RNN Introduction

講者：Isaac

Outline

- Word2vect
- RNN Introduction
- LSTM/GRU Cell

Word2vect

Can machine understand the meaning of words?

"Dog" is close to "Cat"

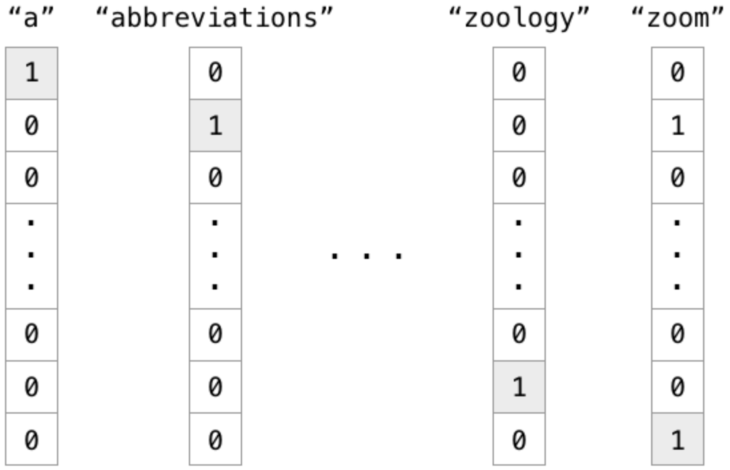
"Love" is close to "Hate"

"Korea" is close to "America"

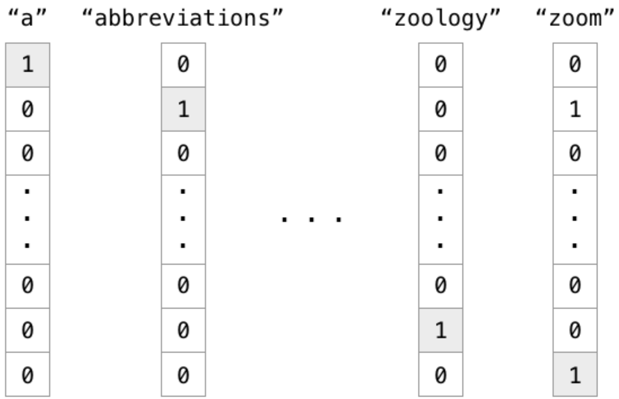
Word representation

What happen if

we let machine read a lot of articles and use one hot encoding on each word



Word representation



Drawbacks

1. Dog is not close to Cat
    - machine can not understand the meaning of word
2. Waste a lot of entity
    - most of entity are zero
    - dimension of vector = vocabulary size (very large)

Word representation

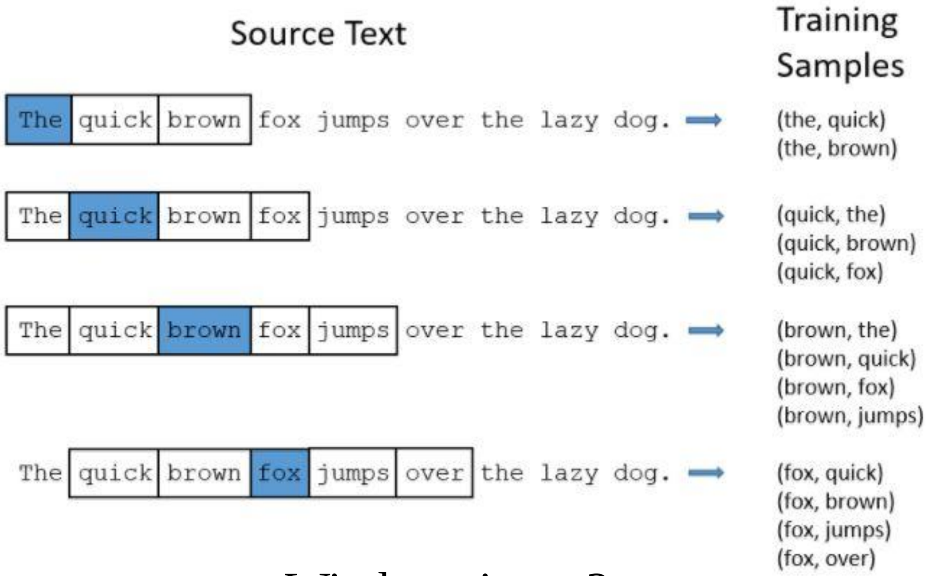
This is why we need better representation of words

Word representation

- We will introduce two method in this course
  - Skip-Gram
  - CBOW

Skip-Gram

Transform many sentences into training pairs



Window size = 2

Skip-Gram

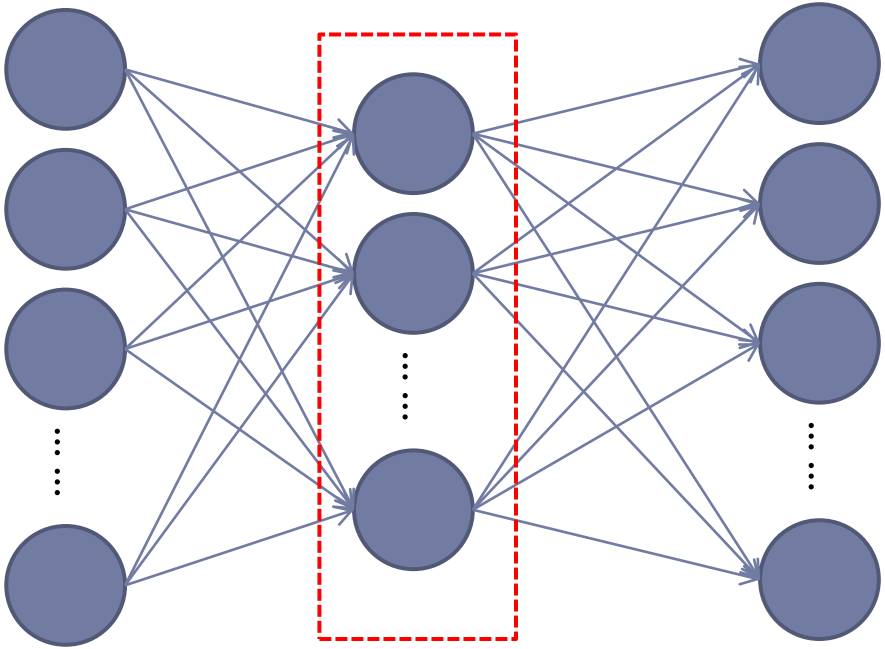
Transform to one hot encoding

(the, quick)

the $\begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$

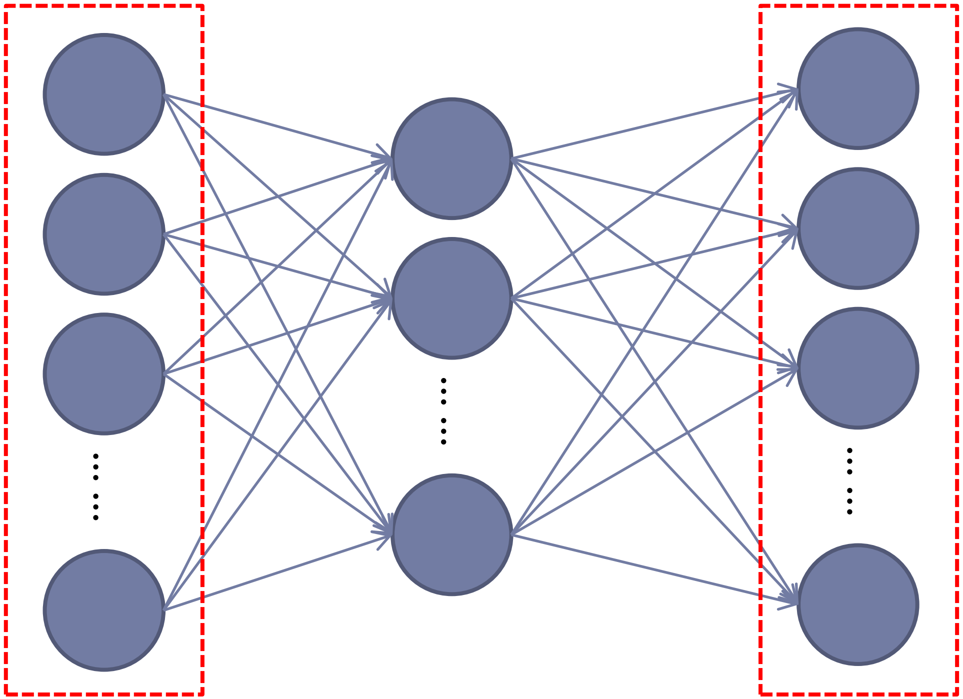
quick $\begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$

Skip-Gram



# of hidden node is less than input node

Skip-Gram



# of input/output node = # of vocabulary

Skip-Gram

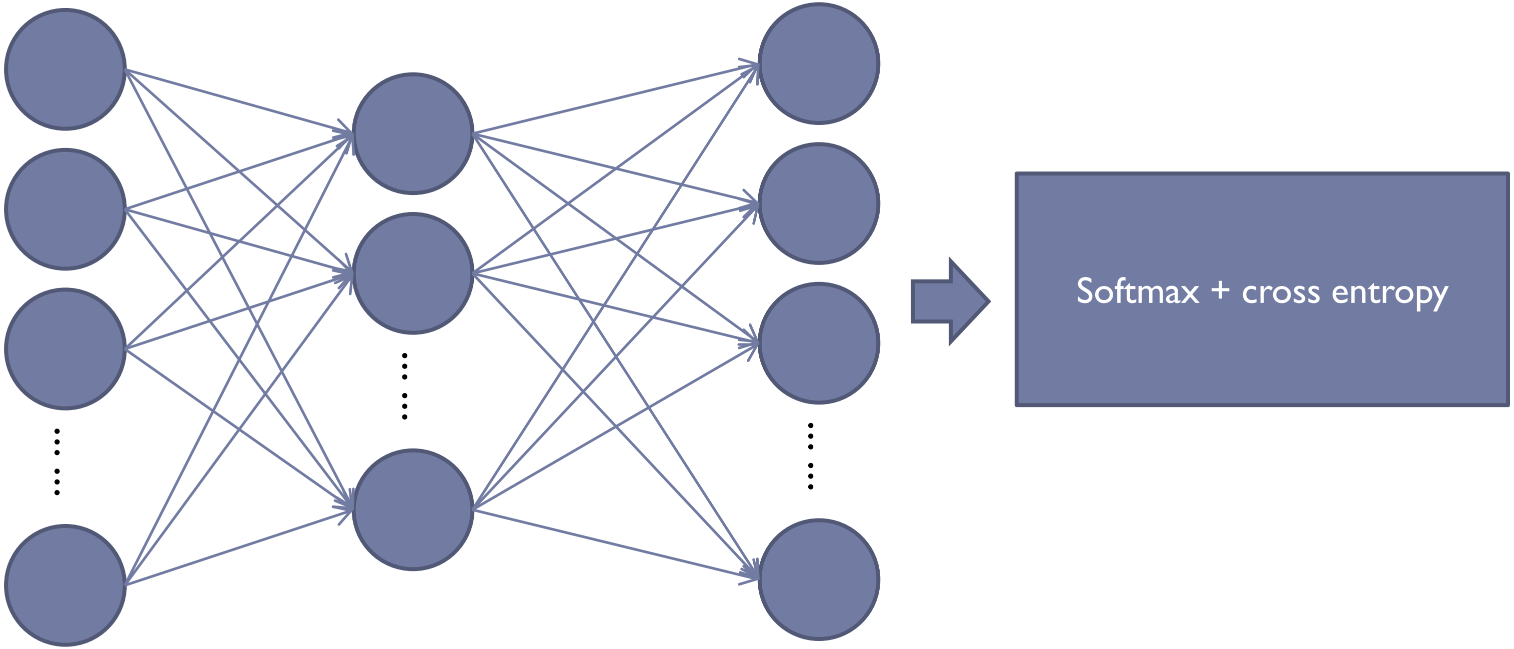
Training pairs

(the, quick)
(the, brown)

(quick, the)
(quick, brown)
(quick, fox)

(brown, the)
(brown, quick)
(brown, fox)
(brown, jumps)

(fox, quick)
(fox, brown)
(fox, jumps)
(fox, over)



Note that there is no activation function on hidden layer

Skip-Gram

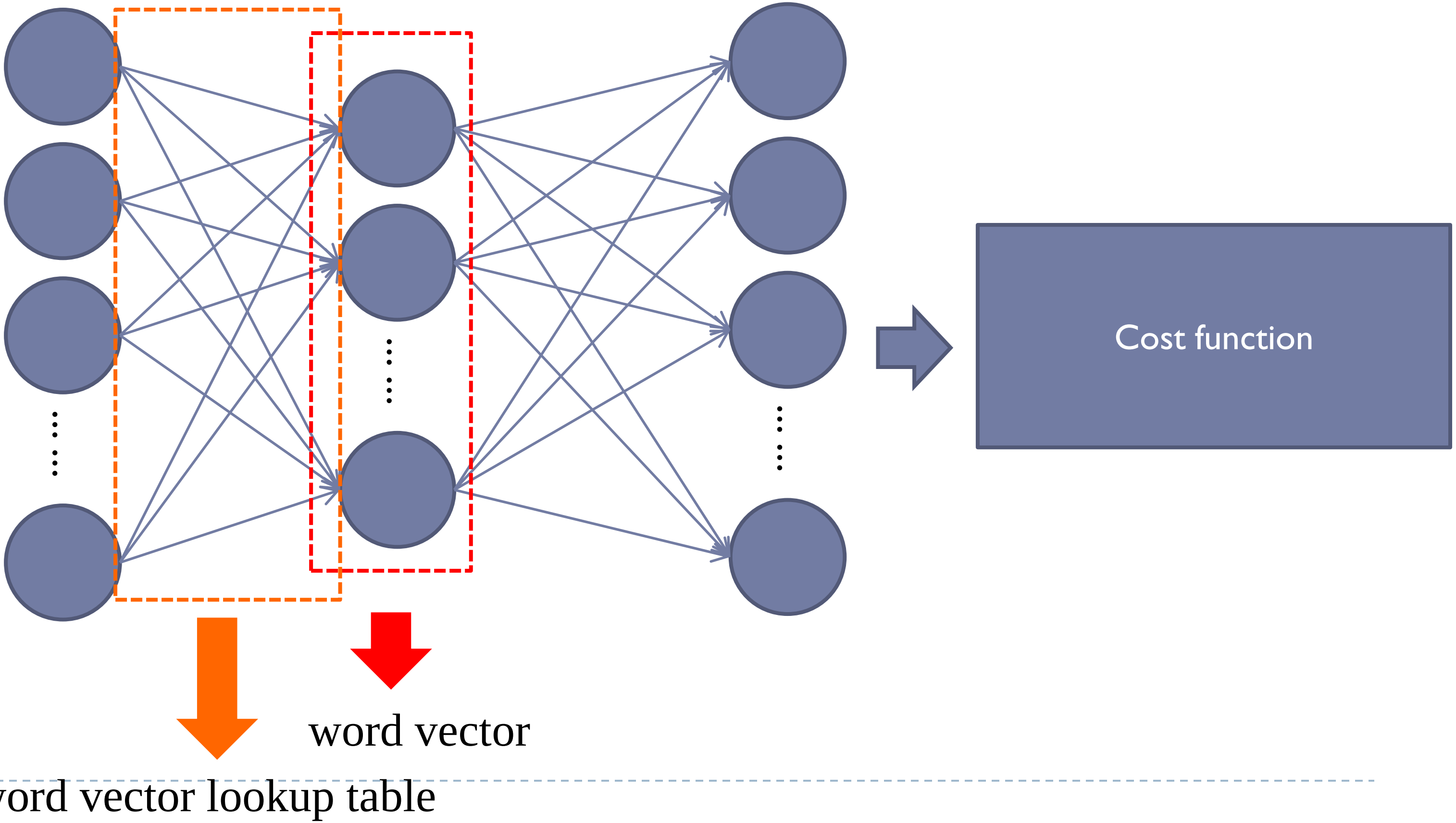
Training pairs

(the, quick)
(the, brown)

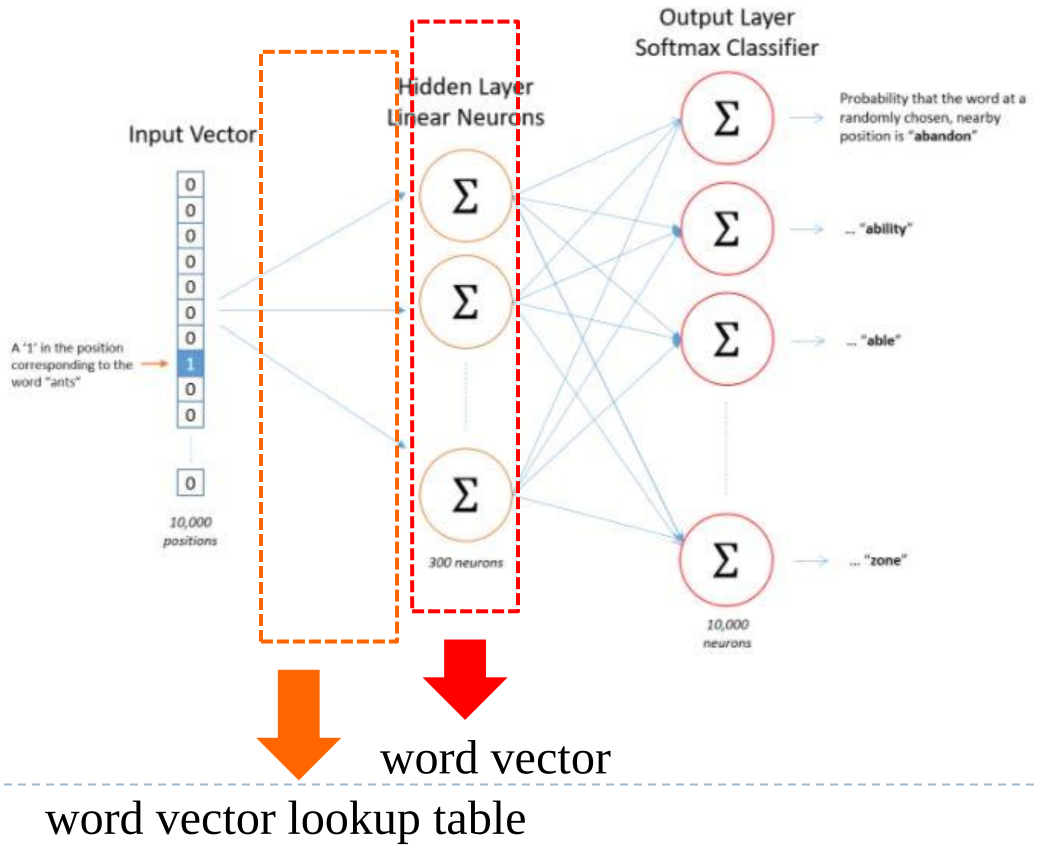
(quick, the)
(quick, brown)
(quick, fox)

(brown, the)
(brown, quick)
(brown, fox)
(brown, jumps)

(fox, quick)
(fox, brown)
(fox, jumps)
(fox, over)

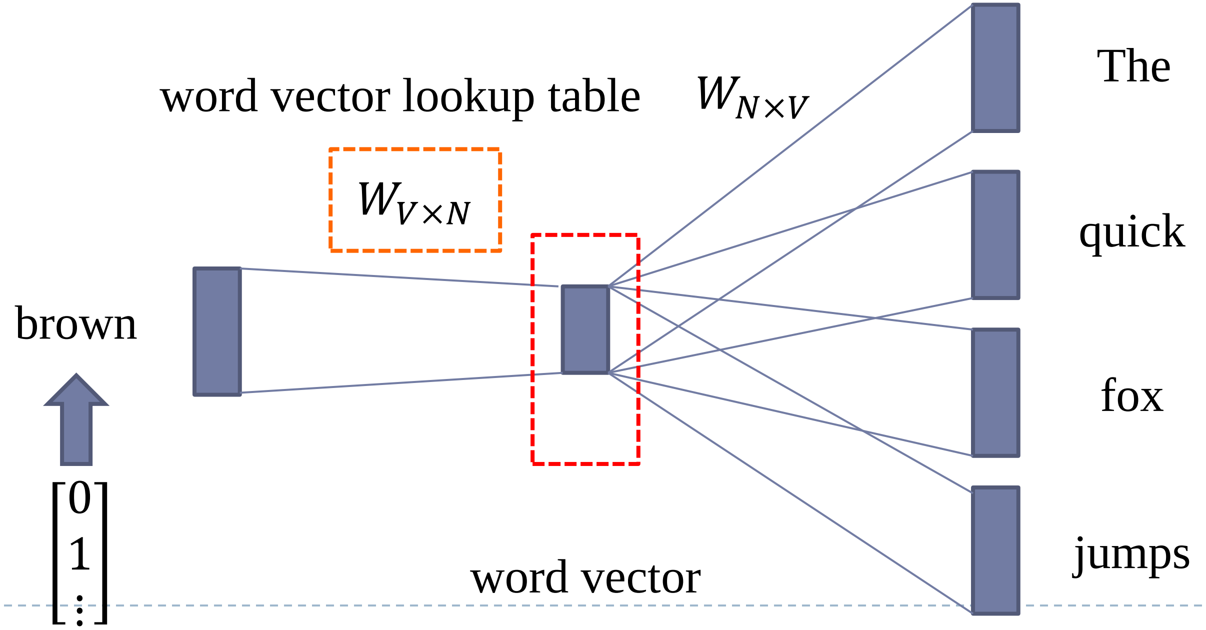


Skip-Gram

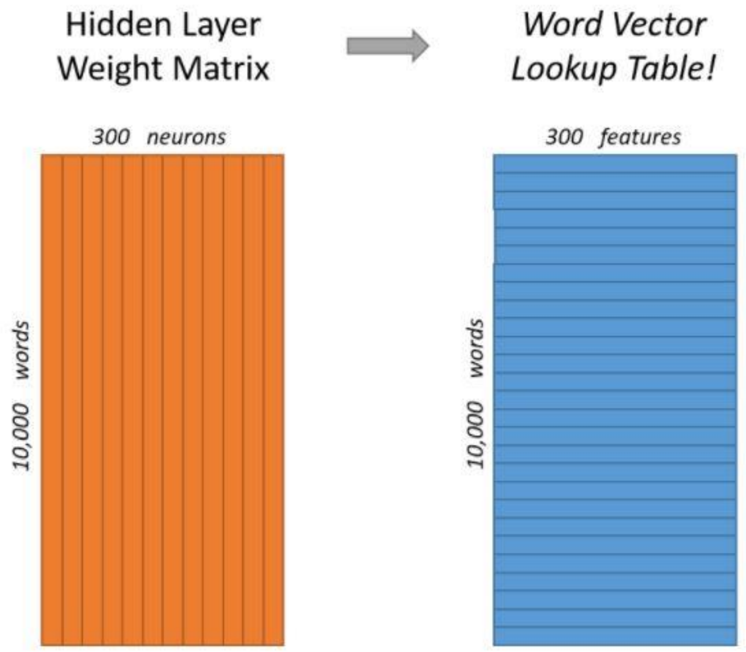

Output Layer
Softmax Classifier
Hidden Layer
Linear Neurons
Input Vector
Probability that the word at a randomly chosen, nearby position is "abandon"
... "ability"
... "able"
... "zone"
A '1' in the position corresponding to the word "ants"
10,000 positions
300 neurons
10,000 neurons
word vector
word vector lookup table

Skip-Gram

The quick brown fox jumps over the lazy dog



Skip-Gram

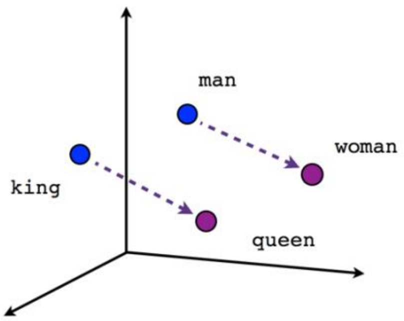


Assume vocabulary = 10000

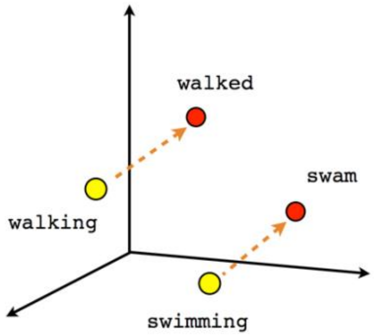
$$[0 \quad 0 \quad 0 \quad 1 \quad 0] \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ 10 & 12 & 19 \\ 11 & 18 & 25 \end{bmatrix} = [10 \quad 12 \quad 19]$$

Small example

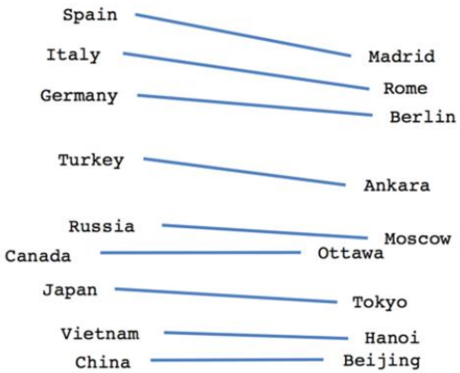
Skip-Gram



Male-Female



Verb tense



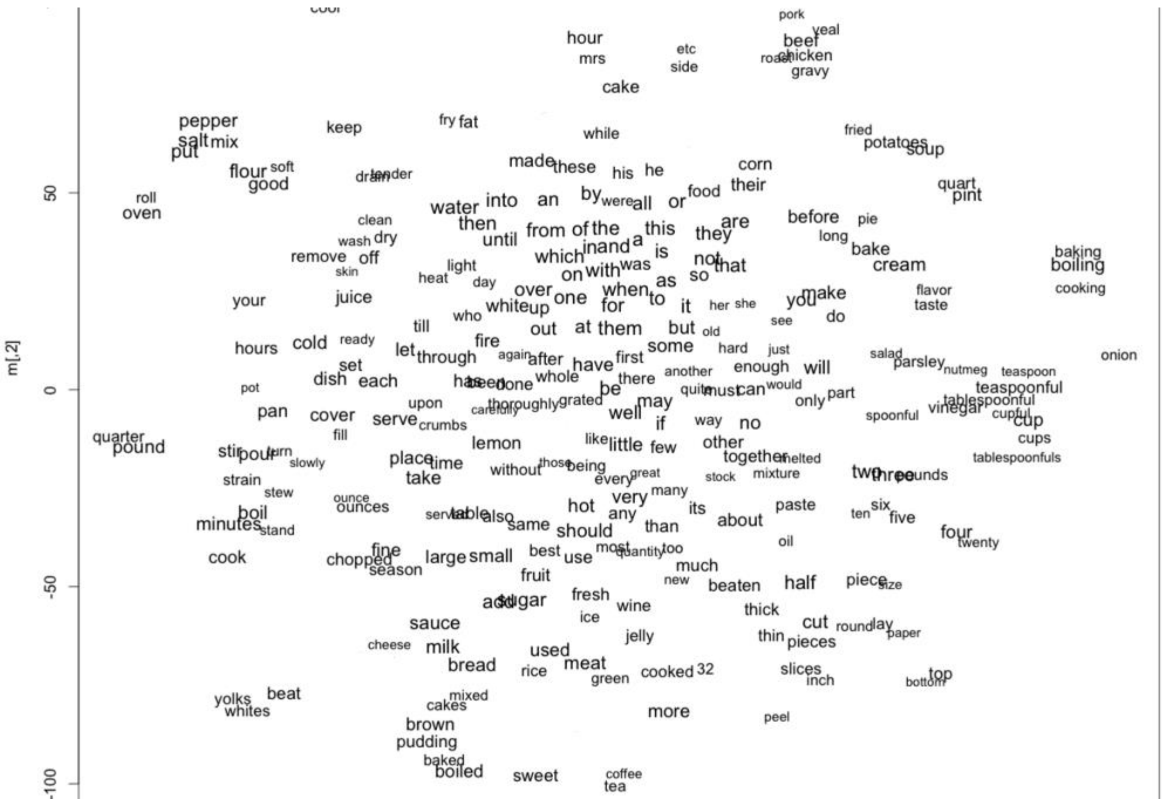
Country-Capital

vector(king) – vector(queen) = vector(man) – vector(woman)

vector(walking) – vector(walked) = vector(swimming) – vector(swam)

vector(Spain) – vector(Italy) = vector(Madrid) – vector(Rome)

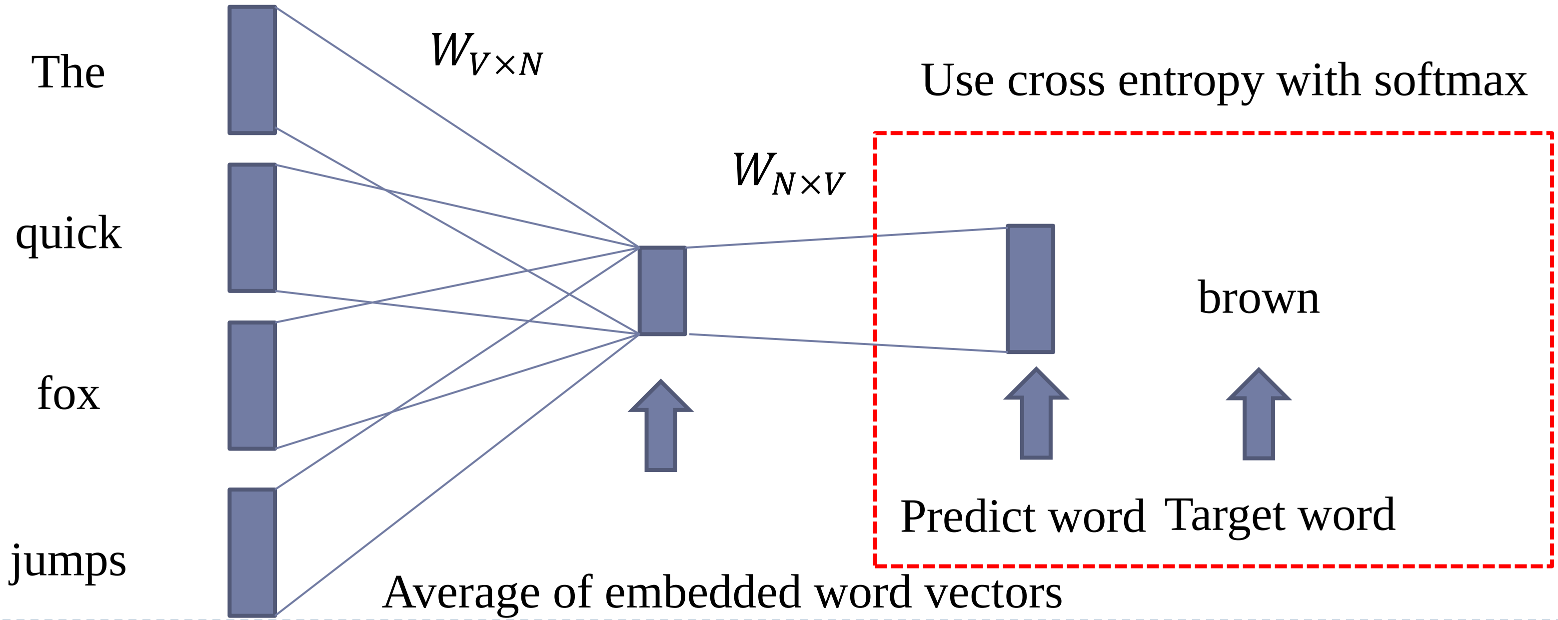
Skip-Gram



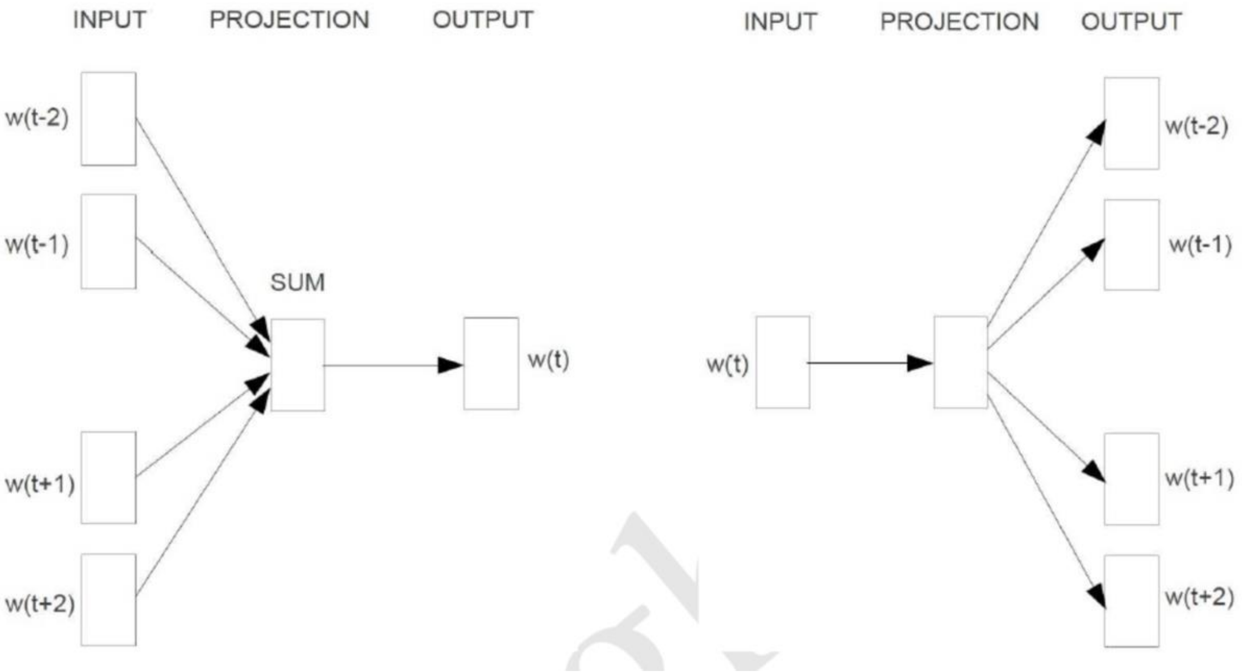
m[,2]
50
0
-50
-100
pork
veal
hour
etc
beef
mrs
chicken
side
gravy
cake
pepper
keep
fry fat
while
fried
mix
potatoes
put
soup
made
these
his
he
corn
flour soft
tender
good
their
quart
pint
roll
water
into
an
by
were
all
or
food
oven
clean
then
before
pie
until
from
of
the
this
they
are
wash
dry
long
bake
which
in
and
a
is
remove
off
not
that
cream
baking
boiling
light
was
skin
heat
on
with
as
so
day
over
when
flavor
cooking
your
juice
one
white
up
for
to
it
her
she
make
you
taste
who
do
till
see
out
at
them
but
old
fire
ready
hours
cold
let
through
some
hard
just
again
after
have
first
salad
onion
set
parsley
enough
will
nutmeg
teaspoon
another
dish
each
whole
there
be
would
teaspoonful
pot
done
can
part
upon
only
grated
may
tablespoonful
thoroughly
vinegar
pan
well
spoonful
cover
serve
cupful
crumbs
cup
way
no
if
quarter
fill
lemon
like
little
few
other
cups
pound
stir
pour
turn
place
together
melted
slowly
time
those
being
tablespoonfuls
without
great
two
three
pounds
strain
take
every
stock
mixture
stew
many
ounce
ounces
very
paste
hot
its
six
ten
boil
table
also
any
five
same
about
minutes
stand
should
than
four
twenty
oil
fine
best
cook
large
small
most
quantity
too
chopped
season
use
much
fruit
new
beaten
half
piece
size
add
sugar
fresh
wine
ice
thick
sauce
cut
round
lay
paper
jelly
thin
pieces
cheese
milk
used
bread
meat
cooked
32
slices
rice
green
inch
top
bottom
beat
yolks
mixed
whites
cakes
more
peel
brown
pudding
baked
boiled
sweet
coffee
tea

CBOW

The quick brown fox jumps over the lazy dog



Skip-Gram V.S. CBOW



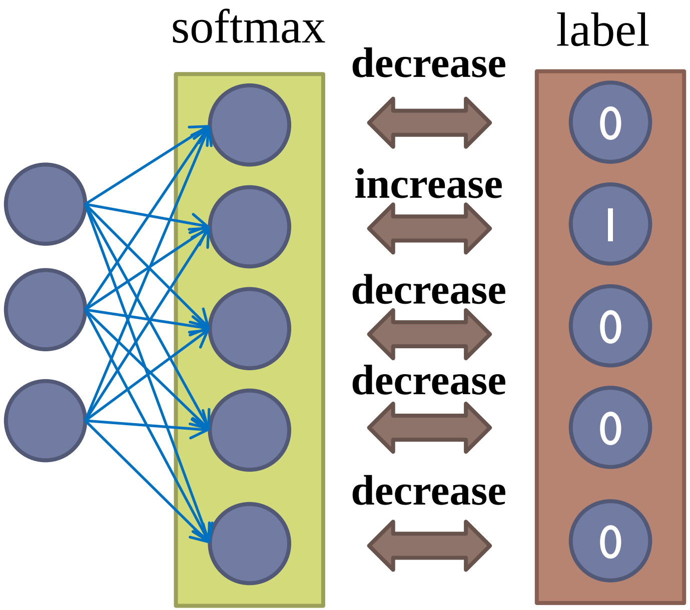
Cost function in Word2vect

- Softmax function
  - Computation expensive
  - Use sampled version of softmax function

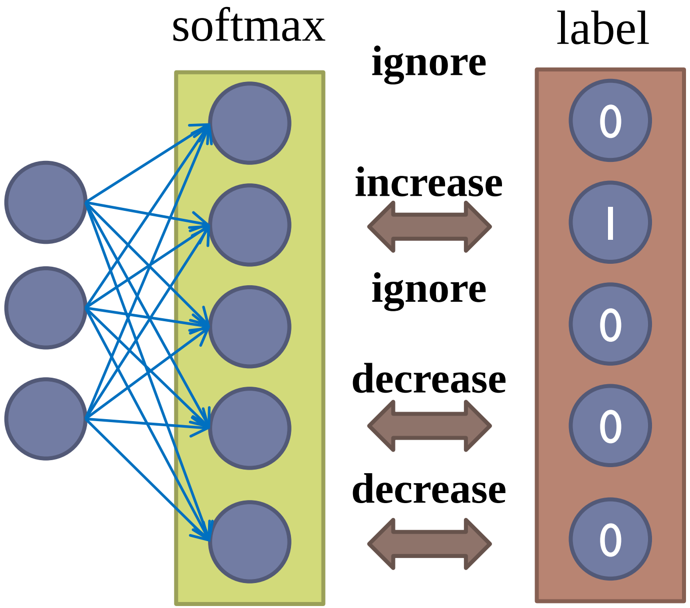
$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \quad \text{for } j = 1, \ldots, K.$$

K = number of vocabulary (very large)

Cost function in Word2vect



Origin method



Sampled method
(randomly select to ignore)

Perplexity

- A measurement of how well a probability distribution predict a sample

$$2^{H(p)} = 2^{-\sum_x p(x) \log_2 p(x)}$$

RNN Introduction

Why RNN?

- Position of words is important
- Slightly change a word may change meaning of whole sentence

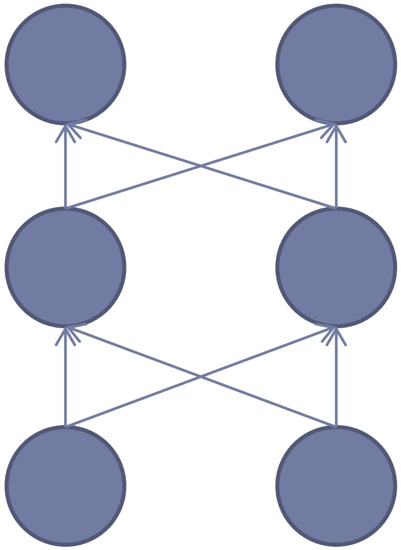
I am handsome	Am I handsome
I will **leave** Taiwan in January	I will **arrive** Taiwan in January

Why RNN?

Lack of sequence concept in ordinary NN

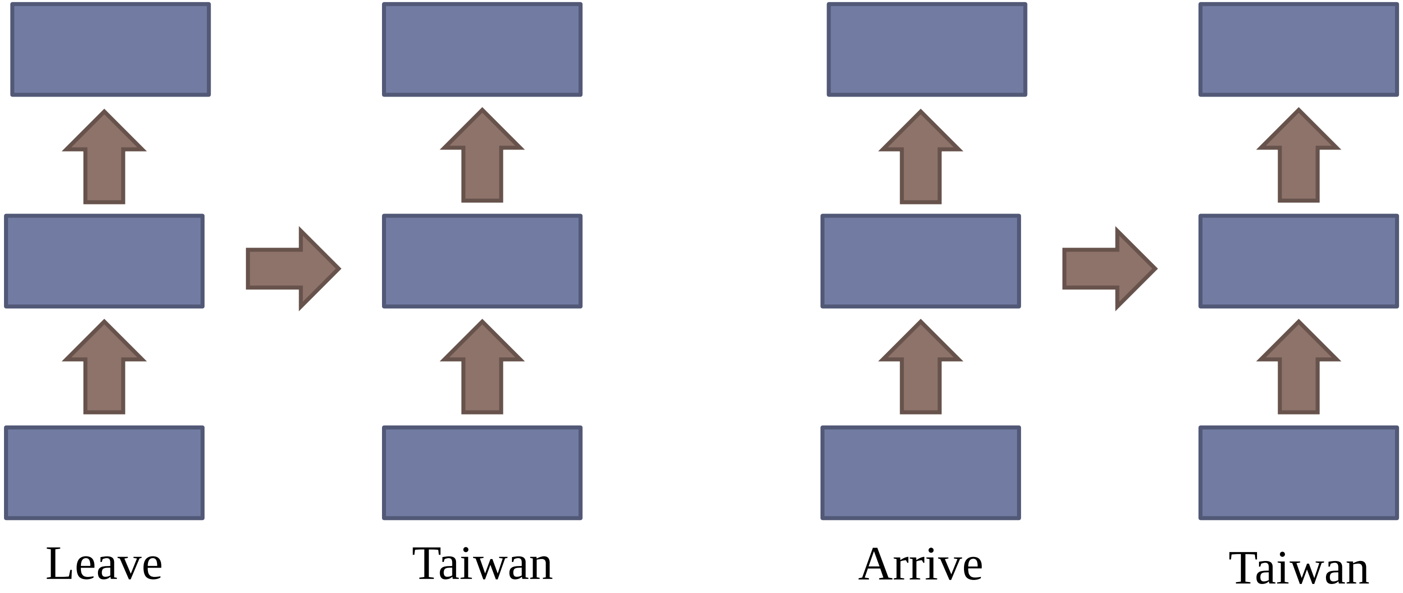
I will **leave** Taiwan on January

I will **arrive** Taiwan on January



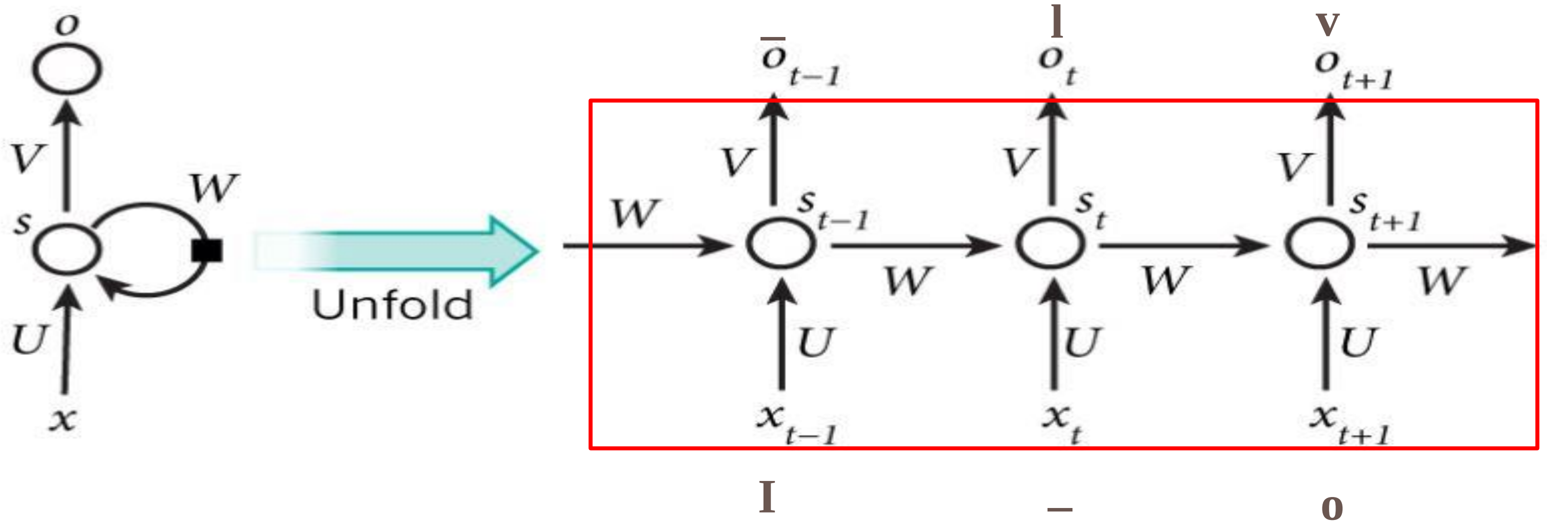
Taiwan

Why RNN?



This type of structure contain sequence concept

Memory in RNN



$$\boxed{s_t = f(Ux_t + Ws_{t-1})}$$

$$o_t = \mathrm{softmax}(Vs_t).$$

Input Xt can be as following:

I love you

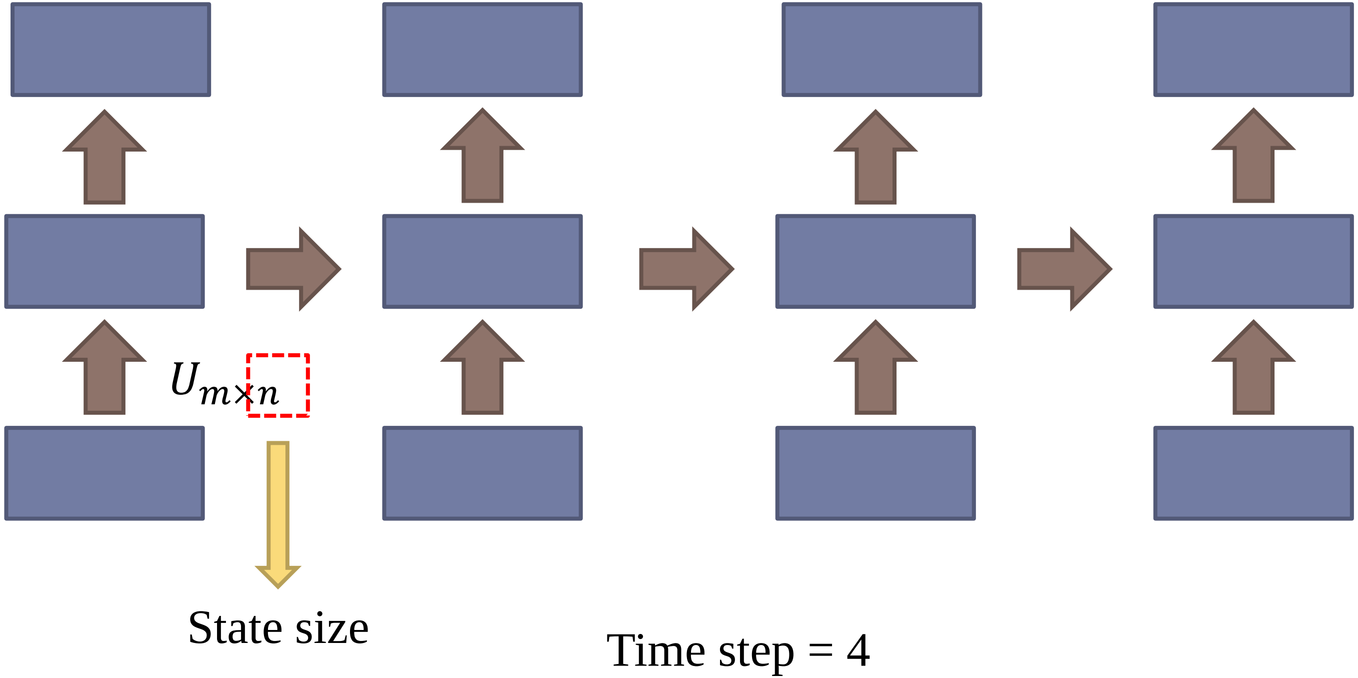
I love you

I love you

I love you (word based)

Etc … …

RNN



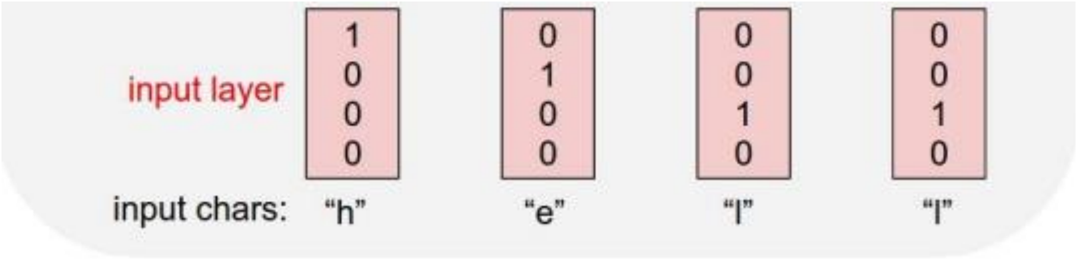
Example

- We would like our computer to read a novel
  - One common solution is feed data into RNN character by character
  - Also called character-level language model

hello, I am Isaac

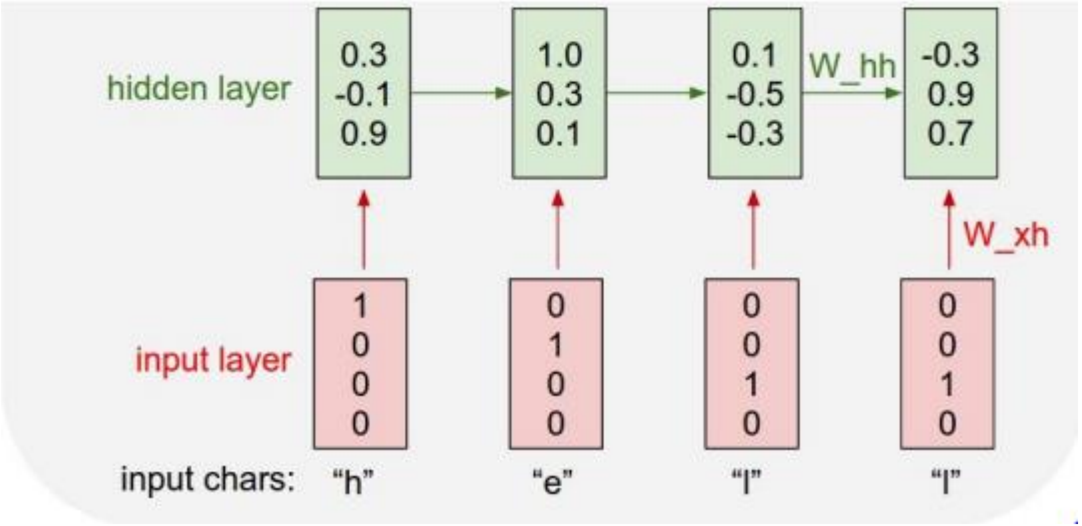
‘h’ ‘e’ ‘l’ ‘l’ ‘o’ ‘,’ ‘ ’ ‘I’ ‘ ‘ ‘a’ ‘m’ ‘ ‘ ‘I’ ‘s’ ‘a’ ‘a’ ‘c’

Example



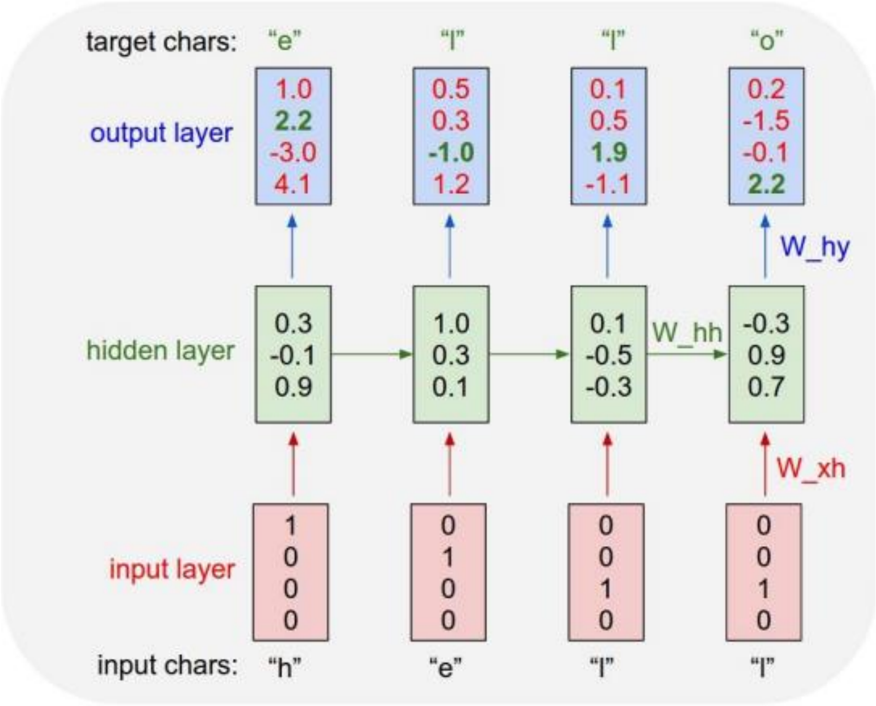
Character-level Language Model

Example



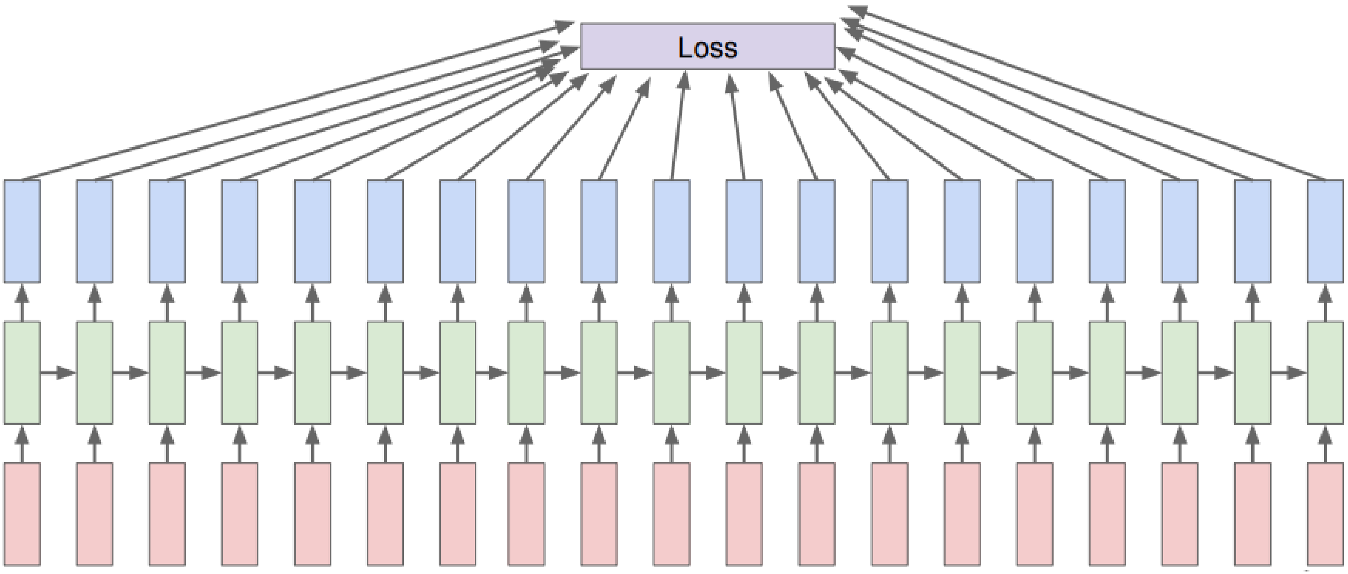
Character-level Language Model

Example



Character-level Language Model

Example



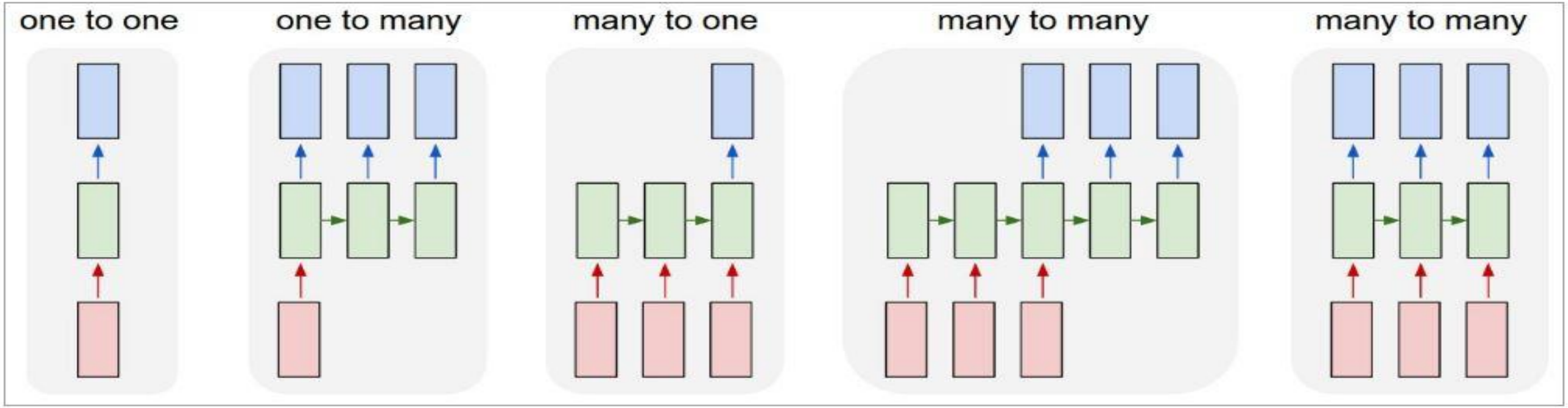
Character-level Language Model

Example

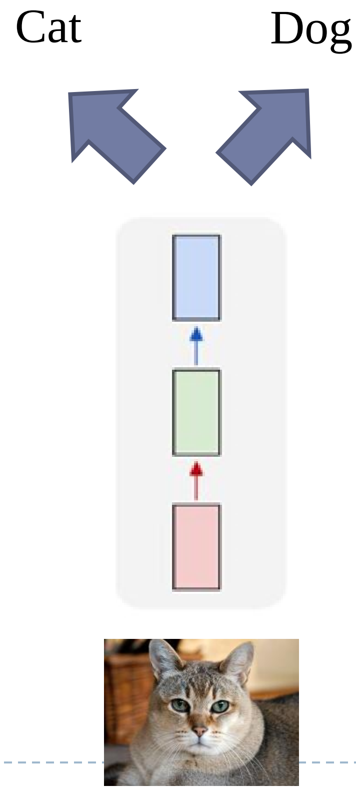
- After training, RNN can learn relationship of characters and even generate it

```
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain 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 nues 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.
```

Different Type of RNN Application



Example



Classification task

One to one (DNN)

Example

This is a cat

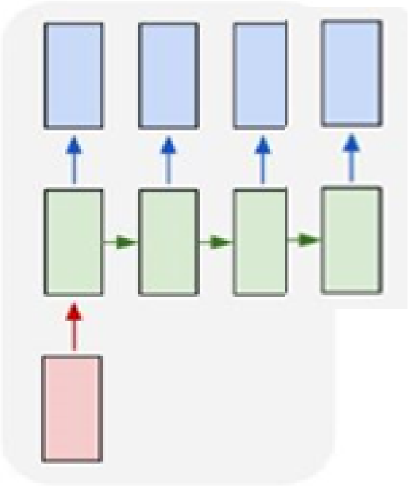
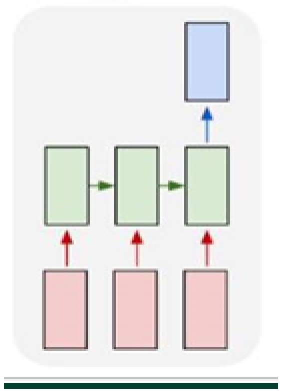


Image Caption

One to many

Example

Good/Bad ?



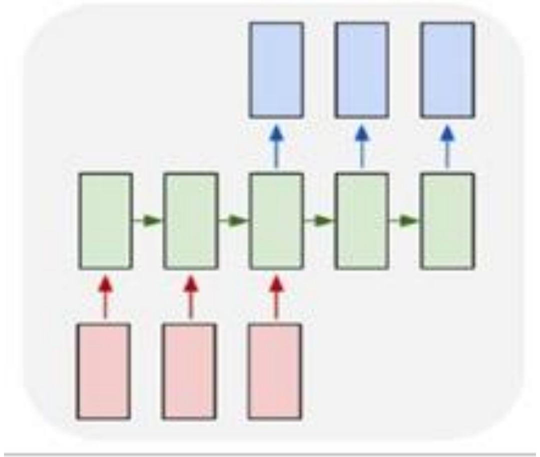
It is good

Movie Review
(positive/negative)

Many to one

Example

I love you

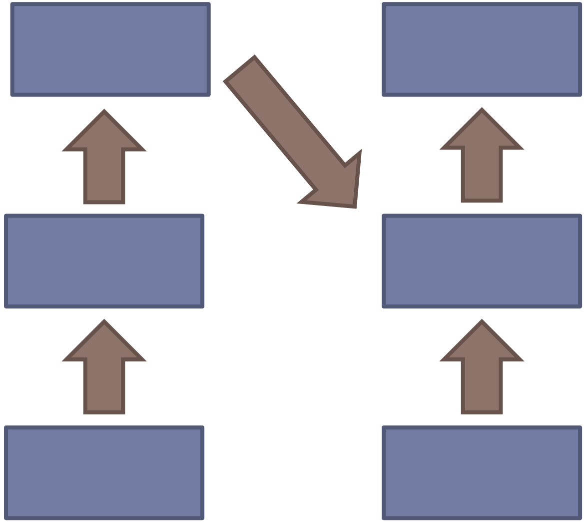


我　愛　你

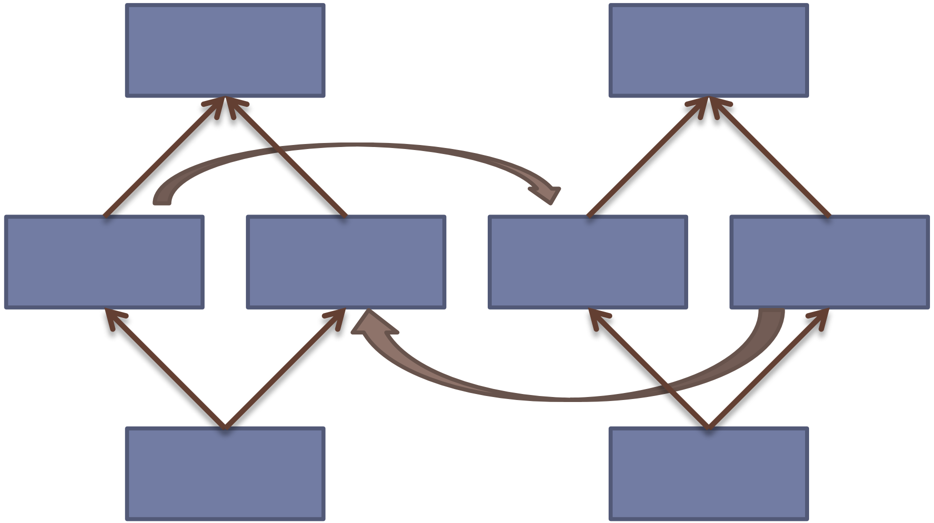
Language translation

Many to many

Different Type of RNN Structure

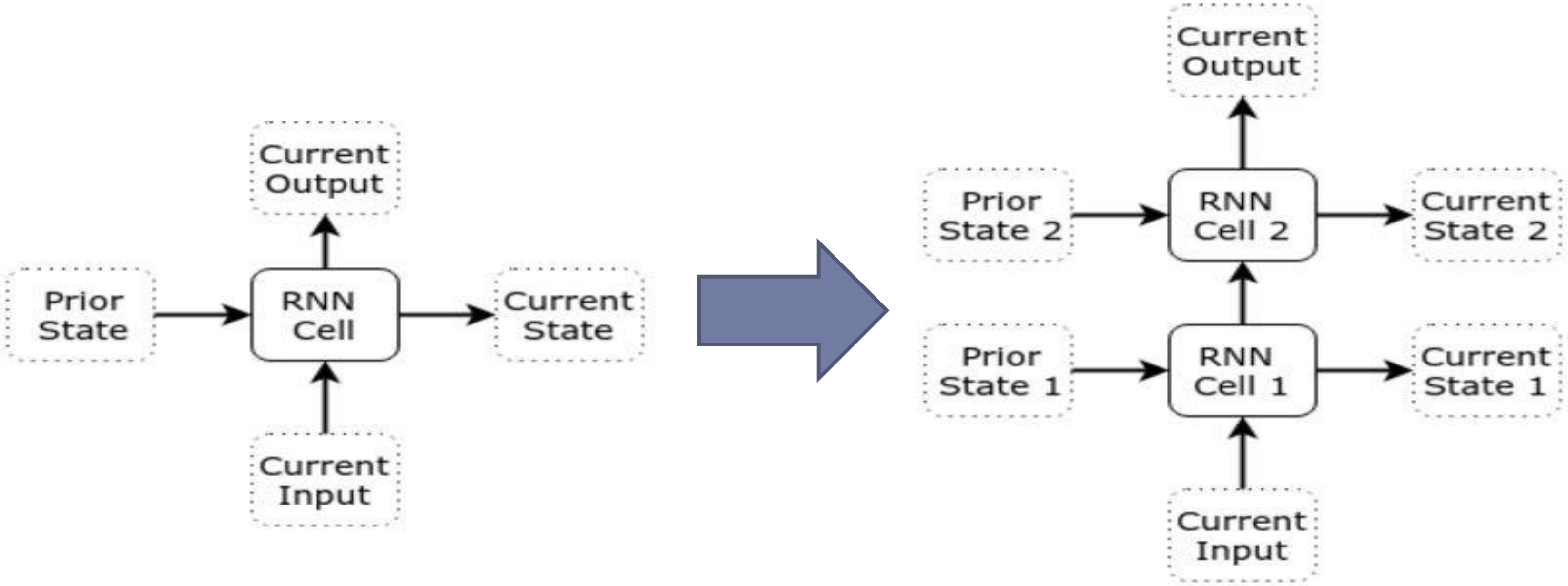


Jordan Network

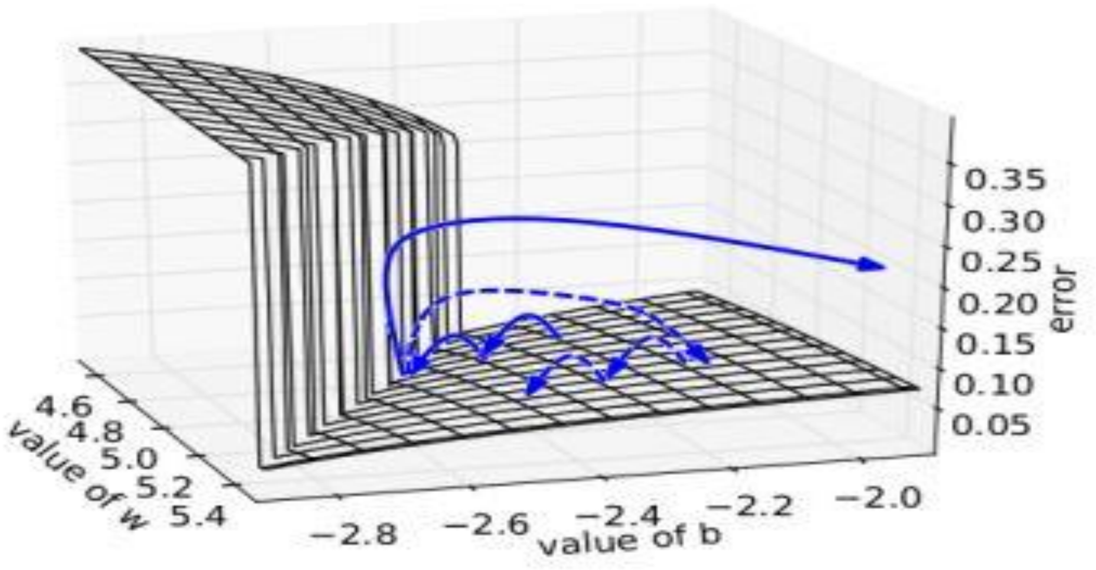


Bidirectional RNN

Go Deep in RNN

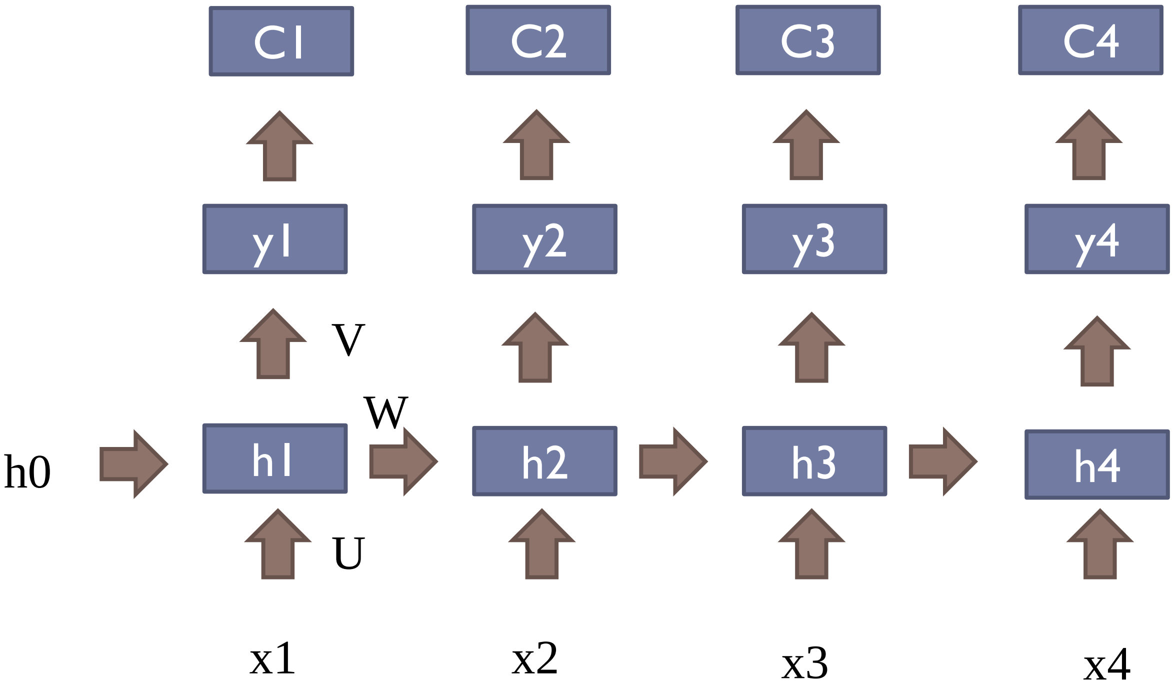


Error surface in RNN



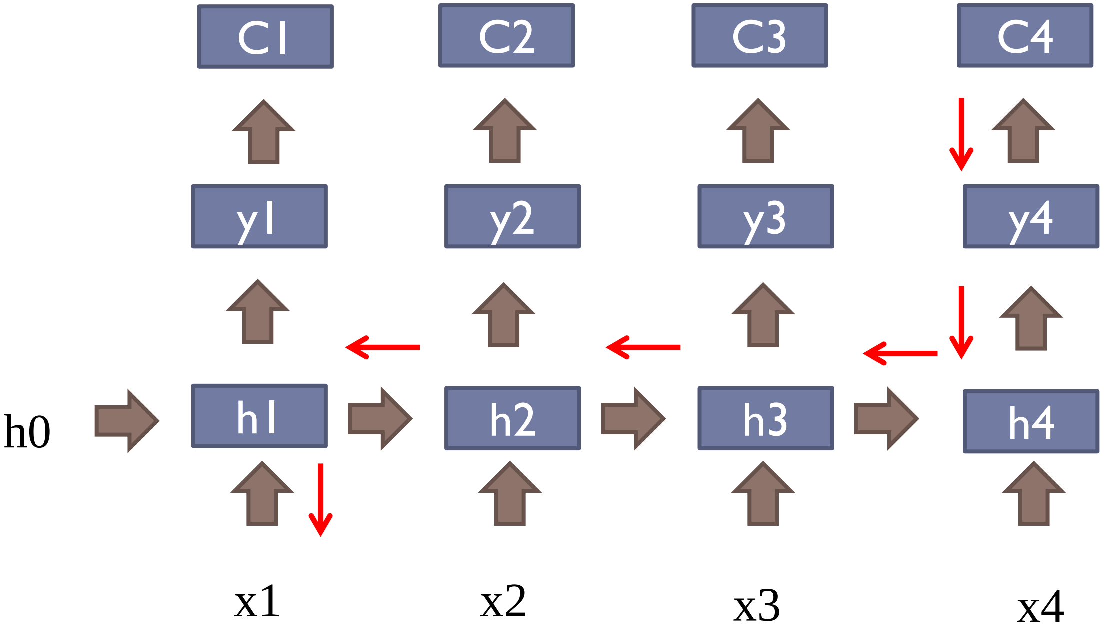
Very flat or very steep on error surface in practice

Learning RNN



BPTT

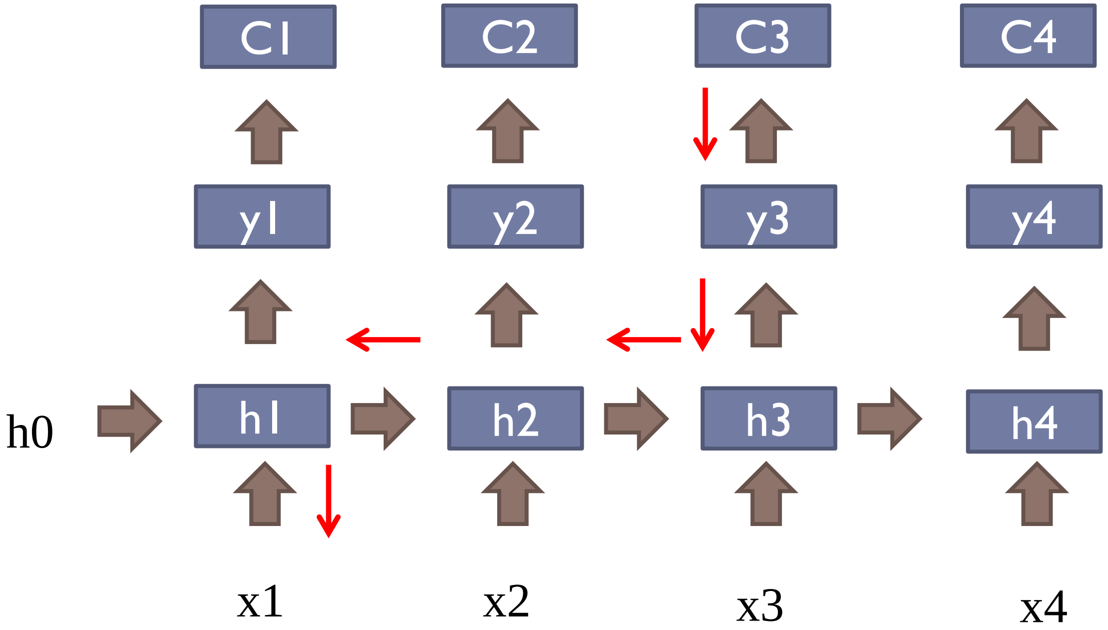
backpropagation through time (BPTT)



$$\frac{\partial C_4}{\partial w} = \frac{\partial C_4}{\partial y4} * \frac{\partial y_4}{\partial h4} * \frac{\partial h_4}{\partial h3} * \frac{\partial h_3}{\partial h2} * \frac{\partial h_2}{\partial h1} * \frac{\partial h_1}{\partial w}$$

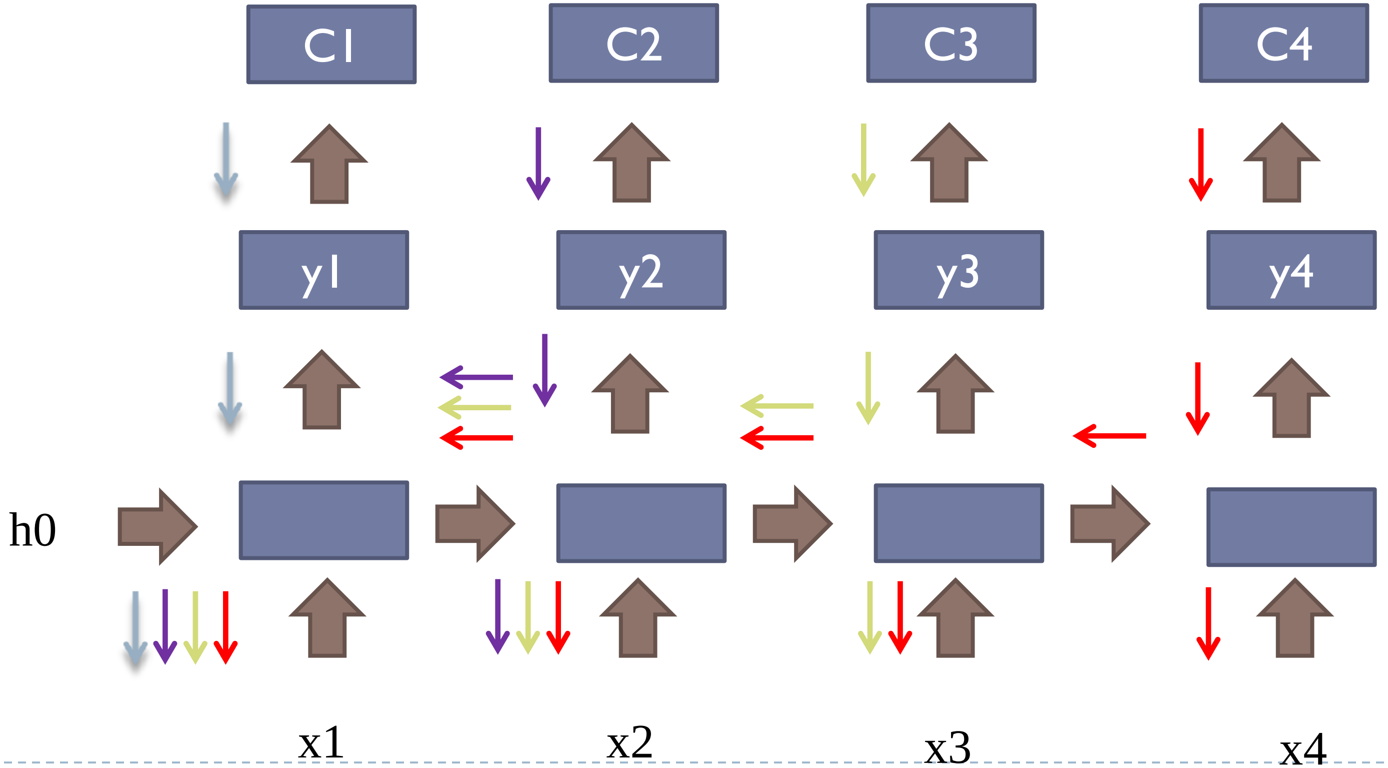
BPTT

backpropagation through time (BPTT)



$$\frac{\partial C_3}{\partial h1} = \frac{\partial C_3}{\partial y3} * \frac{\partial y_3}{\partial h3} * \frac{\partial h_3}{\partial h2} * \frac{\partial h_2}{\partial h1} * \frac{\partial h_1}{\partial w}$$

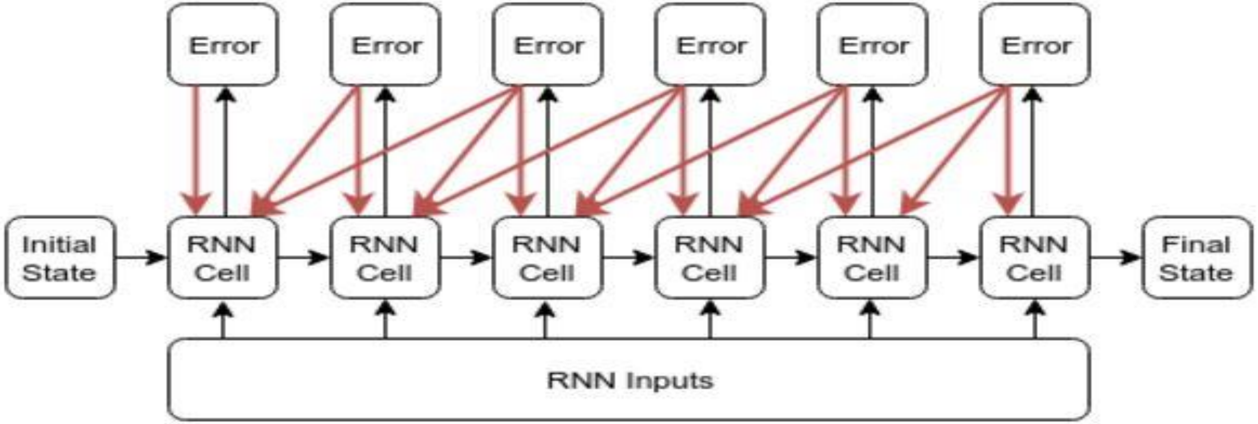
BPTT



Some problem in BPTT

- Inefficiency on propagation under long time step RNN structure
  - Truncated BPTT
- Gradient vanish/exploding problem
  - Main reason why RNN fail in early day
  - New kind of RNN cell

Truncated BPTT



Truncated backpropagation

Gradient Vanish/Exploding problem

$$\frac{\partial C_t}{\partial h1} = \frac{\partial C_t}{\partial y_t} * \frac{\partial y_t}{\partial h_t} * \boxed{\frac{\partial h_t}{\partial h_{t-1}} * \frac{\partial h_{t-1}}{\partial h_{t-2}} * \cdots * \frac{\partial h_3}{\partial h_2} * \frac{\partial h_2}{\partial h_1}}$$

- If very big (gradient exploding)
    - Clip the value

Gradient Vanish/Exploding problem

$$\frac{\partial C_t}{\partial h1} = \frac{\partial C_t}{\partial y_t} * \frac{\partial y_t}{\partial h_t} * \boxed{\frac{\partial h_t}{\partial h_{t-1}} * \frac{\partial h_{t-1}}{\partial h_{t-2}} * \cdots * \frac{\partial h_3}{\partial h_2} * \frac{\partial h_2}{\partial h_1}}$$

- If very small (gradient vanish)
  - need $\frac{\partial h_n}{\partial h_{n-1}}$ to be constant

Gradient Vanish/Exploding problem

How to avoid $\frac{\partial h_n}{\partial h_{n-1}}$ too small ?

If two hidden state is recursive

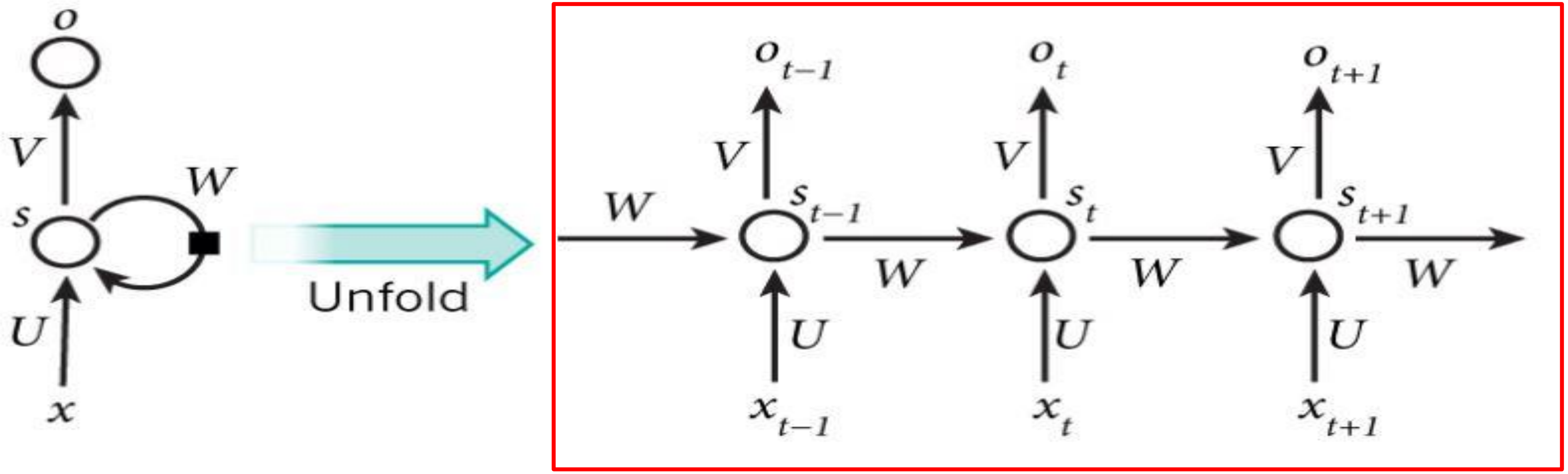
$$h_n = h_{n-1} + \; \ldots \ldots$$

$$\frac{\partial C_t}{\partial h1} = \frac{\partial C_t}{\partial y_t} * \frac{\partial y_t}{\partial h_t} * \boxed{\frac{\partial h_t}{\partial h_{t-1}} * \frac{\partial h_{t-1}}{\partial h_{t-2}} * \cdots * \frac{\partial h_3}{\partial h_2} * \frac{\partial h_2}{\partial h_1}}$$

1

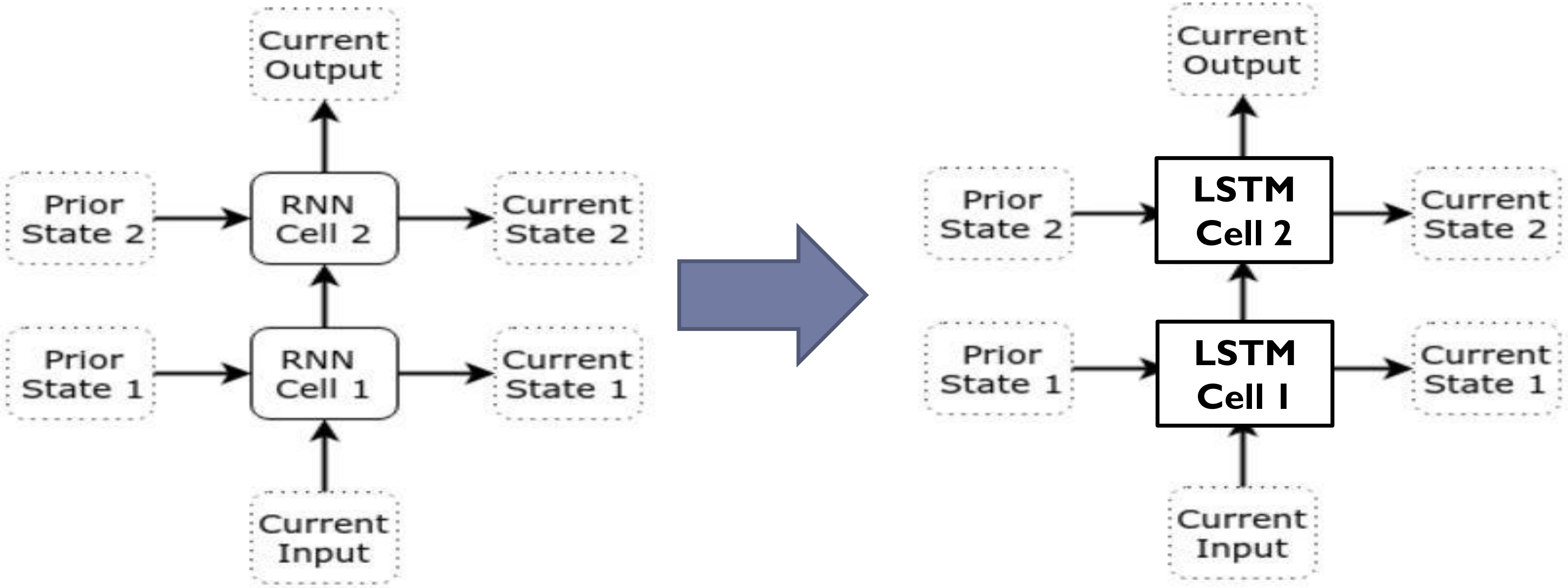
LSTM/GRU Cell

Recall RNN Cell



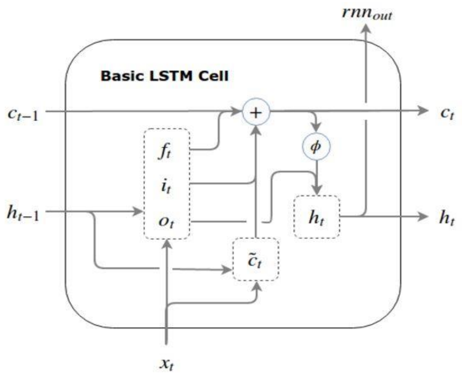
RNN Cell

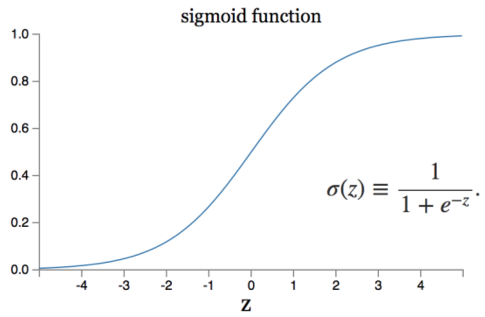
LSTM Cell



Current Output
Prior State 2
RNN Cell 2
Current State 2
Prior State 1
RNN Cell 1
Current State 1
Current Input
Current Output
Prior State 2
LSTM Cell 2
Current State 2
Prior State 1
LSTM Cell 1
Current State 1
Current Input

LSTM





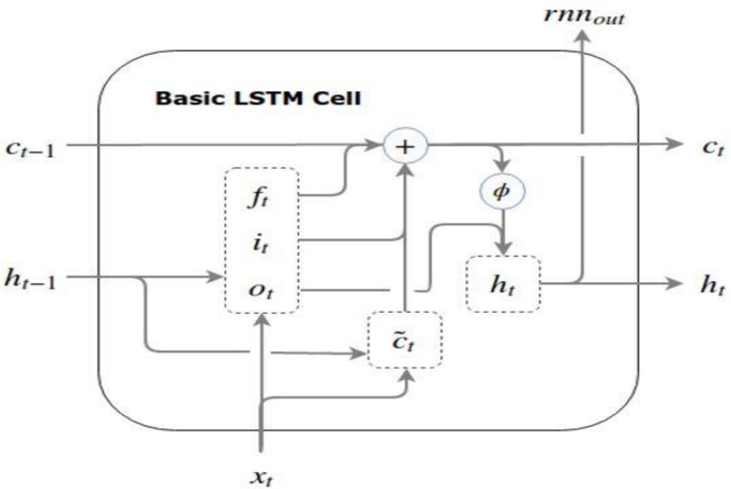
$$i_t = \boxed{\sigma}(W_i h_{t-1} + U_i x_t + b_i)$$
$$o_t = \boxed{\sigma}(W_o h_{t-1} + U_o x_t + b_o)$$
$$f_t = \boxed{\sigma}(W_f h_{t-1} + U_f x_t + b_f)$$

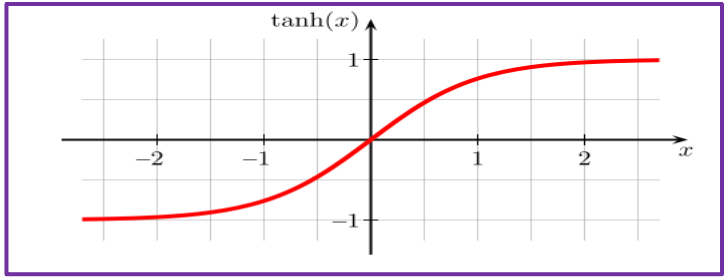
$$\tilde{c}_t = \phi(W h_{t-1} + U x_t + b)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

$$h_t = o_t \odot \phi(c_t)$$

$$\mathrm{rnn}_{out} = h_t$$

LSTM





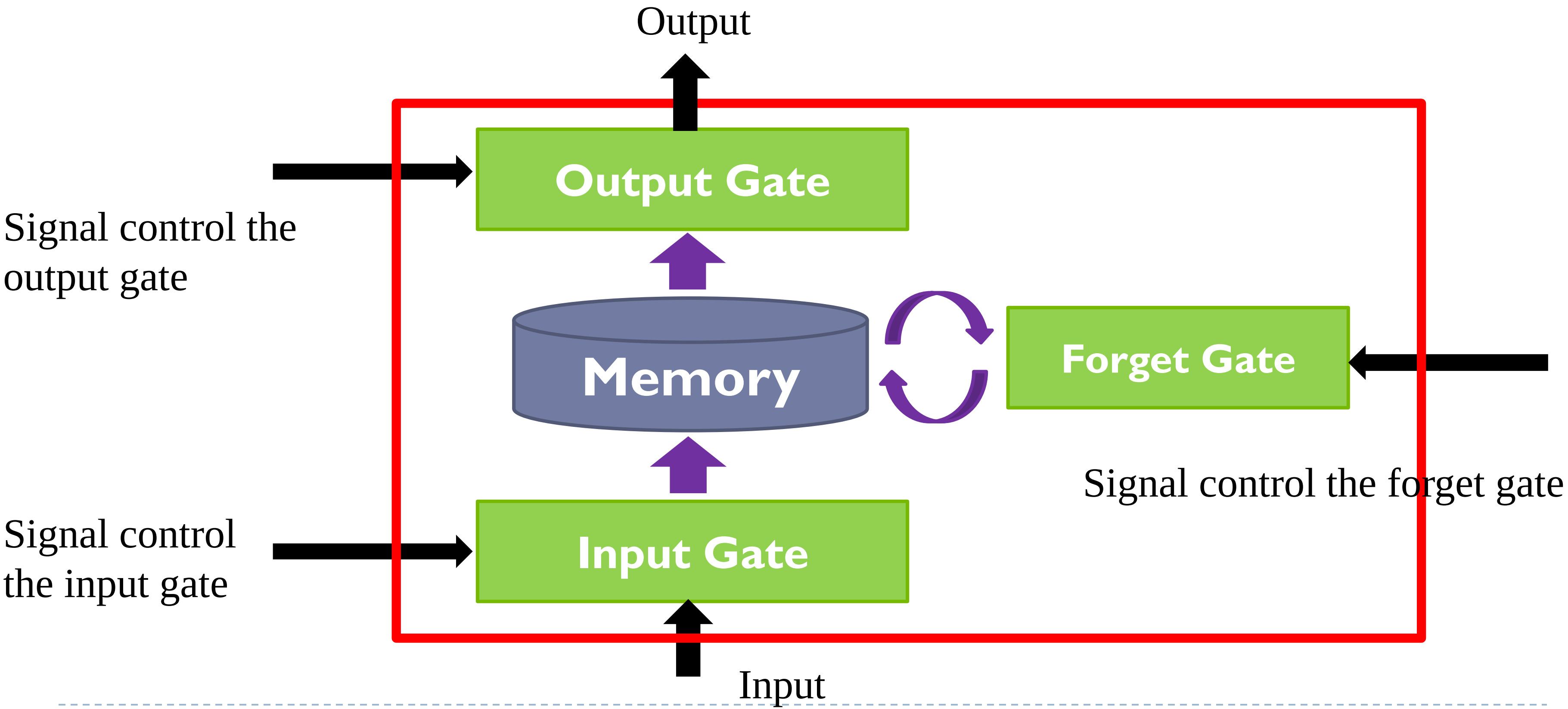
$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i)$$
$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o)$$
$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f)$$

$$\tilde{c}_t = \boxed{\phi}(W h_{t-1} + U x_t + b)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

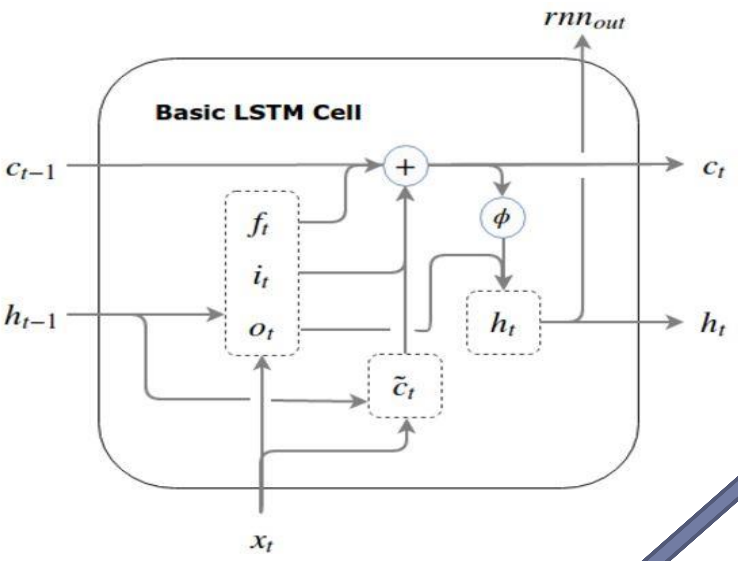
$$h_t = o_t \odot \phi(c_t)$$

$$\mathrm{rnn}_{out} = h_t$$

LSTM Structure



LSTM



$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i)$$
$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o)$$
$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f)$$

$$\tilde{c}_t = \phi(W h_{t-1} + U x_t + b)$$
$$\boxed{c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t}$$

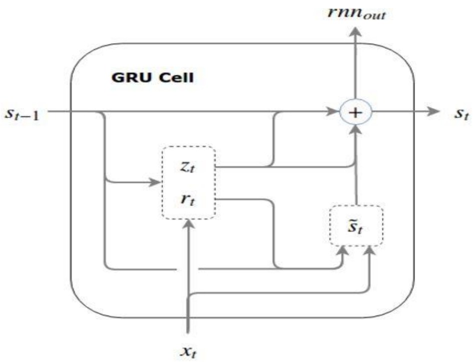
$$h_t = o_t \odot \phi(c_t)$$

Key point of LSTM

$$\mathrm{rnn}_{out} = h_t$$

Varient LSTM

- GRU
  - input gate, forget gate, and output gate are replaced by update gate and reset gate



$$r_t = \sigma(W_r s_{t-1} + U_r x_t + b_r)$$
$$z_t = \sigma(W_z s_{t-1} + U_z x_t + b_z)$$

$$\tilde{s}_t = \phi(W(r_t \odot s_{t-1}) + U x_t + b)$$
$$s_t = z_t \odot s_{t-1} + (1 - z_t) \odot \tilde{s}_t$$

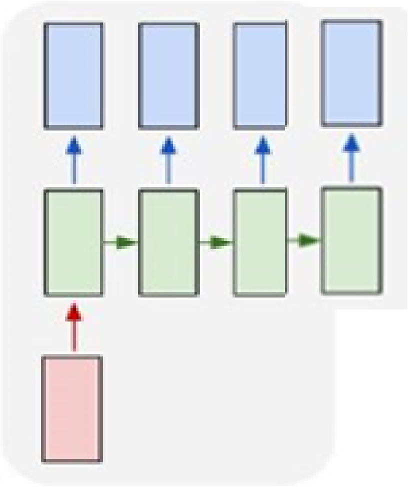
Variant LSTM

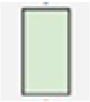
- There are more……
- What kind of LSTM is the best?

?

Example

This is a cat





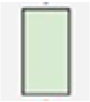
GRU/LSTM cell

Image Caption

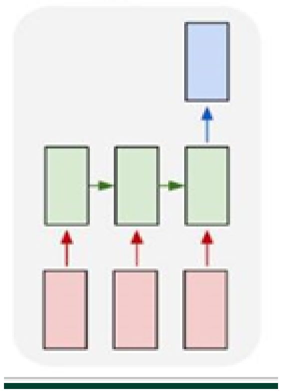
One to many

Example

Good/Bad ?



GRU/LSTM cell



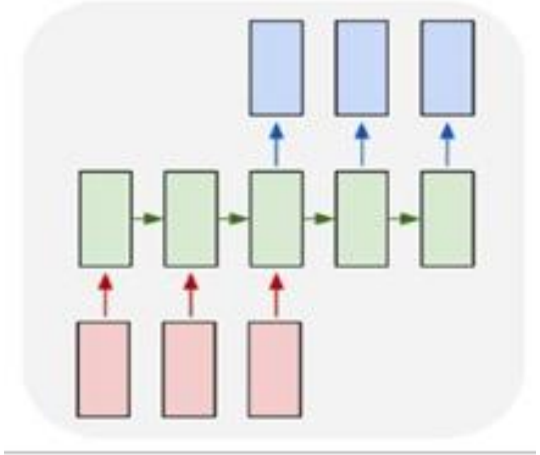
It is good

Movie Review
(positive/negative)

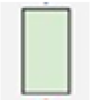
Many to one

Example

I love you



我　愛　你



GRU/LSTM cell

Language translation

Many to many

Summary

- Build your RNN model
  - Observe your application which kind of RNN is suitable
  - Many to one, many to many, ……
- Use LSTM/GRU cell now
  - Perform better than traditional RNN