

Recurrent neural network

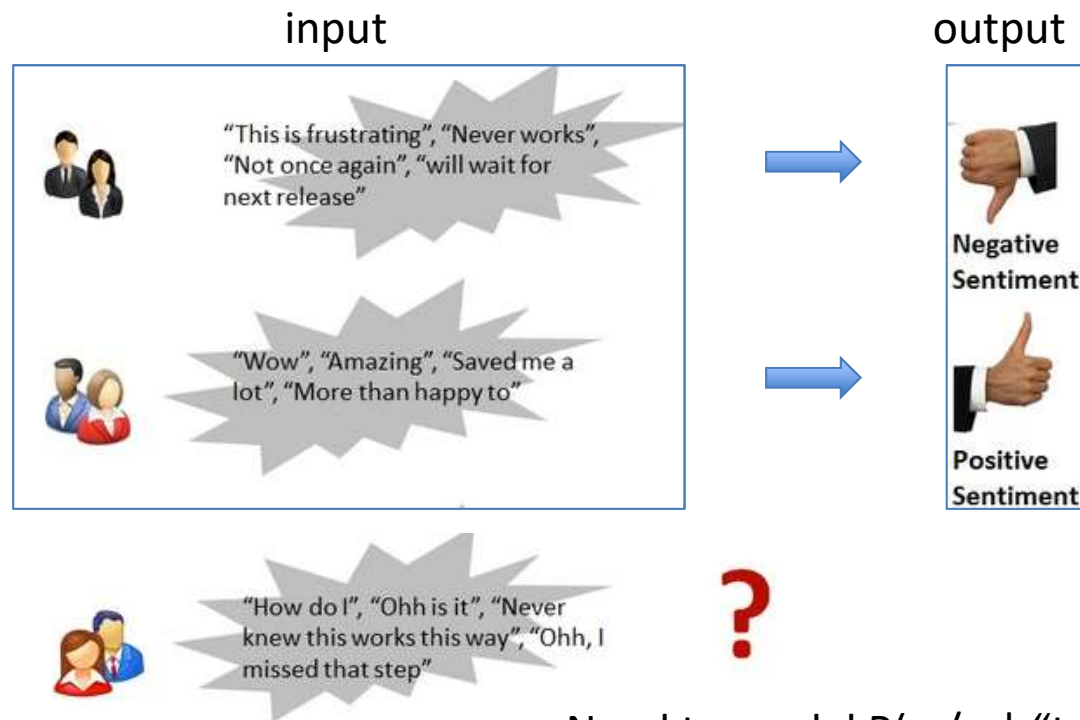
Kyung-Ah Sohn

Ajou University

Applications with sequence data

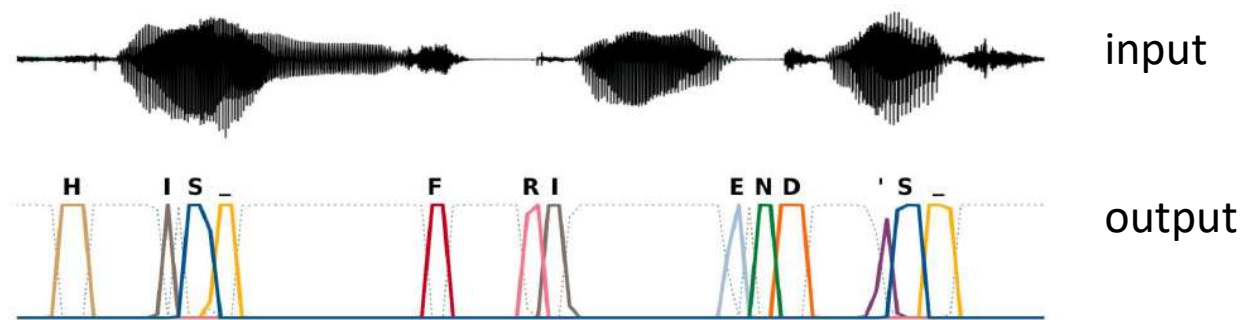
- Stock prediction
- Speech recognition
- Text classification
- Sequence generation

Sentiment analysis / text classification



Need to model $P(+/- \mid \text{"text sequence"})$

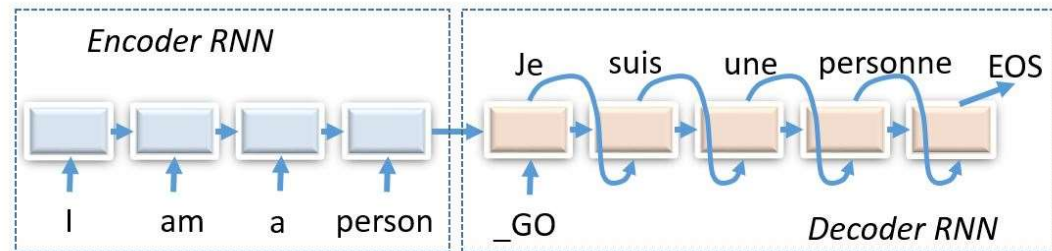
Speech recognition



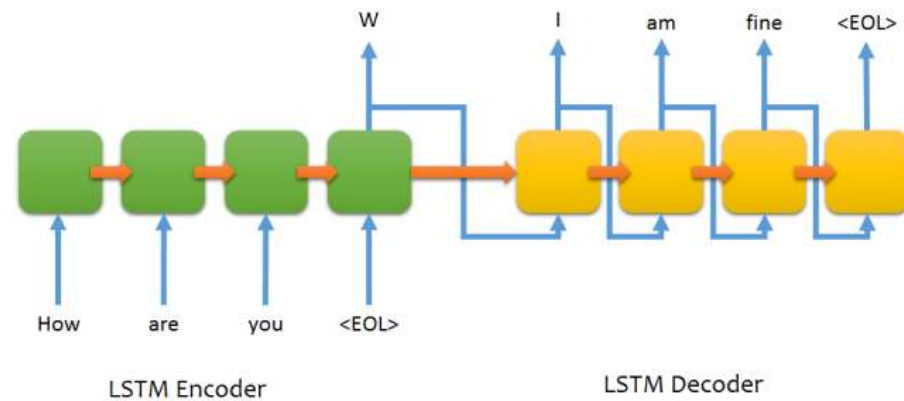
Graves & Jaitly, 2014

Need to model $P(\text{"text sequence"} \mid \text{audio})$

Machine Translation

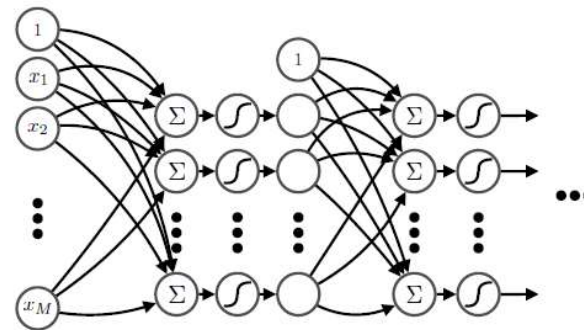


Text generation



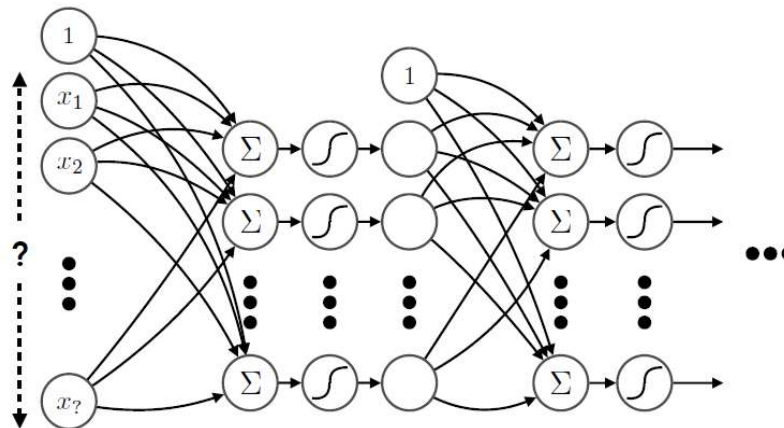
Sequence modeling

- Sequence-based mapping, e.g. $p(\text{"Hey Jude"} \mid \text{audio waveform})$ is difficult to define by hand, need to learn from data
- How to define network architecture?



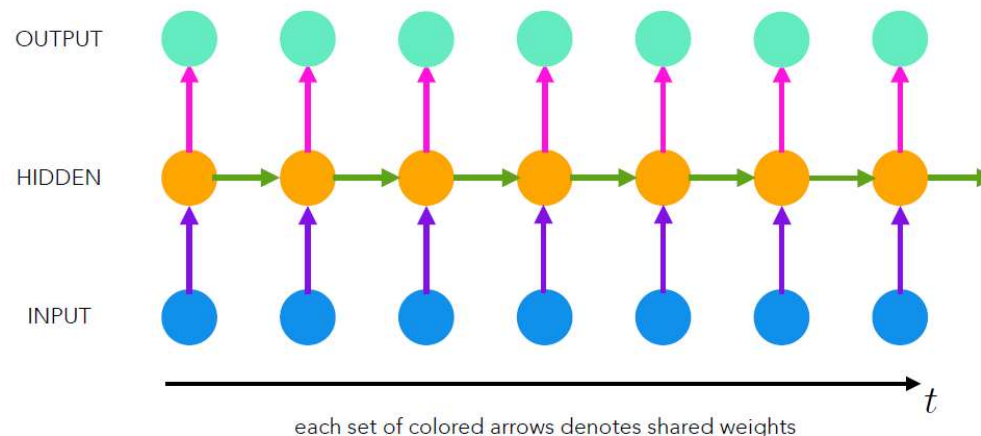
Inputs of variable size

- e.g in sentiment analysis, the input sentences can be of variable length
- Standard neural networks can only handle data of a fixed input size

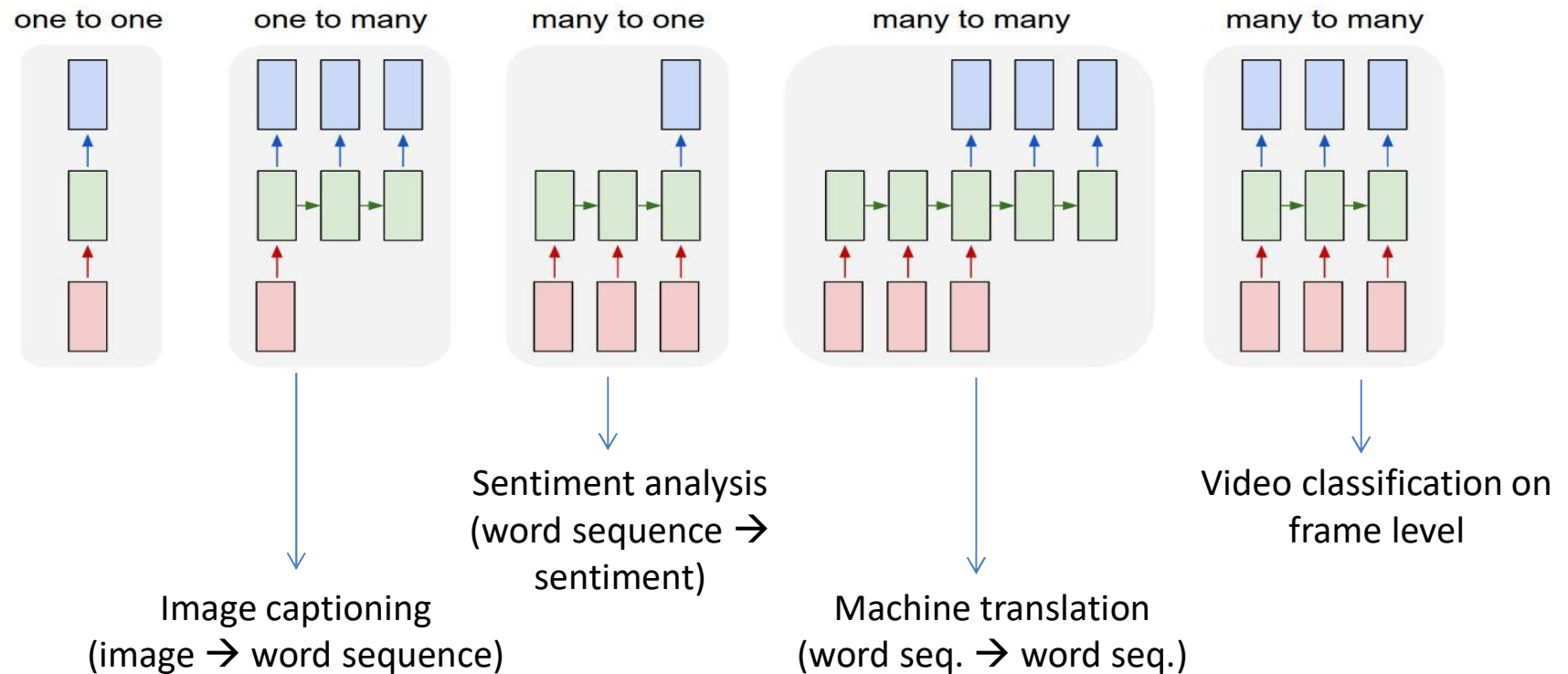


Recurrent Neural Network

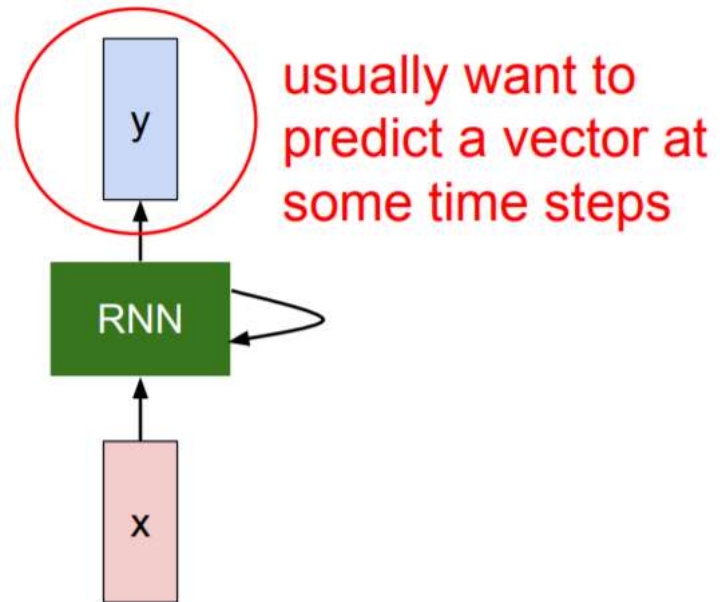
- Use output of the network as a partial input
- Suited for processing sequence data
 - Language: sequence of words
 - Acoustic: sequence of phonemes
 - Movie: sequence of images
 - Genes



RNN: make use of sequential information



Recurrent Neural Network

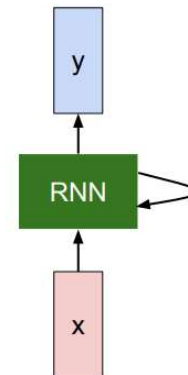


Recurrent neural network

- We can process a sequence of vectors x by applying a recurrence formula at every time step:

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

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



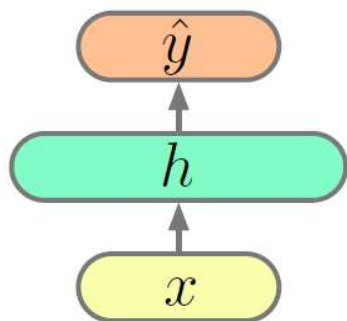
Notice: the same function and the same set of parameters are used at every time step

RNN

Feed-forward network

$$h = g(Vx + c)$$

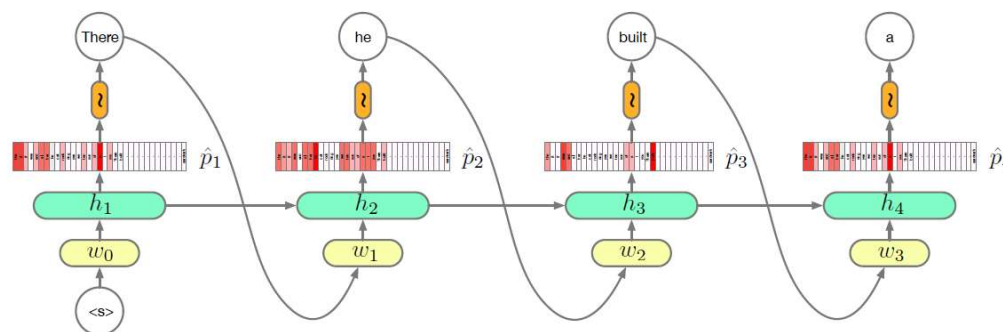
$$\hat{y} = Wh + b$$



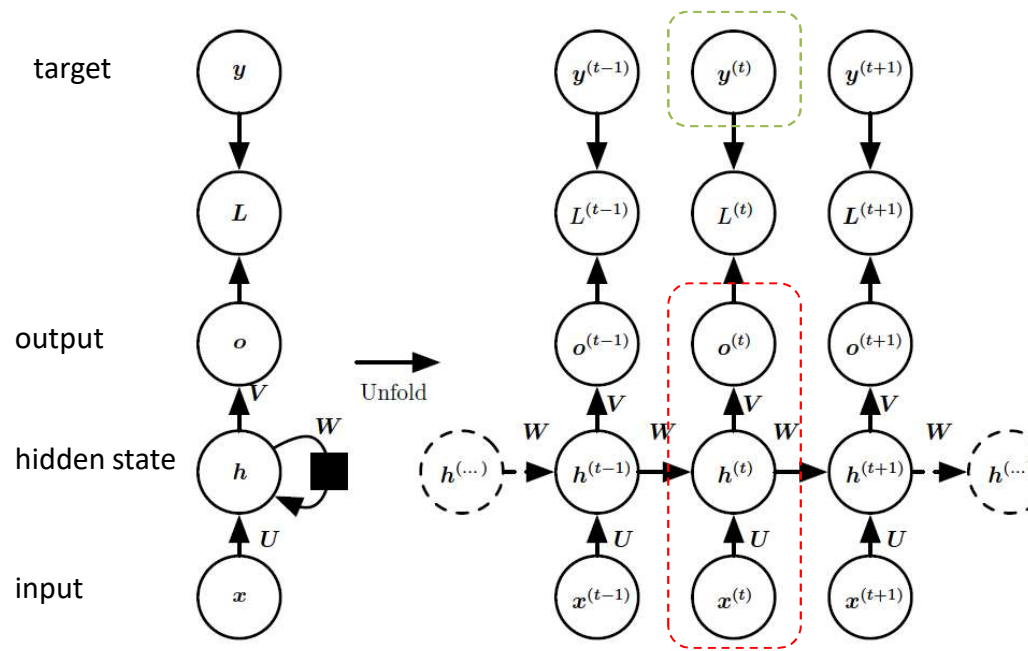
Recurrent network

$$h_n = g(V[x_n; h_{n-1}] + c)$$

$$\hat{y}_n = Wh_n + b$$



Recurrent Hidden Units



$$a^{(t)} = b + \mathbf{W}h^{(t-1)} + \mathbf{U}x^{(t)}$$

$$h^{(t)} = \sigma_1(a^{(t)}) : \text{hidden state}$$

$$o^{(t)} = c + \mathbf{V}h^{(t)}$$

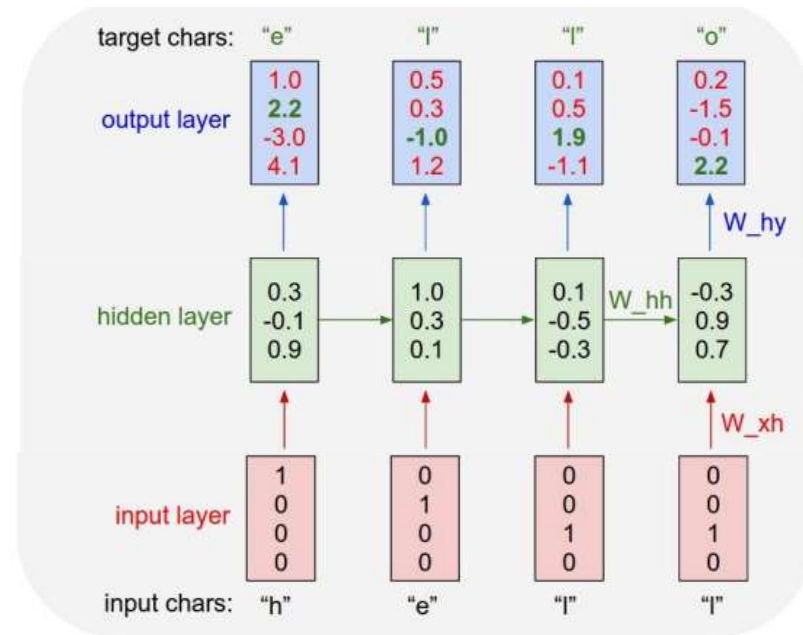
$$\hat{y}^{(t)} = \sigma_2(o^{(t)}) : \text{estimated target}$$

$$y^{(t)} : \text{true target}$$

$$L^{(t)} = \text{loss at } t$$

Example: Character-level Language Model

- Vocabulary
- [h,e,l,o]
- Trained to Predict the next character
- Example training sequence
- “hello”



- The output layer contains confidences the RNN assigns for the next character
- We want the green numbers to be high and red numbers to be low

Word-level language model

We want to train a language model

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

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

We want these probabilities to be high:

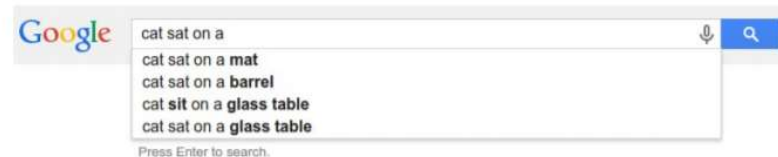
$P(\text{cat} \mid [<S>])$

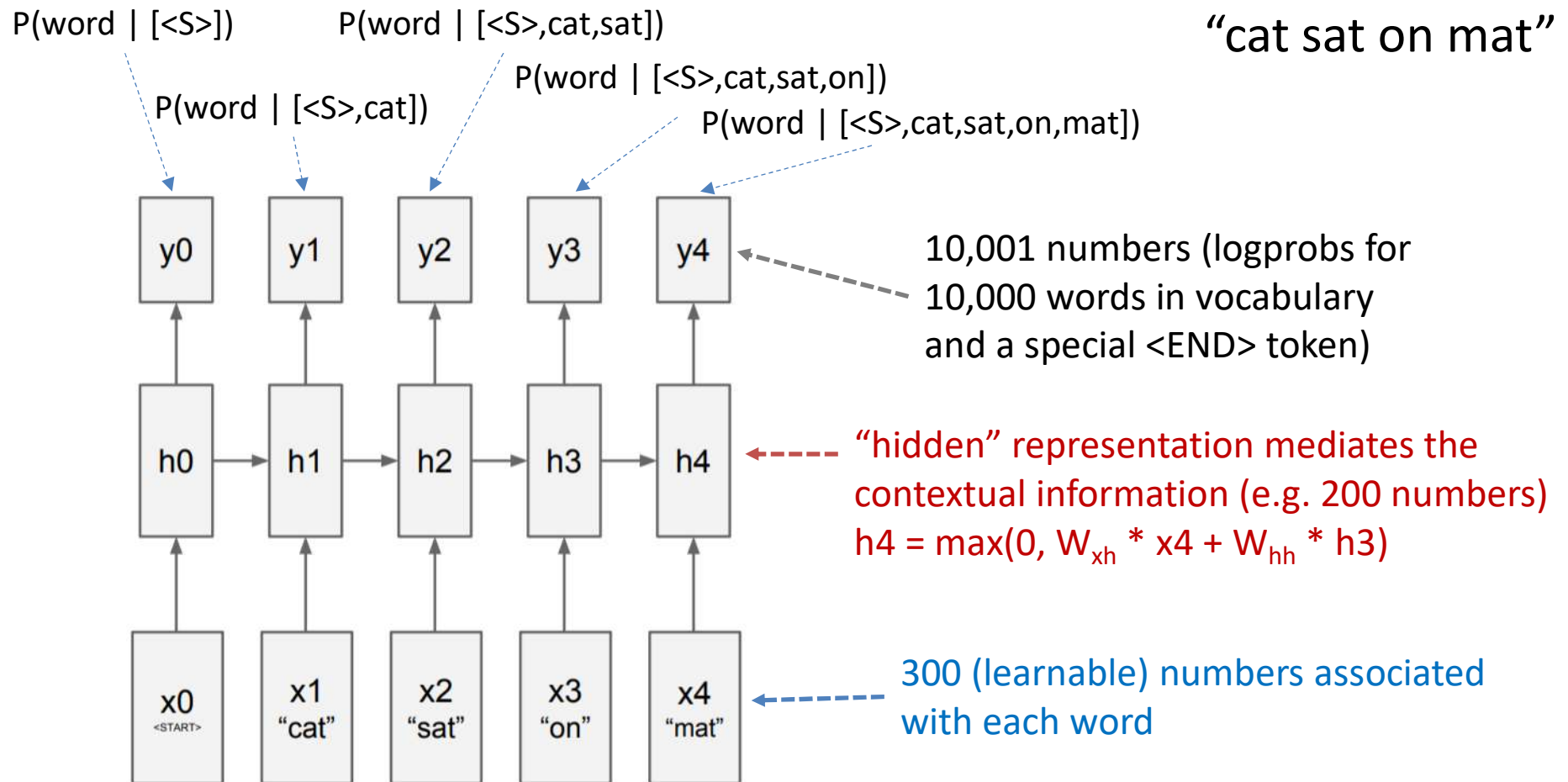
$P(\text{sat} \mid [<S>, \text{cat}])$

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

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

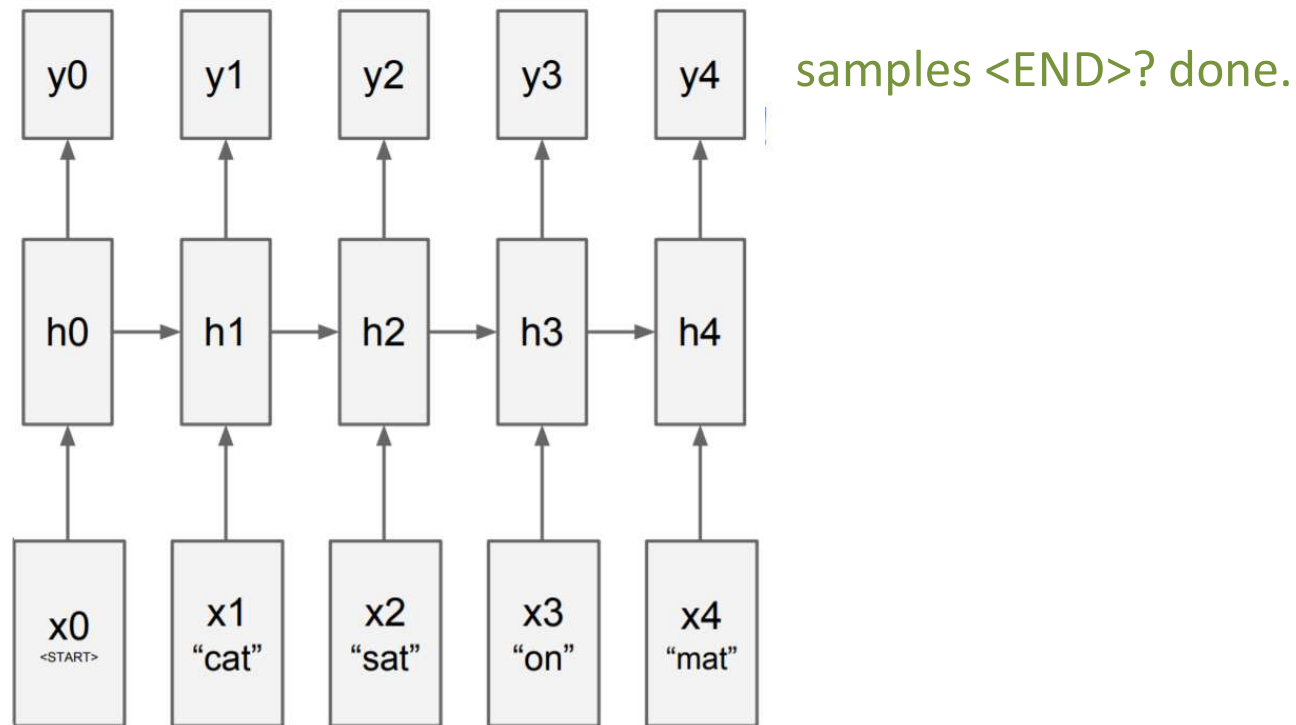
Word-level language model. Similar to:





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

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



Example: char-RNN

at first:

tyntd-iafhatawiaoirdemot lytdws e ,tfti, astai f ogoh eoase rrranbyne 'nhthnee e
plia tklrge 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 shuwv 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 offer.

↓
train more

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

입력: 셰익스피어의 모든 희곡(4.4MB)

RNN 모델: LSTM[512] * 3 layer

PANDARUS:

Alas, I think he shall be come approached and the day
When little strain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:

They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:

Well, your wit is in the care of side and that.

Second Lord:

They would be ruled after this chamber, and
my fair nudes begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.

Clown:

Come, sir, I will make did behold your worship.

VIOLA:

I'll drink it.

VIOLA:

Why, Salisbury must find his flesh and thought
That which I am not apt, not a man and in fire,
To show the reining of the raven and the wars
To grace my hand reproach within, and not a fair are hand,
That Caesar and my goodly father's world;
When I was heaven of presence and our fleets,
We spare with hours, but cut thy council I am great,
Murdered and by thy master's ready there
My power to give thee but so much as hell:
Some service in the noble bondman here,
Would show him to her wine.

KING LEAR:

O, if you were a feeble sight, the courtesy of your law,
Your sight and several breath, will wear the gods
With his heads, and my hands are wonder'd at the deeds,
So drop upon your lordship's head, and your opinion
Shall be against your honour.

입력: LaTeX 문서(16MB)

```
\begin{proof}
We may assume that  $\mathcal{I}$  is an abelian sheaf on  $\mathcal{C}$ .
\item Given a morphism  $\Delta : \mathcal{F} \rightarrow \mathcal{I}$ 
is an injective and let  $\mathcal{Q}$  be an abelian sheaf on  $X$ .
Let  $\mathcal{F}$  be a fibered complex. Let  $\mathcal{C}$  be a category.
\begin{enumerate}
\item \hyperref[setain-construction-phantom]{Lemma}
\label{lemma-characterize-quasi-finite}
Let  $\mathcal{F}$  be an abelian quasi-coherent sheaf on  $\mathcal{C}$ .
Let  $\mathcal{F}$  be a coherent  $\mathcal{O}_X$ -module. Then
 $\mathcal{F}$  is an abelian catenary over  $\mathcal{C}$ .
\item The following are equivalent
\begin{enumerate}
\item  $\mathcal{F}$  is an  $\mathcal{O}_X$ -module.
\end{enumerate}
\end{enumerate}
\end{proof}
```

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 $X_{\text{étale}}$ we have

$$\mathcal{O}_X(\mathcal{F}) = \{ \text{morph}_1 \times_{\mathcal{O}_X} (\mathcal{G}, \mathcal{F}) \}$$

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

Lemma 0.2. *This is an integer 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 $U \subset X$ 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. □

```

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 & 0x00000000ffffffff) & 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 ");
}

```

Generated C code

리눅스 소스 코드(474MB)를 학습한 후, 유사한 C 코드를 생성

<https://medium.com/@samim/obama-rnn-machine-generated-political-speeches-c8abd18a2ea0#.3n1f7i68f>



samim

Follow

Compassionate Alien & Thinker. Game, Music, Creativity + A.I. experimenter. Code magician and Narrative e...
Jun 4, 2015 · 8 min read

Obama-RNN—Machine generated political speeches.

<https://gist.github.com/nylki/1efbaa36635956d35bcc>

```
neural net cooking recipes.txt
1  MMMMM
2
3  MMMMM----- Recipe via Meal-Master (tm) v8.05
4
5      Title: BARBECUE RIBS
6  Categories: Chinese, Appetizers
7      Yield: 4 Servings
8
9      1 pk Seasoned rice
10     1 Beer -- cut into
11     -cubes
12     1 ts Sugar
13     3/4 c Water
14     Chopped finels,
15     -up to 4 tblsp of chopped
16     2 pk Yeast Bread/over
```

<https://highnoongmt.wordpress.com/2015/05/22/lisls-stis-recurrent-neural-networks-for-folk-music-generation/>

Lisl's Stis.

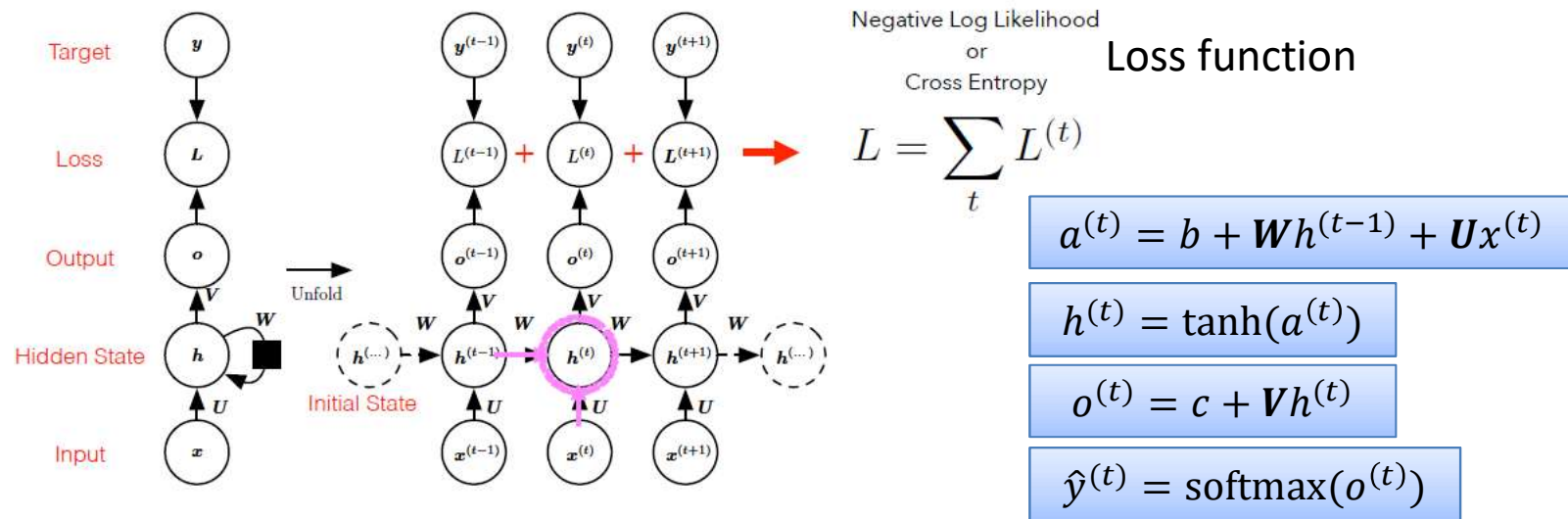


Perhaps the pickup is more of a grace note, but it is clear that the 3/8 time signature is not correct. The key signature works, and the IV-V-I resolution is good with the octave jump down. Here is another, named "Quirch cathp'3b (The Nille L' theys Lags Bollue's)".

Quirch cathp'3b
The Nille L' theys Lags Bollue's



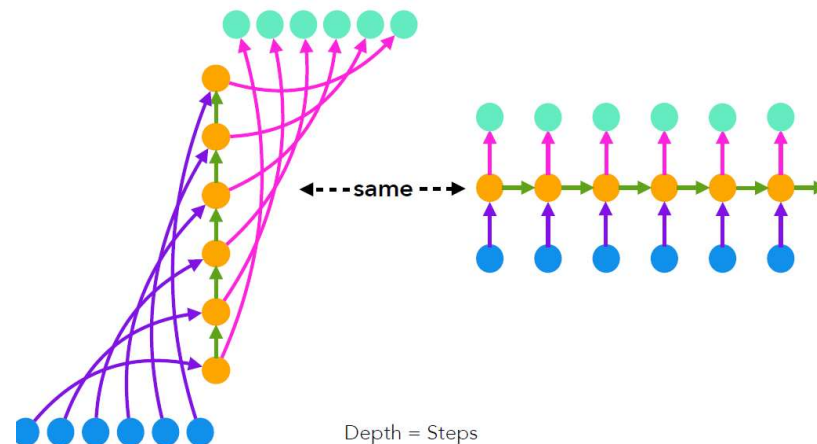
Training RNN



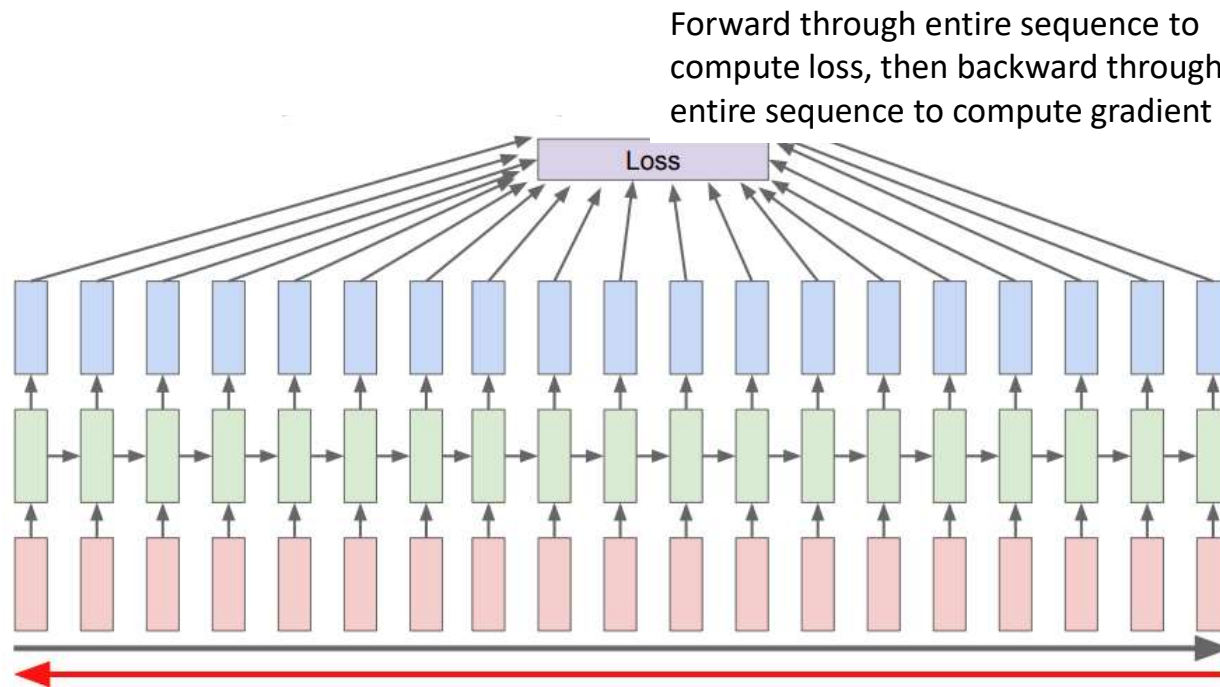
Note: need to calculate gradient of b, c, W, U, V

Training RNN

- Basic recurrent networks are also a *special case* of standard neural networks with *skip connections* and *shared weights*
- Therefore we can use standard backpropagation

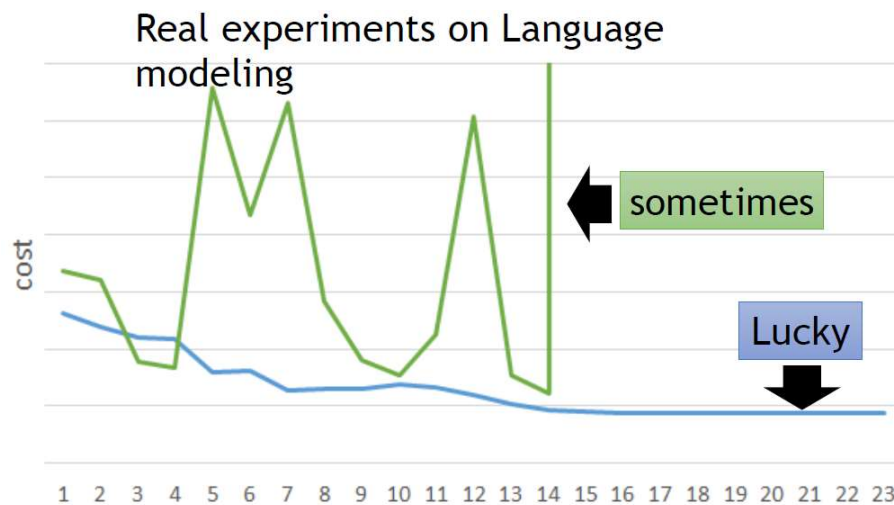


BackPropagation Through Time (BPTT)



RNN is hard to train

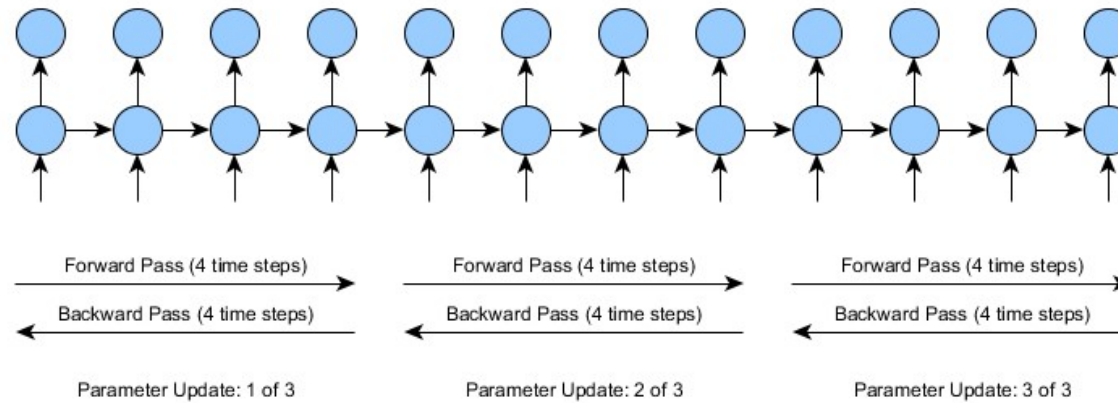
- RNN-based network is not always easy to learn



- Vanishing gradient problem
- Exploding gradient problem

Truncated BPTT

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




Practical measure

- Exploding gradients
 - Truncated BPTT (how many steps to BP)
 - Clip gradient at threshold
 - RMSProp to adjust learning rate
- Vanishing gradient (harder to detect)
 - Weight initialization (Identity Matrix for W)
 - ReLU as Activation
 - LSTM, GRU

Using Gate & Extra Memory

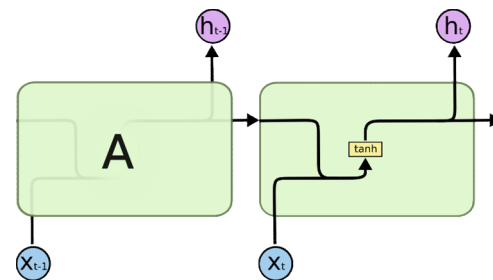
- Introducing gate
 - A vector
 - 1 or 0
 - If 1, info will be kept
 - If 0, info will be flushed
 - *Sigmoid* is an intuitive selection
 - To be learned by Neural Network



The diagram shows a gate operation. At the top, a red horizontal line is crossed by a blue circle with a red diagonal slash. Below this, the operation is represented as a vector multiplication:

$$\begin{array}{c} \text{before} \\ \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_j \\ a_n \\ \cdot \end{bmatrix} \end{array} \odot \begin{array}{c} \text{gate} \\ \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ \cdot \end{bmatrix} \end{array} = \begin{array}{c} \text{after} \\ \begin{bmatrix} a_1 \\ 0 \\ a_3 \\ 0 \\ 0 \\ \cdot \end{bmatrix} \end{array}$$

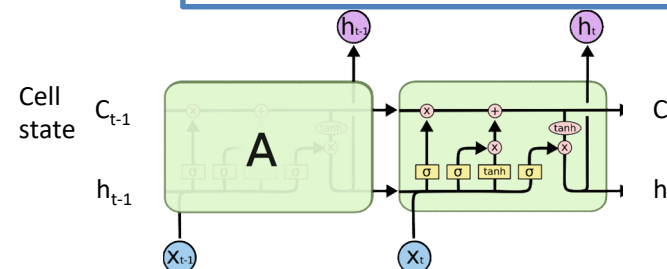
Long Short-Term Memory



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



e.g. $f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$



Three gates:

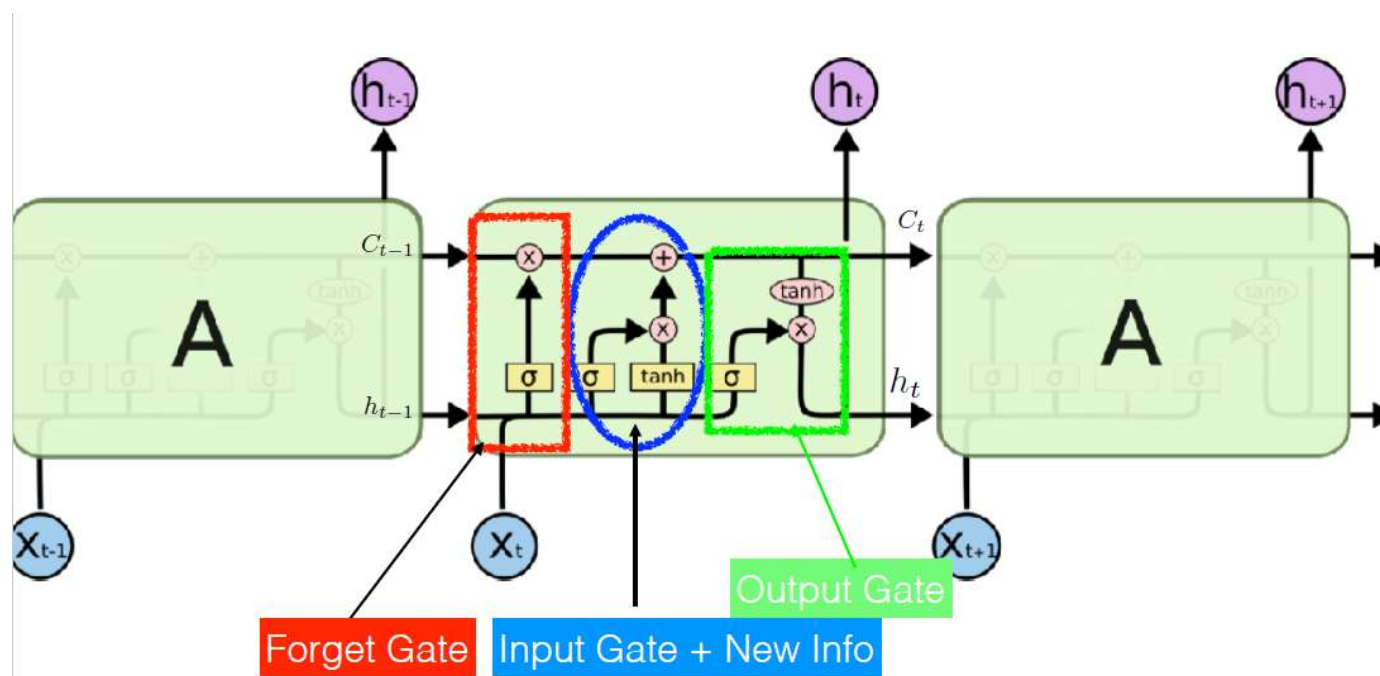
- Input
- forget
- output

$$\begin{aligned} f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f) && \text{forget gate} \\ i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i) && \text{input gate} \\ o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o) && \text{output gate} \end{aligned}$$

$$\begin{aligned} c_t &= f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) && \text{new cell memory} \\ h_t &= o_t \circ \sigma_h(c_t) && \text{new hidden} \end{aligned}$$

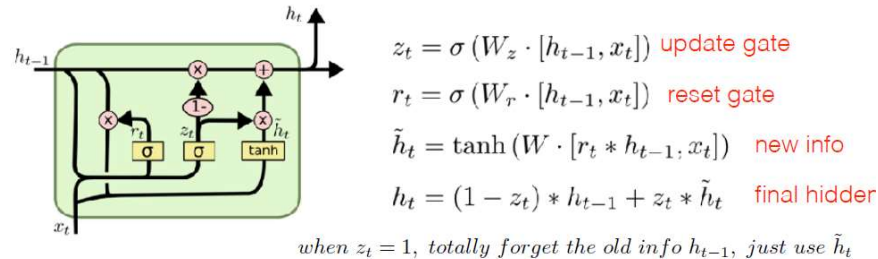
- Multiply the old state by f_t , forgetting the things we decided to forget
- Then add $i_t * g_t$, the new candidate values, scaled by how much we decided to update each state value.

LSTM



Gated Recurrent Units (GRU)

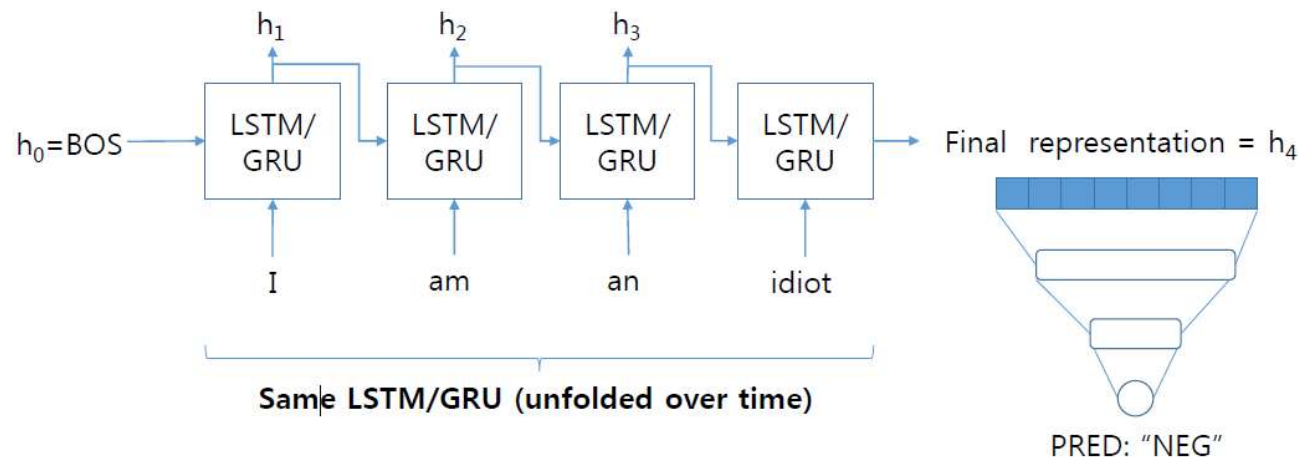
- LSTM is good but seems redundant
 - Do we need so many gates?
 - Do we need Hidden State And Cell Memory to remember?



- Combines forget and input gate into Update Gate
 - Merges Cell state and hidden state
- With fewer parameters than LSTM, faster to train

How LSTM/GRUs process sequential input

- Suppose we are classifying sentences into {POS, NEG} classes
 - Need to represent sentences as vectors
 - Sequentially feed in words into the LSTM and take the final hidden representation



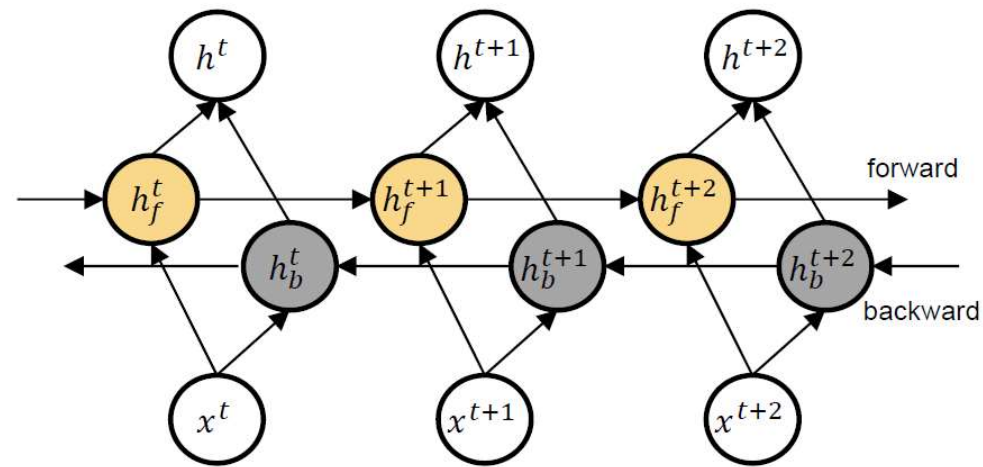
Advice on Training (gated) RNN

- Use an LSTM or GRU: *it makes your life so much simpler!*
- Initialize recurrent matrices to be orthogonal
- Initialize other matrices with a sensible (small) scale
- Initialize forget gate bias to 1: *default to remembering*
- Use adaptive learning rate algorithms: *Adam, AdaDelta, ...*
- Clip the norm of the gradient
- Either only dropout vertically or learn how to do it right
- *Be patient!*

[Saxe et al., ICLR2014;
Ba, Kingma, ICLR2015;
Zeiler, arXiv2012;

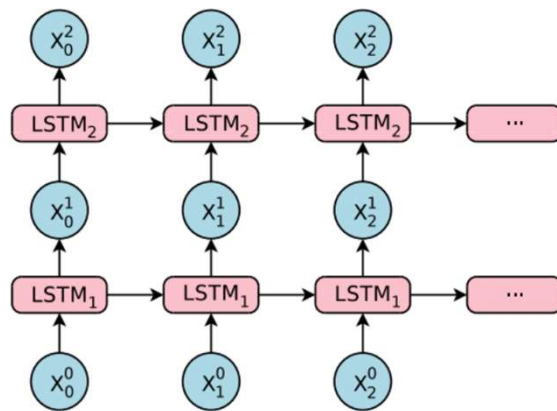
source: Oxford Deep Learning for NLP 2017

Bi-directional RNN

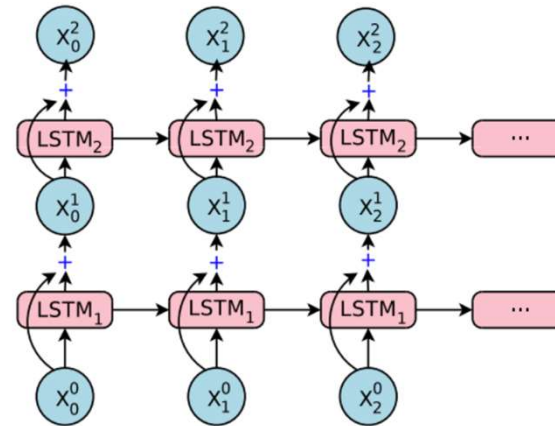


Speech recognition, handwriting recognition

Deep RNN

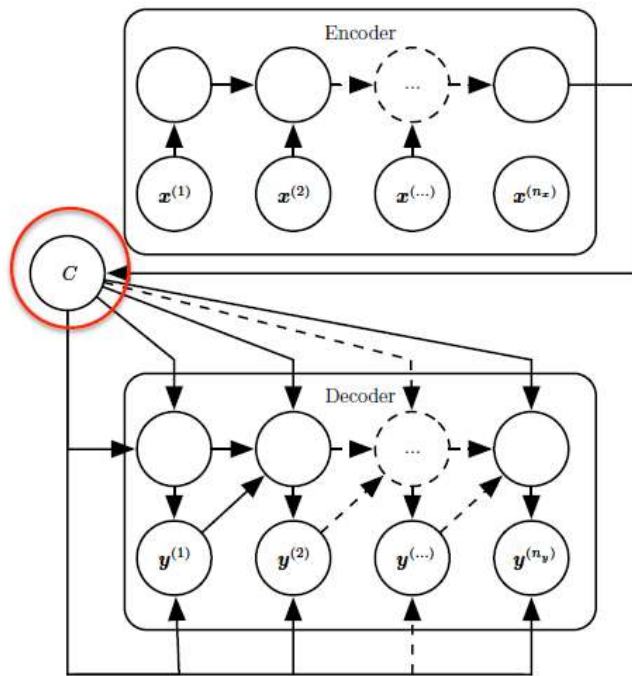


Usual stacking



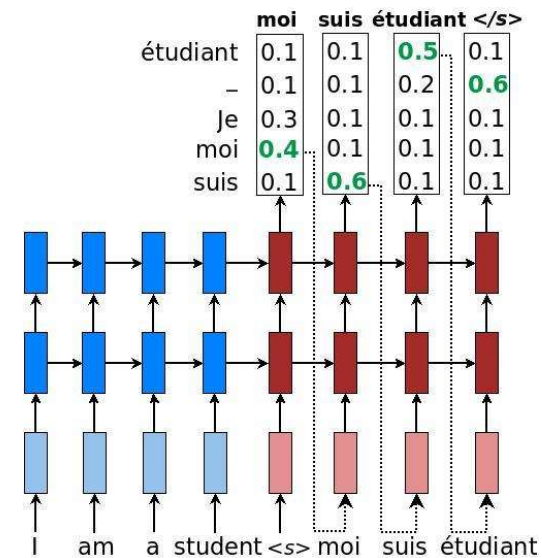
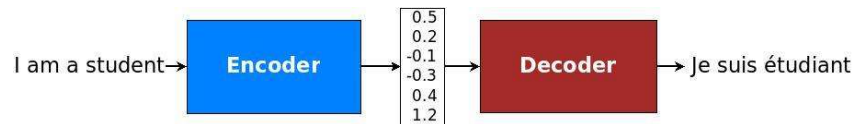
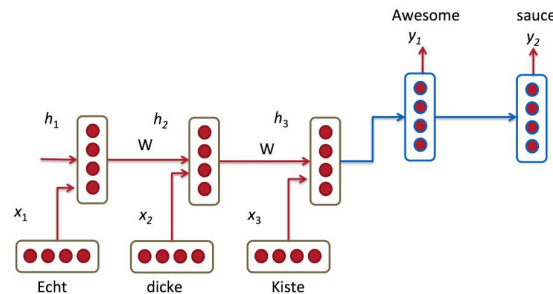
Residual connection

Sequence to sequence with RNN

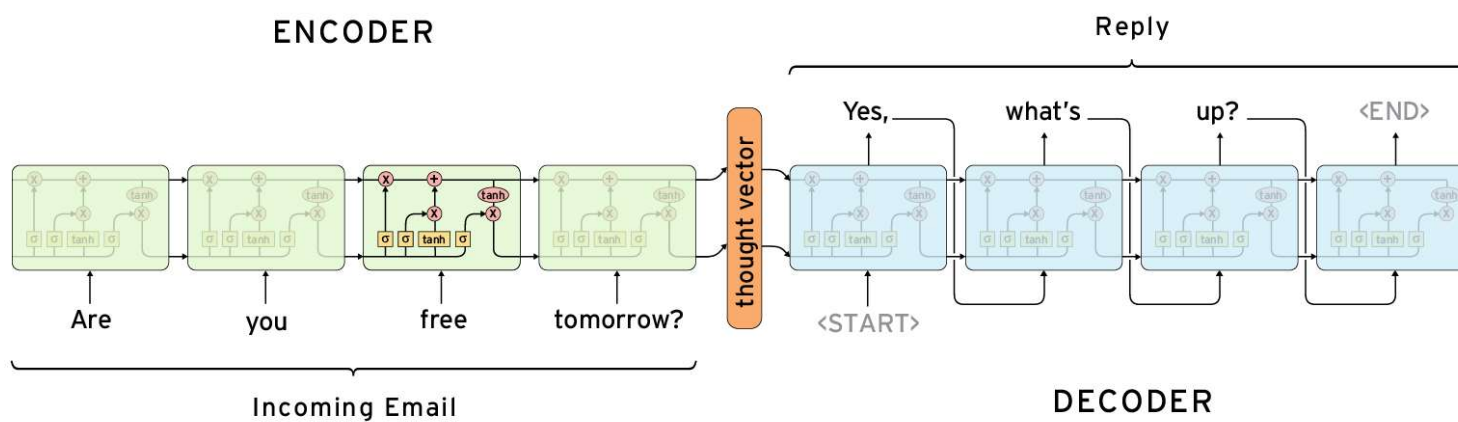


To learn the context Variable **C** which represents a semantic summary of the input sequence, and for later decoder RNN

Neural Machine Translation: seq2seq



Sequence to sequence



Google AI Experiment: Quick, Draw!

- <https://experiments.withgoogle.com/ai/quick-draw>



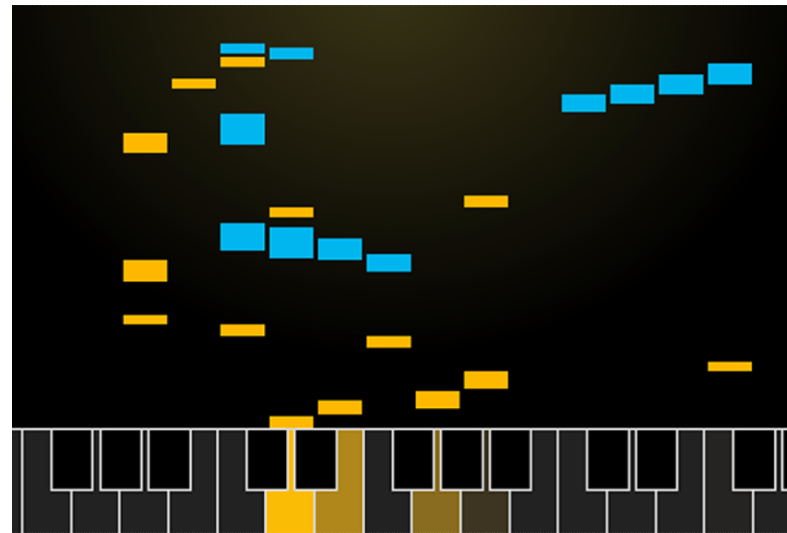
머신 러닝 기술이 학습을 통해 낙서를 인식할 수 있을까요?

여러분의 그림으로 머신 러닝의 학습을 도와주세요. Google은 머신 러닝 연구를 위해 [세계 최대의 낙서 데이터 세트](#)를 오픈소스로 공유합니다

시작하기

Google AI Experiments: AI Duet

- <https://experiments.withgoogle.com/ai/ai-duet>



MULTIMODAL APPLICATIONS

Image captioning

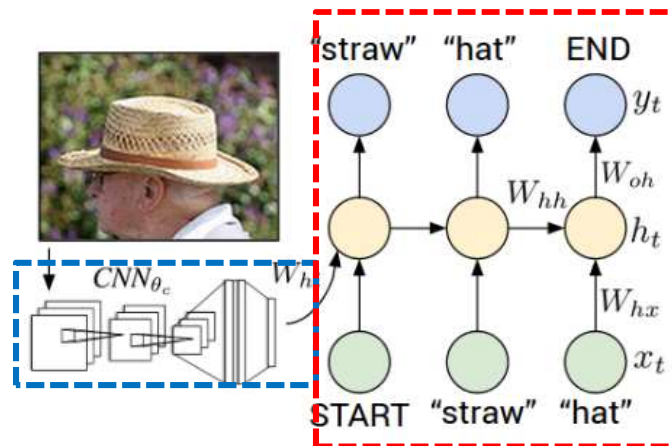
[image \rightarrow captions]

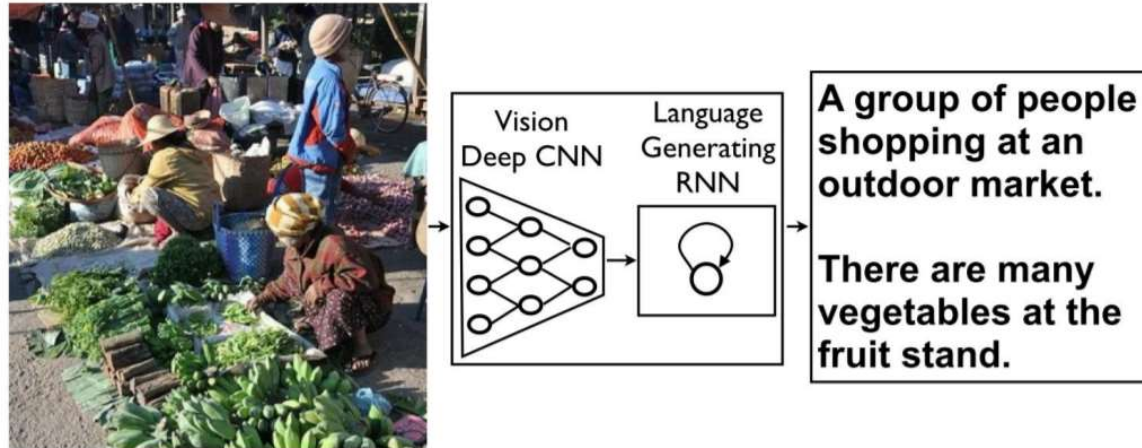
- **Encoder**

- CNN as an encoder: the input image is given to CNN to extract the features
- The last hidden state of the CNN is connected to the Decoder.

- **Decoder**

- RNN does language modelling up to the word level
- The first time step receives the encoded output from the encoder and also the <START> vector.





[Vinyals et al., "Show and Tell: A Neural Image Caption Generator", CVPR 2015]

Datasets

- [Common Objects in Context \(COCO\)](#). A collection of more than 120K with descriptions
- [Flickr 8K](#). A collection of 8 thousand described images taken from flickr.com.
- [Flickr 30K](#). A collection of 30 thousand described images taken from flickr.com.
- [Exploring Image Captioning Datasets](#), 2016

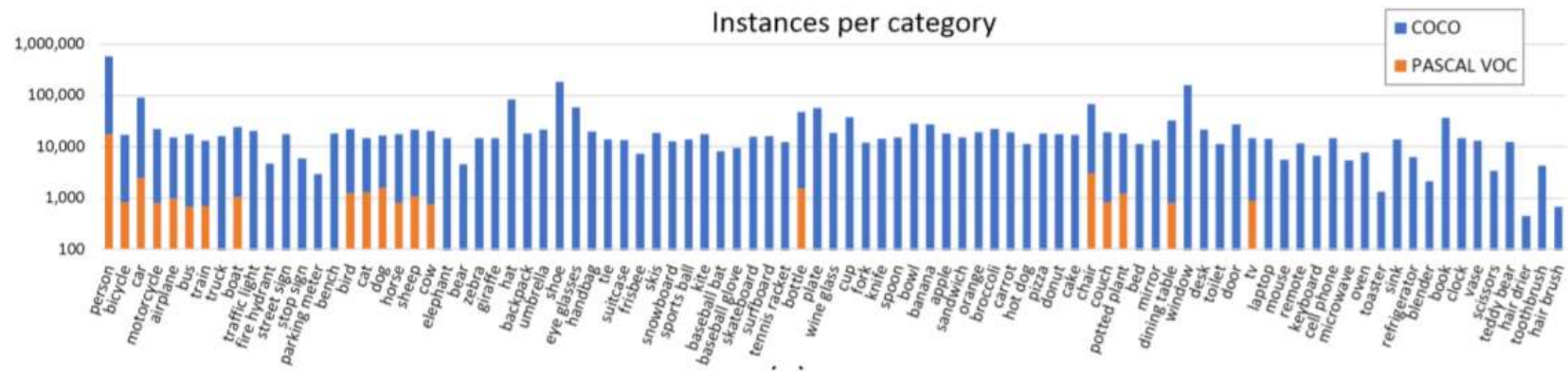
Flickr 8k, Flickr 30k



- A biker in red rides in the countryside.
- A biker on a dirt path.
- A person rides a bike off the top of a hill and is airborne.
- A person riding a bmx bike on a dirt course.
- The person on the bicycle is wearing red.

- 8k images in Flickr8k, >30k images in Flickr30k, with 5 descriptions per image.
- 21% images have static verbs like sit, stand, wear, look or no verbs.

Microsoft CoCo [Tsung-Yi Lin et al. 2014]



- 120k train + validation images
- Instance level segmentations labels with 91 object classes and 2.5M labelled instances.
- Standard benchmark for image caption generation task.

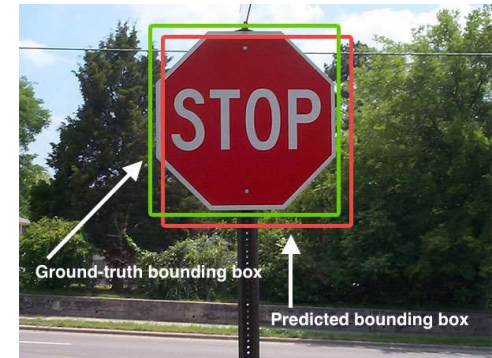
Metrics

- Image measures
- Text measures
 - Automatic measures
 - Human based measures

Image measures

- IoU (or Jaccard Index)
- Precision, Recall, F1 measure

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$



$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \cdot P \cdot R}{P + R}$$

Text metrics

- BLEU (Bilingual evaluation understudy)
 - Based on n-gram based precision
 - A measure of fluency rather than semantic similarity between two sentences
- Rouge (Recall Oriented Understudy of Gisting Evaluation)
- METEOR (Metric for Evaluation of Translation with Explicit ORdering)

Text metrics

- Human based measures
 - Measuring quality of a single best result
 - Success@k = % image sentence pairs for which at least one relevant result is found in the top-k list.
 - R-precision = average % of relevant items in the top-k list

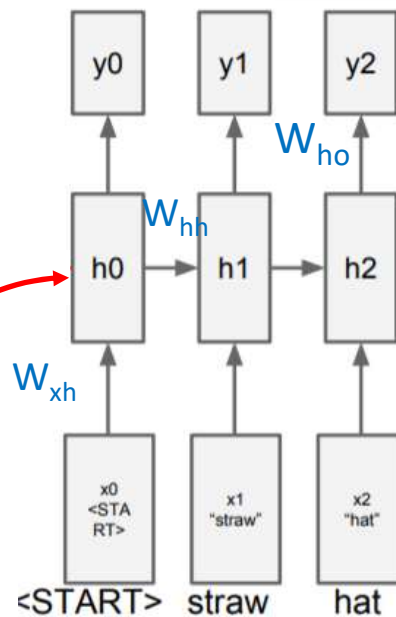
CNN



“straw hat”

training example

RNN

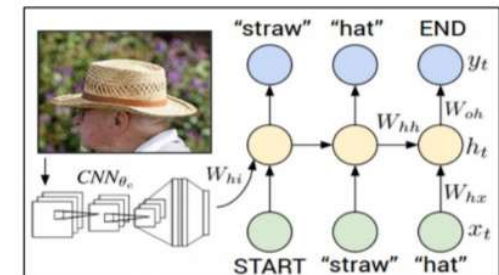


before:

$$h0 = \max(0, W_{xh} * x0)$$

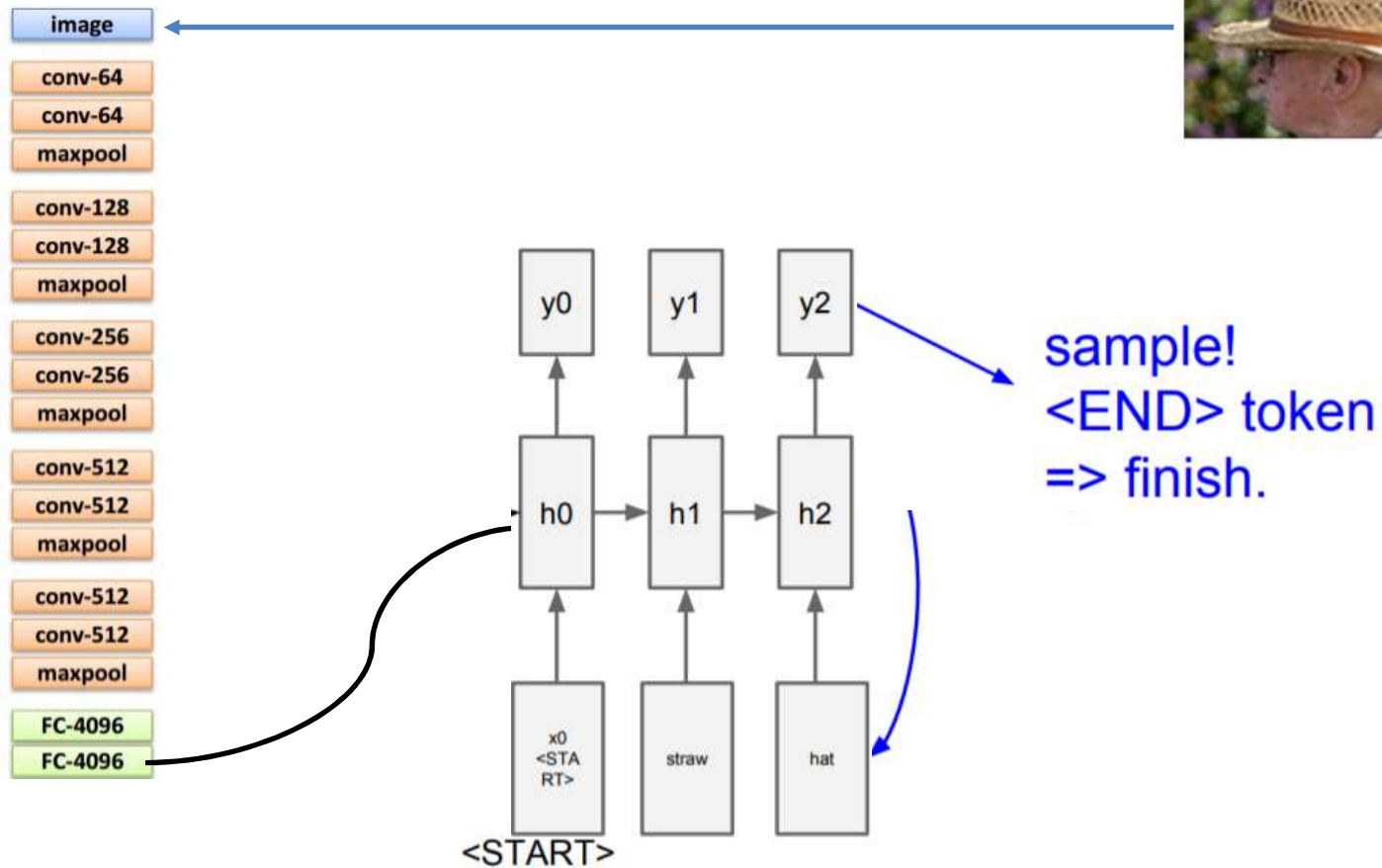
now:

$$h0 = \max(0, W_{xh} * x0 + W_{ih} * v)$$



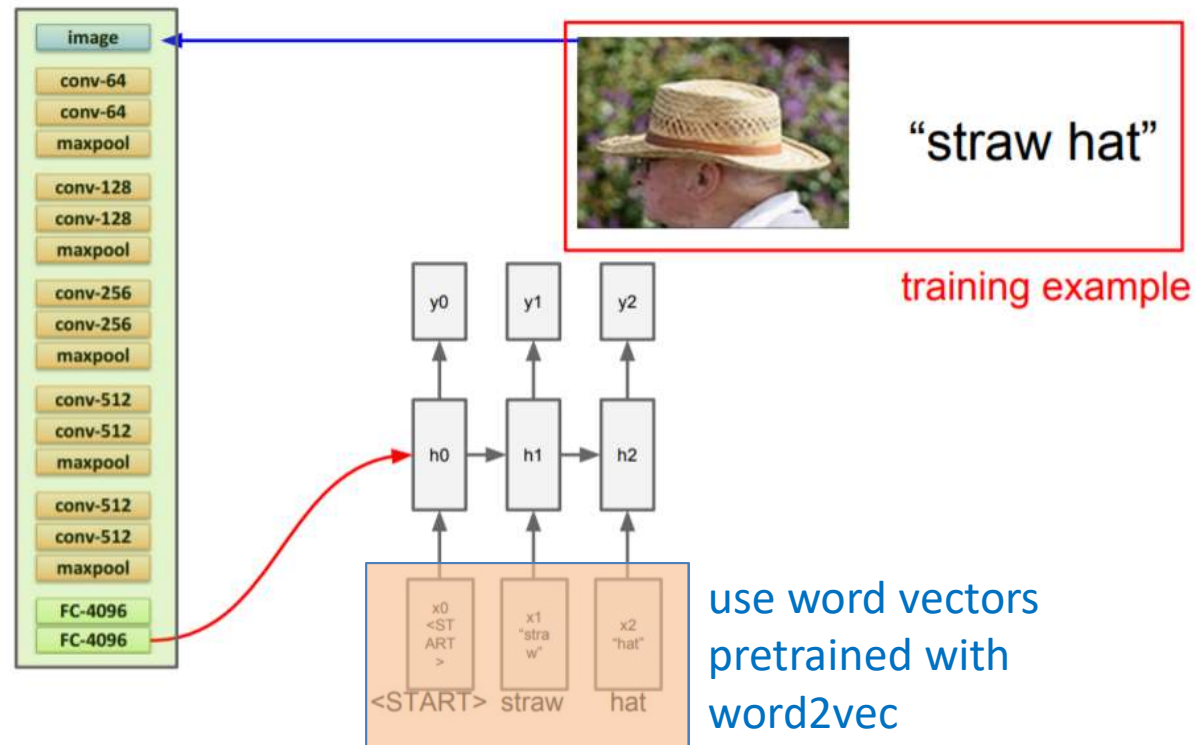


test image



+ Transfer learning

use weights
pretrained from
ImageNet





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



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



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

Visual Question Answering

- Task - Given an image and a question, answer the question (<http://www.visualqa.org/>)



What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?




Does it appear to be rainy?
Does this person have 20/20 vision?

Visual QA

- Real images
 - 200k MS COCO images
 - 600k questions
 - 6M answers
 - 1.8M plausible answers
- Abstract images
 - 50k scenes
 - 150k questions
 - 1.5M answers
 - 450k plausible answers

8653. COCO_train2014_000000458914

Image On/Off



Open-Ended/Multiple-Choice/Ground-Truth/Common-Sense

Q: Are these veggies or fruits?


Ground Truth Answers:

(1) fruits	(6) fruit
(2) fruits	(7) fruits
(3) fruits	(8) fruits
(4) fruits	(9) fruits
(5) fruits	(10) fruits

Q: What is in the white bowl?

Ground Truth Answers:

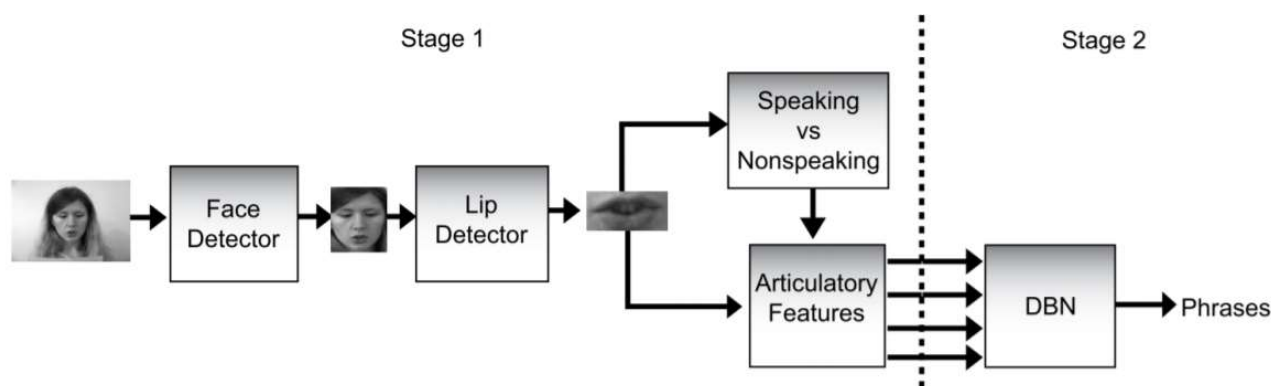
(1) strawberries	(6) strawberries
(2) strawberries	(7) strawberry
(3) strawberry	(8) strawberries
(4) strawberries	(9) strawberries
(5) fruits	(10) strawberries



Is this person expecting company?
What is just under the tree?

Visual Speech Recognition - Lipreading

- Vision → Language
- easier to evaluate
- Difficult problem as the mapping from a viseme and a phoneme is ambiguous (many sounds look the same on the lips)



[Saenko et al., 2005; Bear and Harvey 2016]

Speech synthesis

- Text → Sound
- Many intermediate building blocks
- End to end training approaches are becoming popular
- Works best for synthesizing a particular person's speech (with lots of training data for that individual)
- Very difficult to evaluate

Summary

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

