The lecture will be recorded

https://www.video.ethz.ch/lectures/d-infk/2020/spring/263-3710-00L

user: hil-20s
pw: 3Ewk6Fv

Sequence Modelling

Recurrent Neural Networks

Machine Perception

Otmar Hilliges

26 March 2020

[Some slides adopted from Karpathy, De Freitas, Goodfellow]

Advanced Interactive Technologies

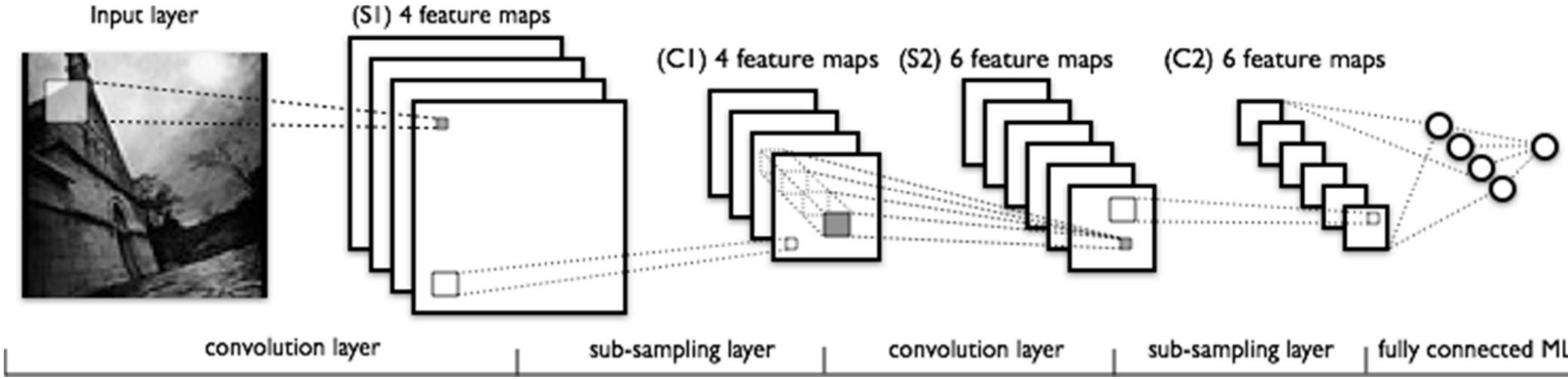
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

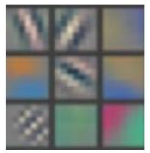
Last Week

CNNs – continued

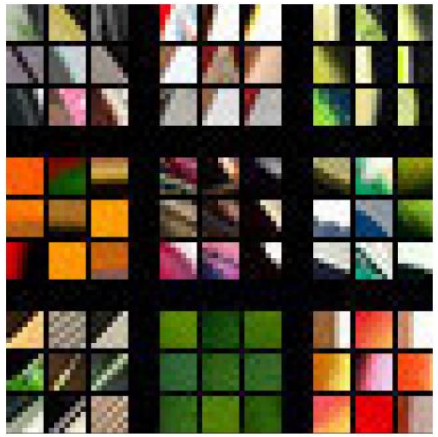
Fully convolutional networks

Some advanced vision tasks (<- we didn’t get to this, more next week with Siyu Tang)

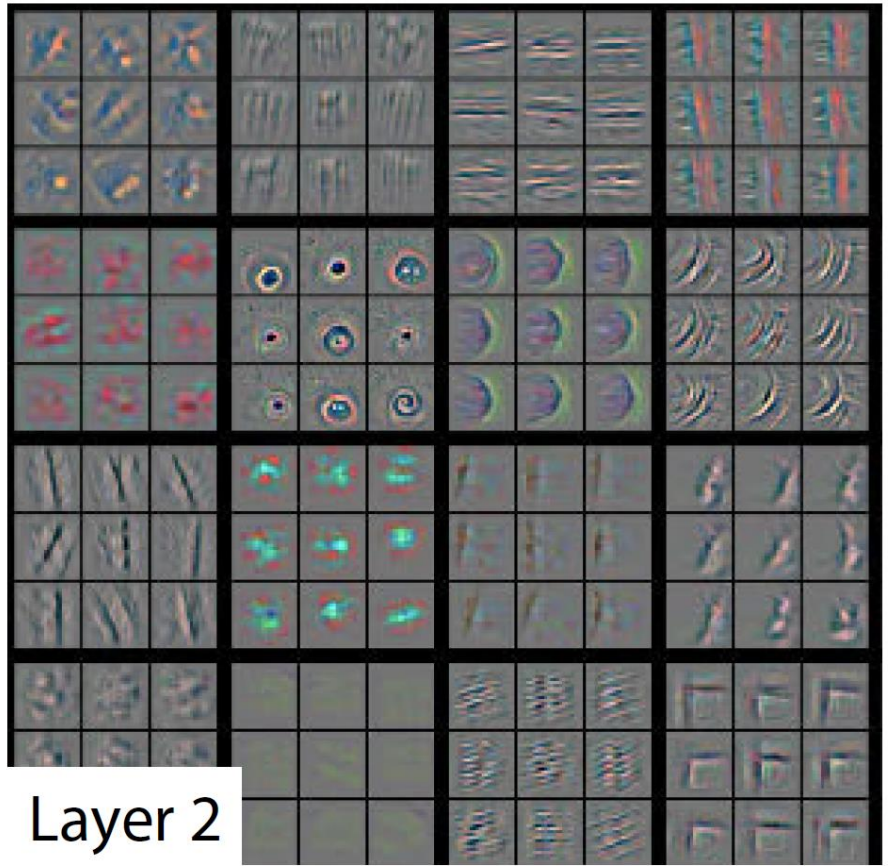




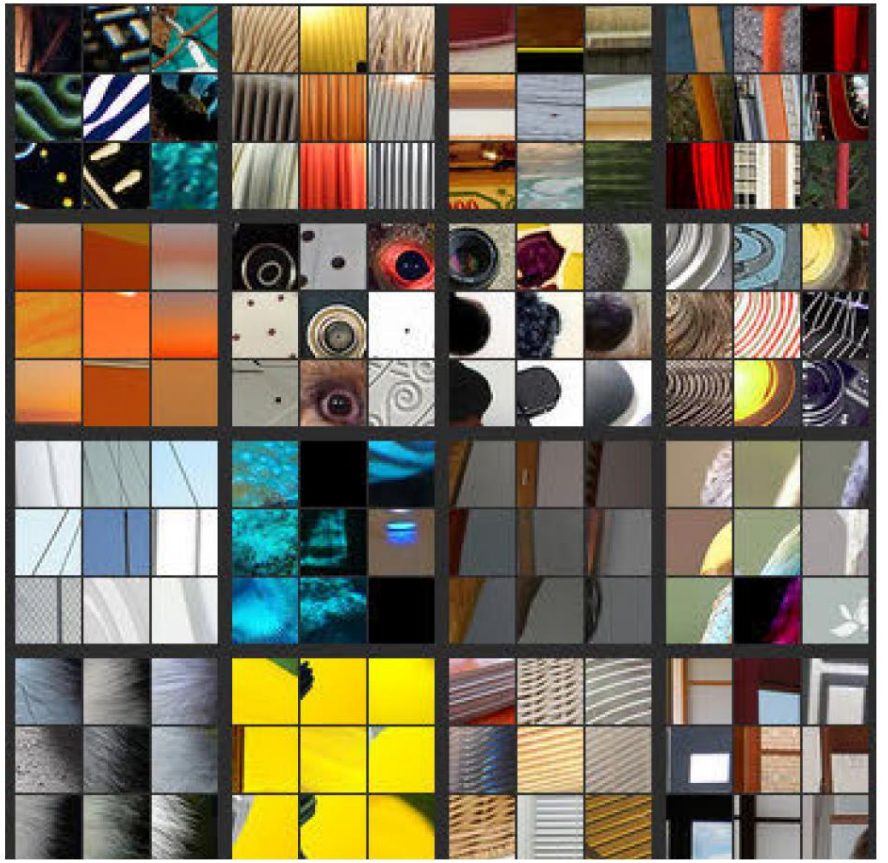
Layer 1

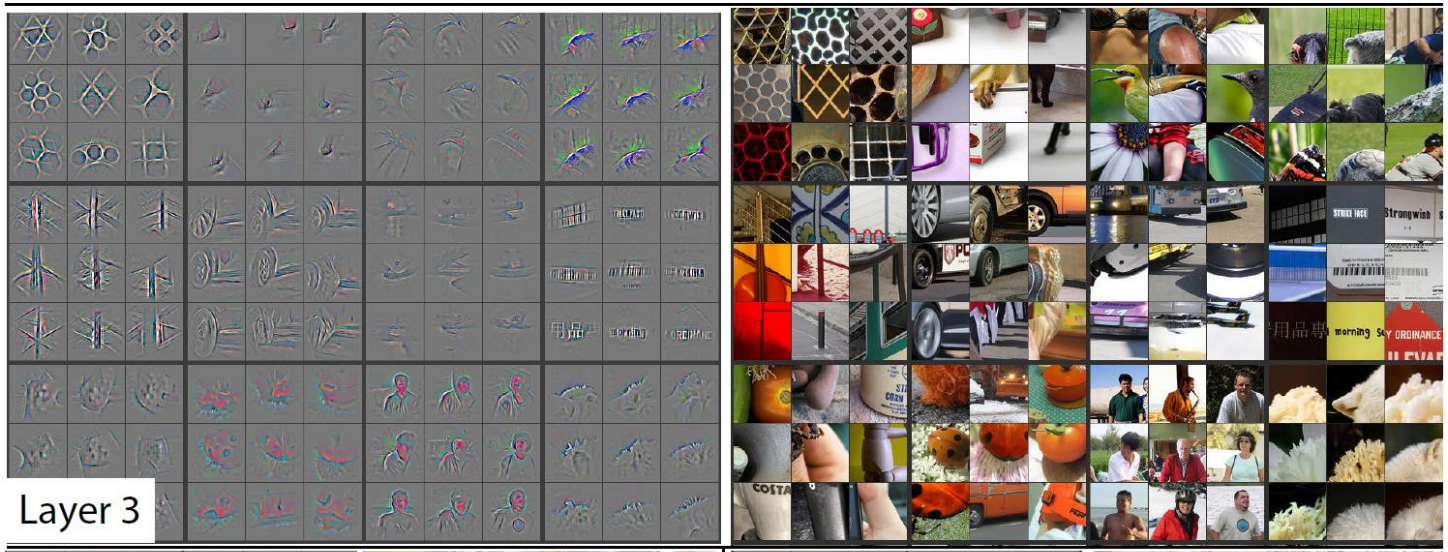


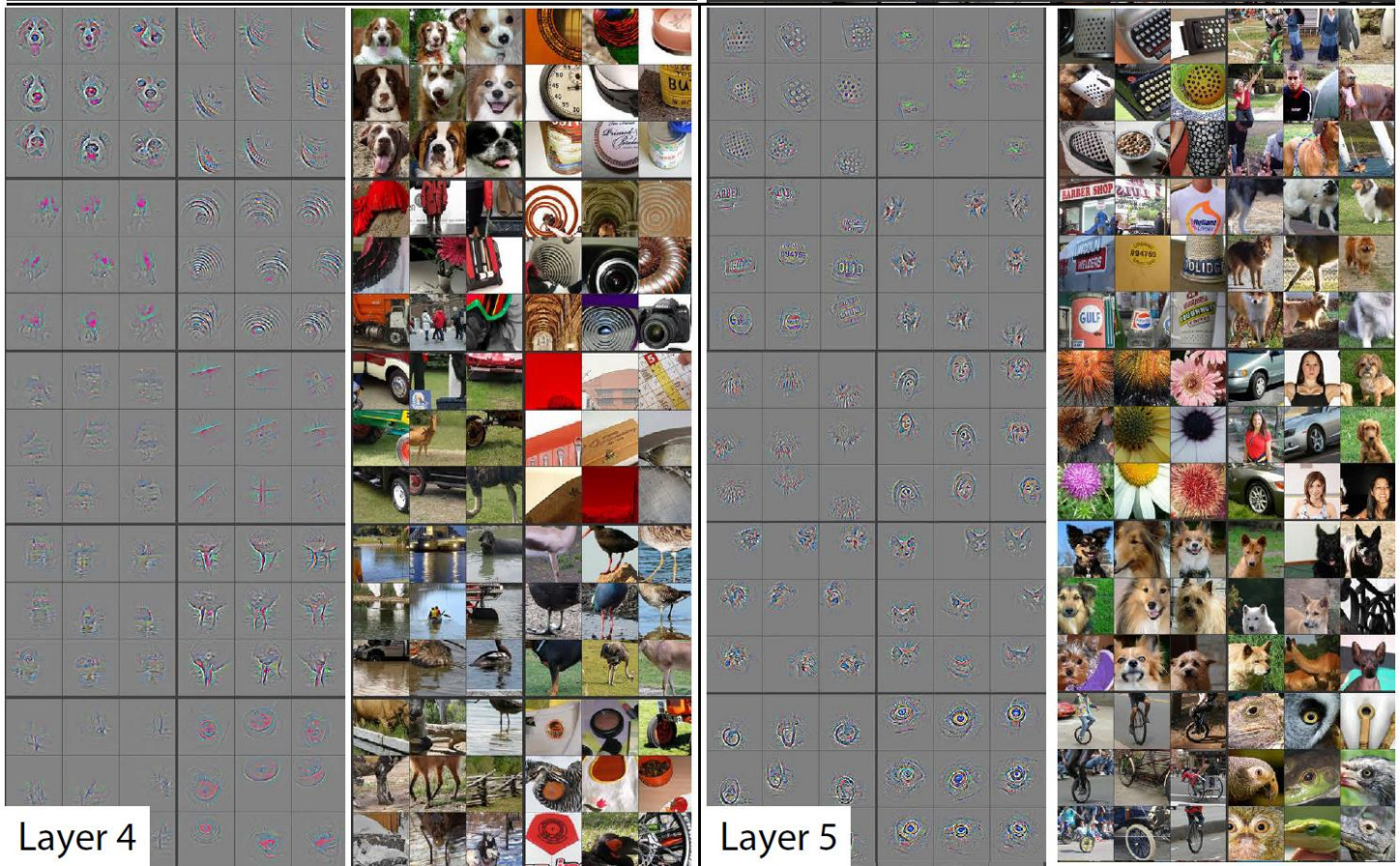
[Zeiler & Fergus]



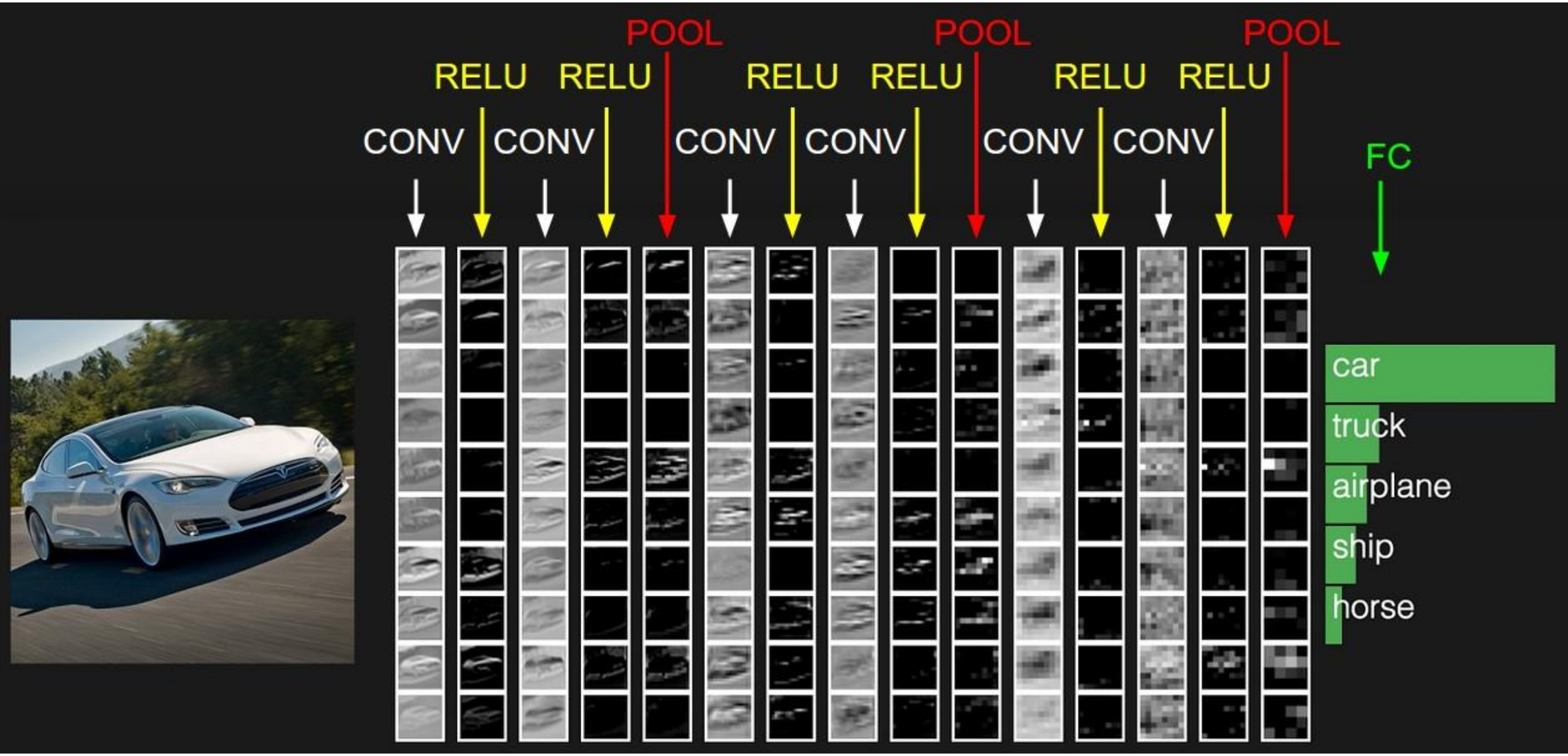
Layer 2







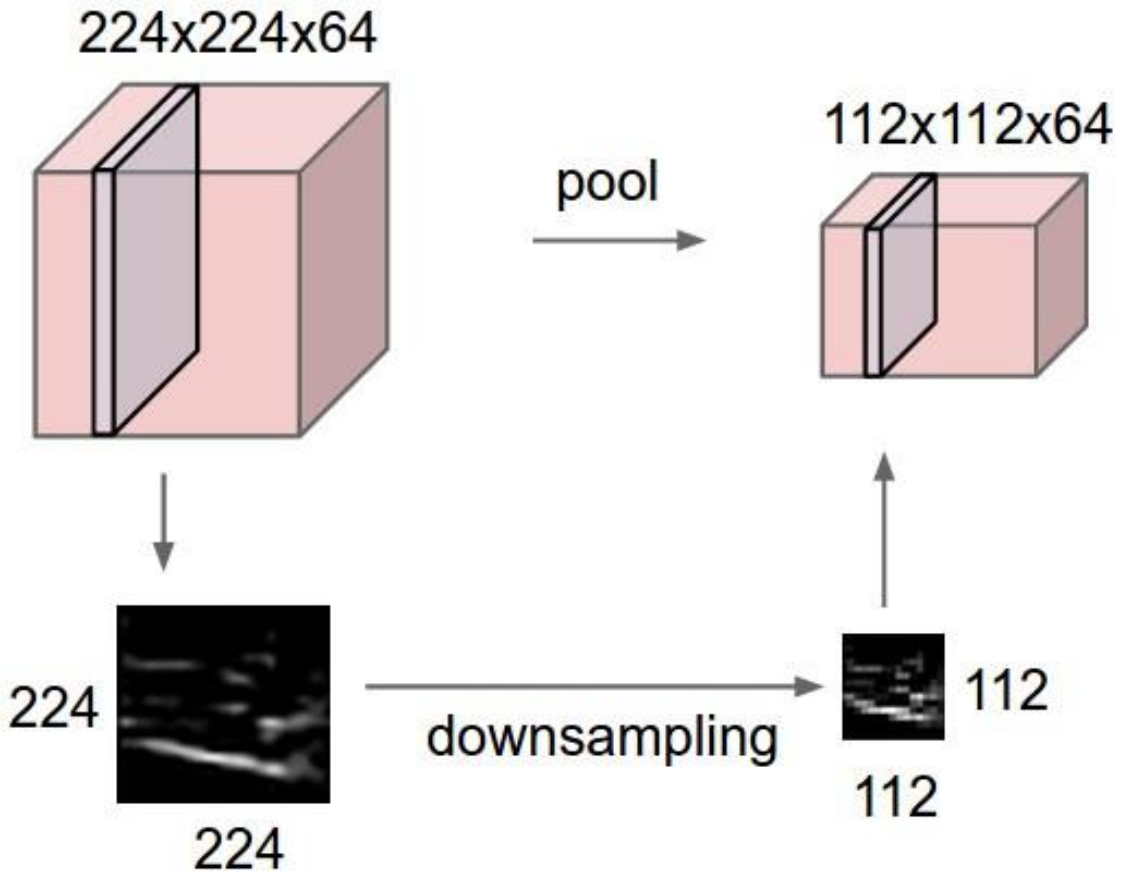
[Zeiler & Fergus]



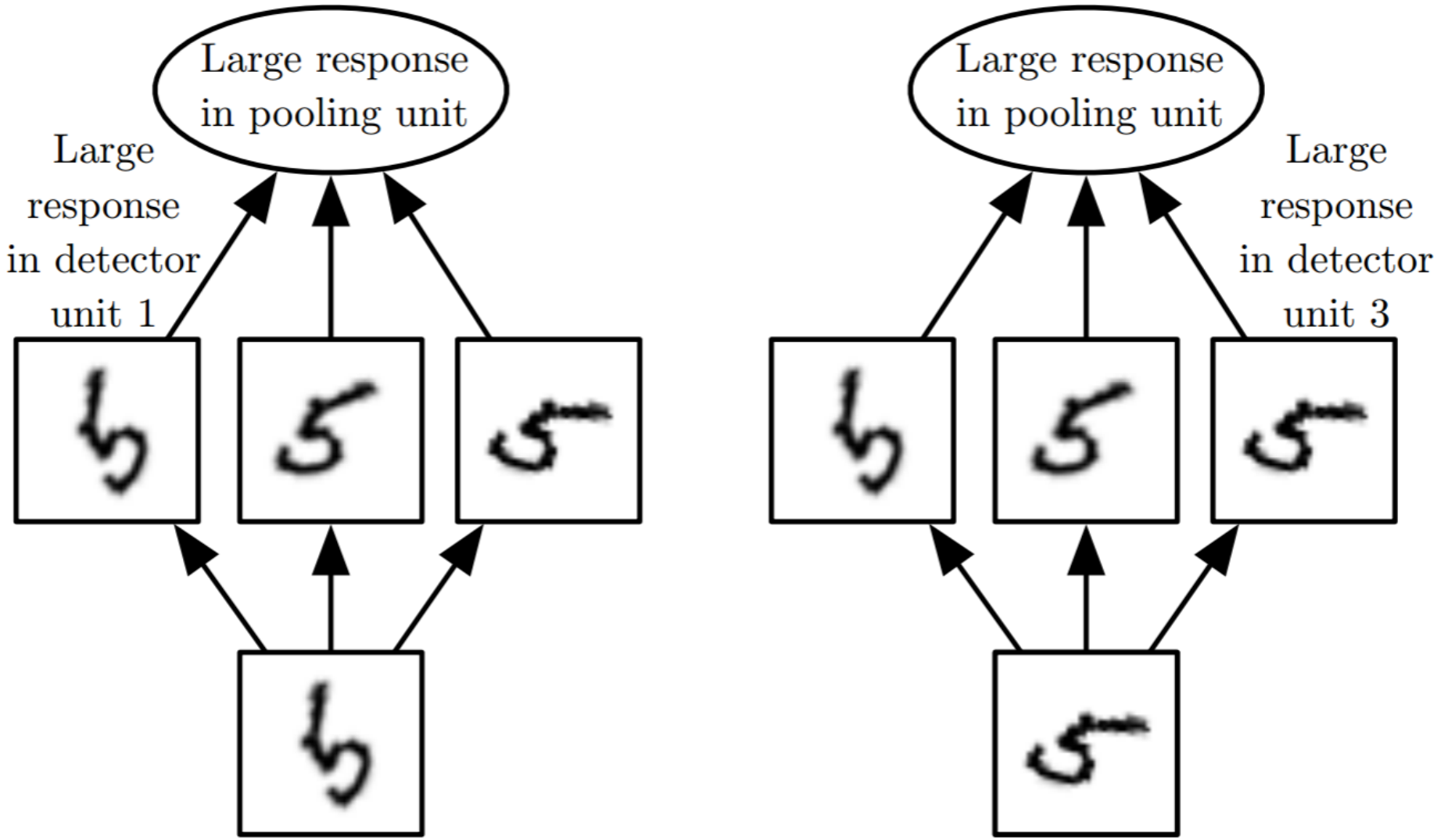
[Slides adopted from: Li, Karpathy]

Pooling layer

- makes the representations smaller and more manageable
- operates over each activation map independently:



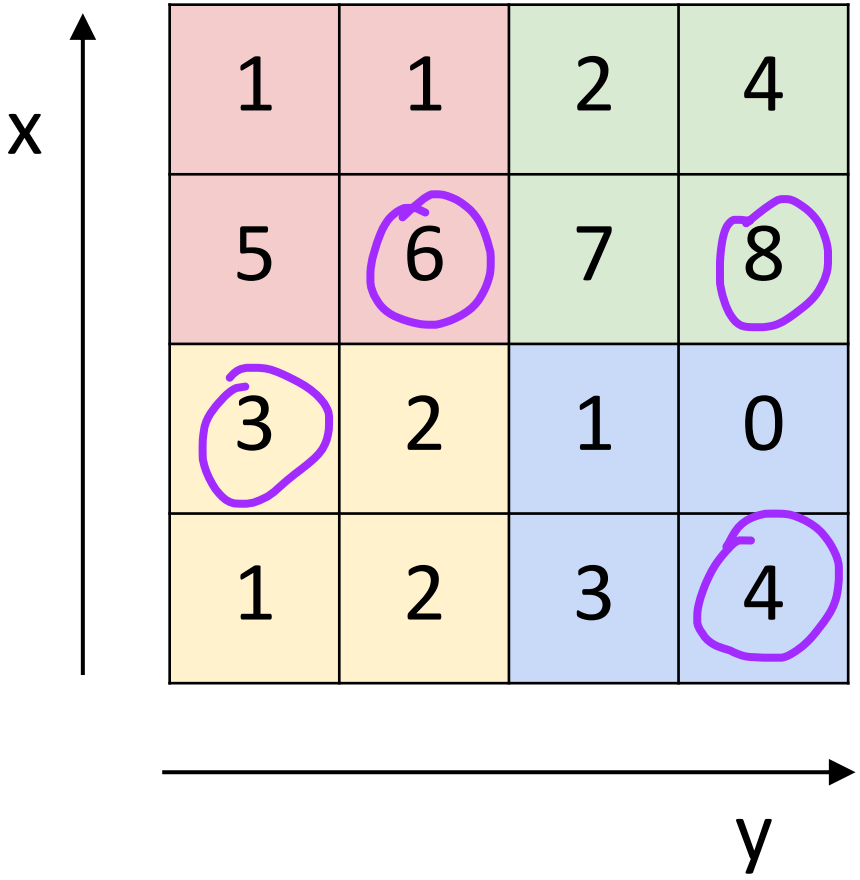
[Slides adopted from: Li, Karpathy]



[image source: deeplearningbook.org]

MAX POOLING

Single depth slice



max pool with 2x2 filters
and stride 2

6	8
3	4

[Slides adopted from: Li, Karpathy]

Max pooling layer

Foiwasd pass:

$$z^{(l)} = \max\{z_k^{(l)}\}$$

Backwasd pass:

$$\frac{\partial z^{(l)}}{\partial z^{(l-1)}} = \begin{cases} 1, & \text{if } k = \operatorname{argmax}\{z_k^{(l-1)}\} \\ 0, & \text{otherwise} \end{cases}$$

$$\delta^{(l-1)} = \{\delta^{(l)}\}_{k^*}, \text{ where } k^* = \operatorname{argmax}\{z_k^{(l-1)}\}$$

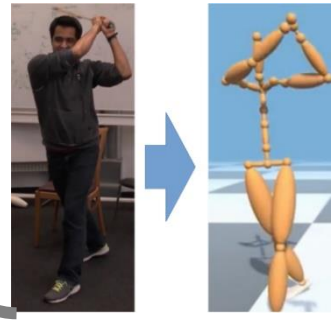
This week

Sequence modelling with Neural Networks

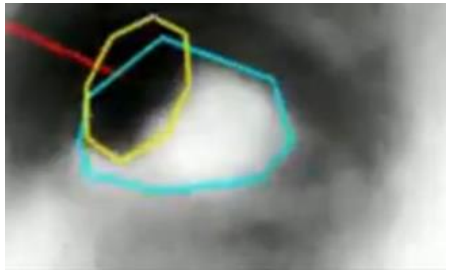
- Recurrent Neural Networks
- The vanishing and exploding gradients problem
- LSTM (augmentation of RNN)

Some Advanced Tasks

Body Pose Estimation

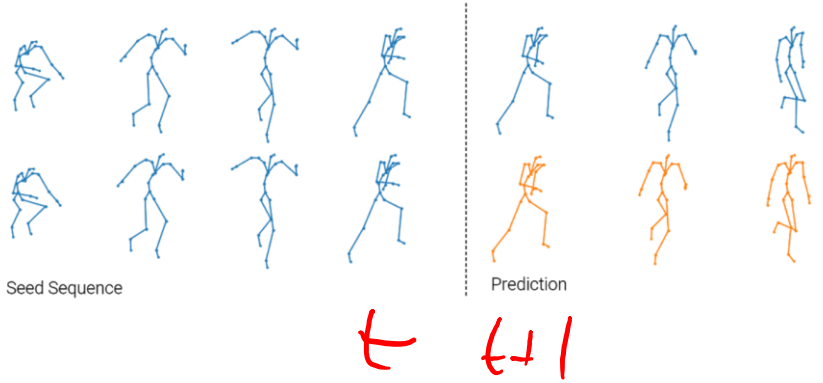


Eye Gaze Estimation



Dynamic Gesture Recognition

Human Motion Prediction



Dynamical system

$$s^t = f(s^{t-1}; \theta)$$

$$s^{(\dots)} \xrightarrow{f} s^{t-1} \xrightarrow{f} s^t \xrightarrow{f} s^{t+1} \rightarrow \dots$$

f propagates state s through time

For $\pi = \{1, 2, 3\}$ (finite horizon)

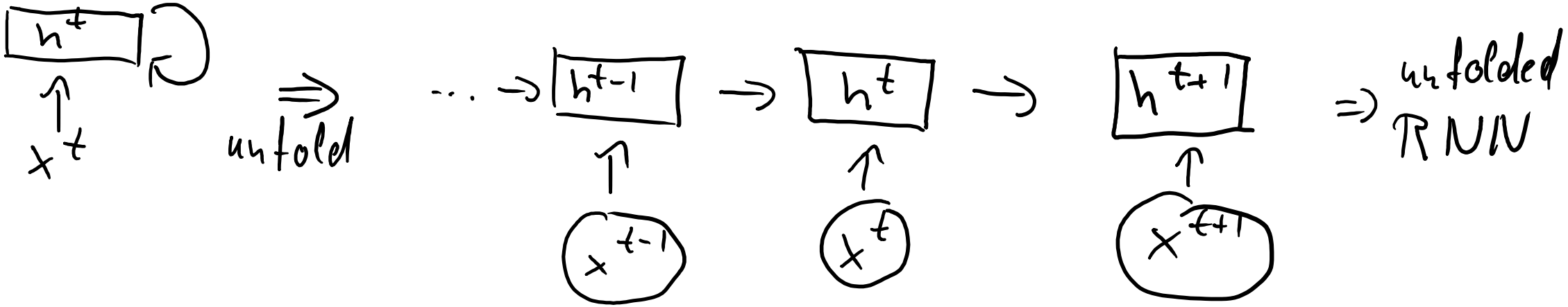
$$s^3 = f(\underbrace{f(s^1; \theta)}_{s^2}; \theta) \Rightarrow \text{unrolled recurrence}$$

Dynamical system with inputs

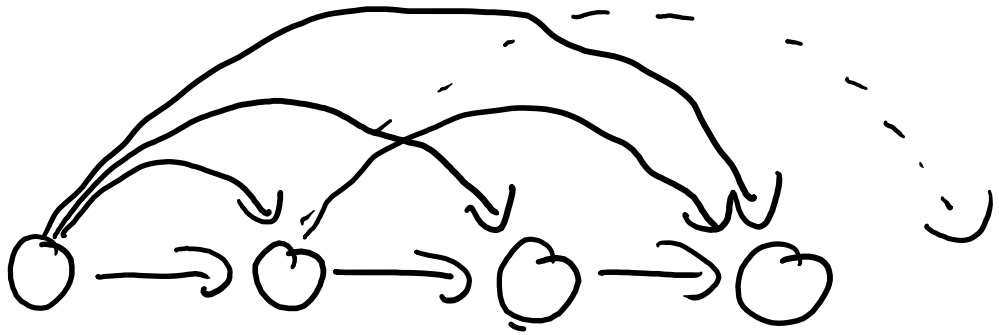
$$s^t = f(s^{t-1}, x^t; \theta)$$

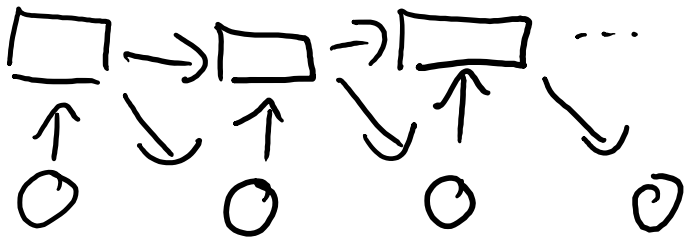
Rename state s^t to 'hidden' state h^t:

$$h^t = f(h^{t-1}, x^t; \theta)$$



Two ways to represent dyn. sys:





$s^t = g^t(x^1 \dots x^{t-1})$

$h^t = f(h^{t-1}, x^t; \theta)$

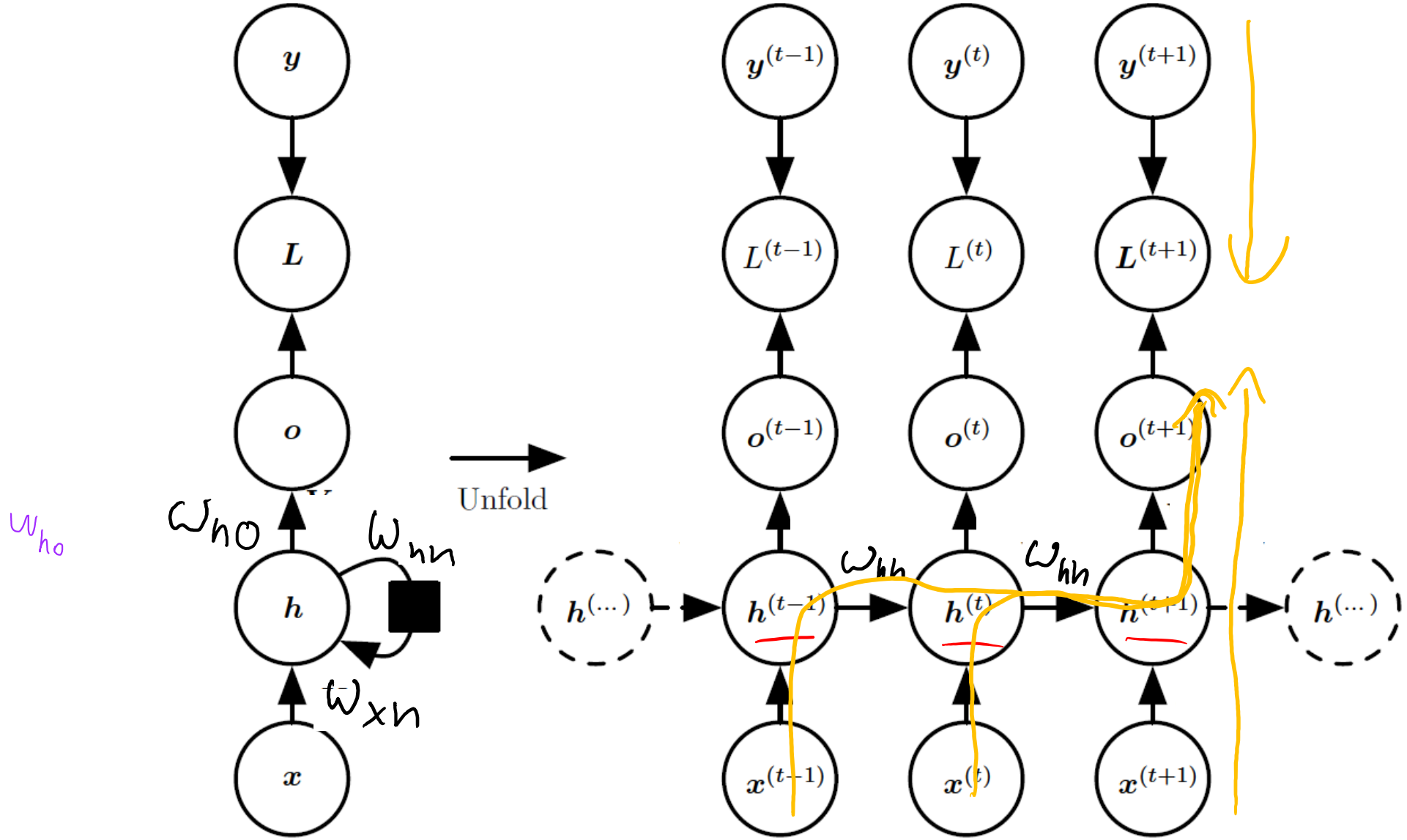
1 function for each t

variable sequence length

same transition function

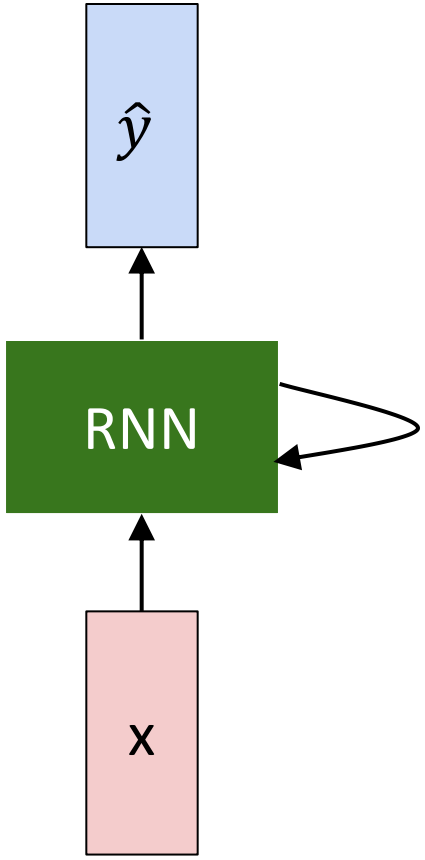
same params @ $\forall_t$

RNNs – a more complete example



(Vanilla) Recurrent Neural Network

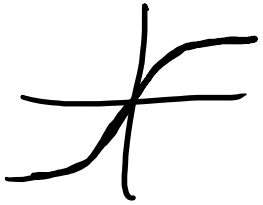
The state consists of a single *"hidden"* vector **h**:



sign

$$\hat{y}^t = W_{hy} h^t$$

tanh



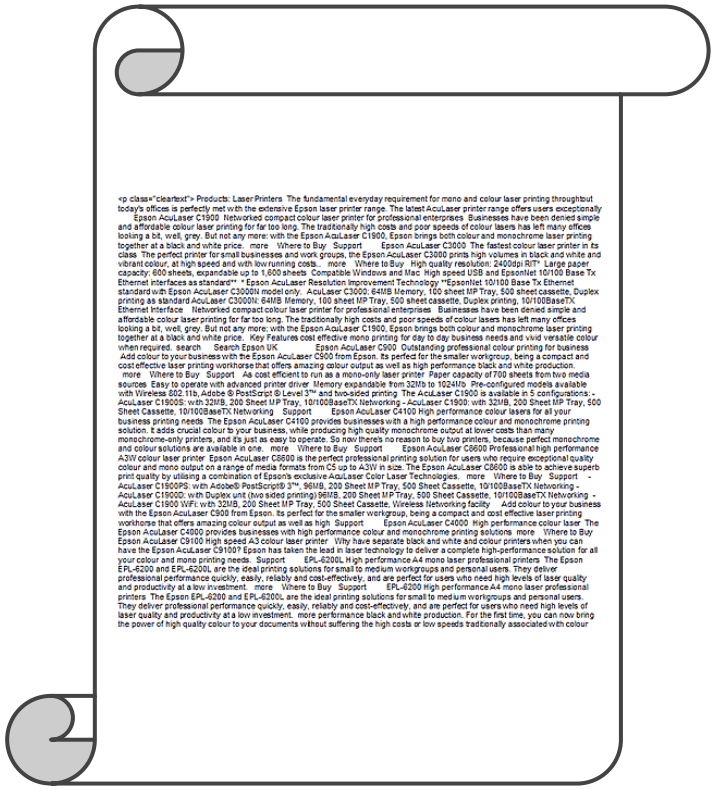
$$h^t = tanh(W_{hh} h^{t-1} + W_{xh} x^t)$$

$$h^t = f(h^{t-1}, x^t, W)$$

[Slides adopted from: Li, Karpathy]

The unreasonable effectiveness of RNNs

http://karpathy.github.io/2015/05/21/rnn-effectiveness/



min-char-rnn.py gist: 112 lines of Python

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

at first:

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

train more

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

train more

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

train more

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


[Slides adopted from: Li, Karpathy]

For $\bigoplus_{n=1,\ldots,m}$ where $\mathcal{L}_{m_\bullet} = 0$, 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 $U \to T$ is a separated algebraic space.

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

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

and the compariciely in the fibre product covering we have to prove the lemma generated by $\coprod Z \times_U U \to V$. Consider the maps M along the set of points Sch_{fppf} and $U \to U$ 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 $Sh(G)$ such that $\mathrm{Spec}(R') \to S$ 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 $\mathcal{O}_{X,x'} \to \mathcal{O}'_{X',x'}$ is separated. By Algebra, Lemma ?? we can define a map of complexes $\mathrm{GL}_{S'}(x'/S'')$ and we win. □

To prove study we see that $\mathcal{F}|_U$ is a covering of $\mathcal{X}'$, 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

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

is a unique morphism of algebraic stacks. Note that

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

and

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

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. □

The result for prove any open covering follows from the less of Example ??. It may replace S by $X_{spaces,étale}$ 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* $\underline{Proj}_X(\mathcal{A}) = \mathrm{Spec}(B)$ *over* U *compatible with the complex*

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

When in this case of to show that $\mathcal{Q} \to \mathcal{C}_{Z/X}$ *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* $S = \mathrm{Spec}(R)$ *and* $Y = \mathrm{Spec}(R)$.

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

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

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* $i_0 \leq \mathfrak{p}$ *is a subset of* $\mathcal{J}_{n,0} \circ \overline{A}_2$ *works.*

Lemma 0.3. *In Situation ??. Hence we may assume* $\mathfrak{q}' = 0$.

Proof. We will use the property we see that $\mathfrak{p}$ is the mext 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. □

[Slides adopted from: Li, Karpathy]

Proof. Omitted. □

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 $X_{\acute{e}tale}$ we have

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

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

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

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

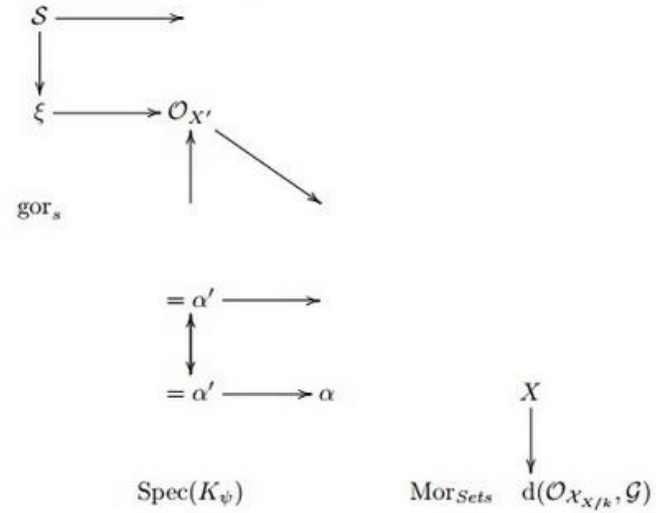
be a morphism of algebraic spaces over S and Y.

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. □

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



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 $X = \mathrm{Spec}(R)$ 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. □

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

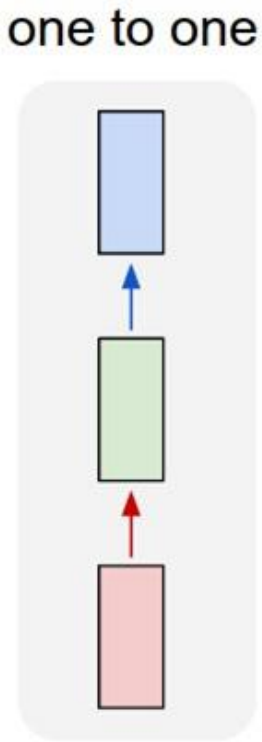
$$\mathcal{O}_{X,x} \longrightarrow \mathcal{F}_{\overline{x}} \quad \text{-1}(\mathcal{O}_{X_{\acute{e}tale}}) \longrightarrow \mathcal{O}_{X_\ell}^{-1}\mathcal{O}_{X_\lambda}(\mathcal{O}_{X_\eta}^{\overline{v}})$$

is an isomorphism of covering of $\mathcal{O}_{X_i}$. 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. □

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

[Slides adopted from: Li, Karpathy]

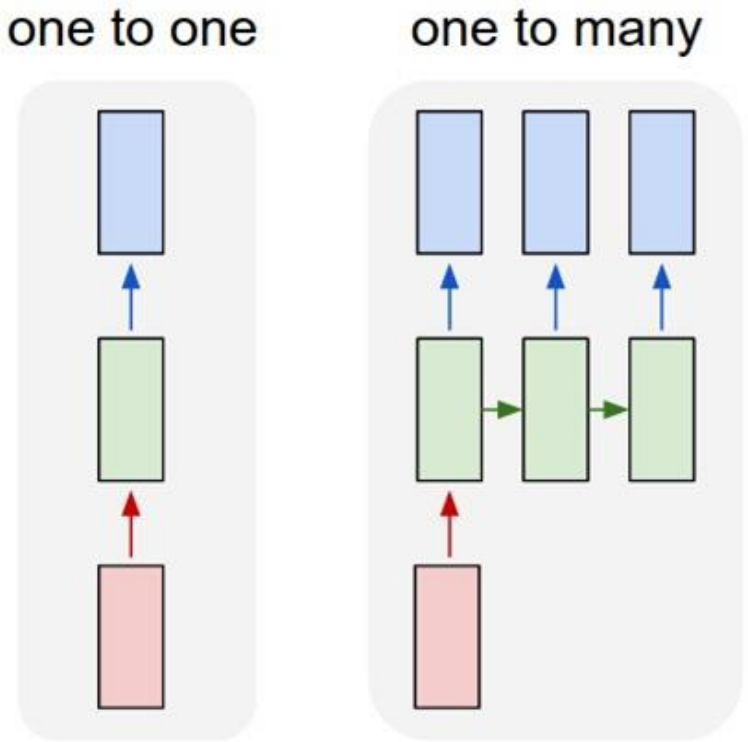
Recurrent Networks offer a lot of flexibility:



Vanilla Neural Networks

[Slides adopted from: Li, Karpathy]

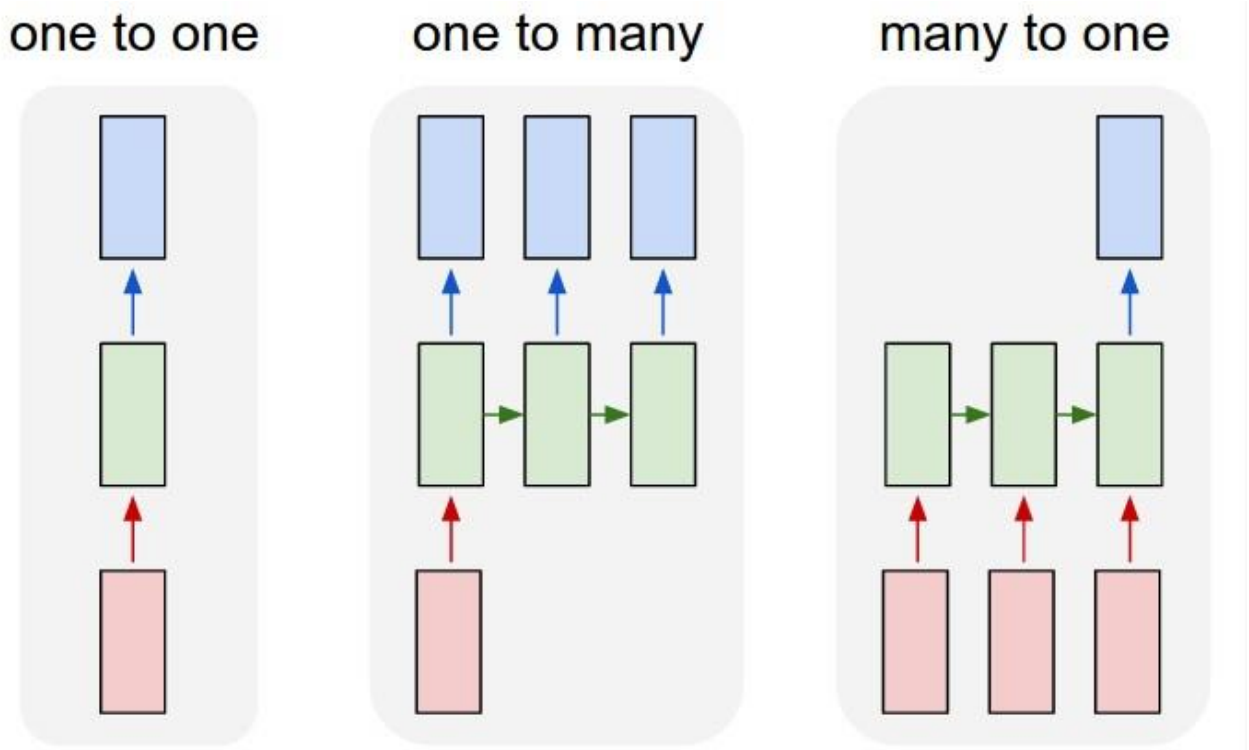
Recurrent Networks offer a lot of flexibility:



e.g. **Image Captioning**
image -> sequence of words

[Slides adopted from: Li, Karpathy]

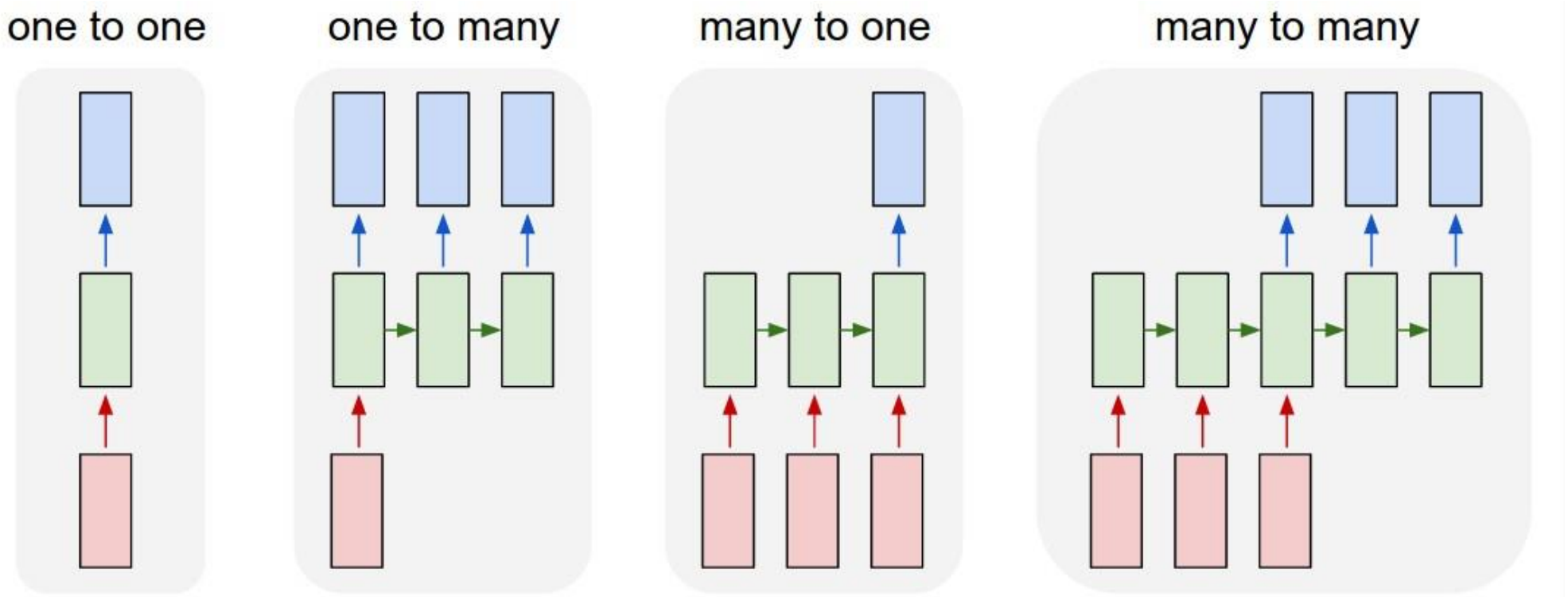
Recurrent Networks offer a lot of flexibility:



e.g. **Sentiment Classification**
sequence of words -> sentiment

[Slides adopted from: Li, Karpathy]

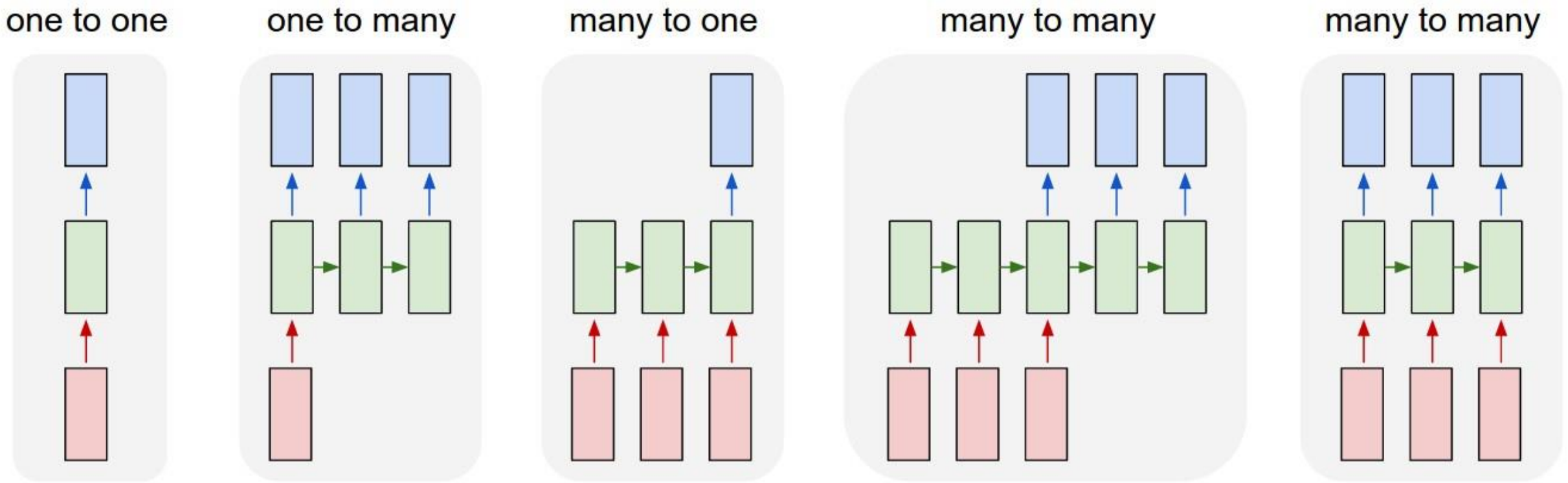
Recurrent Networks offer a lot of flexibility:



e.g. **Machine Translation**
seq of words -> seq of words

[Slides adopted from: Li, Karpathy]

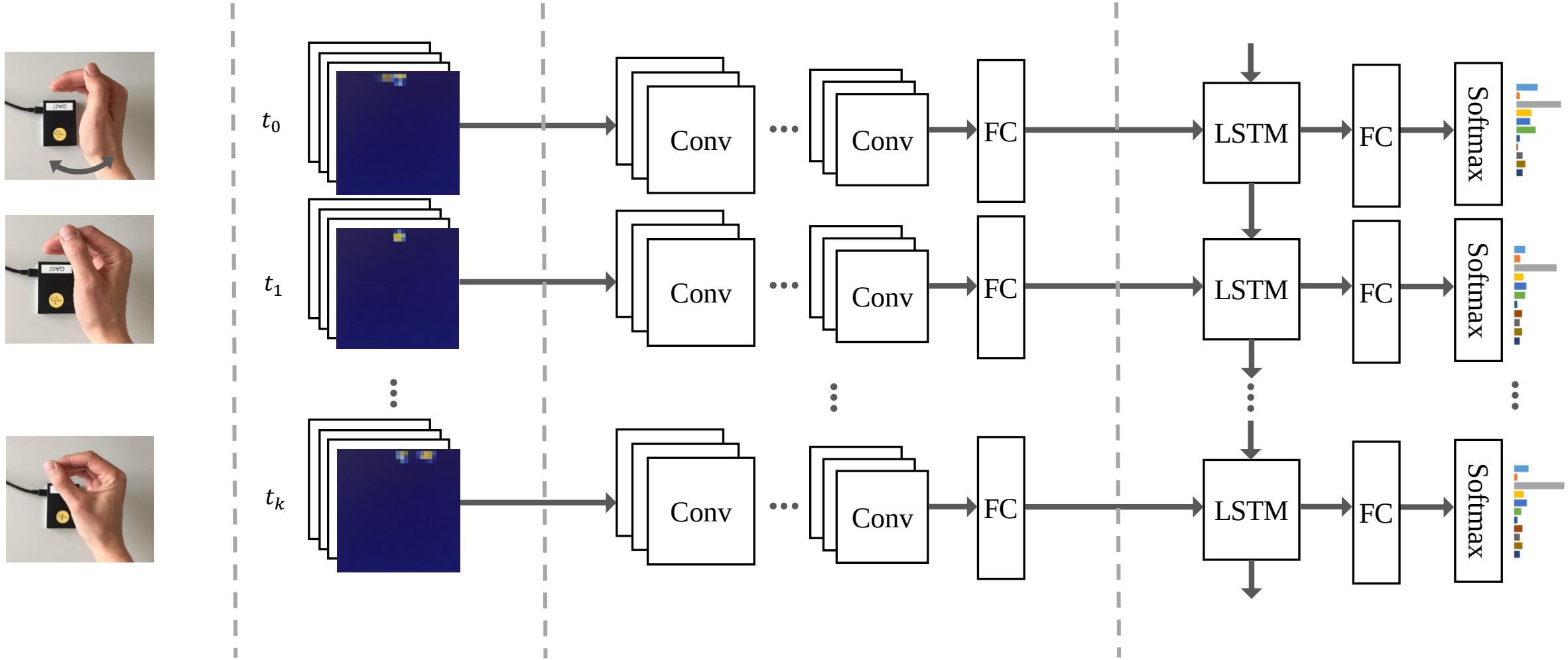
Recurrent Networks offer a lot of flexibility:

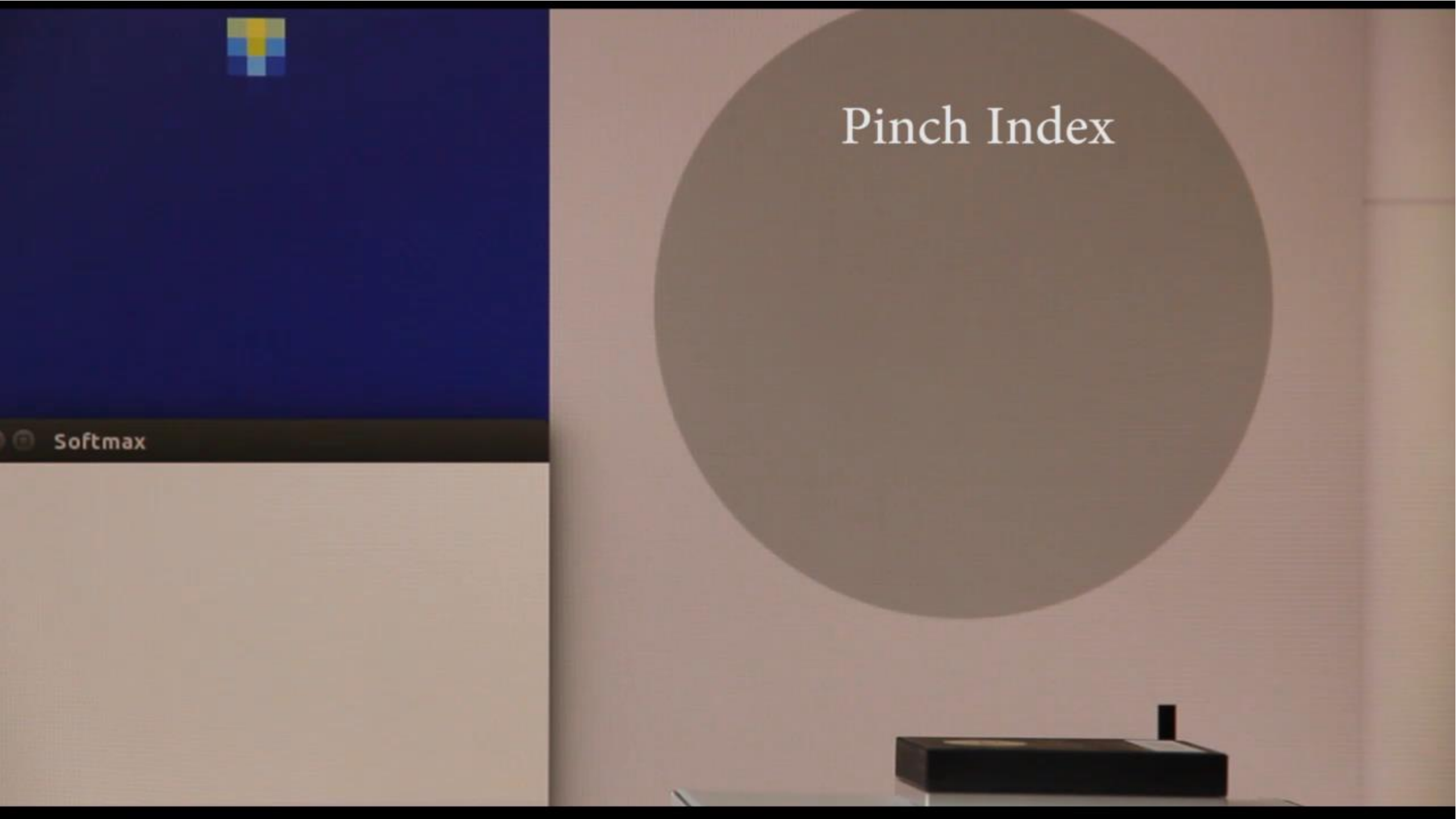


e.g. **Video classification on frame level**

[Slides adopted from: Li, Karpathy]

Network Architecture




Softmax
Pinch Index

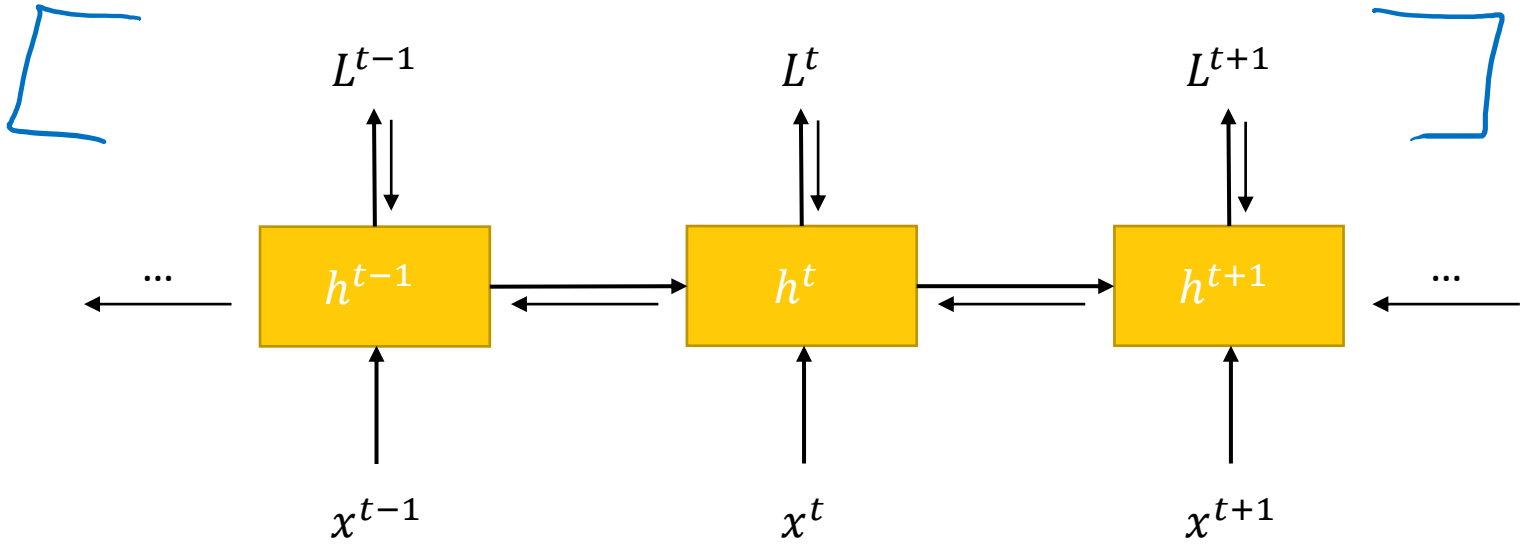
Backprop through time (BPTT)

$$h^t = f(h^{t-1}, x^t; W)$$

$$\hat{y}^t = W_{hy} h^t$$

$$L^t = \|\hat{y}^t - y^t\|^2$$

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{S} \frac{\partial L^t}{\partial W}$$



Intuition: treat unrolled recurrent model as multi-layer network (with unbounded number of layers) and perform backprop

Exploding and vanishing gradients

$$h^t = W^T h^{t-1}$$

$$h^t = (W^t)^T h^1$$

power method

[power method]

if $\exists$ eigen decomposition of W:

exist

$$W = Q \Lambda Q^T$$

then:

$$h^t = \left((Q \Lambda Q^T)^t\right) h^1$$

$$= (Q^T \Lambda^t Q) h^1$$

if $\lambda < 1$, λ^t vanish

$\lambda > 1$, λ^t goes to infinity

diag eigenvalues

Backprop through time (BPTT)

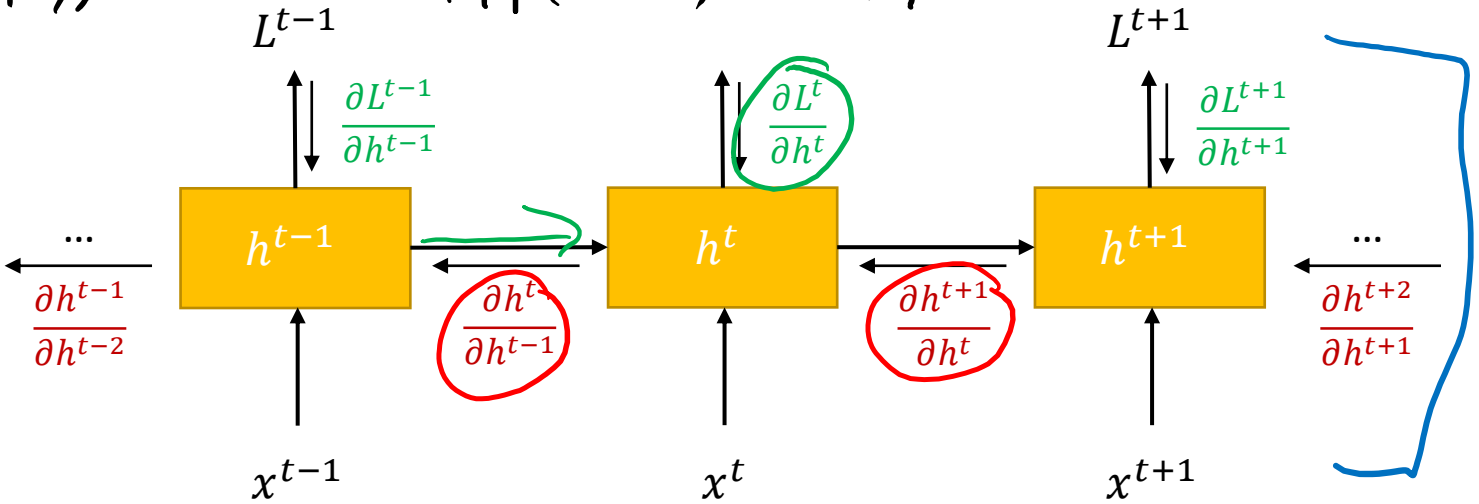
Bengio et al, "*Learning long-term dependencies with gradient descent is difficult*", IEEE Transactions on Neural Networks, 1994
Pascanu et al, "*On the difficulty of training recurrent neural networks*", ICML 2013

$h^t = f(h^{t-1}, x^t; W)$ $= h^t = \tanh(W_{hh} h^{t-1} + W_{xh} x^t + b)$

Ignore bias here for brevity

$\hat{y}^t = W_{hy} h^t$ $\Rightarrow$ assume $L^t(f(h^t)) \Rightarrow h^t = W_{hh} f(h^{t-1}) + W_{xh} x^t$

$L^t = \|\hat{y}^t - y^t\|^2$



$$\frac{\partial L}{\partial W} = \sum_{t=1}^{S} \frac{\partial L^t}{\partial W}$$

$$\frac{\partial L^t}{\partial W} = \sum_{k=1}^{t} \frac{\partial L^t}{\partial y^t} \frac{\partial y^t}{\partial h^t} \frac{\partial h^t}{\partial h^k} \frac{\partial^+ h^k}{\partial W}$$

immediate derivative = row i $\sigma(h^{k-1})$

multivariate function

$f(x(t), y(t))$

assume h^{k-1} is constant

$$\frac{df}{dt} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt}$$

Vanishing / Exploding Gradient (BPTT)

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

$$\frac{\partial L^t}{\partial \mathrm{W}} = \sum_{k=1}^{t} \frac{\partial L^t}{\partial y^t} \frac{\partial y^t}{\partial h^t} \frac{\partial h^t}{\partial h^k} \frac{\partial^+ h^k}{\partial W}$$

$$\frac{\partial h^t}{\partial h^k} = \prod_{i=k+1}^{t} \frac{\partial h^i}{\partial h^{i-1}} = \prod_{i=k+1}^{t} W_{hh}^T diag[f'(h^{i-1})]$$

$$\forall_i, \left\| \frac{\partial h^i}{\partial h^{i-1}} \right\| \leq \|W_{hh}^T\| \|diag[f'(h^{i-1})]\| < \frac{1}{\gamma}\gamma < 1$$

$$\left\| \frac{\partial h^t}{\partial h^k} \right\| < (\eta)^{t-k}$$

Vanishing gradients – Proof

Let λ_1 be the largest singular value of W_{hh};

$\| \text{diag}\, f'(h^{i-1}) \| < \gamma \in \mathbb{R}$

Case I: It is sufficient for $\lambda_1 < \frac{1}{\gamma}$ for gradients to vanish

$\forall i \quad \| \frac{\partial h^i}{\partial h^{i-1}} \| \leq \| W_{hh}^T \| \, \| \text{diag}\, f'(h^{i-1}) \| \leq \frac{1}{\gamma} \gamma < 1$

Let $\eta \in \mathbb{R}$ such that $\forall i, \| \frac{\partial h^i}{\partial h^{i-1}} \| \leq \eta < 1$ (η exists because of)

By induction over i:

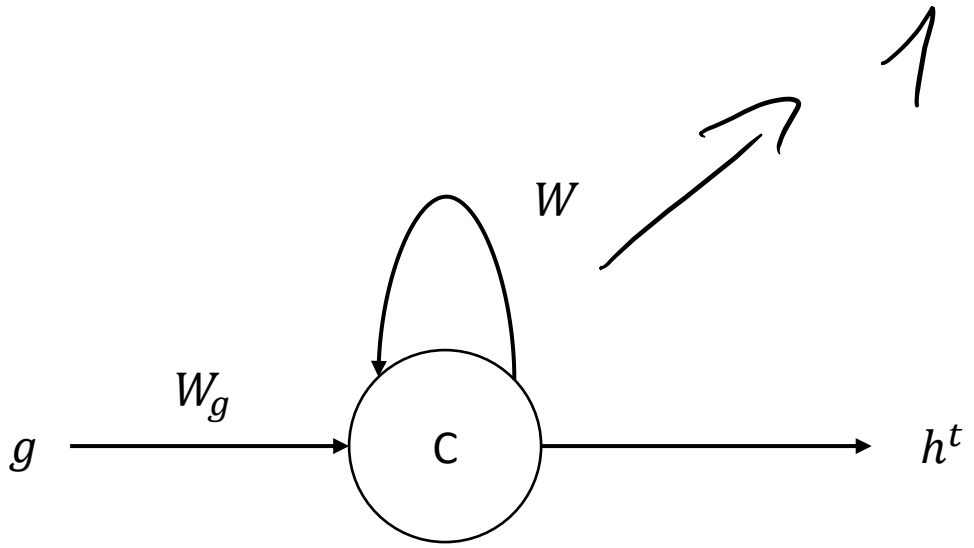
$\| \prod_{i=k+1}^{t} \frac{\partial h^i}{\partial h^{i-1}} \| \leq \eta^{(t-k)}$ $\Rightarrow$ if $\lambda_1 < \frac{1}{\gamma}$ ($\frac{1}{\gamma}$) $\Rightarrow$ long term components will vanish $\square$

Case II: inverted Case 1 $\square$ $\Rightarrow$ if $\lambda_1 > \frac{1}{\gamma}$ l.t. components will explode

Naïve Solution

$c_t = Wc^{t-1} + W_g g^t$

$h_t = \tanh(c^t)$



remember need to read, write & erase (with W=1 we can only add)

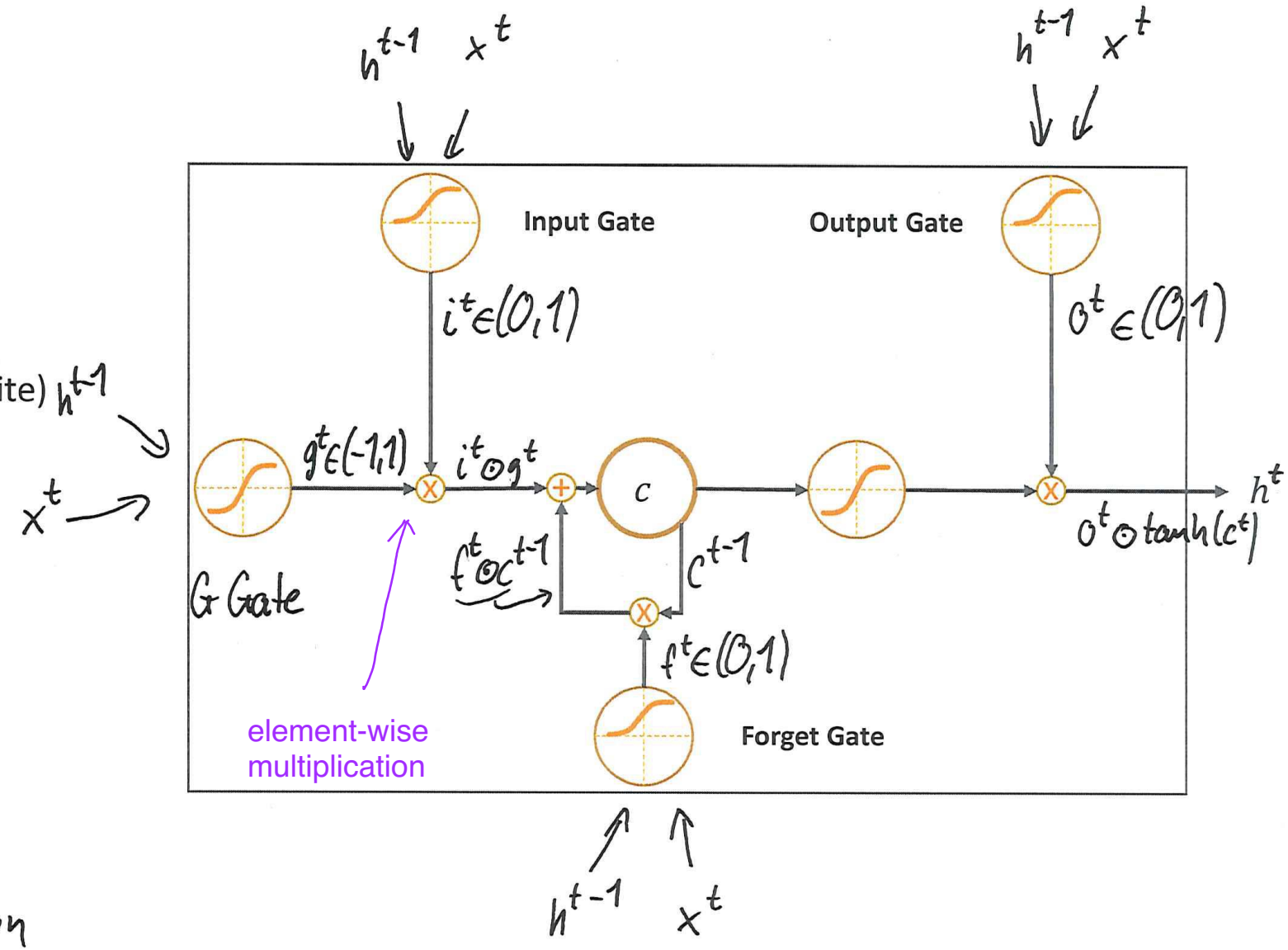
LSTM [Hochreiter et al. 1997]

Input Gate: Scales input to cell (read)

Forget Gate: Scales old cell values (reset)

Output Gate: Scales output from cell (write)

Gate Gate



$$c^t = f^t \odot c^{t-1} + i^t \odot g^t$$

$$h^t = o^t \odot \tanh(c^t)$$

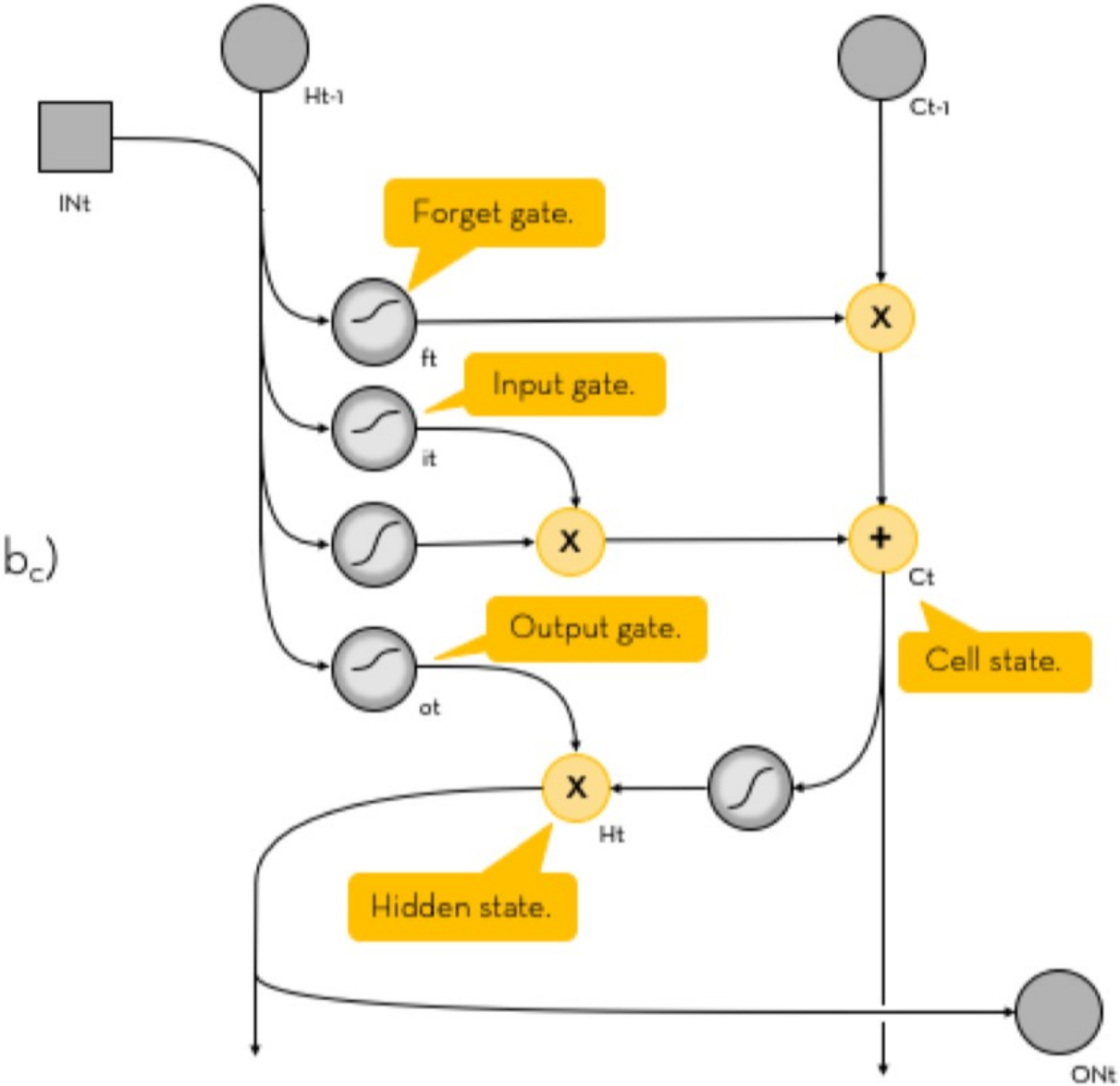
$=$

element wise multiplication

The inner-workings of a LSTM cell
LSTMs use 3 gates to control the flow of information

$$f_t = \text{sigmoid}(U_f * X_t + W_f * H_{t-1} + b_f)$$
$$i_t = \text{sigmoid}(U_i * X_t + W_i * H_{t-1} + b_i)$$
$$o_t = \text{sigmoid}(U_o * X_t + W_o * H_{t-1} + b_o)$$

$$C_t = f_t * C_{t-1} + i_t * \text{tanH}(U_c * X_t + W_c * H_{t-1} + b_c)$$
$$H_t = o_t * \text{tanH}(C_t)$$



http://colah.github.io/posts/2015-08-Understanding-LSTMs/

RNN:

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

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

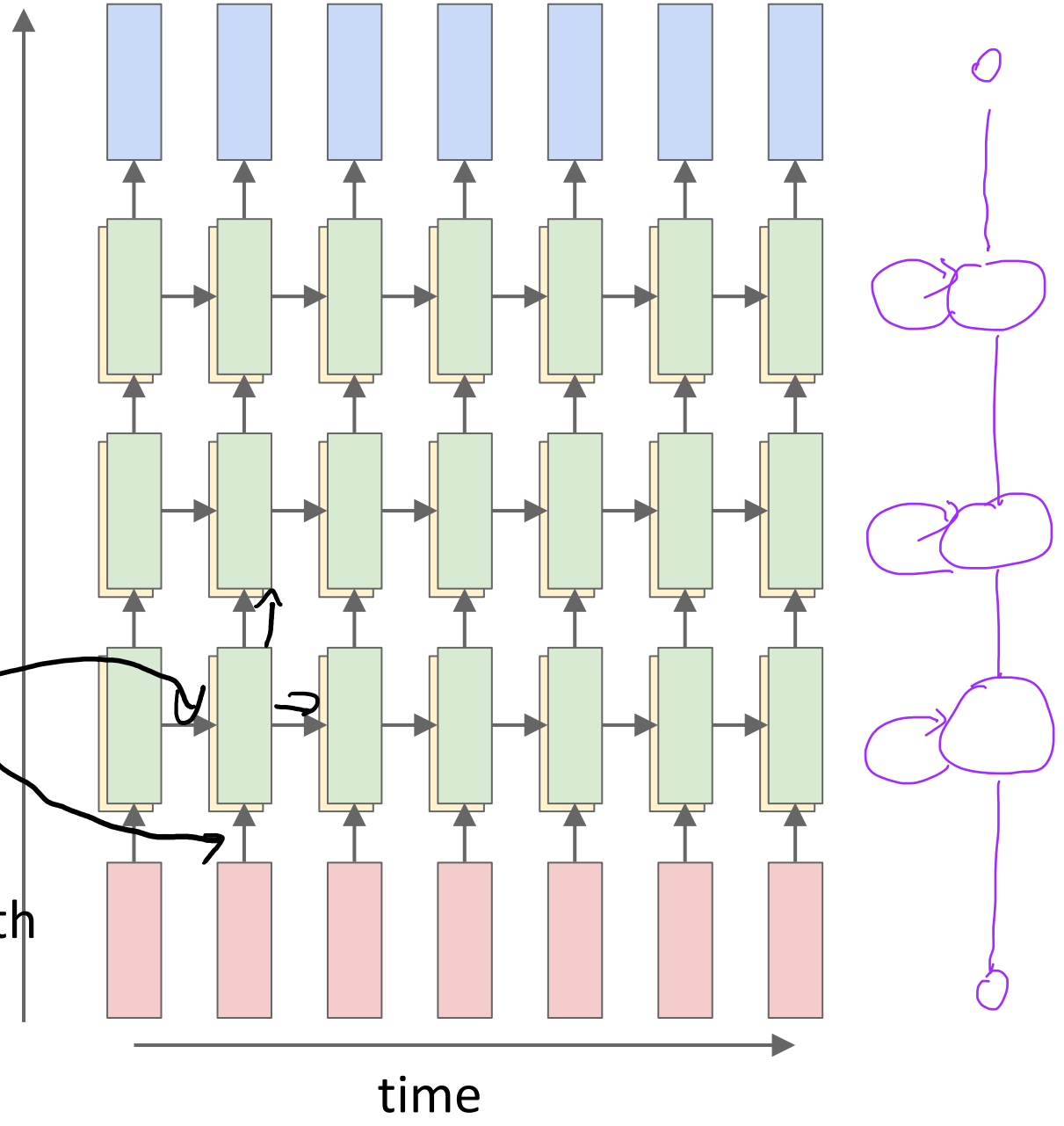
LSTM:

$W^l \ [4n \times 2n]$

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} W^l \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}$$

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

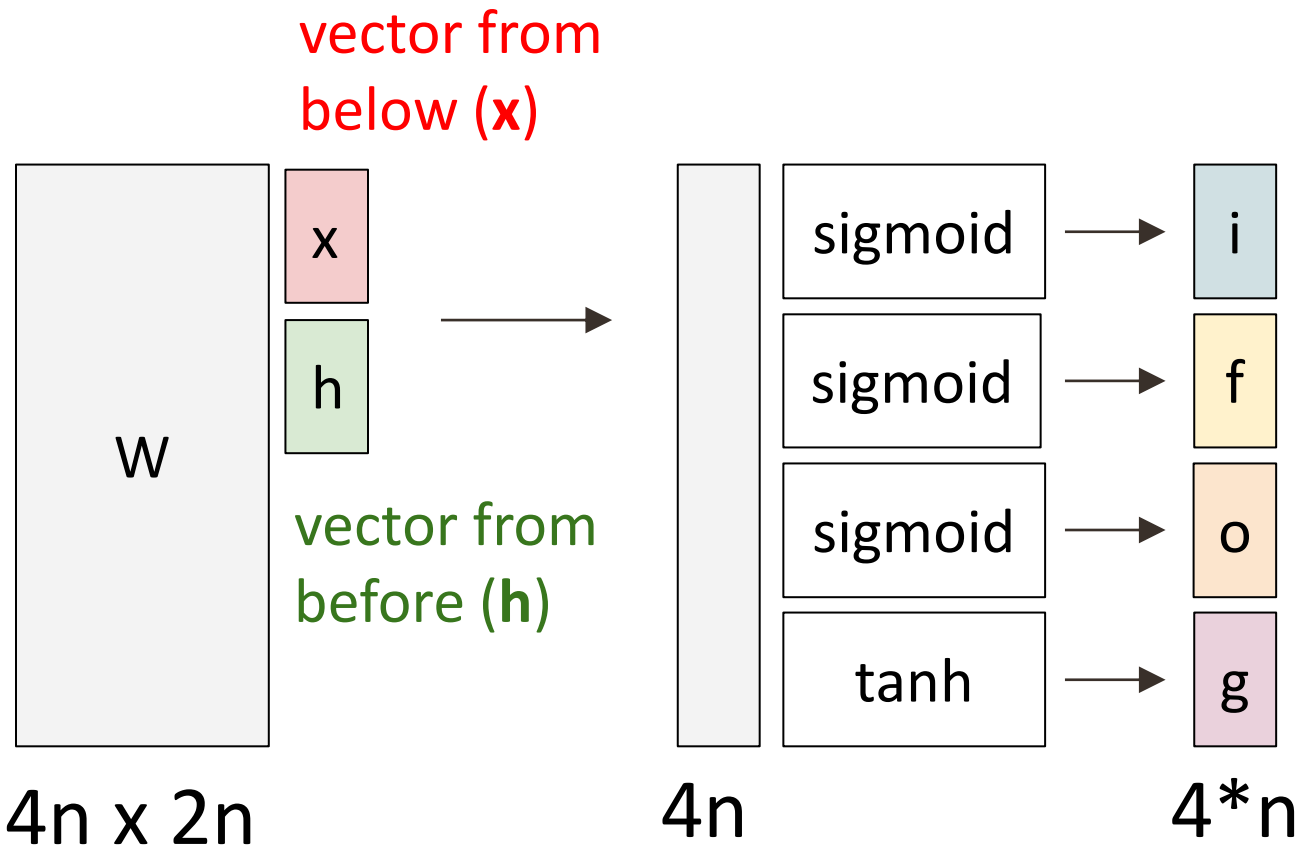
$$h_t^l = o \odot \tanh(c_t^l)$$



[Slides adopted from: Li, Karpathy]

Long Short Term Memory (LSTM)

[Hochreiter et al., 1997]



$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} W^l \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}$$

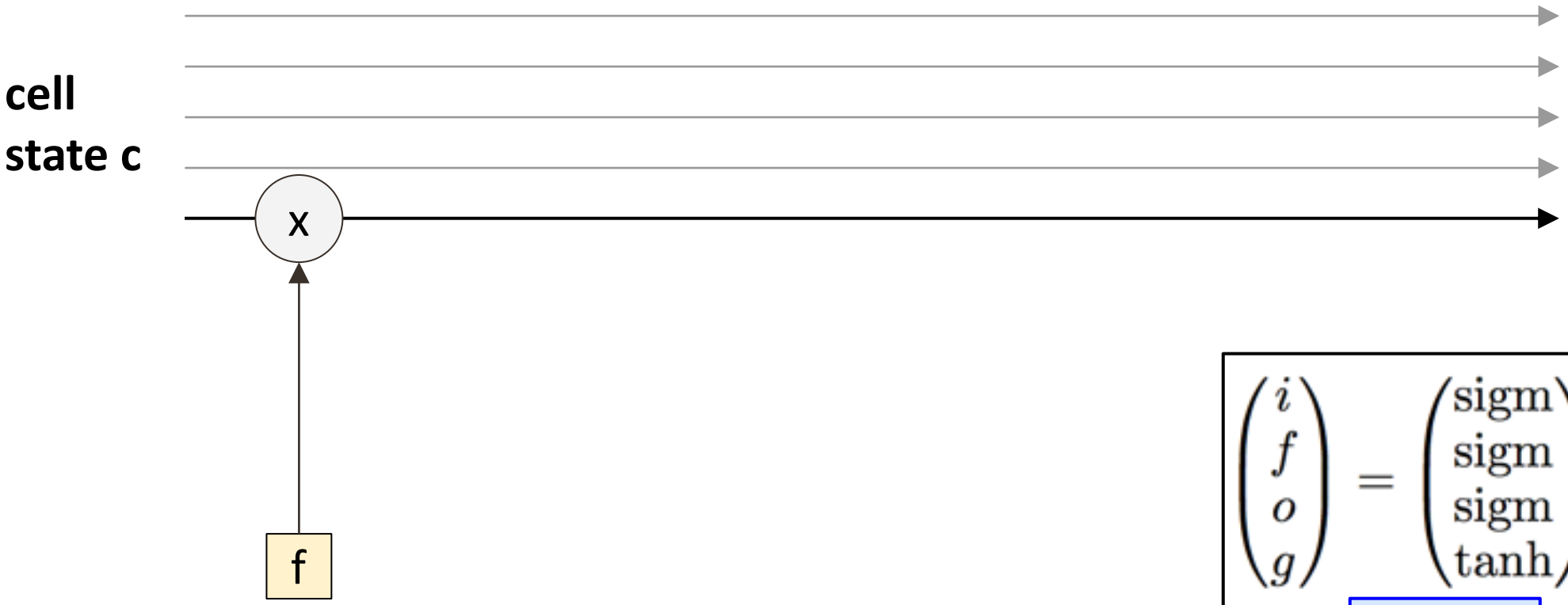
$$c_t^l = f \odot c_{t-1}^l + i \odot g$$

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

[Slides adopted from: Li, Karpathy]

Long Short Term Memory (LSTM)

[Hochreiter et al., 1997]



$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} W^l \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}$$

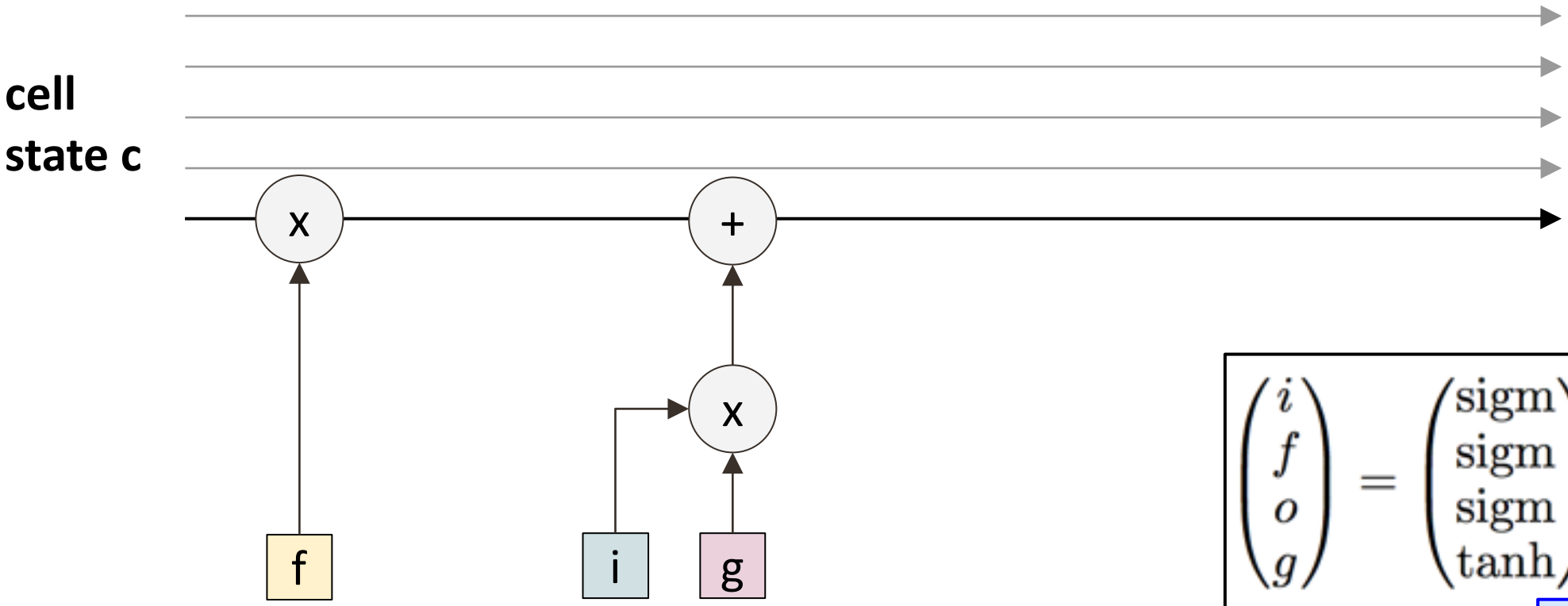
$$c_t^l = f \odot c_{t-1}^l + i \odot g$$

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

[Slides adopted from: Li, Karpathy]

Long Short Term Memory (LSTM)

[Hochreiter et al., 1997]



$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} W^l \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}$$

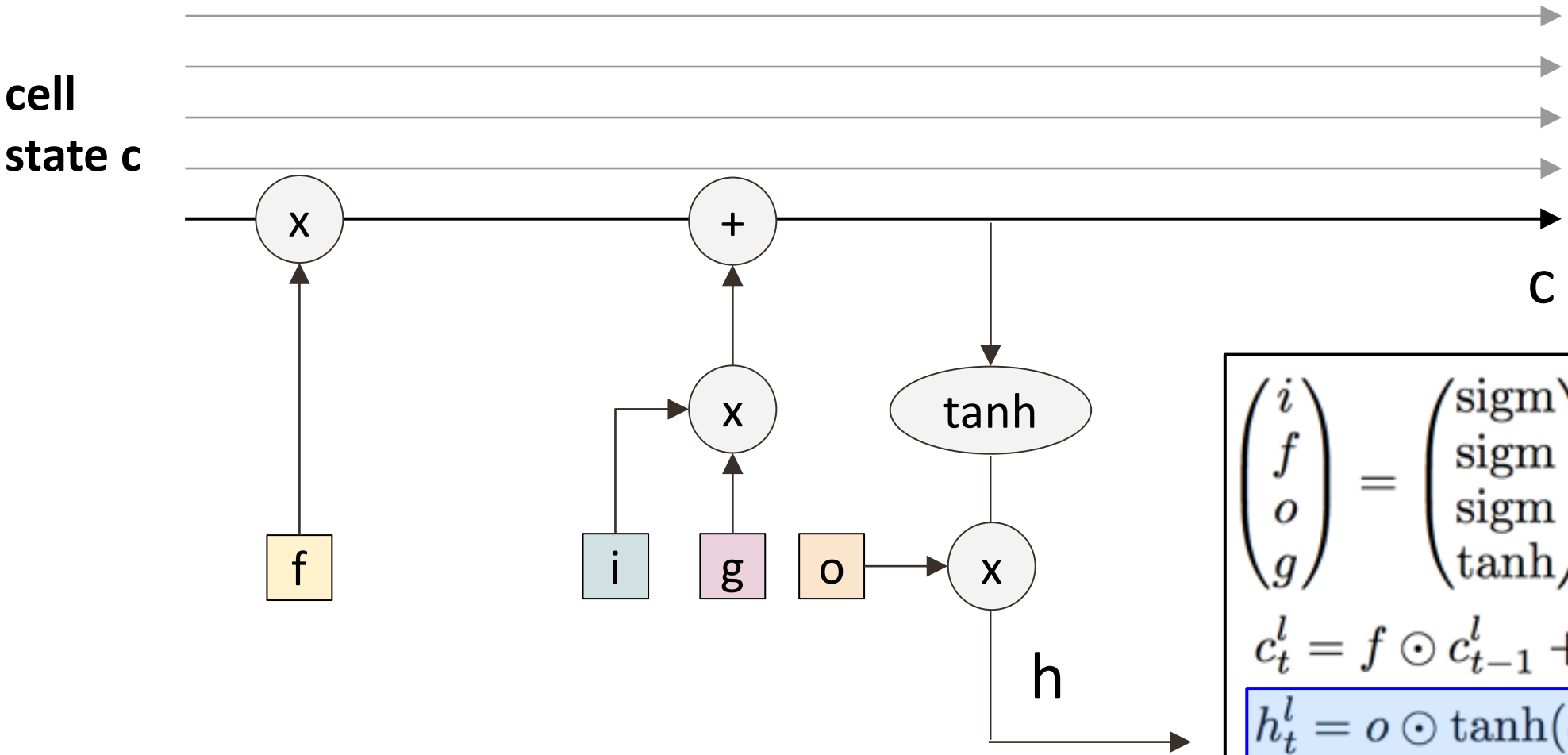
$$c_t^l = f \odot c_{t-1}^l + i \odot g$$

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

[Slides adopted from: Li, Karpathy]

Long Short Term Memory (LSTM)

[Hochreiter et al., 1997]



$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} W^l \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}$$

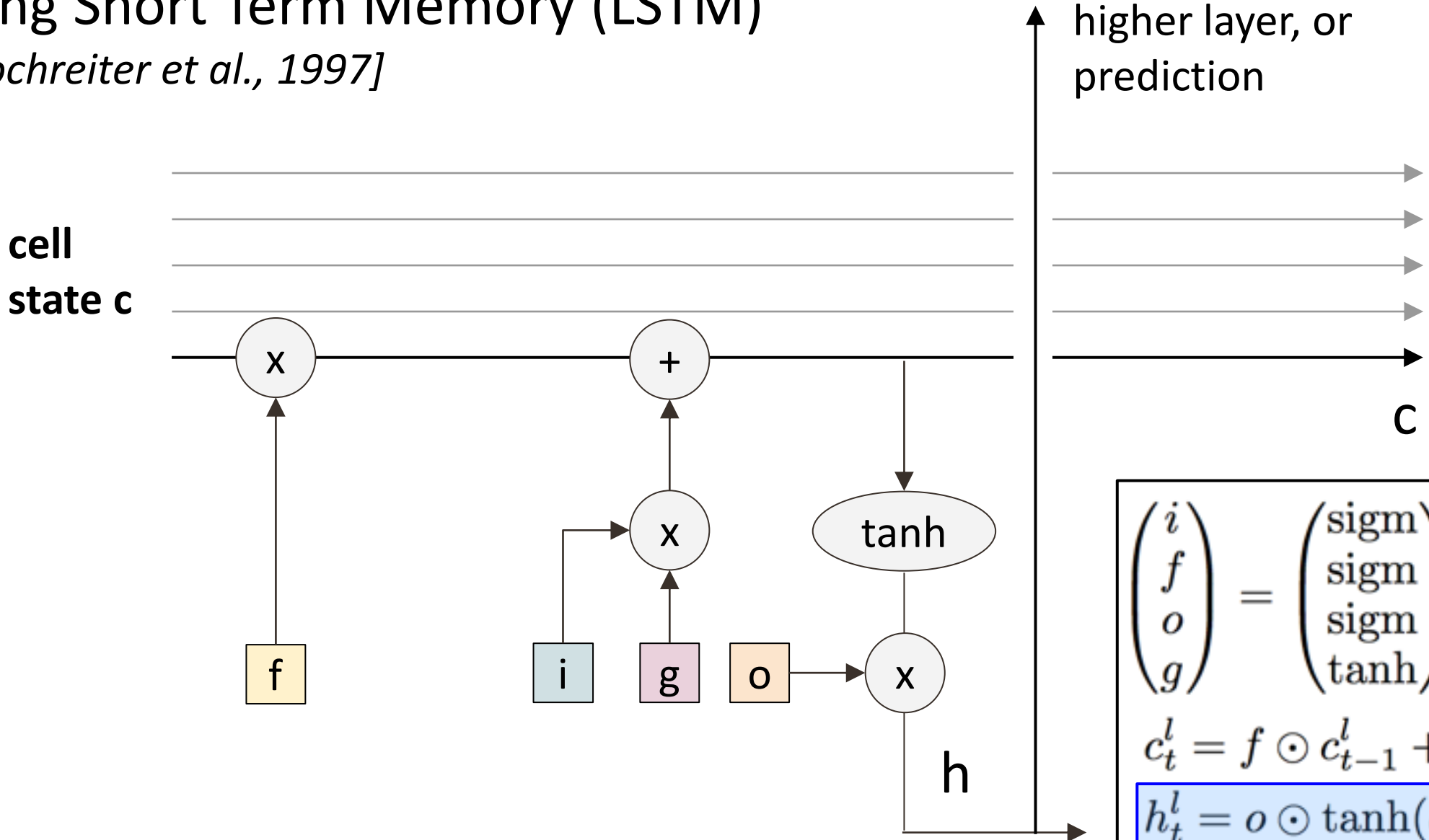
$$c_t^l = f \odot c_{t-1}^l + i \odot g$$

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

[Slides adopted from: Li, Karpathy]

Long Short Term Memory (LSTM)

[Hochreiter et al., 1997]



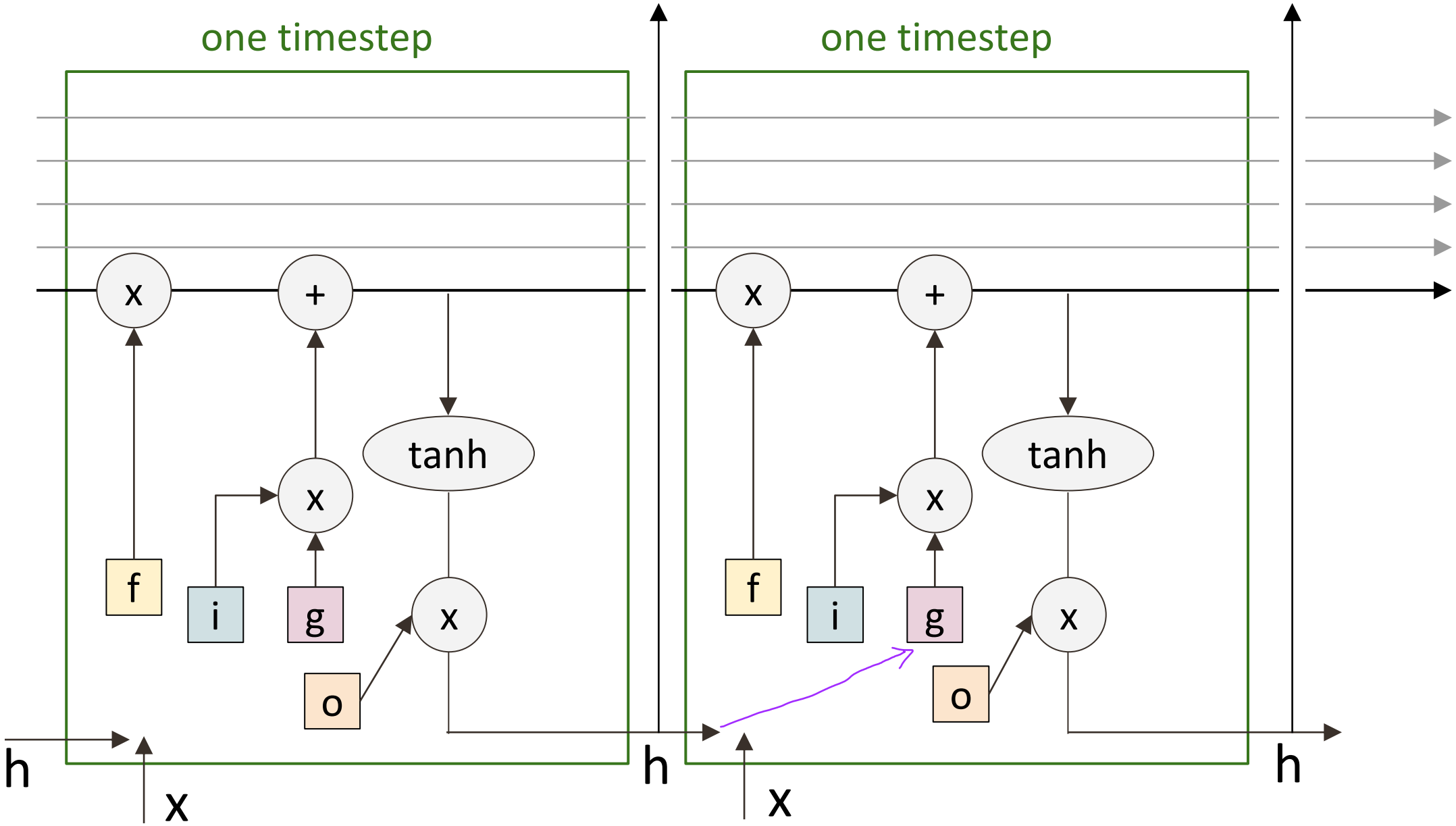
$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} W^l \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}$$

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

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

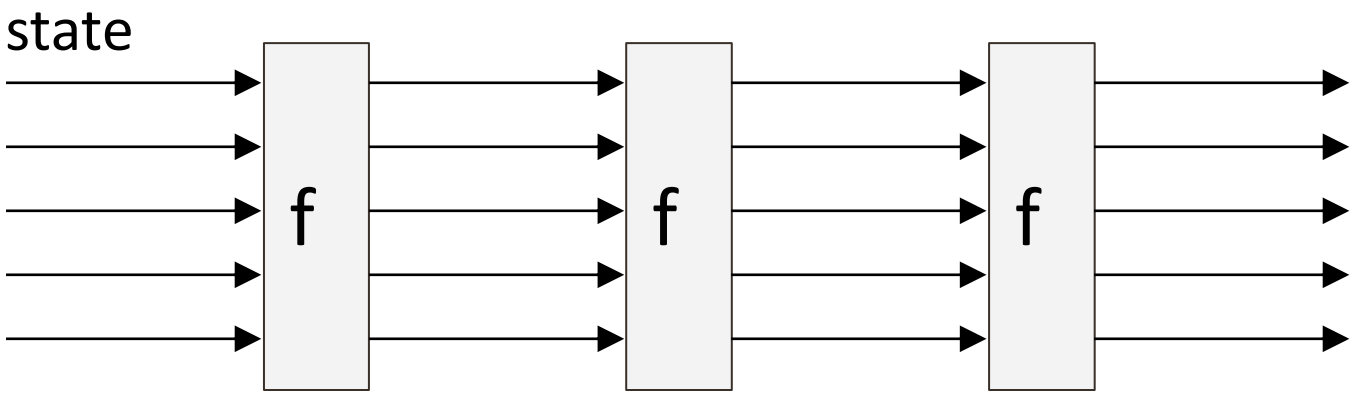
[Slides adopted from: Li, Karpathy]

LSTM



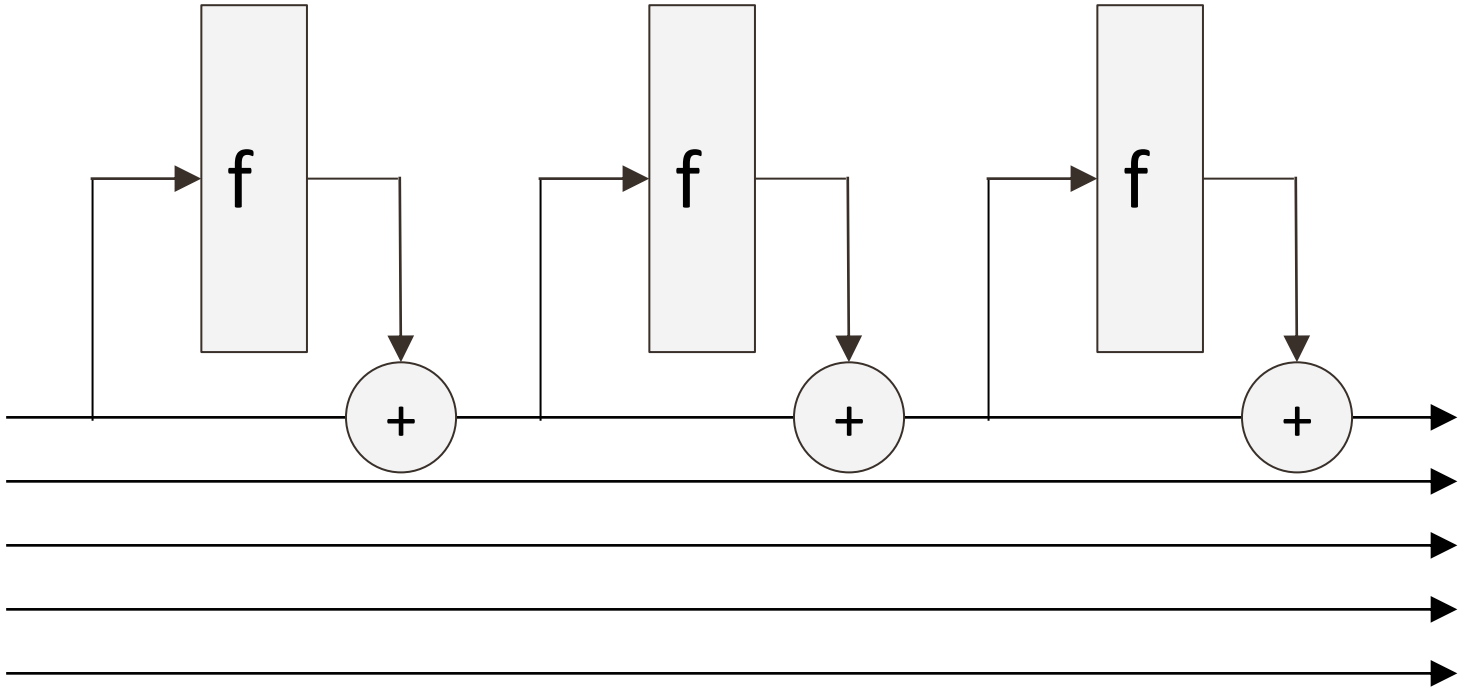
[Slides adopted from: Li, Karpathy]

RNN



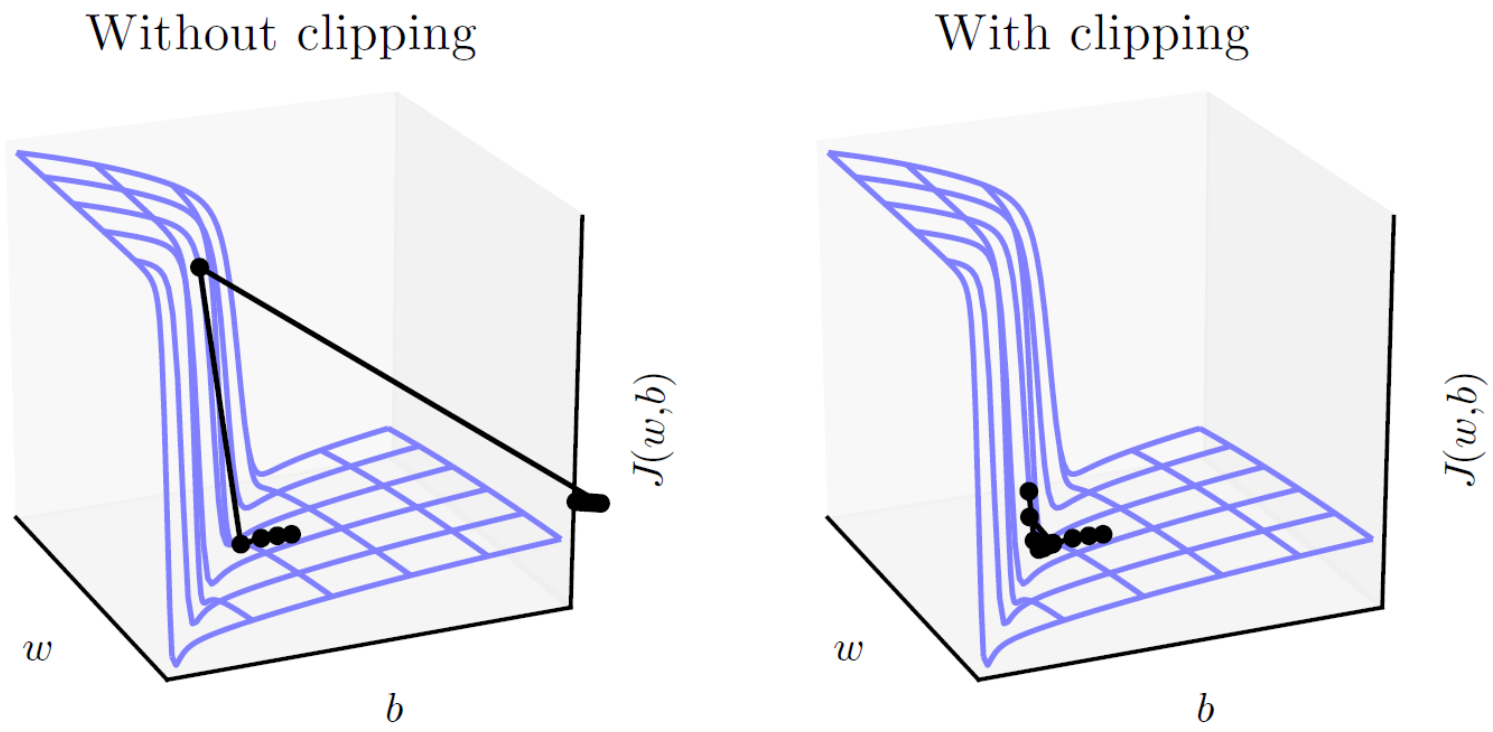
LSTM

(ignoring
forget gates)



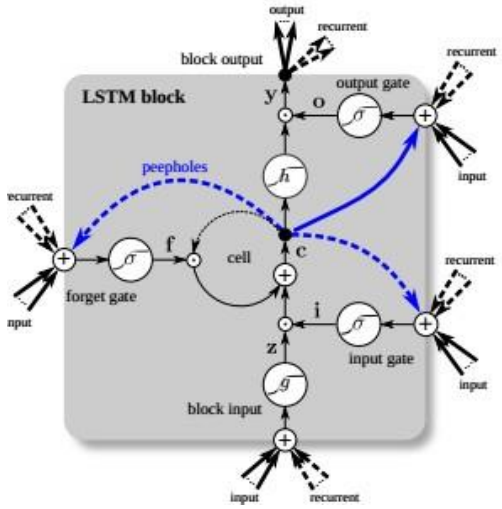
[Slides adopted from: Li, Karpathy]

Gradient Clipping



[Goodfellow 2016]

LSTM variants and friends



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

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

$$
\begin{aligned}
r_t &= \text{sigm}\left(W_{\text{xr}} x_t + W_{\text{hr}} h_{t-1} + b_{\text{r}}\right) \\
z_t &= \text{sigm}(W_{\text{xz}} x_t + W_{\text{hz}} h_{t-1} + b_{\text{z}}) \\
\tilde{h}_t &= \tanh(W_{\text{xh}} x_t + W_{\text{hh}} (r_t \odot h_{t-1}) + b_{\text{h}}) \\
h_t &= z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t
\end{aligned}
$$

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

MUT1:

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

MUT2:

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

MUT3:

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

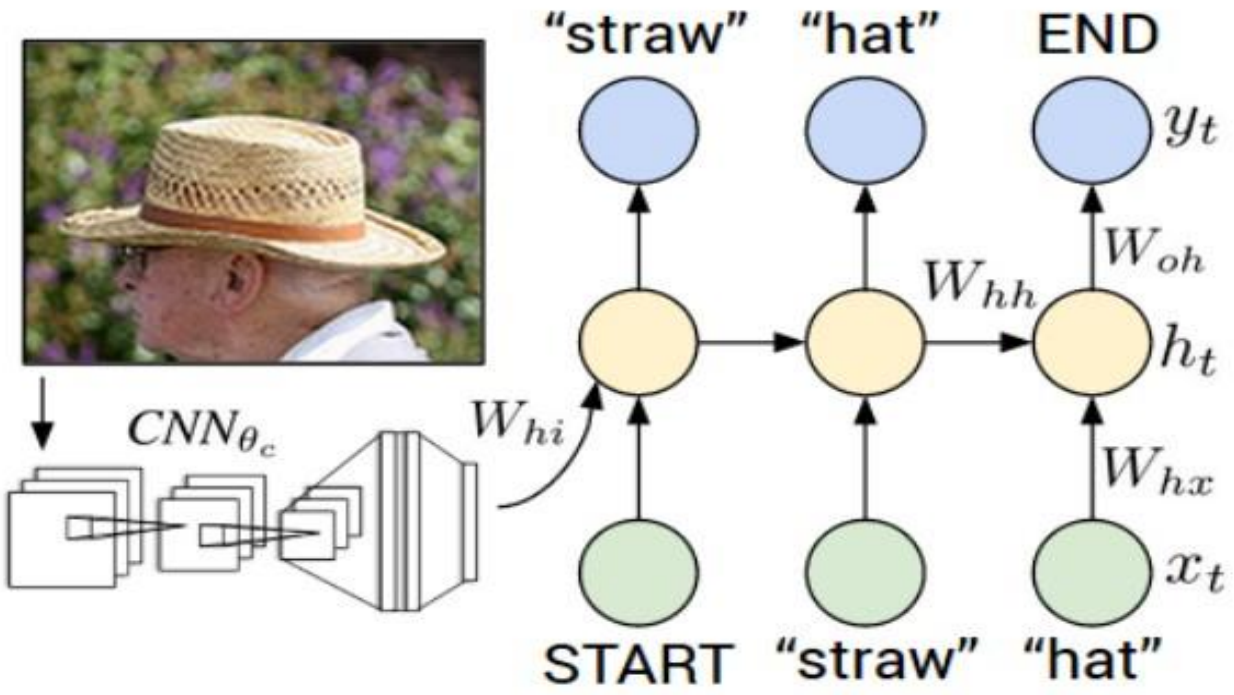
Next Week

Taught by Siyu Tang

CNNs in computer vision

Fully convolutional CNNs

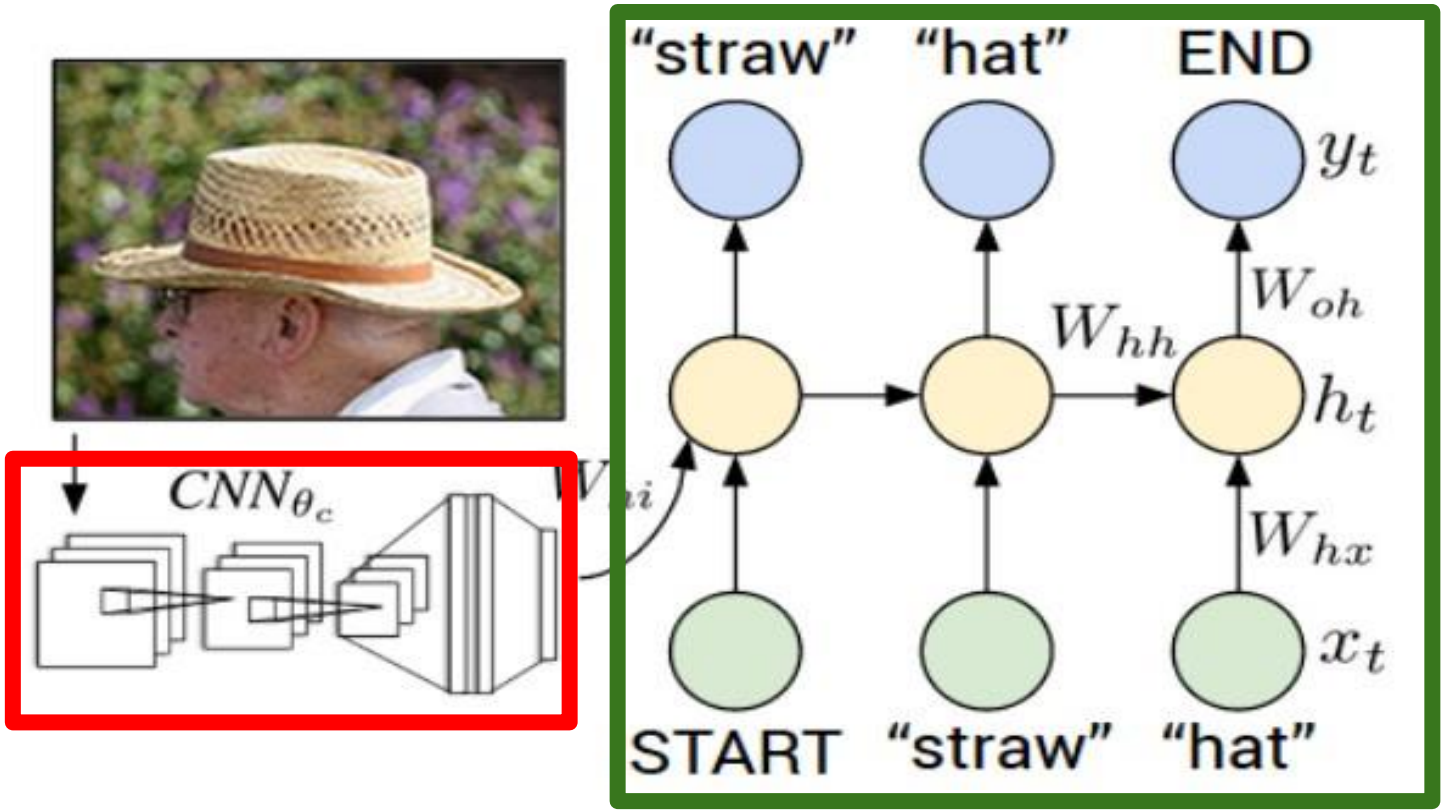
Image Captioning



[Explain Images with Multimodal Recurrent Neural Networks, Mao et al.]
[Deep Visual-Semantic Alignments for Generating Image Descriptions, Karpathy and Fei-Fei]
[Show and Tell: A Neural Image Caption Generator, Vinyals et al.]
[Long-term Recurrent Convolutional Networks for Visual Recognition and Description, Donahue et al.]
[Learning a Recurrent Visual Representation for Image Caption Generation, Chen and Zitnick]

[Slides adopted from: Li, Karpathy]

Recurrent Neural Network



Convolutional Neural Network

[Slides adopted from: Li, Karpathy]

test image

image

conv-64

conv-64

maxpool

conv-128

conv-128

maxpool

conv-256

conv-256

maxpool

conv-512

conv-512

maxpool

conv-512

conv-512

maxpool

FC-4096

FC-4096

FC-1000

softmax

test image

[Slides adopted from: Li, Karpathy]

image

conv-64

conv-64

maxpool

conv-128

conv-128

maxpool

conv-256

conv-256

maxpool

conv-512

conv-512

maxpool

conv-512

conv-512

maxpool

FC-4096

FC-4096

FC-1000

softmax

X

test image

[Slides adopted from: Li, Karpathy]

image

conv-64

conv-64

maxpool

conv-128

conv-128

maxpool

conv-256

conv-256

maxpool

conv-512

conv-512

maxpool

conv-512

conv-512

maxpool

FC-4096

FC-4096

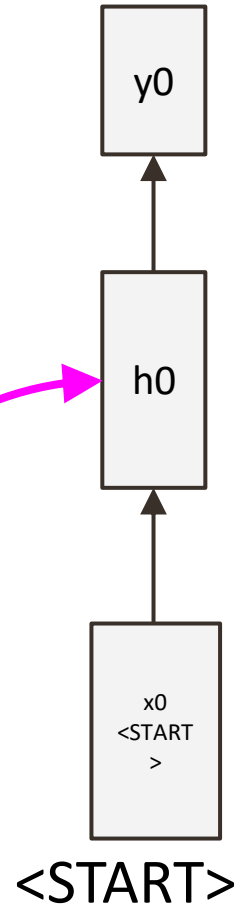
test image

x0
<START
>

<START>

[Slides adopted from: Li, Karpathy]

test image



Wih

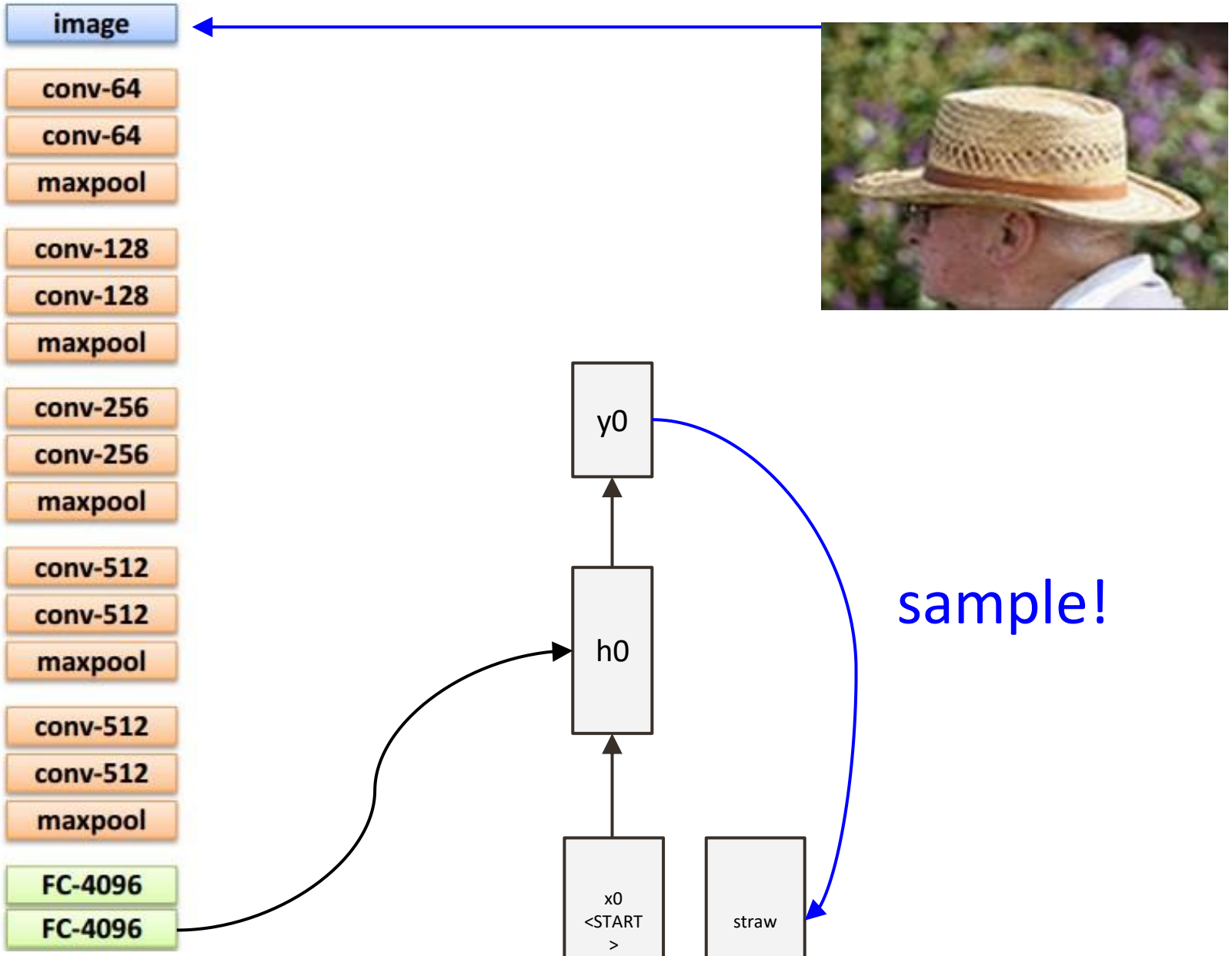
before:

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

now:

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

[Slides adopted from: Li, Karpathy]



[Slides adopted from: Li, Karpathy]

image

conv-64
conv-64
maxpool

conv-128
conv-128
maxpool

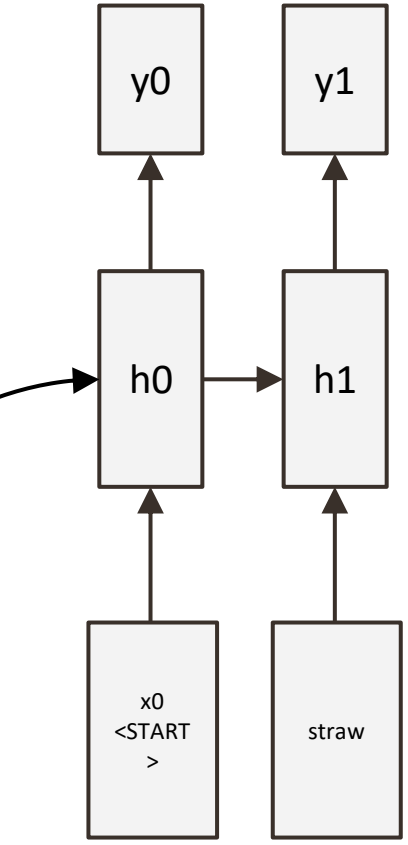
conv-256
conv-256
maxpool

conv-512
conv-512
maxpool

conv-512
conv-512
maxpool

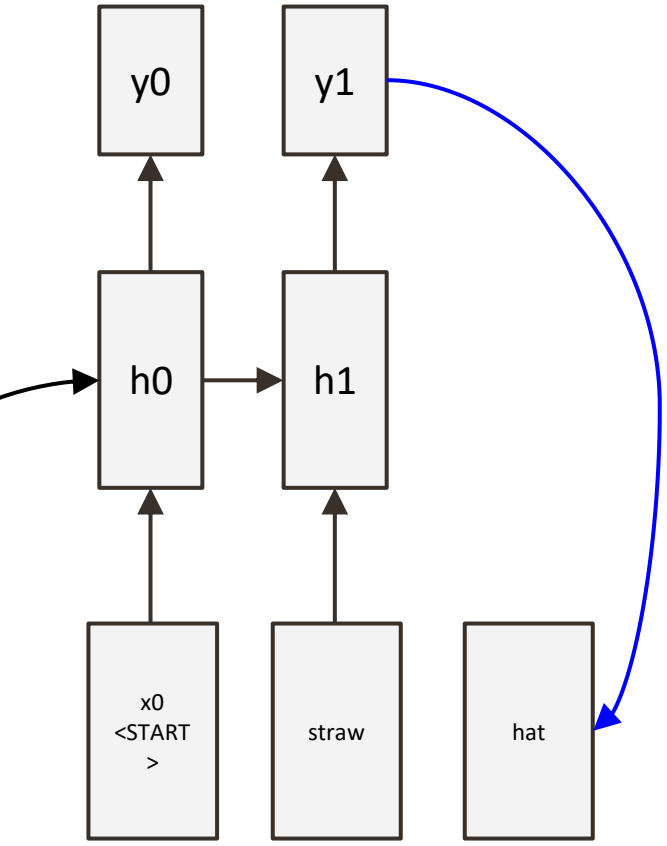
FC-4096
FC-4096

test image



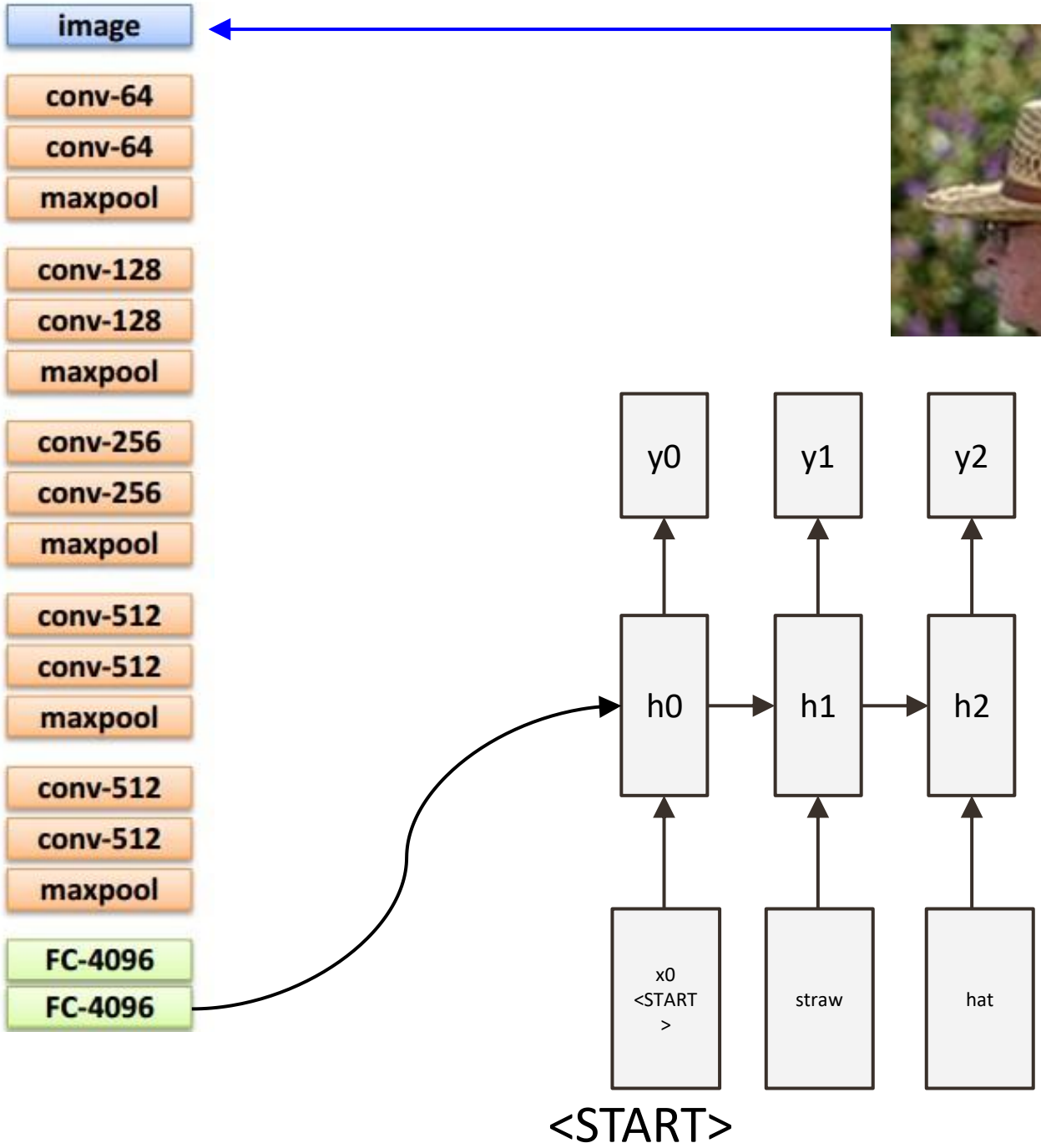
[Slides adopted from: Li, Karpathy]

test image



sample!

[Slides adopted from: Li, Karpathy]



test image

[Slides adopted from: Li, Karpathy]

image

conv-64

conv-64

maxpool

conv-128

conv-128

maxpool

conv-256

conv-256

maxpool

conv-512

conv-512

maxpool

conv-512

conv-512

maxpool

FC-4096

FC-4096

test image

y0 y1 y2

h0 h1 h2

x0 <START>

straw

hat

<START>

sample <END> token => finish.

[Slides adopted from: Li, Karpathy]

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

[Slides adopted from: Li, Karpathy]

"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."

[Slides adopted from: Li, Karpathy]

"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."

"a young boy is holding a baseball bat."

"a cat is sitting on a couch with a remote control."

"a woman holding a teddy bear in front of a mirror."

"a horse is standing in the middle of a road."