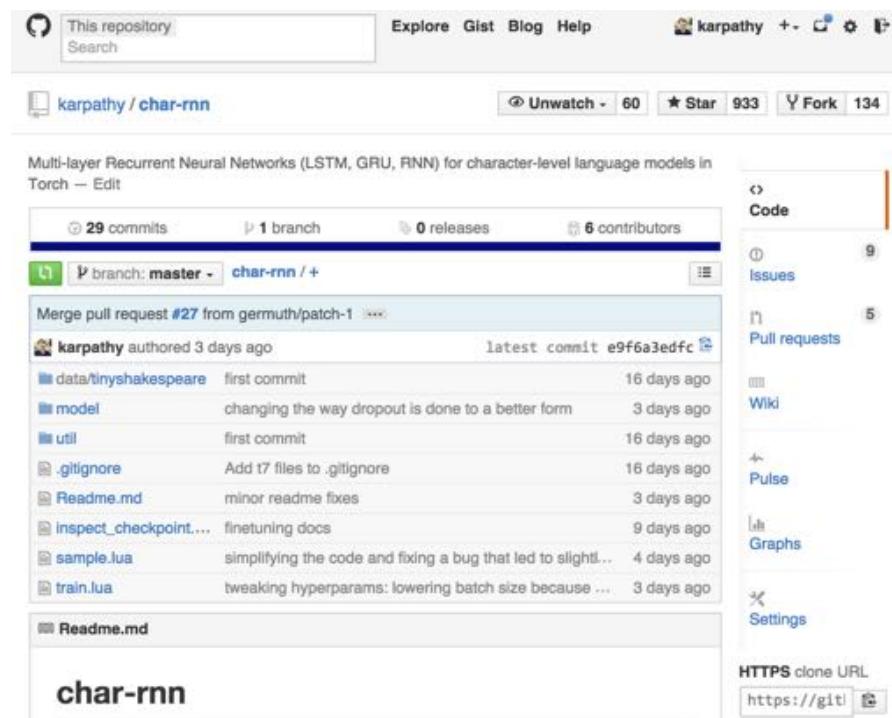
is an isomorphism of covering of \mathcal{O}_{X_ℓ} . If \mathcal{F} is the unique element of \mathcal{F} such that X is an isomorphism.

The property \mathcal{F} is a disjoint union of Proposition ?? and we can filtered set of presentations of a scheme \mathcal{O}_X -algebra with \mathcal{F} are opens of finite type over S .
If \mathcal{F} is a scheme theoretic image points. \square

If \mathcal{F} is a finite direct sum \mathcal{O}_{X_λ} is a closed immersion, see Lemma ??.
This is a sequence of \mathcal{F} is a similar morphism.

slide credit: Andrej Karpathy

Try it yourself: **char-rnn** on Github (uses Torch7)



slide credit: Andrej Karpathy

Cooking Recipes

Title: BASIC CHEESE WINGS:
Categories: Desserts
Yield: 6 Servings

3 Eggs
2 tb Chopped fresh curry
-or cooking spray
1 c Water; cooked
2 Lemons minced mushrooms
3 oz Sweet cooked rice
1/2 Onion; chopped
3 c Butter, melted
2 ts Soy sauce
1 ts Cinnamon
2 md Sugar or food coloring;
-stems cored bowl
2 tb Salt and freshly grated
1/4 ts Ground ginger
1/2 c Flour
1 tb Water; fresh parsley
1 c Water (or or)
1 Clove garlic, minced

Preheat oven to 350F. Combine sugar, salt, baking soda, celery and sugar. Add the chicken broth well. Add the cornstarch to the pan; cool. Add the olive oil, oil, and basil or cooking spray. Pour the onions until melted.

Obama Speeches

Good afternoon. God bless you.

The United States will step up to the cost of a new challenges of the American people that will share the fact that we created the problem. They were attacked and so that they have to say that all the task of the final days of war that I will not be able to get this done. The promise of the men and women who were still going to take out the fact that the American people have fought to make sure that they have to be able to protect our part. It was a chance to stand together to completely look for the commitment to borrow from the American people. And the fact is the men and women in uniform and the millions of our country with the law system that we should be a strong stretches of the forces that we can afford to increase our spirit of the American people and the leadership of our country who are on the Internet of American lives.

Thank you very much. God bless you, and God bless the United States of America.

slide credit: Andrej Karpathy

RNN Bible @RNN_Bible
Random bible verses generated using Recurrent Neural Networks (char-rnn).

TWEETS 80 FOLLOWING 1 FOLLOWERS 51

Tweets [Tweets & replies](#)

RNN Bible @RNN_Bible · 3h
32:22 And they shall be the children of Israel, and they that shall come upon us, that they may be their God.

RNN Bible @RNN_Bible · 7h
2:11 Therefore shall they see thy chastisement for them, they shall live: I will sing praise to thee in the night thy servant.

RNN Bible @RNN_Bible · 11h
8:26 And they set the book of the law which Michal the Baptist came near to Man.

slide credit: Andrej Karpathy

The screenshot shows the GitHub repository page for `torvalds/linux`. The top navigation bar includes links for 'Explore', 'Gist', 'Blog', and 'Help'. On the right, there's a user profile for `karpathy` with options to '+', fork, settings, and a copy icon. The repository header shows 'Watch' (3,711), 'Star' (23,054), and 'Fork' (9,141). The main content area displays the 'Linux kernel source tree' with statistics: 520,037 commits, 1 branch, 420 releases, and 5,039 contributors. A dropdown menu shows the current branch is 'master'. Below this, a list of recent commits is shown, all authored by `torvalds` and merged into the 'master' branch. The commits are categorized by file type (Documentation, arch, block, crypto, drivers, firmware, fs, include, init, ipc) and include details like commit message, target date, and author. To the right of the commit list are links for 'Code', 'Pull requests' (74), 'Pulse', and 'Graphs'. At the bottom, there's an 'HTTPS clone URL' field containing `https://github.com/torvalds/linux`, a note about cloning with HTTPS or SSH, and buttons for 'Clone in Desktop' and 'Download ZIP'.

Linux kernel source tree

520,037 commits 1 branch 420 releases 5,039 contributors

branch: master / +

Merge branch 'drm-fixes' of git://people.freedesktop.org/~airlied/linux ...

torvalds authored 9 hours ago latest commit 4b1706927d

Category	Commit Message	Date
Documentation	Merge git://git.kernel.org/pub/scm/linux/kernel/git/nab/target-pending	6 days ago
arch	Merge branch 'x86-urgent-for-linus' of git://git.kernel.org/pub/scm/l...	a day ago
block	block: discard bdi_unregister() in favour of bdi_destroy()	9 days ago
crypto	Merge git://git.kernel.org/pub/scm/linux/kernel/git/heribert/crypto-2.6	10 days ago
drivers	Merge branch 'drm-fixes' of git://people.freedesktop.org/~airlied/linux	9 hours ago
firmware	firmware/ihex2fw.c: restore missing default in switch statement	2 months ago
fs	vfs: read file_handle only once in handle_to_path	4 days ago
include	Merge branch 'perf-urgent-for-linus' of git://git.kernel.org/pub/scm/...	a day ago
init	init: fix regression by supporting devices with major:minor:offset fo...	a month ago
ipc	ipc: fix race between NewRequest and old->limit lowered on the same file	a month ago

HTTPS clone URL
`https://github.com/torvalds/linux`

You can clone with **HTTPS**, **SSH**, or **Subversion**.

Clone in Desktop **Download ZIP**

slide credit: Andrej Karpathy

```
static void do_command(struct seq_file *m, void *v)
{
    int column = 32 << (cmd[2] & 0x80);
    if (state)
        cmd = (int)(int_state ^ (in_8(&ch->ch_flags) & Cmd) ? 2 : 1);
    else
        seq = 1;
    for (i = 0; i < 16; i++) {
        if (k & (1 << i))
            pipe = (in_use & UMXTHREAD_UNCCA) +
                ((count & 0x00000000fffff8) & 0x0000000f) << 8;
        if (count == 0)
            sub(pid, ppc_md.kexec_handle, 0x20000000);
        pipe_set_bytes(i, 0);
    }
    /* Free our user pages pointer to place camera if all dash */
    subsystem_info = &of_changes[PAGE_SIZE];
    rek_controls(offset, idx, &soffset);
    /* Now we want to deliberately put it to device */
    control_check_polarity(&context, val, 0);
    for (i = 0; i < COUNTER; i++)
        seq_puts(s, "policy ");
}
```

Learning from Linux Source Code

slide credit: Andrej Karpathy

```
static void do_command(struct seq_file *m, void *v)
{
    int column = 32 << (cmd[2] & 0x80);
    if (state)
        cmd = (int)(int_state ^ (in_8(&ch->ch_flags) & Cmd) ? 2 : 1);
    else
        seq = 1;
    for (i = 0; i < 16; i++) {
        if (k & (1 << i))
            pipe = (in_use & UMXTHREAD_UNCCA) +
                ((count & 0x00000000fffff8) & 0x000000f) << 8;
        if (count == 0)
            sub(pid, ppc_md.kexec_handle, 0x20000000);
        pipe_set_bytes(i, 0);
    }
    /* Free our user pages pointer to place camera if all dash */
    subsystem_info = &of_changes[PAGE_SIZE];
    rek_controls(offset, idx, &soffset);
    /* Now we want to deliberately put it to device */
    control_check_polarity(&context, val, 0);
    for (i = 0; i < COUNTER; i++)
        seq_puts(s, "policy ");
}
```



Linus Torvalds

Shared publicly - May 24, 2015

I'm not a fan of traditional AI (rule building and LISP/prolog etc), but am just waiting for neural nets to take over.

This is a great example, and training on the Linux kernel sources just explains so much.

Sigh.

David Alan Gilbert originally shared:

This is wonderful; neural network setup producing wonderful results; one example is training it on

[Read more](#)



slide credit: Andrej Karpathy

```
/*
 * Copyright (c) 2006-2010, Intel Mobile Communications. All rights reserved.
 *
 * This program is free software; you can redistribute it and/or modify it
 * under the terms of the GNU General Public License version 2 as published by
 * the Free Software Foundation.
 *
 * This program is distributed in the hope that it will be useful,
 * but WITHOUT ANY WARRANTY; without even the implied warranty of
 * MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the
 *
 * GNU General Public License for more details.
 *
 * You should have received a copy of the GNU General Public License
 * along with this program; if not, write to the Free Software Foundation,
 * Inc., 675 Mass Ave, Cambridge, MA 02139, USA.
 */

#include <linux/kexec.h>
#include <linux/errno.h>
#include <linux/io.h>
#include <linux/platform_device.h>
#include <linux/multi.h>
#include <linux/ckevent.h>

#include <asm/io.h>
#include <asm/prom.h>
#include <asm/e820.h>
#include <asm/system_info.h>
#include <asm/seteew.h>
#include <asm/pgproto.h>
```

slide credit: Andrej Karpathy

```
#include <asm/io.h>
#include <asm/prom.h>
#include <asm/e820.h>
#include <asm/system_info.h>
#include <asm/seteew.h>
#include <asm/pgproto.h>

#define REG_PG      vesa_slot_addr_pack
#define PFM_NOCOMP  AFSR(0, load)
#define STACK_DDR(type)      (func)

#define SWAP_ALLOCATE(nr)      (e)
#define emulate_sigs()  arch_get_unaligned_child()
#define access_rw(TST)  asm volatile("movd %%esp, %0, %3" : : "r" (0));    \
    if (__type & DO_READ)

static void stat_PC_SEC __read_mostly offsetof(struct seq_argsqueue, \
    pC>[1]);

static void
os_prefix(unsigned long sys)
{
#ifdef CONFIG_PREEMPT
    PUT_PARAM_RAID(2, sel) = get_state_state();
    set_pid_sum((unsigned long)state, current_state_str(),
                (unsigned long)-1->lr_full; low;
}

```

Yoav Goldberg n-gram experiments

Order 10 ngram model on Shakespeare:

APEMANTUS:

Ho, ho! I laugh to see your beard!

BOYET:

Madam, in great extremes of passion as she
discovers it.

PAROLLES:

By my white head and her wit
Values itself: to the sepulchre!
With this, my lord,
That I have some business: let's away.

First Keeper:

Forbear to murder: and wilt thou not say he lies,
And lies, and let the devil would have said, sir, their speed
Hath been balm to heal their woes,

But on Linux:

```
~~/*
 * linux/kernel/time.c
 * Please report this on hardware.
 */
void irq_mark_irq(unsigned long old_entries, eval);

/*
 * Divide only 1000 for ns^2 -> us^2 conversion values
don't overflow:
    seq_puts(m, "\tttramp: %pS",
            (void *)class->contending_point
]++;  

    if (likely(t->flags & WQ_UNBOUND)) {
        /*
         * Update inode information. If the
         * slowpath and sleep time (abs or rel)
        * @rmtp: remaining (either due
        * to consume the state of ring buffer size. */
        header_size - size, in bytes, of the chain.
        */
        BUG_ON(!error);
    } while (cgrp) {
        if (old) {
            if (kdb_continue_catastrophic;
#endif
```

Overview Today's Lecture

- Data Preprocessing
 - ▶ Activation functions
 - ▶ Batch normalization
- Recurrent Neural Networks (RNNs)
 - ▶ Motivation & flexibility of RNNs
 - ▶ Language modeling
 - including “unreasonable effectiveness of RNNs”
 - ▶ RNNs for image description / captioning
 - ▶ Standard RNN and a particularly successful RNN:
Long Short Term Memory (LSTM)
 - including “visualizations of RNN cells”

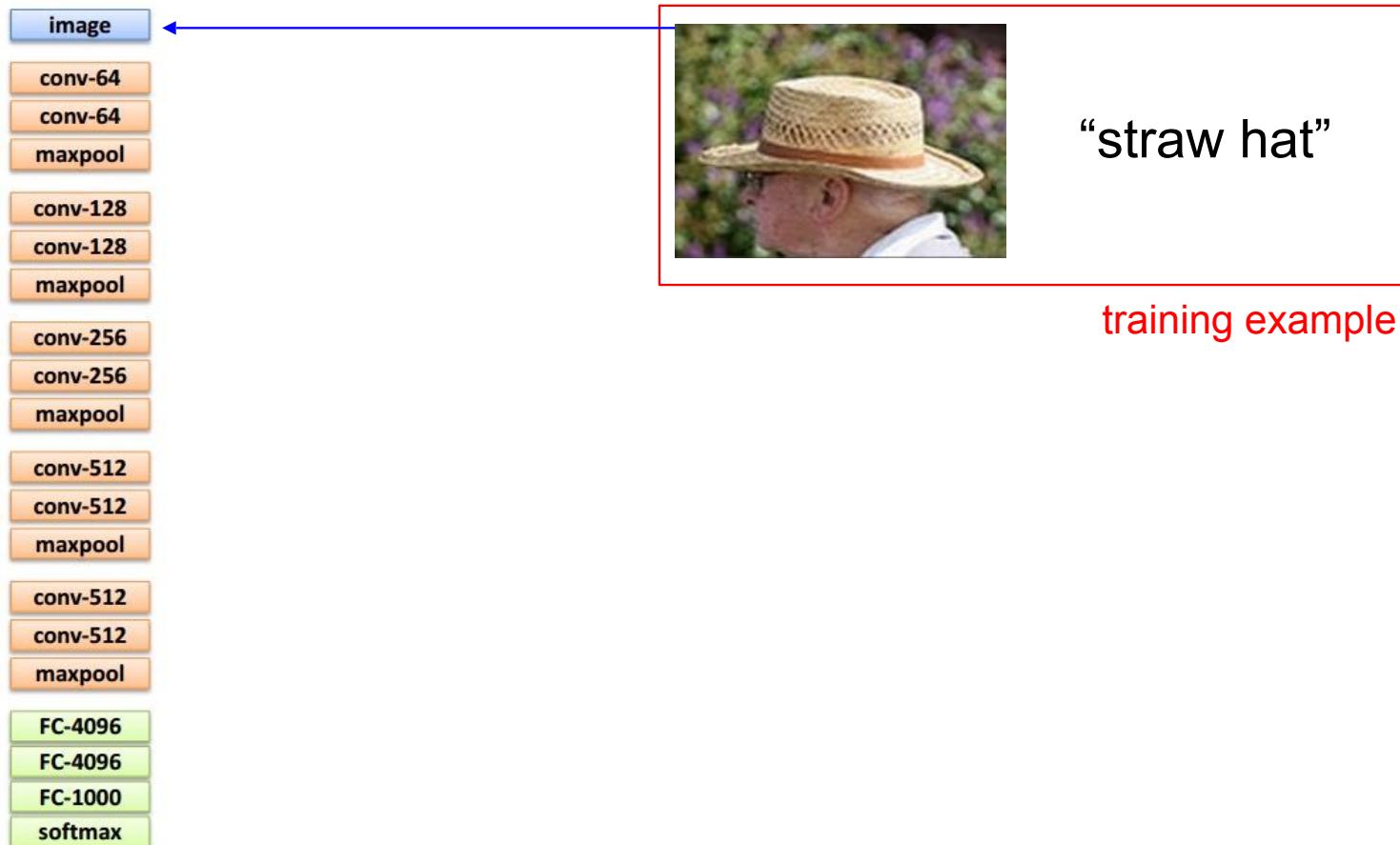
slide credit: Andrej Karpathy



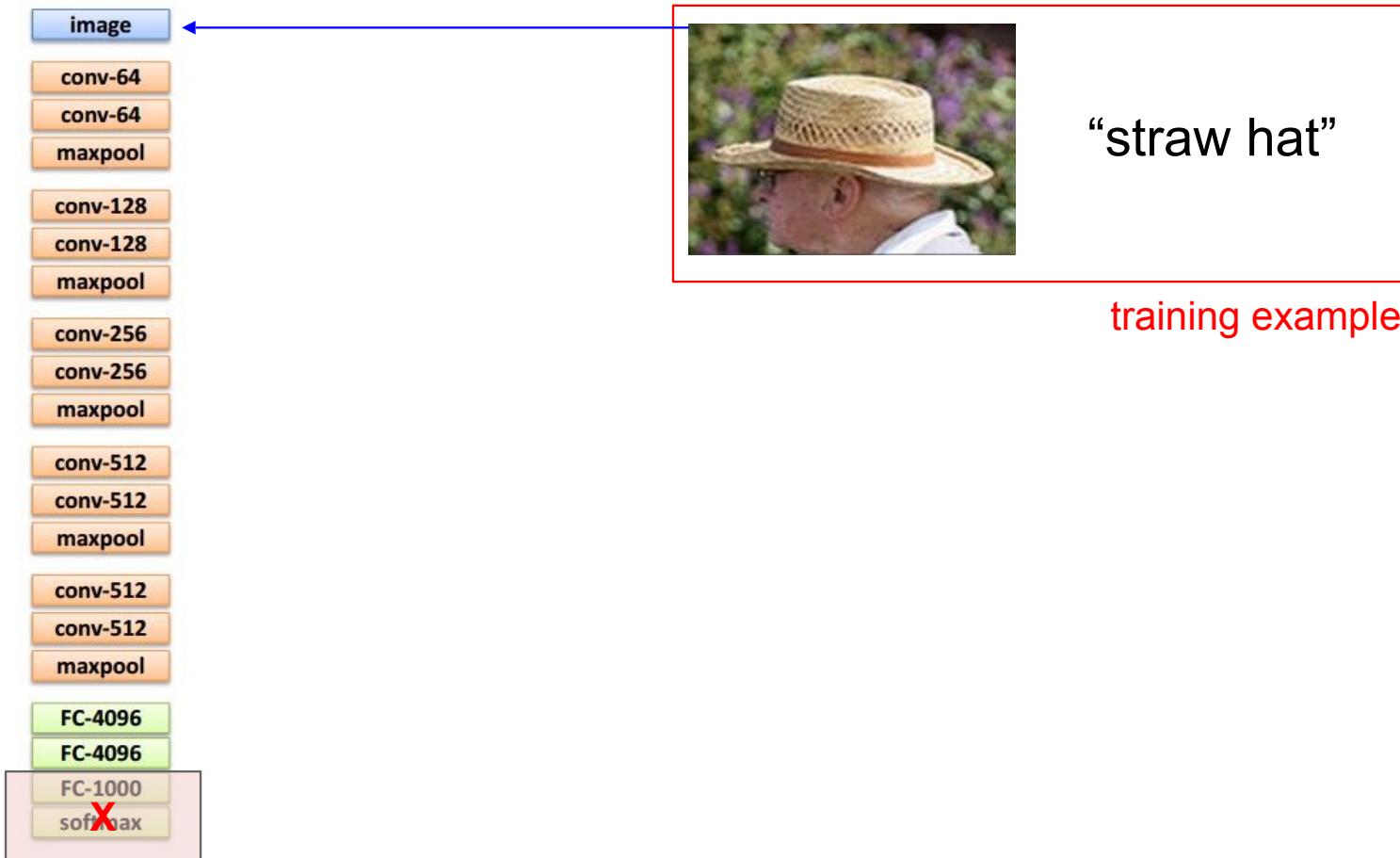
“straw hat”

training example

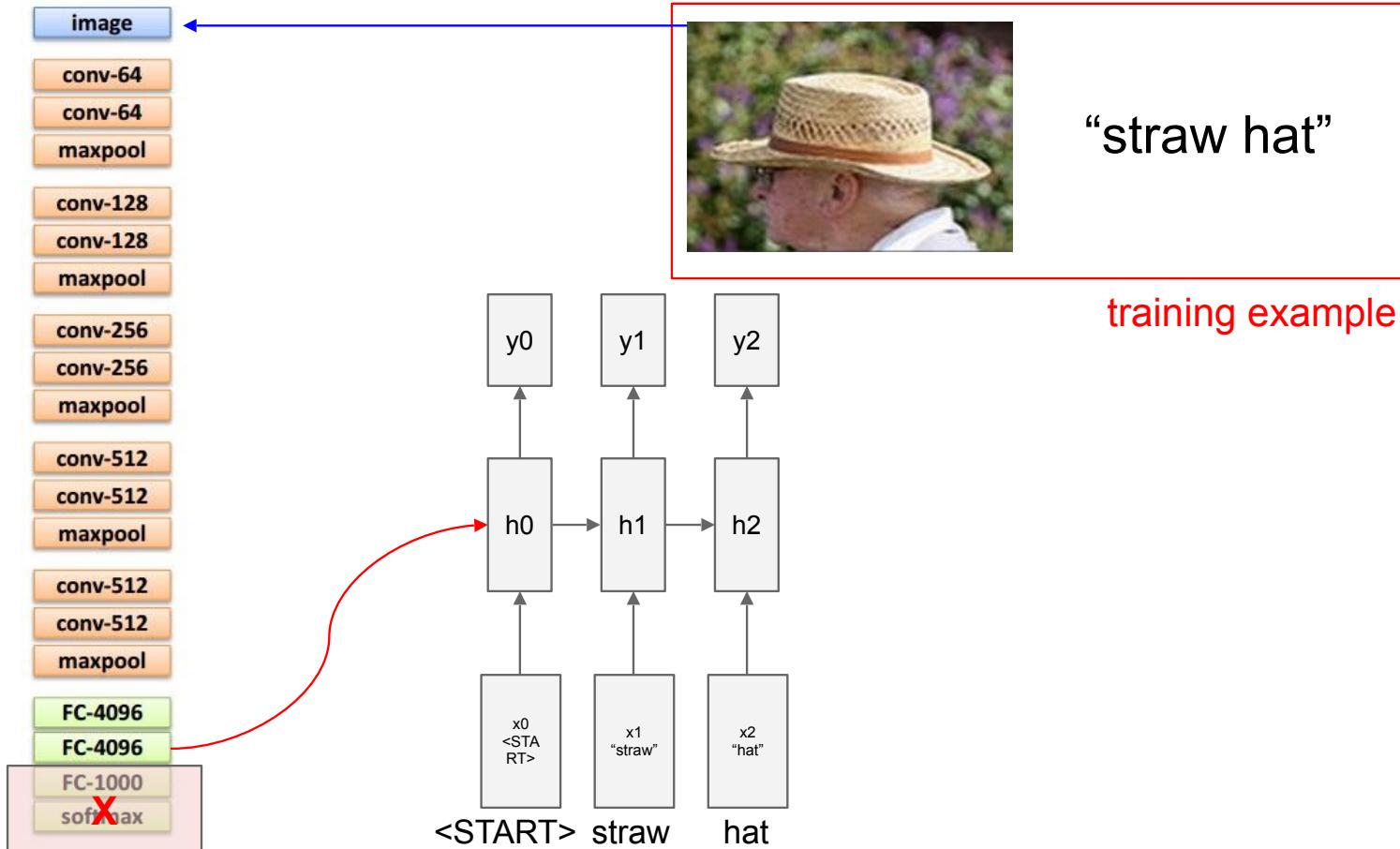
slide credit: Andrej Karpathy



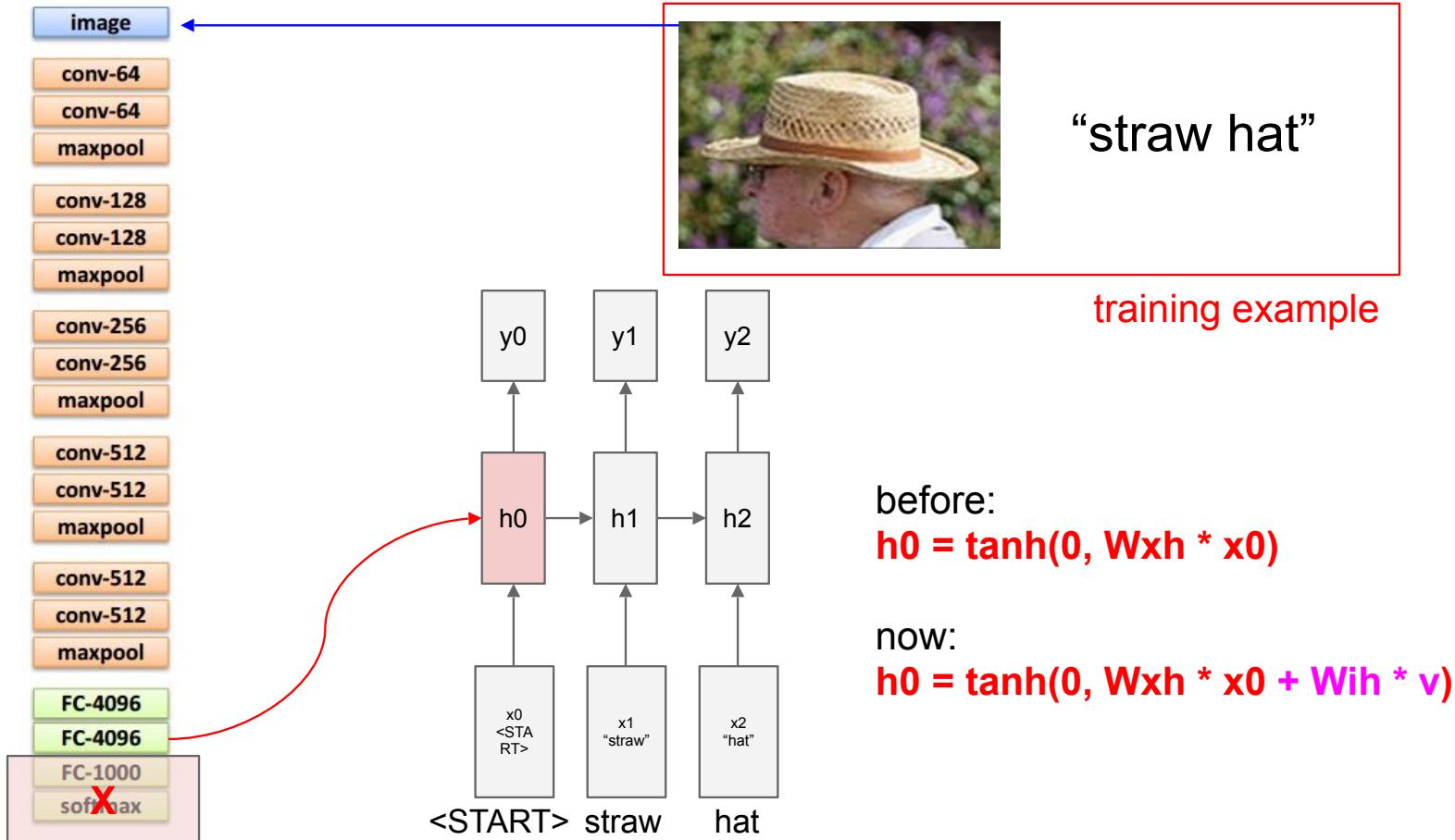
slide credit: Andrej Karpathy



slide credit: Andrej Karpathy



slide credit: Andrej Karpathy

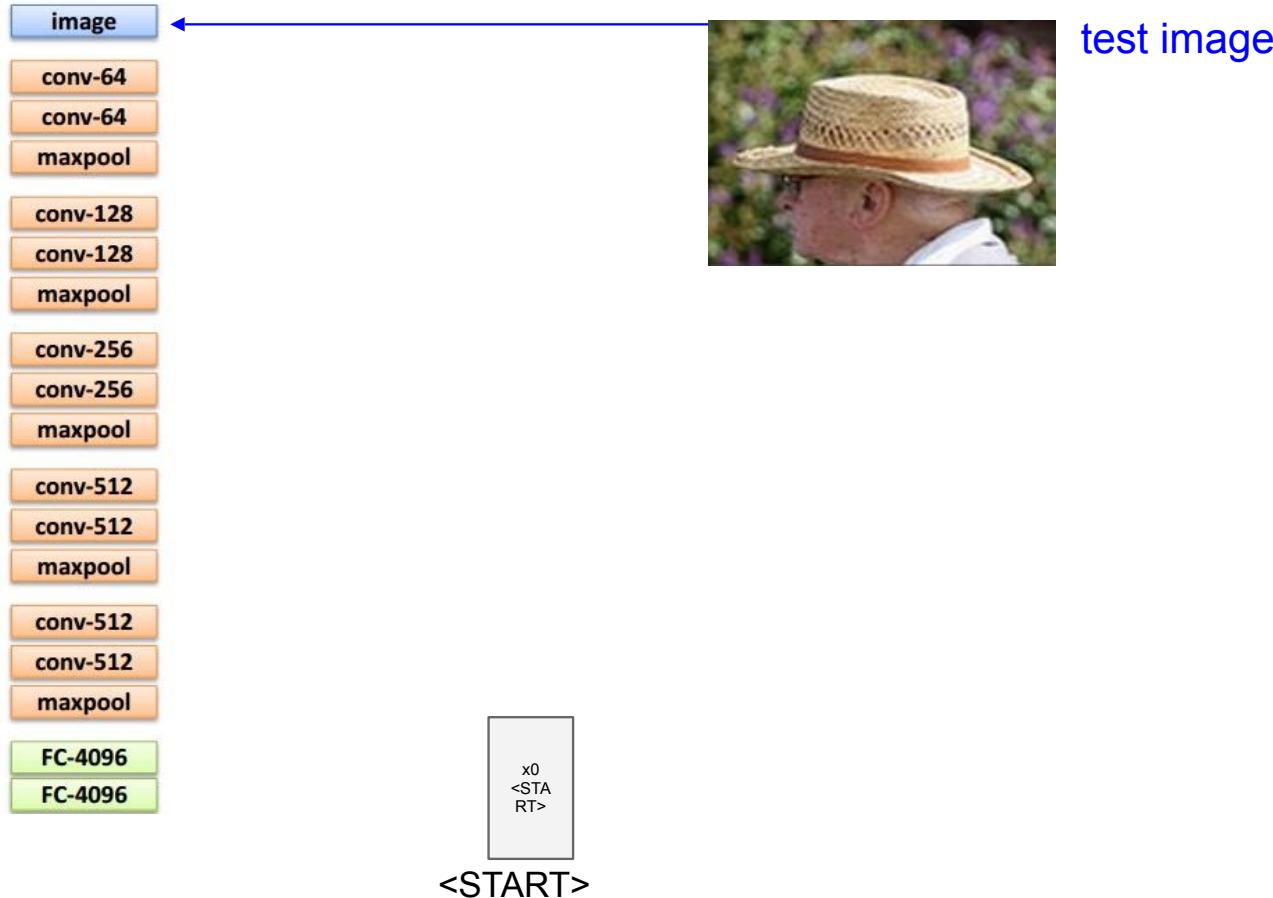


slide credit: Andrej Karpathy

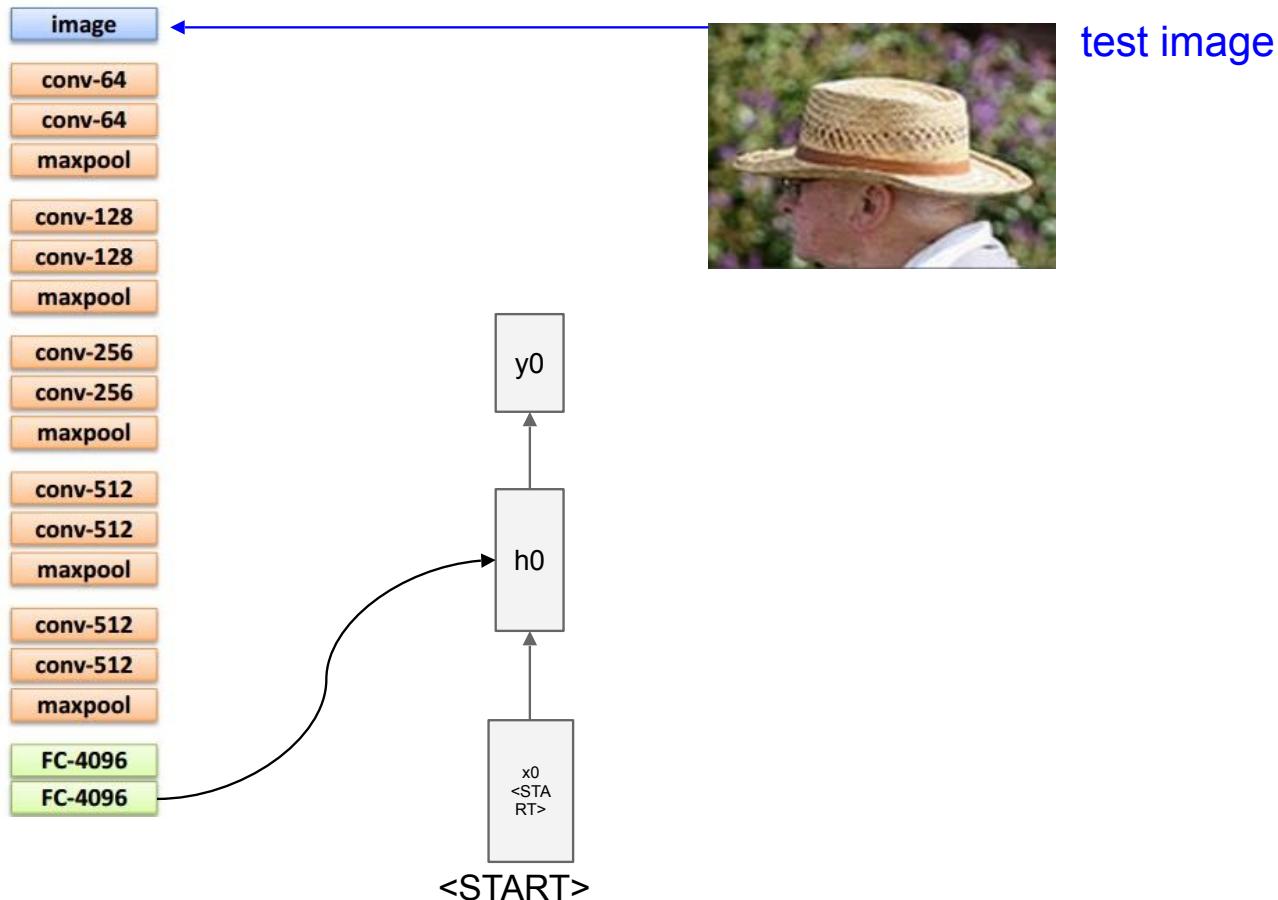


test image

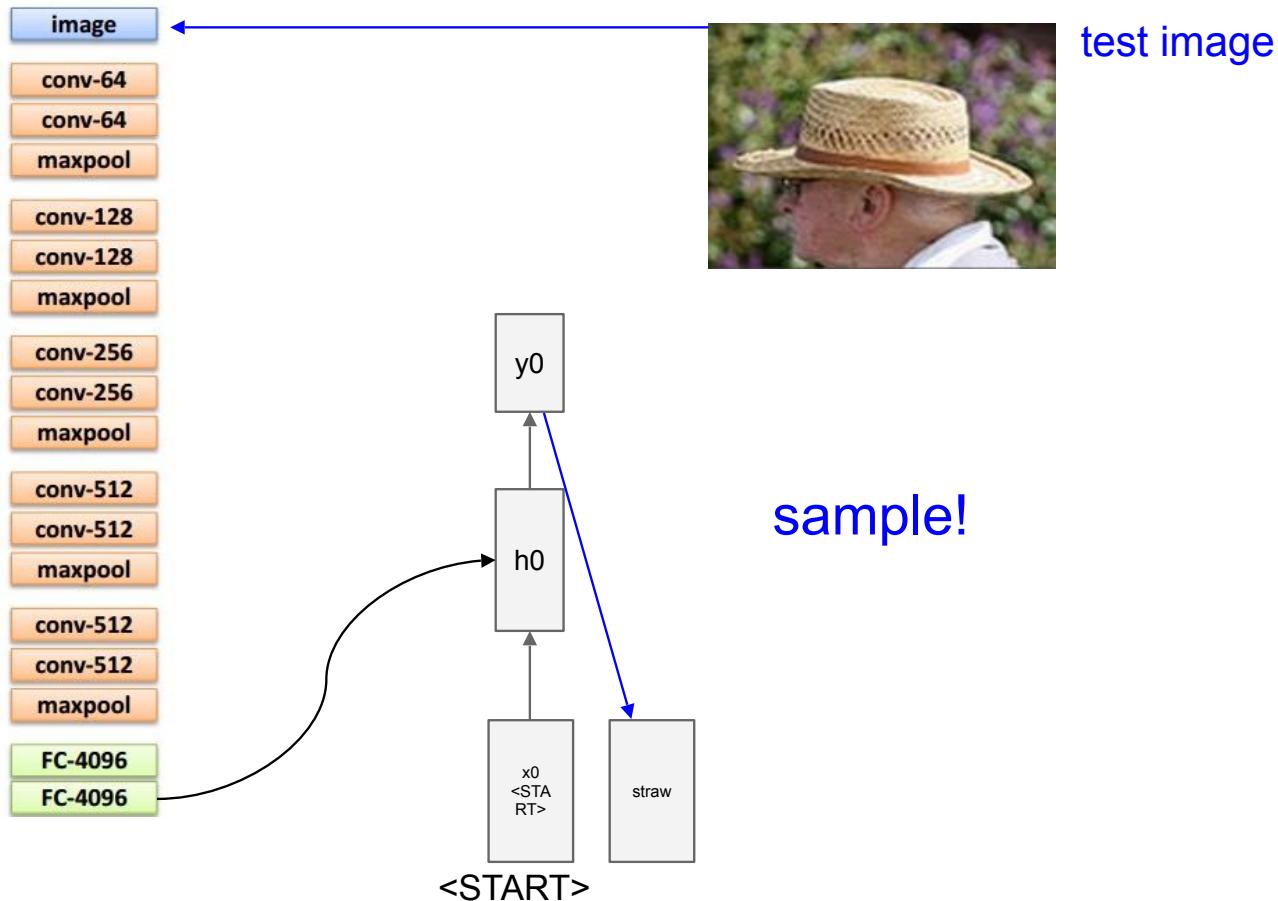
slide credit: Andrej Karpathy



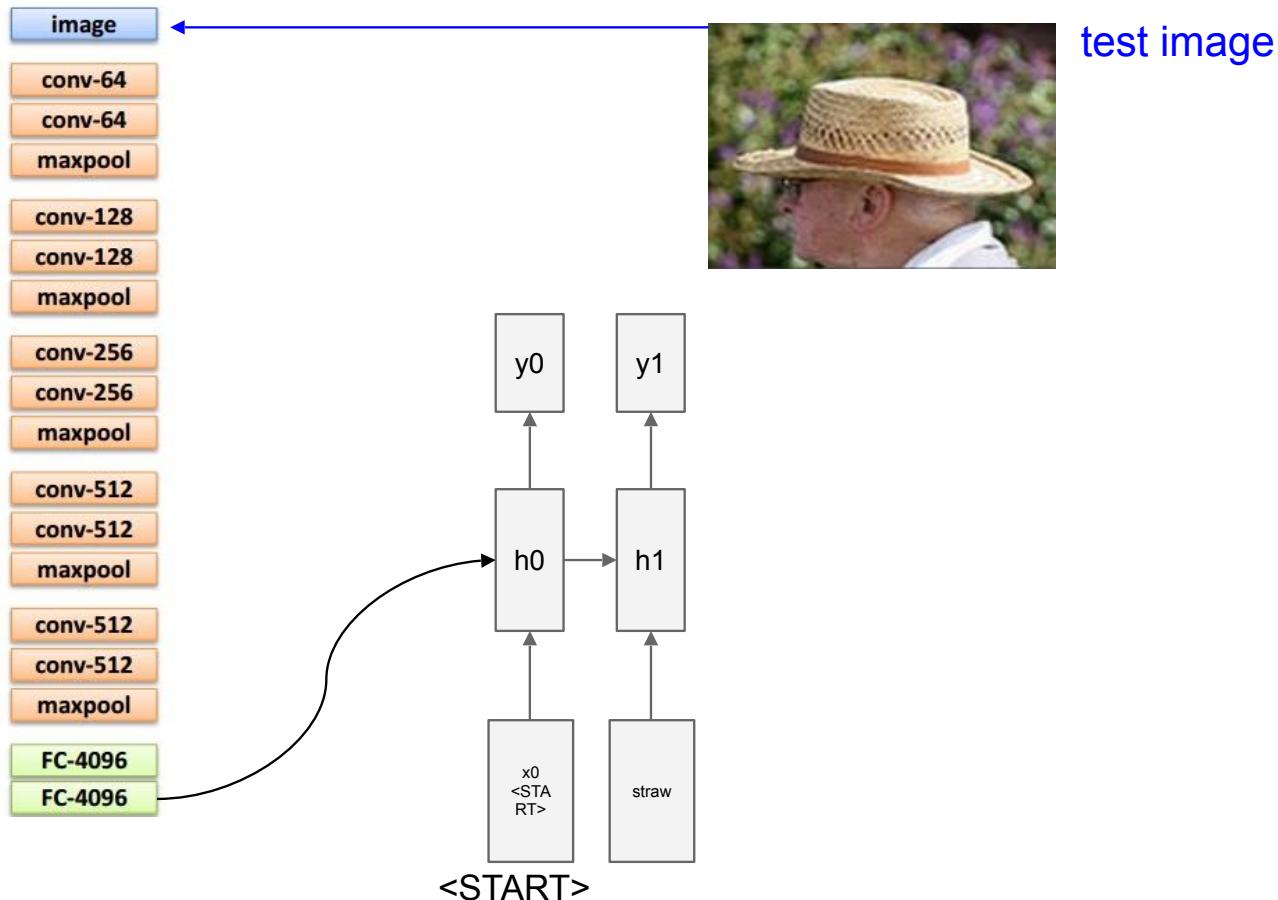
slide credit: Andrej Karpathy



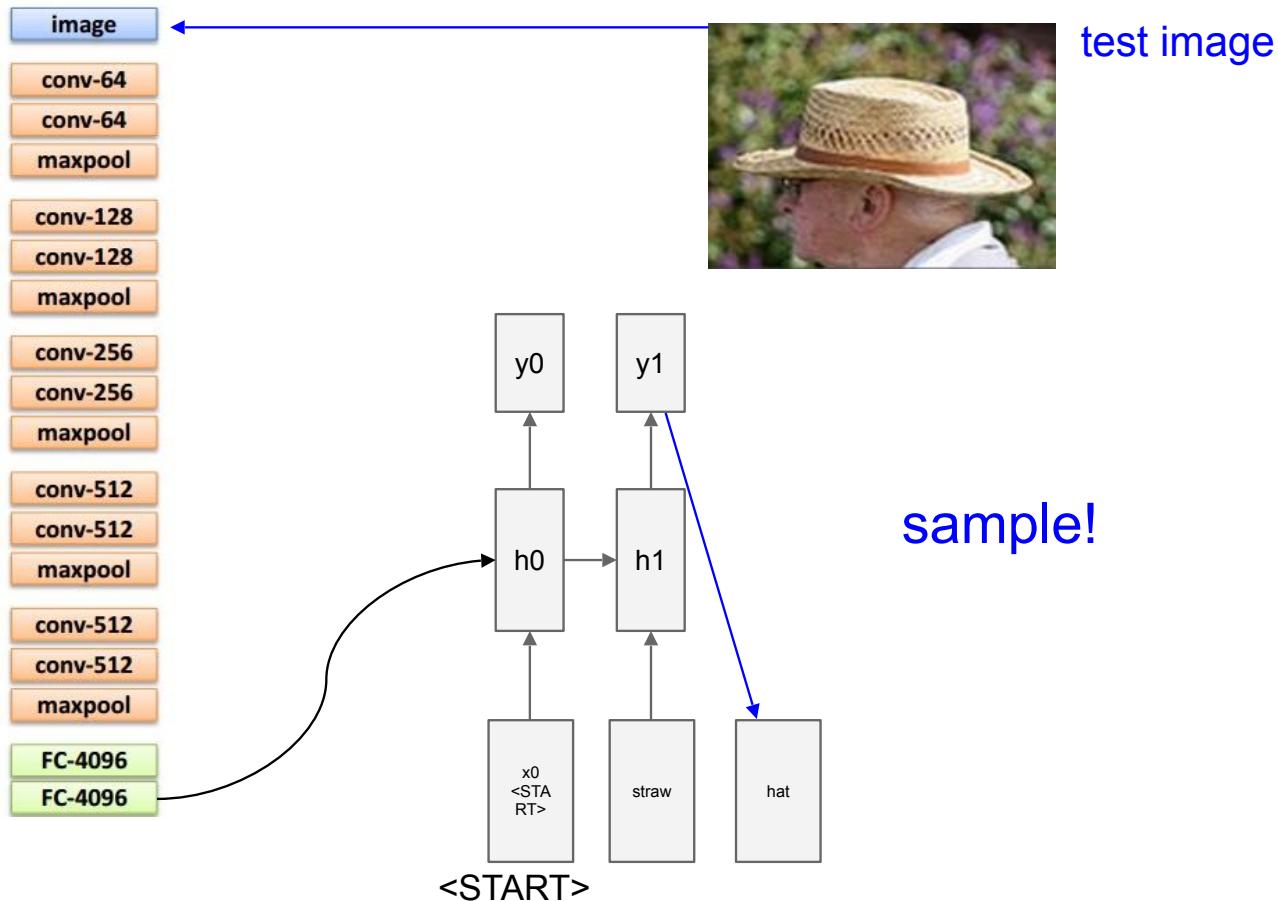
slide credit: Andrej Karpathy



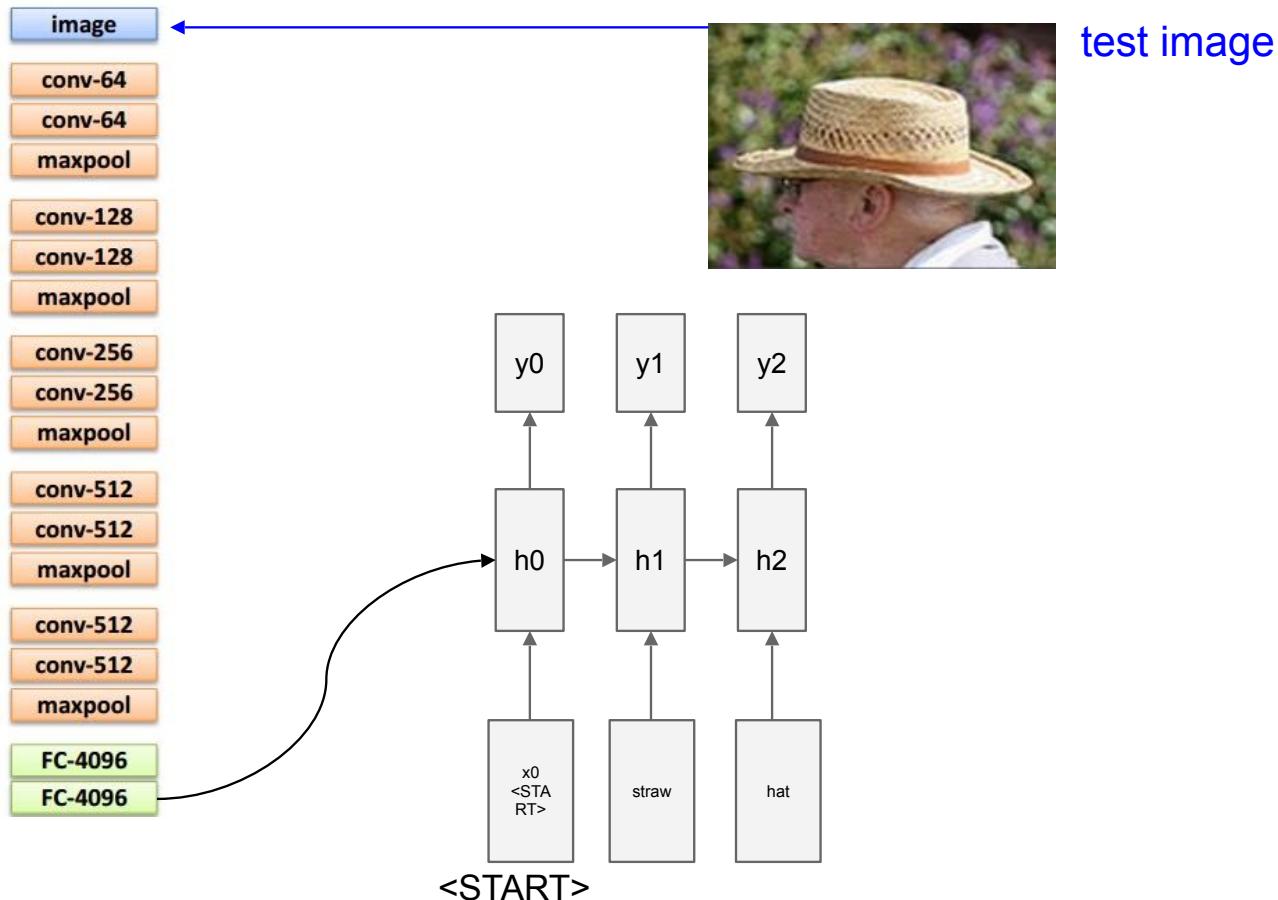
slide credit: Andrej Karpathy



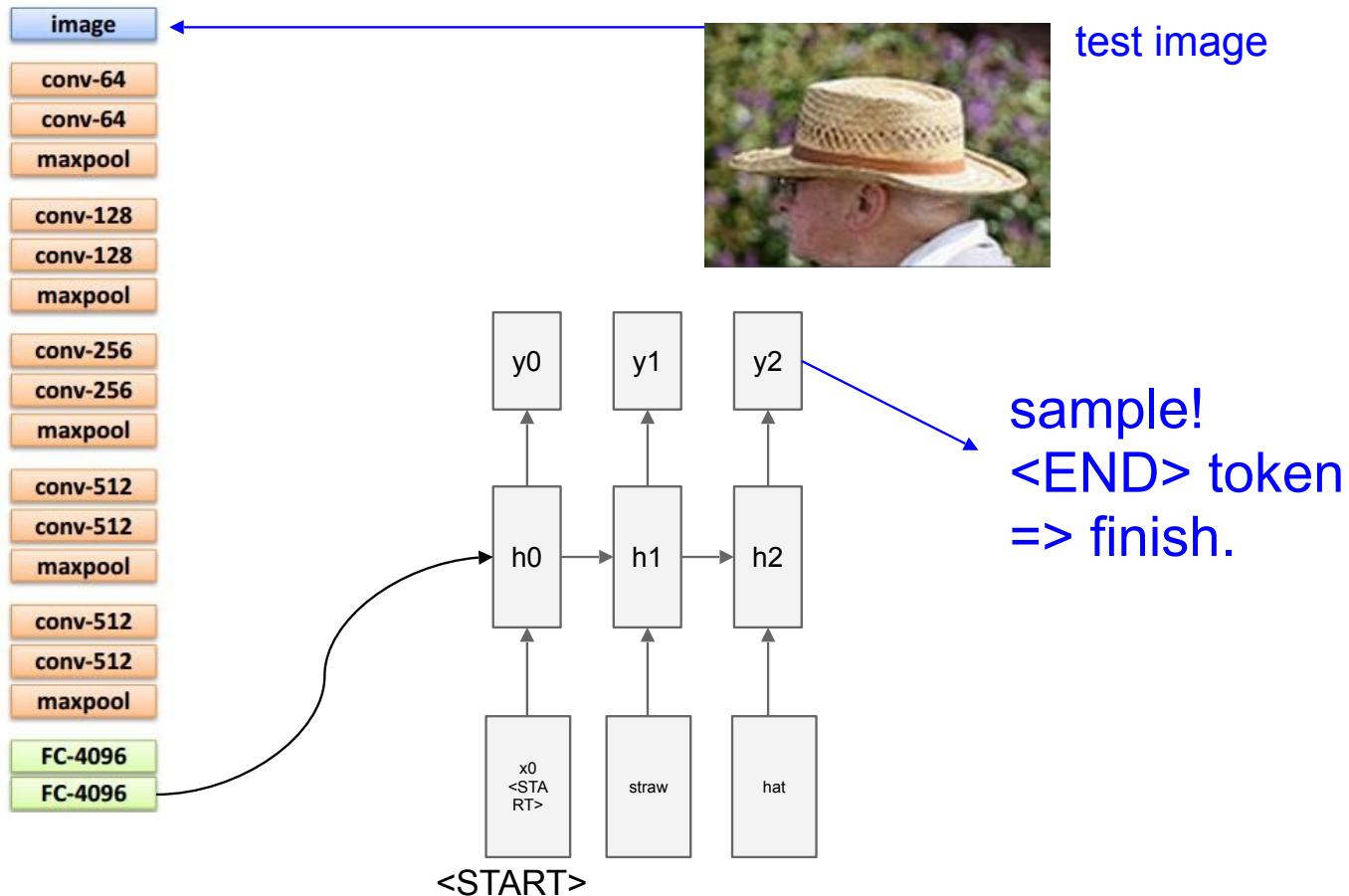
slide credit: Andrej Karpathy



slide credit: Andrej Karpathy



slide credit: Andrej Karpathy



slide credit: Jeff Donahue

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

slide credit: Jeff Donahue

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

slide credit: Jeff Donahue

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

slide credit: Jeff Donahue

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

slide credit: Jeff Donahue

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

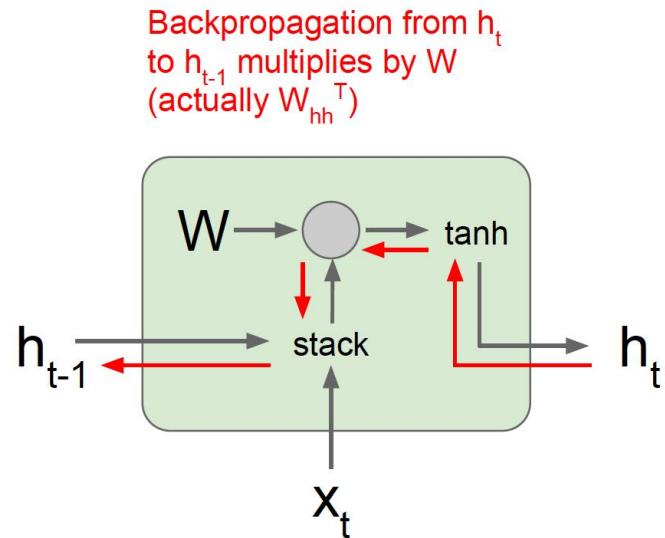


a horse is standing in the middle of a road
logprob: -10.34

Vanilla RNN...

Vanilla RNN Gradient Flow

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



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

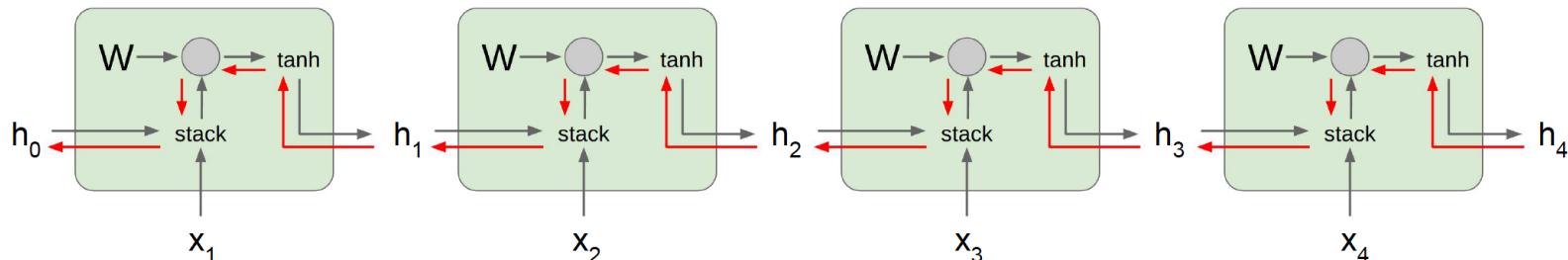
slide credit: Fei-Fei, Justin Johnson, Serena Yeung



Vanilla RNN...

Vanilla RNN Gradient Flow

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



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

Largest singular value > 1 :
Exploding gradients

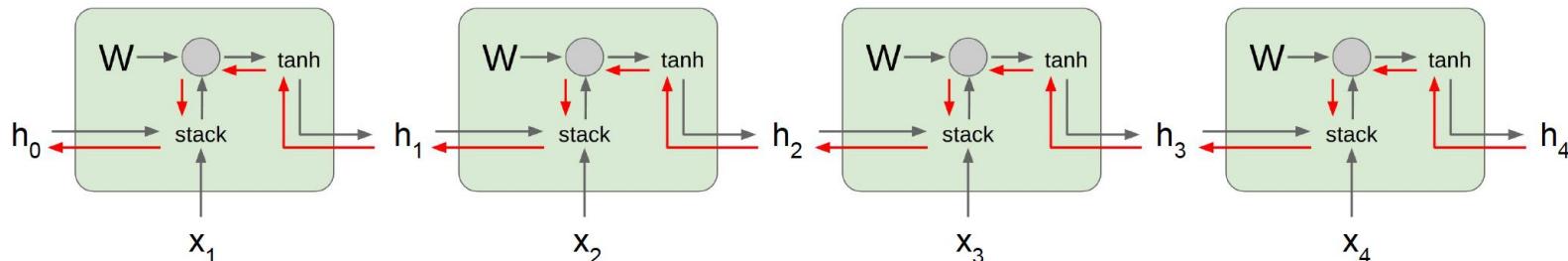
Largest singular value < 1 :
Vanishing gradients

slide credit: Fei-Fei, Justin Johnson, Serena Yeung

Vanilla RNN...

Vanilla RNN Gradient Flow

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



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

Largest singular value > 1 :
Exploding gradients

Largest singular value < 1 :
Vanishing gradients

Gradient clipping: Scale gradient if its norm is too big

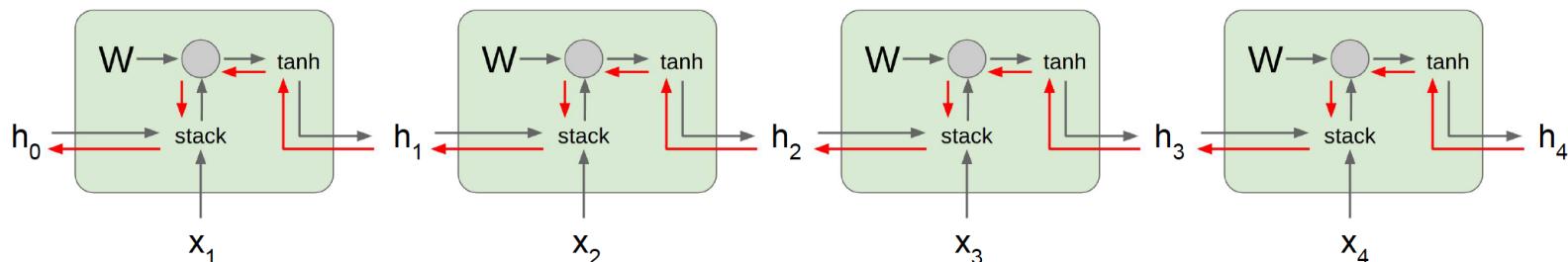
```
grad_norm = np.sum(grad * grad)
if grad_norm > threshold:
    grad *= (threshold / grad_norm)
```

slide credit: Fei-Fei, Justin Johnson, Serena Yeung

Vanilla RNN...

Vanilla RNN Gradient Flow

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



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

Largest singular value > 1 :
Exploding gradients

Largest singular value < 1 :
Vanishing gradients → Change RNN architecture

slide credit: Fei-Fei, Justin Johnson, Serena Yeung

Long Short Term Memory (LSTM)

Vanilla RNN

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

LSTM

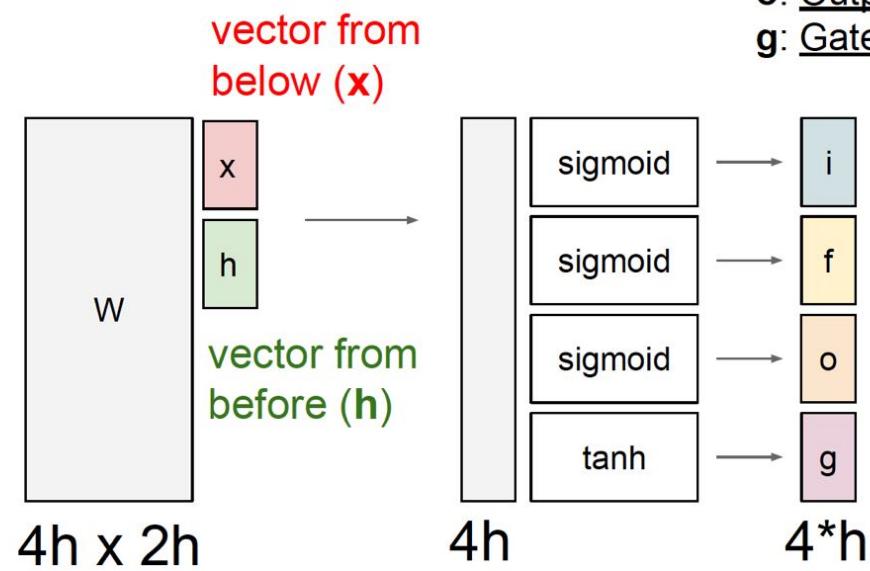
$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}$$
$$c_t = f \odot c_{t-1} + i \odot g$$
$$h_t = o \odot \tanh(c_t)$$

Long Short Term Memory (LSTM)
[Hochreiter et al., 1997]

slide credit: Fei-Fei, Justin Johnson, Serena Yeung

Long Short Term Memory (LSTM)

Long Short Term Memory (LSTM)
[Hochreiter et al., 1997]



- i: Input gate, whether to write to cell
- f: Forget gate, Whether to erase cell
- o: Output gate, How much to reveal cell
- g: Gate gate (?), How much to write to cell

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

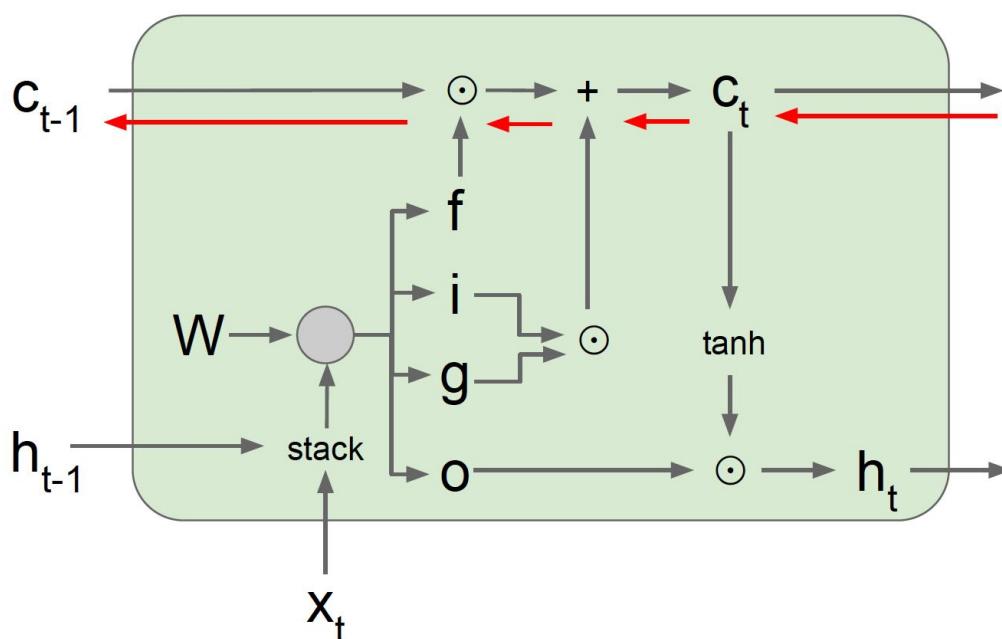
$$c_t = f \odot c_{t-1} + i \odot g$$
$$h_t = o \odot \tanh(c_t)$$

slide credit: Fei-Fei, Justin Johnson, Serena Yeung



Long Short Term Memory (LSTM)

Long Short Term Memory (LSTM)
[Hochreiter et al., 1997]



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

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

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

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

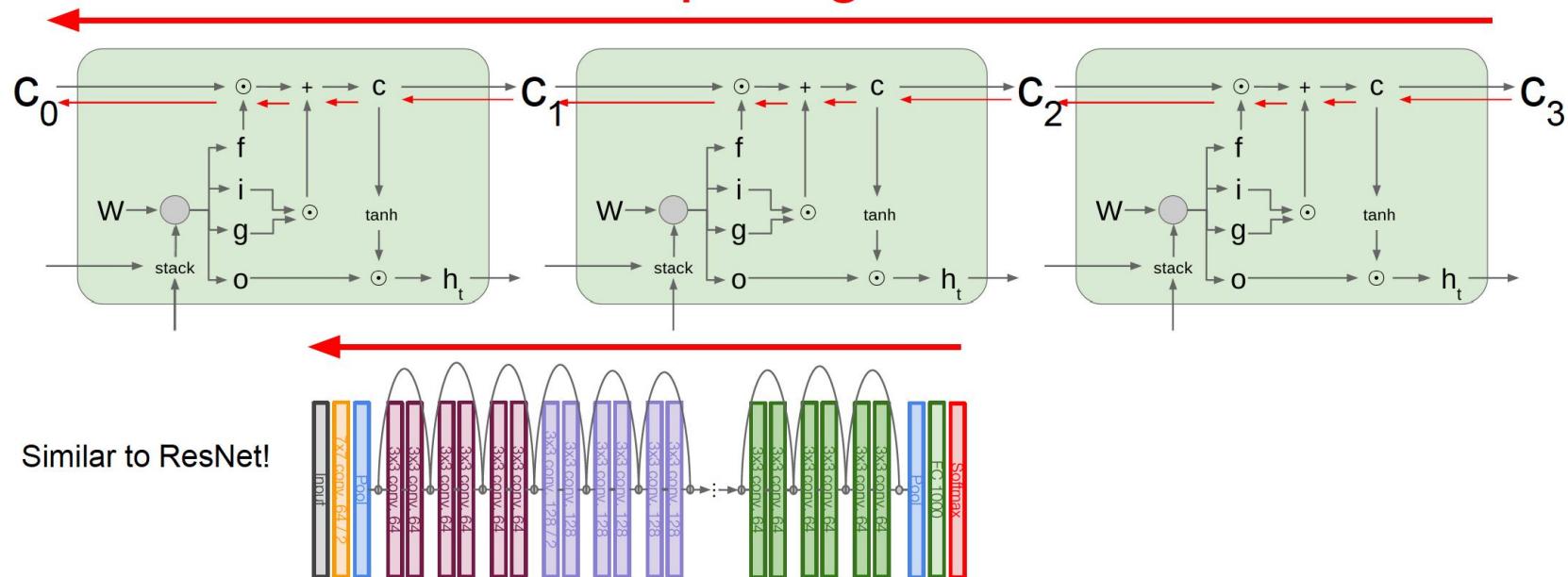
slide credit: Fei-Fei, Justin Johnson, Serena Yeung



Long Short Term Memory (LSTM): Gradient Flow

Long Short Term Memory (LSTM)
[Hochreiter et al., 1997]

Uninterrupted gradient flow!



slide credit: Fei-Fei, Justin Johnson, Serena Yeung



Alternatives

Other RNN Variants

GRU [*Learning phrase representations using rnn encoder-decoder for statistical machine translation, Cho et al. 2014*]

$$\begin{aligned} r_t &= \sigma(W_{xr}x_t + W_{hr}h_{t-1} + b_r) \\ z_t &= \sigma(W_{xz}x_t + W_{hz}h_{t-1} + b_z) \\ \tilde{h}_t &= \tanh(W_{xh}x_t + W_{hh}(r_t \odot h_{t-1}) + b_h) \\ h_t &= z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t \end{aligned}$$

[*LSTM: A Search Space Odyssey, Greff et al., 2015*]

[*An Empirical Exploration of Recurrent Network Architectures, Jozefowicz et al., 2015*]

MUT1:

$$\begin{aligned} z &= \text{sigm}(W_{xz}x_t + b_z) \\ r &= \text{sigm}(W_{xr}x_t + W_{hr}h_t + b_r) \\ h_{t+1} &= \tanh(W_{hh}(r \odot h_t) + \tanh(x_t) + b_h) \odot z \\ &\quad + h_t \odot (1 - z) \end{aligned}$$

MUT2:

$$\begin{aligned} z &= \text{sigm}(W_{xz}x_t + W_{hz}h_t + b_z) \\ r &= \text{sigm}(x_t + W_{hr}h_t + b_r) \\ h_{t+1} &= \tanh(W_{hh}(r \odot h_t) + W_{xh}x_t + b_h) \odot z \\ &\quad + h_t \odot (1 - z) \end{aligned}$$

MUT3:

$$\begin{aligned} z &= \text{sigm}(W_{xz}x_t + W_{hz}\tanh(h_t) + b_z) \\ r &= \text{sigm}(W_{xr}x_t + W_{hr}h_t + b_r) \\ h_{t+1} &= \tanh(W_{hh}(r \odot h_t) + W_{xh}x_t + b_h) \odot z \\ &\quad + h_t \odot (1 - z) \end{aligned}$$

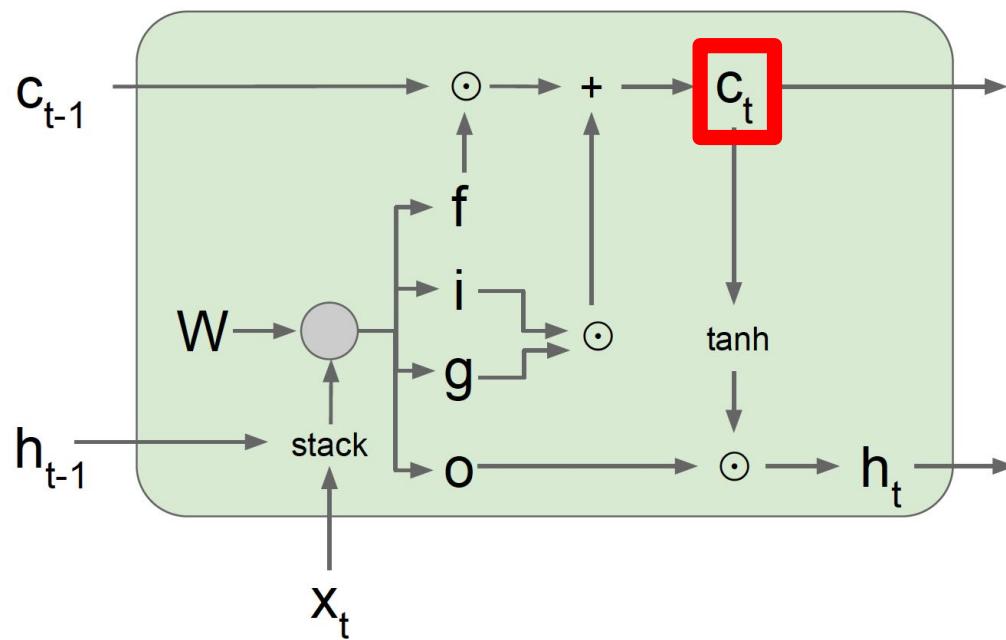
Summary

- RNNs allow a lot of flexibility in architecture design
- Vanilla RNNs are simple but don't work very well
- Common to use LSTM or GRU: their additive interactions improve gradient flow
- Backward flow of gradients in RNN can explode or vanish. Exploding is controlled with gradient clipping. Vanishing is controlled with additive interactions (LSTM)
- Better/simpler architectures are a hot topic of current research
- Better understanding (both theoretical and empirical) is needed.

slide credit: Fei-Fei, Justin Johnson, Serena Yeung



Hunting interpretable cells



$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}$$
$$c_t = f \odot c_{t-1} + i \odot g$$
$$h_t = o \odot \tanh(c_t)$$

Visualizing and Understanding Recurrent Networks
Andrej Karpathy*, Justin Johnson*, Li Fei-Fei (on [arXiv.org](https://arxiv.org/))

Hunting interpretable cells

```
/* Unpack a filter field's string representation from user-space
 * buffer. */
char *audit_unpack_string(void **bufp, size_t *remain, size_t len)
{
    char *str;
    if (!*bufp || (len == 0) || (len > *remain))
        return ERR_PTR(-EINVAL);
    /* Of the currently implemented string fields, PATH_MAX
     * defines the longest valid length.
    */
```

Hunting interpretable cells

```
"You mean to imply that I have nothing to eat out of.... On the  
contrary, I can supply you with everything even if you want to give  
dinner parties," warmly replied Chichagov, who tried by every word he  
spoke to prove his own rectitude and therefore imagined Kutuzov to be  
animated by the same desire.  
  
Kutuzov, shrugging his shoulders, replied with his subtle penetrating  
smile: "I meant merely to say what I said."
```

quote detection cell

Hunting interpretable cells

Cell sensitive to position in line:

```
The sole importance of the crossing of the Berezina lies in the fact  
that it plainly and indubitably proved the fallacy of all the plans for  
cutting off the enemy's retreat and the soundness of the only possible  
line of action--the one Kutuzov and the general mass of the army  
demanded--namely, simply to follow the enemy up. The French crowd fled  
at a continually increasing speed and all its energy was directed to  
reaching its goal. It fled like a wounded animal and it was impossible  
to block its path. This was shown not so much by the arrangements it  
made for crossing as by what took place at the bridges. When the bridges  
broke down, unarmed soldiers, people from Moscow and women with children  
who were with the French transport, all--carried on by vis inertiae--  
pressed forward into boats and into the ice-covered water and did not,  
surrender.
```

line length tracking cell

Hunting interpretable cells

```
static int __dequeue_signal(struct sigpending *pending, sigset_t *mask,
                           siginfo_t *info)
{
    int sig = next_signal(pending, mask);
    if (sig) {
        if (current->notifier) {
            if (sigismember(current->notifier_mask, sig)) {
                if (!!(current->notifier)(current->notifier_data)) {
                    clear_thread_flag(TIF_SIGPENDING);
                    return 0;
                }
            }
            collect_signal(sig, pending, info);
        }
    }
    return sig;
}
```

if statement cell

Hunting interpretable cells

```
/* Duplicate LSM field information.  The lsm_rule is opaque, so
 * re-initialized. */
static inline int audit_dupe_lsm_field(struct audit_field *df,
                                       struct audit_field *sf)
{
    int ret = 0;
    char *lsm_str;
    /* our own copy of lsm_str */
    lsm_str = kstrdup(sf->lsm_str, GFP_KERNEL);
    if (unlikely(!lsm_str))
        return -ENOMEM;
    df->lsm_str = lsm_str;
    /* our own (refreshed) copy of lsm_rule */
    ret = security_audit_rule_init(df->type, df->op, df->lsm_str,
                                    (void **) &df->lsm_rule);
    /* Keep currently invalid fields around in case they
     * become valid after a policy reload. */
    if (ret == -EINVAL) {
        pr_warn("audit rule for LSM \\'%s\\' is invalid\n",
                df->lsm_str);
        ret = 0;
    }
    return ret;
}
```

quote/comment cell

Hunting interpretable cells

```
#ifdef CONFIG_AUDITSYSCALL
static inline int audit_match_class_bits(int class, u32 *mask)
{
    int i;
    if (classes[class]) {
        for (i = 0; i < AUDIT_BITMASK_SIZE; i++)
            if (mask[i] & classes[class][i])
                return 0;
    }
    return 1;
}
```

code depth cell

Hunting interpretable cells

```
char *audit_unpack_string(void **bufp, size_t *remain, si
{
    char *str;
    if (!*bufp || (len == 0) || (len > *remain))
        return ERR_PTR(-EINVAL);
    /* of the currently implemented string fields, PATH_MAX
     * defines the longest valid length.
     */
    if (len > PATH_MAX)
        return ERR_PTR(-ENAMETOOLONG);
    str = kmalloc(len + 1, GFP_KERNEL);
    if (unlikely(!str))
        return ERR_PTR(-ENOMEM);
    memcpy(str, *bufp, len);
    str[len] = 0;
    *bufp += len;
    *remain -= len;
    return str;
}
```

something interesting cell
(not quite sure what)