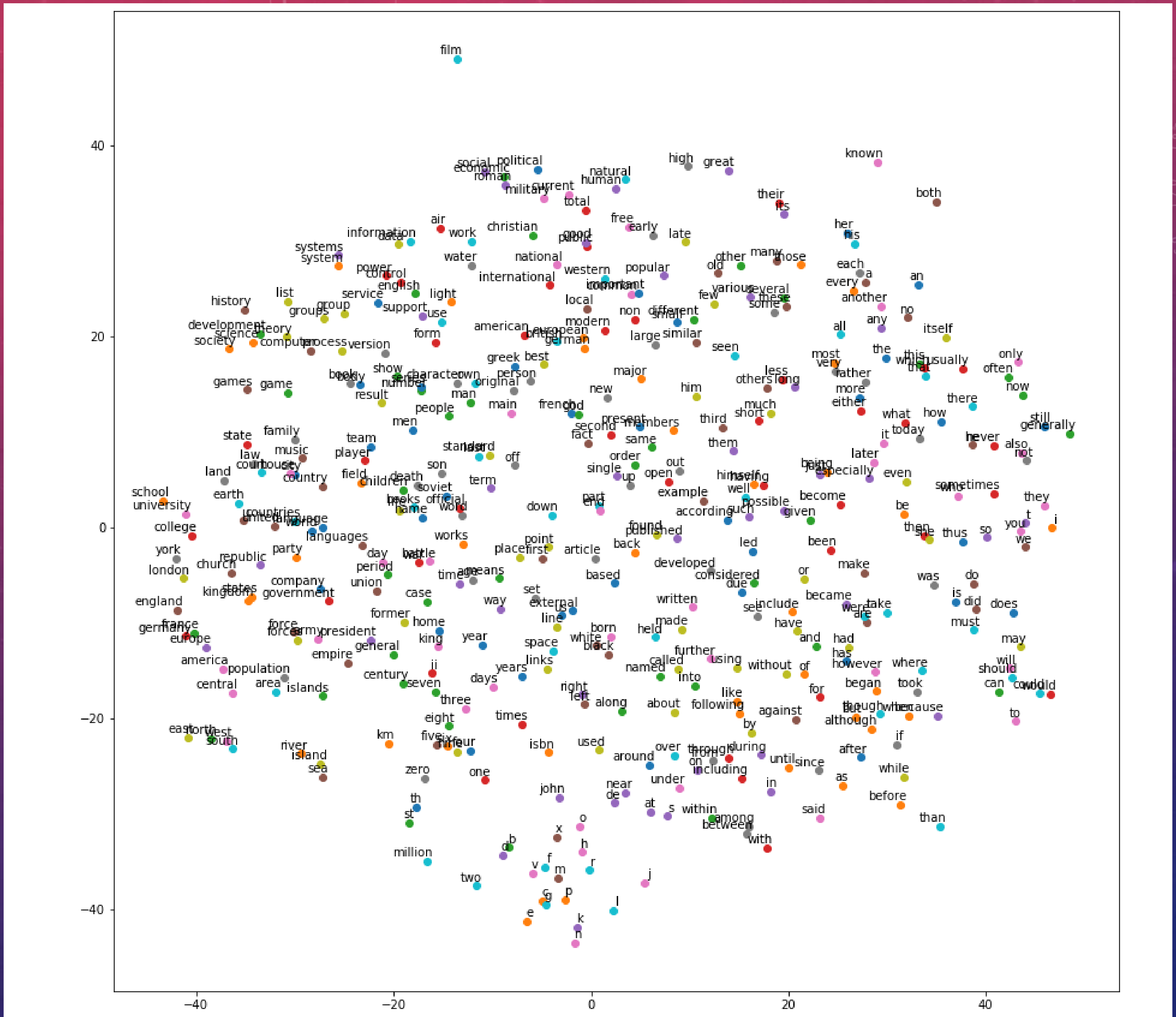
ADVANCED DEEP LEARNING MODELS FOR NATURAL LANGUAGE PROCESSING

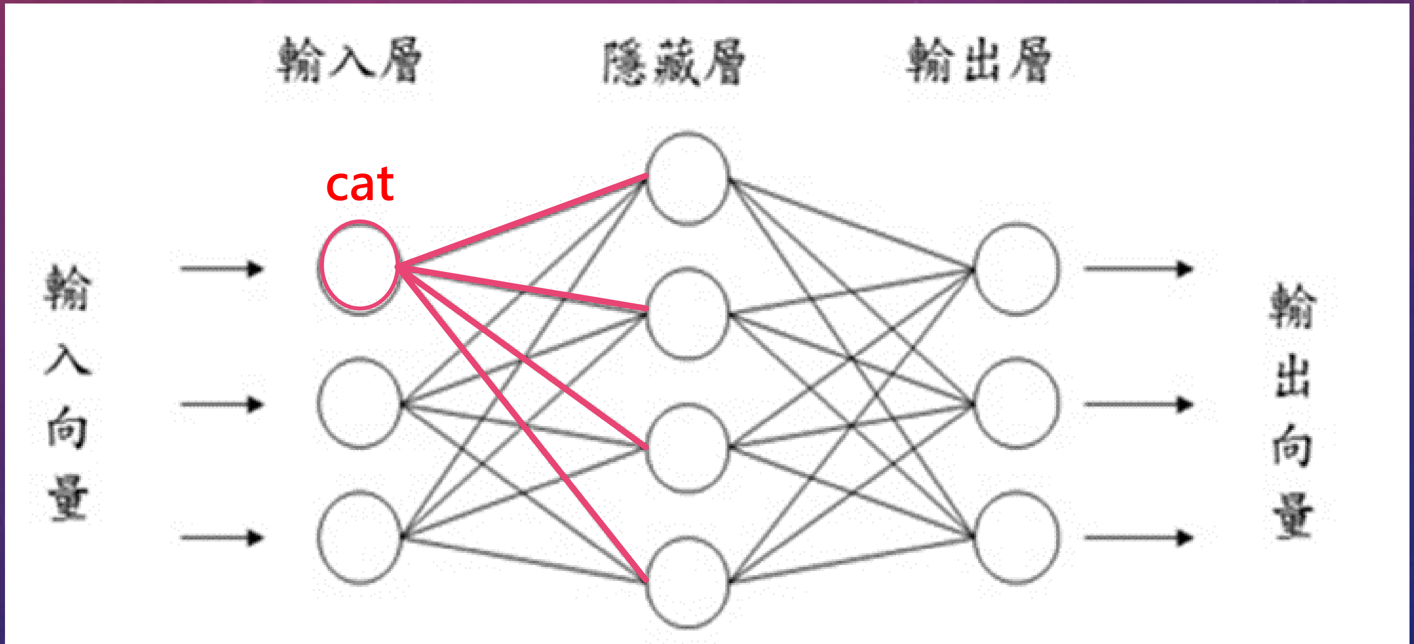
自然語言處理的進階深度模型

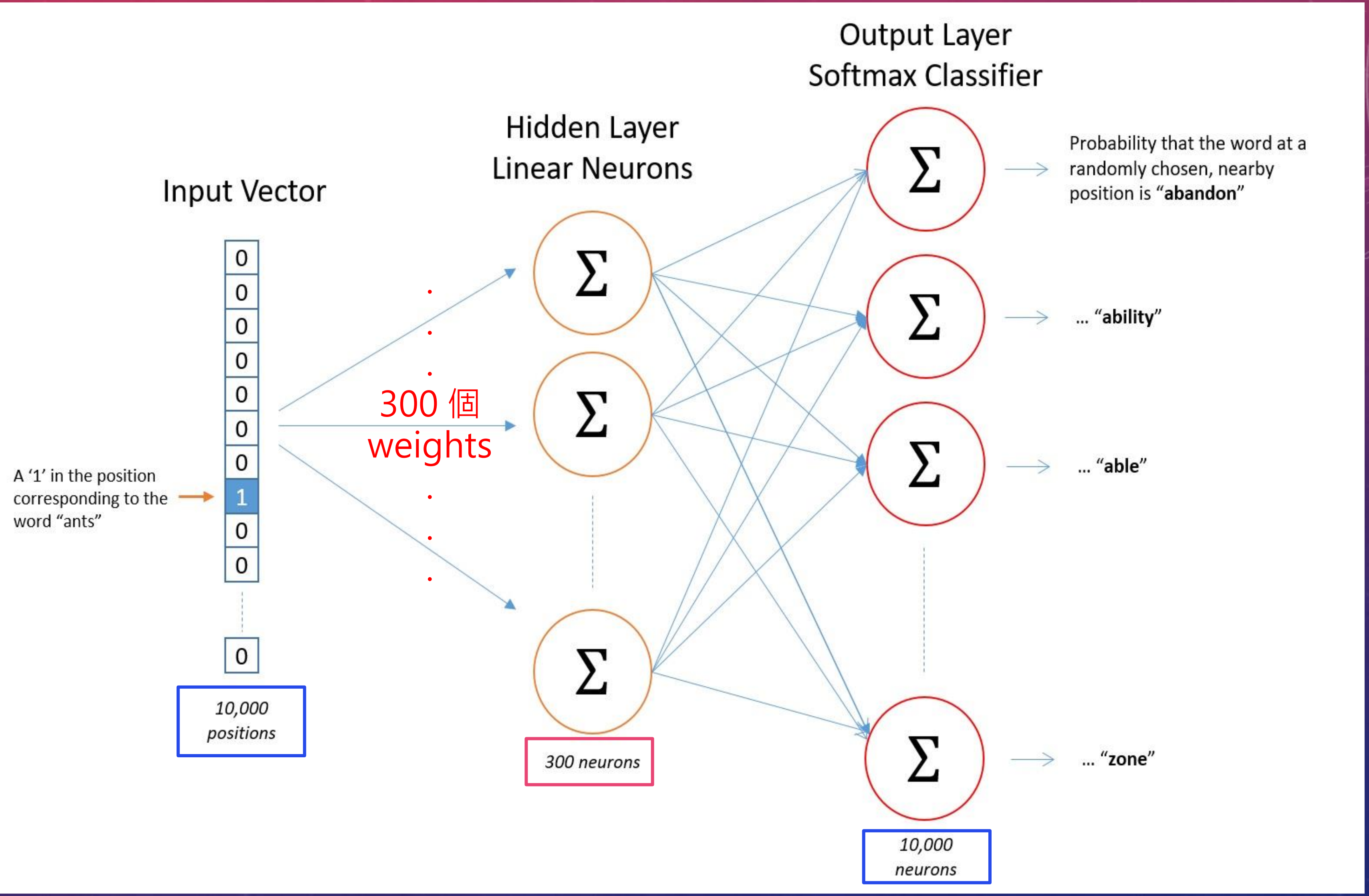
張家瑋 博士
國立臺中科技大學資訊工程系助理教授

REVIEW


film
social
political
economic
roman
military
current
natural
human
total
high
great
known
their
its
both
air
work
christian
information
data
free
early
late
her
his
systems
system
water
national
other
many
those
each
power
control
english
international
western
popular
old
every
a
an
history
list
service
light
local
various
several
few
these
some
another
group
groups
support
use
non
different
small
no
any
all
itself
development
theory
science
society
computer
process
version
american
european
british
german
modern
large
similar
seen
form
only
greek
best
the
most
very
this
usually
often
show
book
body
character
own
person
major
rather
that
games
game
number
original
new
him
others
long
more
now
result
man
either
there
people
main
french
god
much
short
what
how
still
generally
men
present
members
second
third
fact
it
today
state
family
team
standard
same
them
never
also
not
music
player
law
off
later
land
country
field
death
son
order
out
being
especially
even
children
soviet
term
single
up
open
himself
having
sometimes
school
earth
well
example
become
who
they
university
books
official
world
down
part
end
possible
be
according
such
given
countries
language
name
found
published
then
thus
so
you
i
college
languages
works
point
place
first
article
back
led
been
we
york
republic
party
day
war
battle
developed
make
london
church
period
time
means
based
considered
or
was
do
company
union
set
due
is
did
does
kingdom
states
government
case
written
became
england
way
external
see
include
were
take
are
former
line
made
have
must
france
germany
force
forces
army
president
home
born
held
europe
king
year
white
black
space
further
and
had
may
empire
general
links
using
has
america
population
years
named
called
without
of
however
where
will
should
ii
into
central
area
century
seven
days
right
left
like
for
began
took
can
could
would
islands
three
along
about
following
against
though
but
because
to
times
eight
by
although
east
north
west
south
five
six
four
isbn
used
if
river
island
km
over
through
during
after
around
from
on
until
since
while
sea
zero
one
including
as
before
john
near
de
under
in
th
at
s
within
among
between
said
than
st
o
x
with
million
h
b
d
f
r
v
m
j
two
c
g
p
l
e
k
n
40
20
0
−20
−40
−40
−20
0
20
40

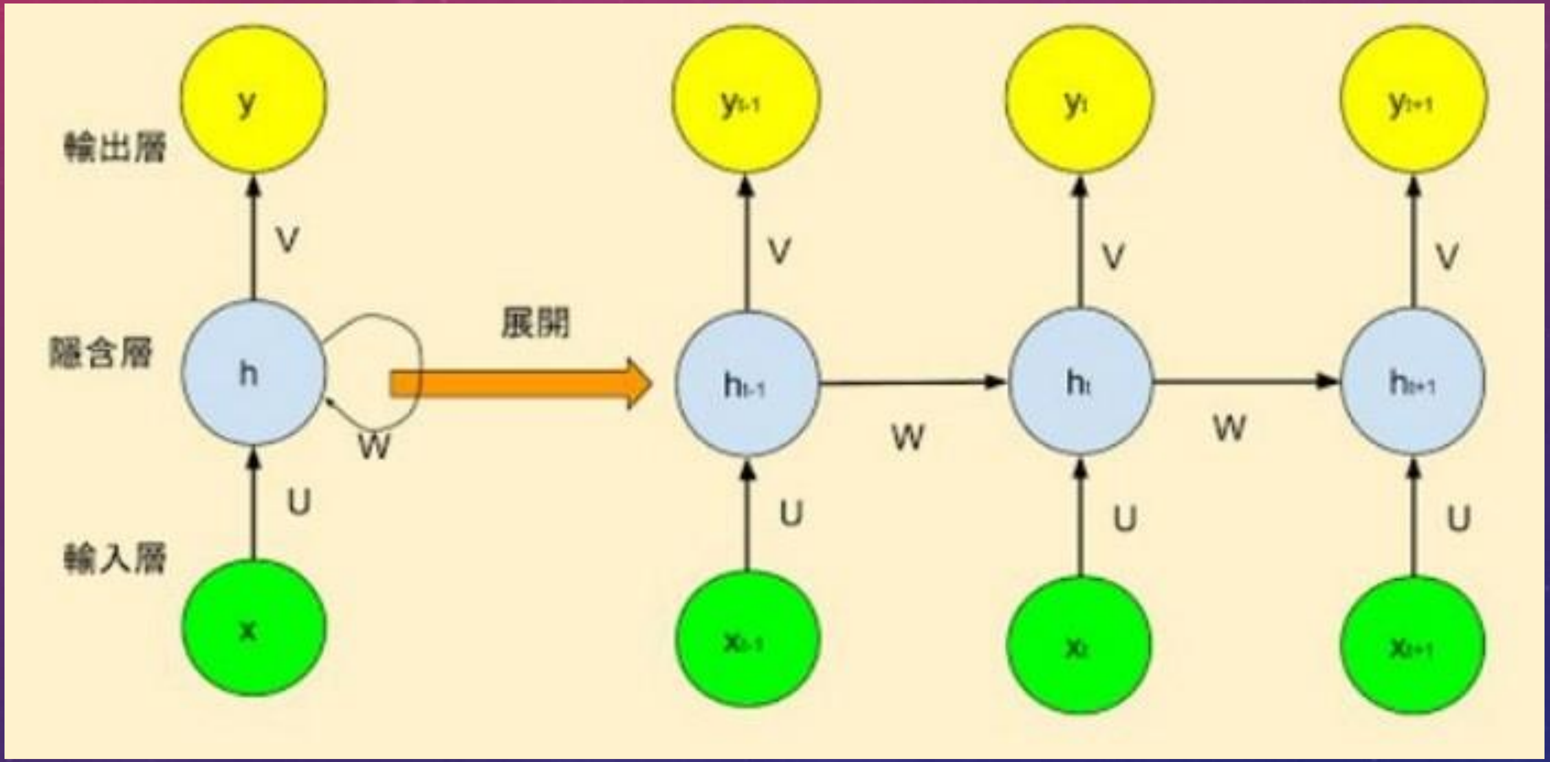
```
The -> [0. 1. 0. 0. 0. 0. 0. 0. 0. 0.]
cat -> [0. 0. 1. 0. 0. 0. 0. 0. 0. 0.]
jump -> [0. 0. 0. 1. 0. 0. 0. 0. 0. 0.]
over -> [0. 0. 0. 0. 1. 0. 0. 0. 0. 0.]
the -> [0. 0. 0. 0. 0. 1. 0. 0. 0. 0.]
dog -> [0. 0. 0. 0. 0. 0. 1. 0. 0. 0.]
The -> [0. 1. 0. 0. 0. 0. 0. 0. 0. 0.]
dog -> [0. 0. 0. 0. 0. 0. 1. 0. 0. 0.]
ate -> [0. 0. 0. 0. 0. 0. 0. 1. 0. 0.]
my -> [0. 0. 0. 0. 0. 0. 0. 0. 1. 0.]
homework -> [0. 0. 0. 0. 0. 0. 0. 0. 0. 1.]
```




Input Vector
0
0
0
0
0
0
0
1
0
0
0
10,000 positions
A '1' in the position corresponding to the word "ants"
300 個 weights
Hidden Layer
Linear Neurons
300 neurons
Output Layer
Softmax Classifier
Probability that the word at a randomly chosen, nearby position is "**abandon**"
… "**ability**"
… "**able**"
… "**zone**"
10,000 neurons

RECURRENT NEURAL NETWORK, RNN

RNN STRUCTURE



*tanh is usually used to be the activation function in RNN

PROBLEMS OF RNN

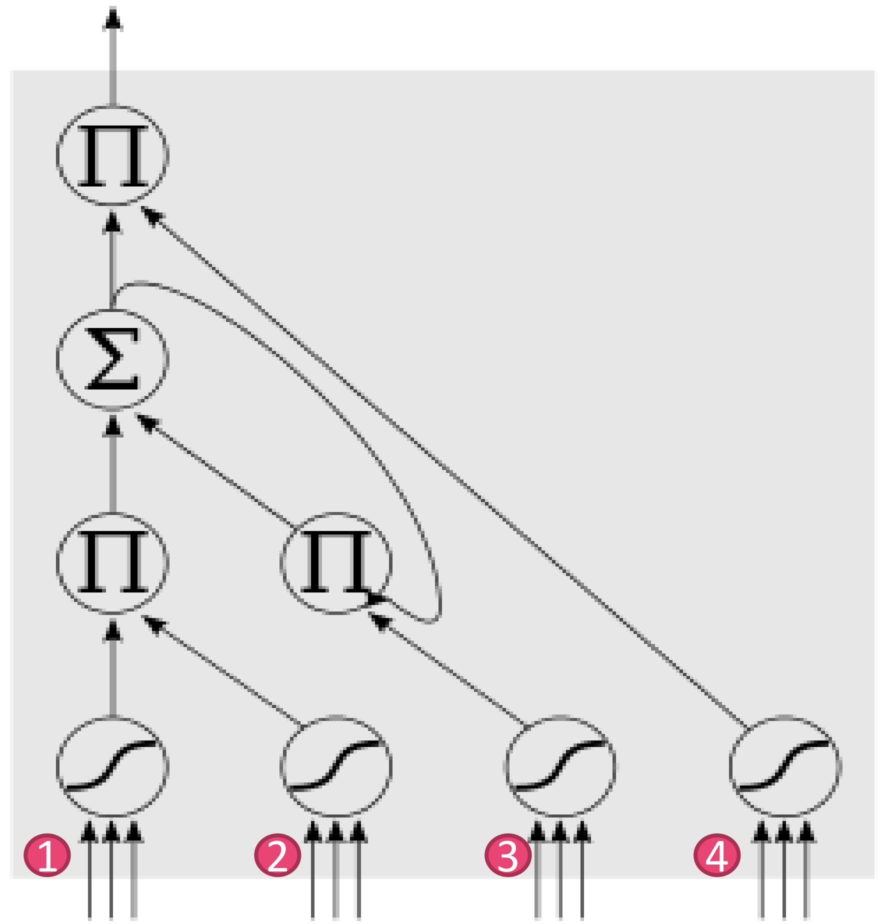
- RNN 對 short-term 敏感
- RNN 對 long-term 容易遺忘
- 越前面的字越會遺忘... 這就是RNN的梯度消失...

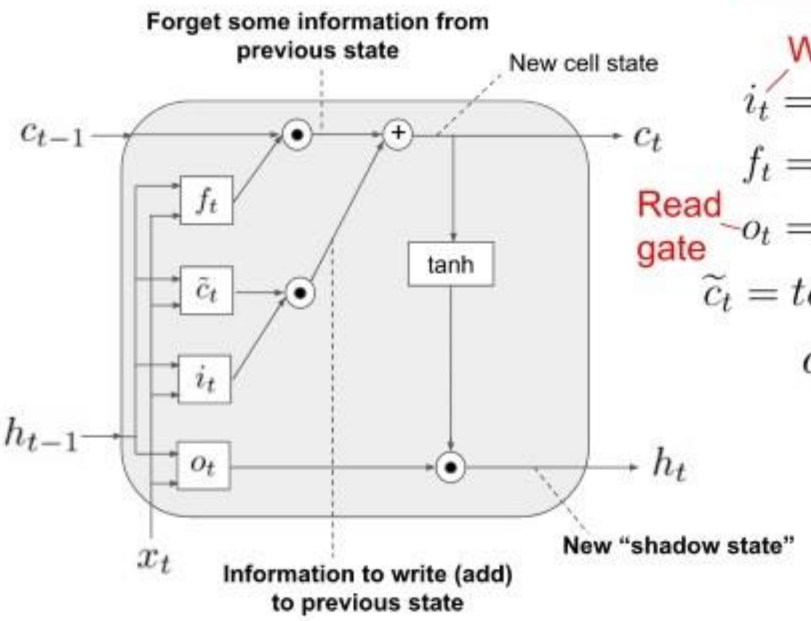
$$h_t = W * h_{t-1} + U * x_t + b$$
$$h_t = (W^2 * h_{t-2} + W * U * x_{t-1} + W * b) + U * x_t + b$$
$$\dots$$
$$h_t = (W^t * h_0 + \dots) + U * x_t + b$$

公式 h_0 對 h_t 的影響力為 w 的 t 次方，通常，w 會小於 1，離目標字越遠的字詞經過越多次傳遞(連乘)... 影響力幾乎不見了

LONG SHORT TERM MEMORY, LSTM

LSTM STRUCTURE





Conceptually, we lose information

We use shadow state to calculate gates

LSTM equations

Write gate

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i)$$ Input gate

$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f)$$ Forget gate

Read gate

$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o)$$ Output gate

$$\tilde{c}_t = tanh(W h_{t-1} + U x_t + b)$$ Memory cell candidate

$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t$$ Memory cell

$$h_t = o_t \circ tanh(c_t)$$ Shadow state

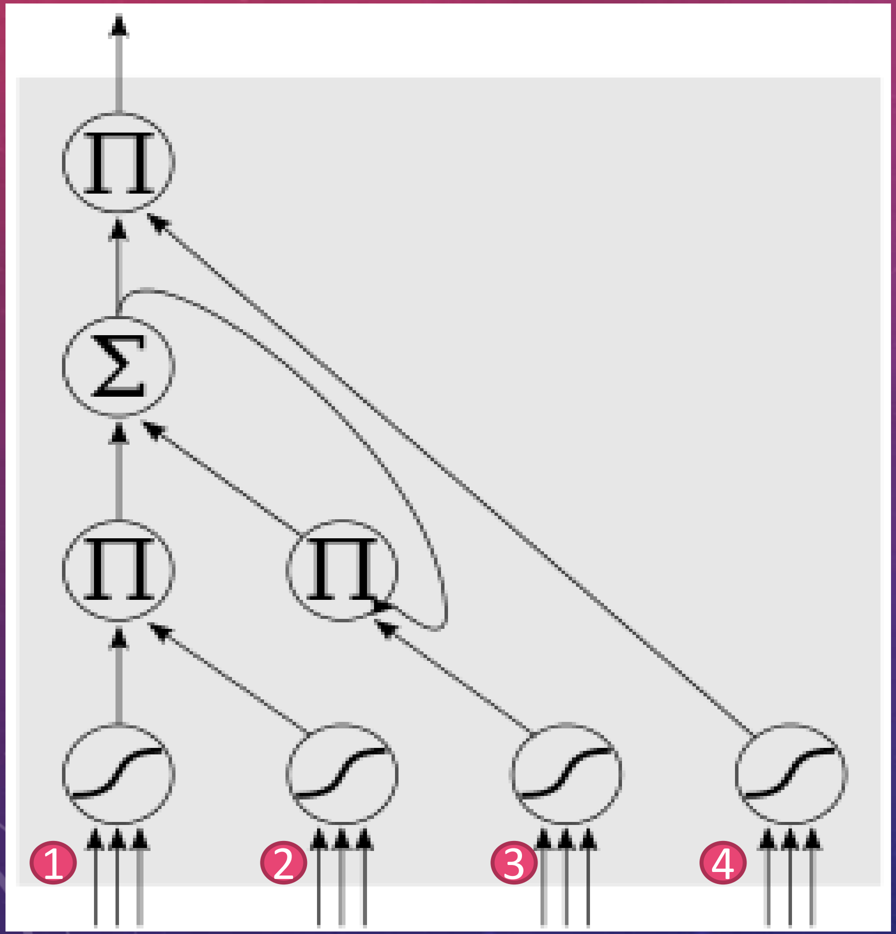
$$y_t = h_t$$ Cell Output

Read occurs after writing

142

deepsystems.ai

LSTM STRUCTURE



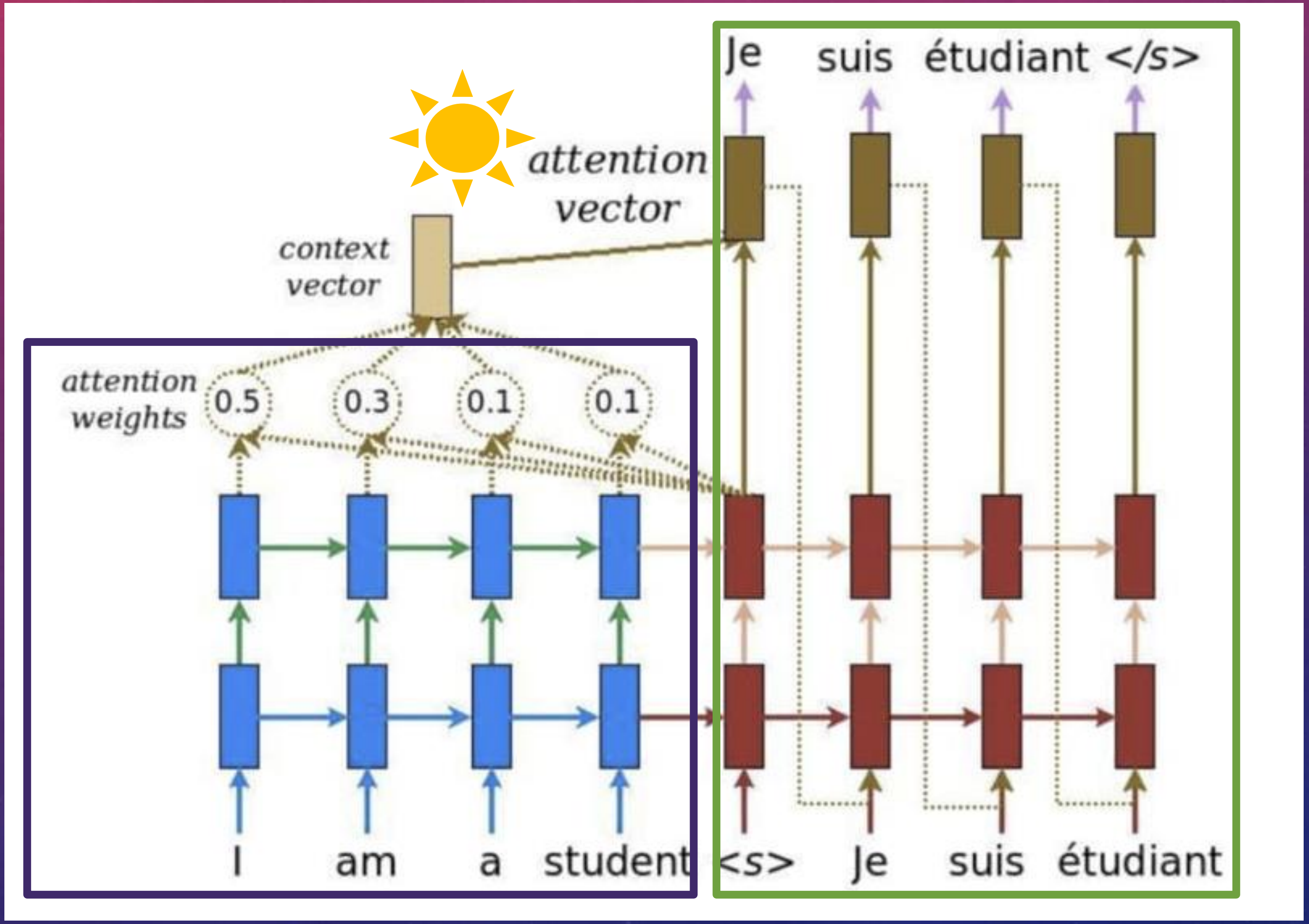
1. 純的 input，下面三個gate會影響這個input能否被記住或作為輸出

---

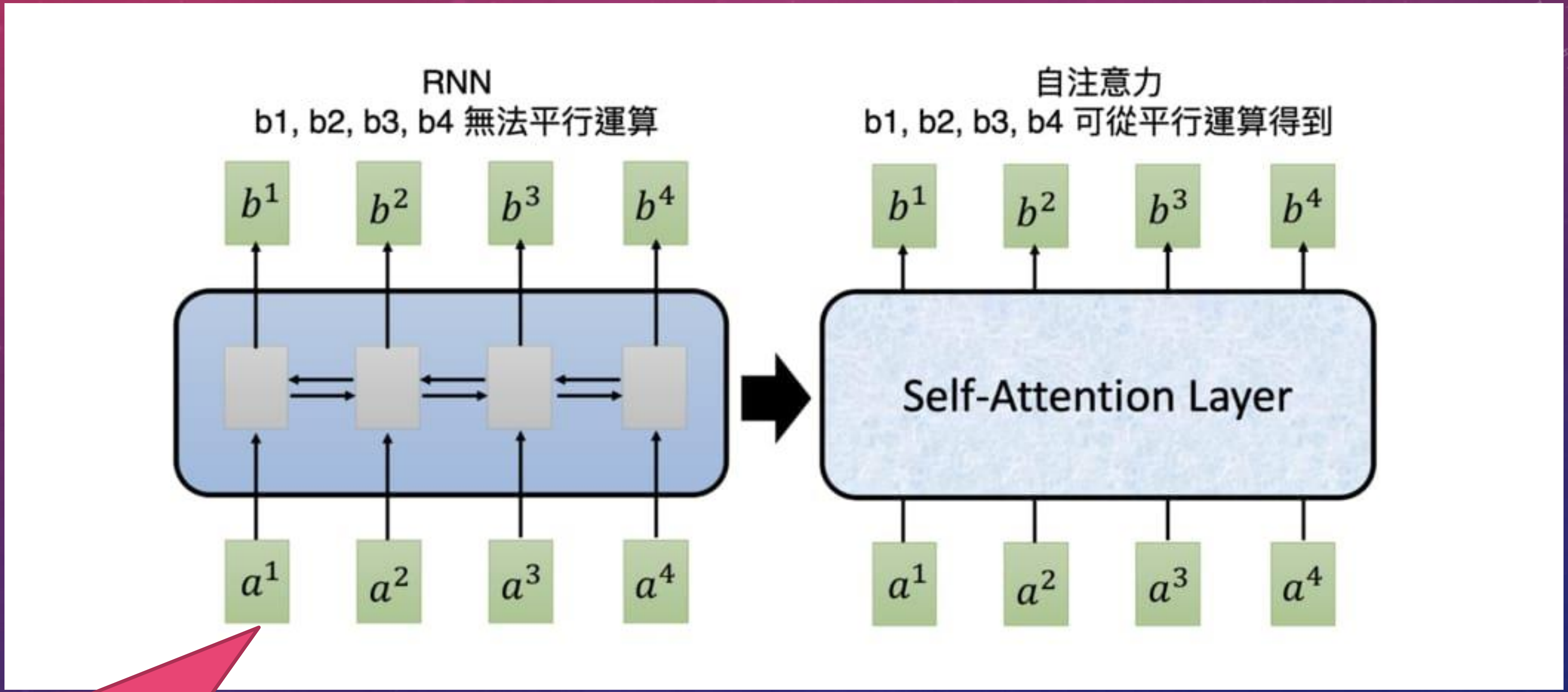
2. Input gate，如果值為0就擋住，不給進下一層 (sigmoid)
3. Forget gate，如果值近似0，就把區塊裡記住的值忘掉 (sigmoid)
4. Output gate，決定在區塊記憶中的input是否能輸出 (sigmoid)

TRANSFORMER

WHAT IS ATTENTION?

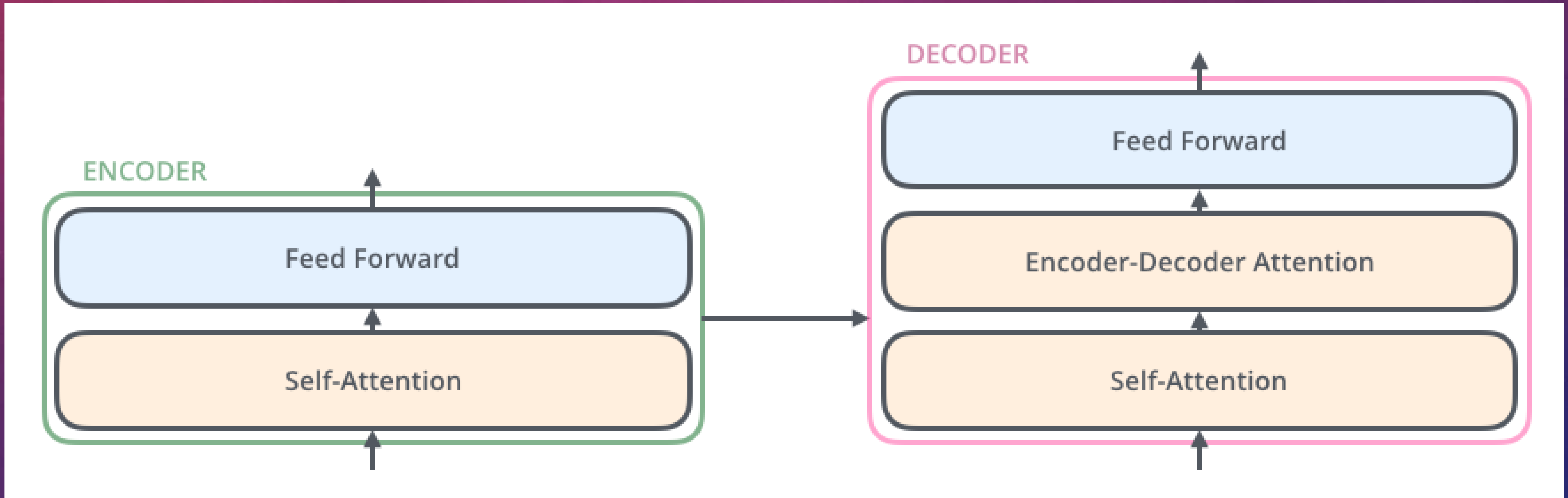


WHY

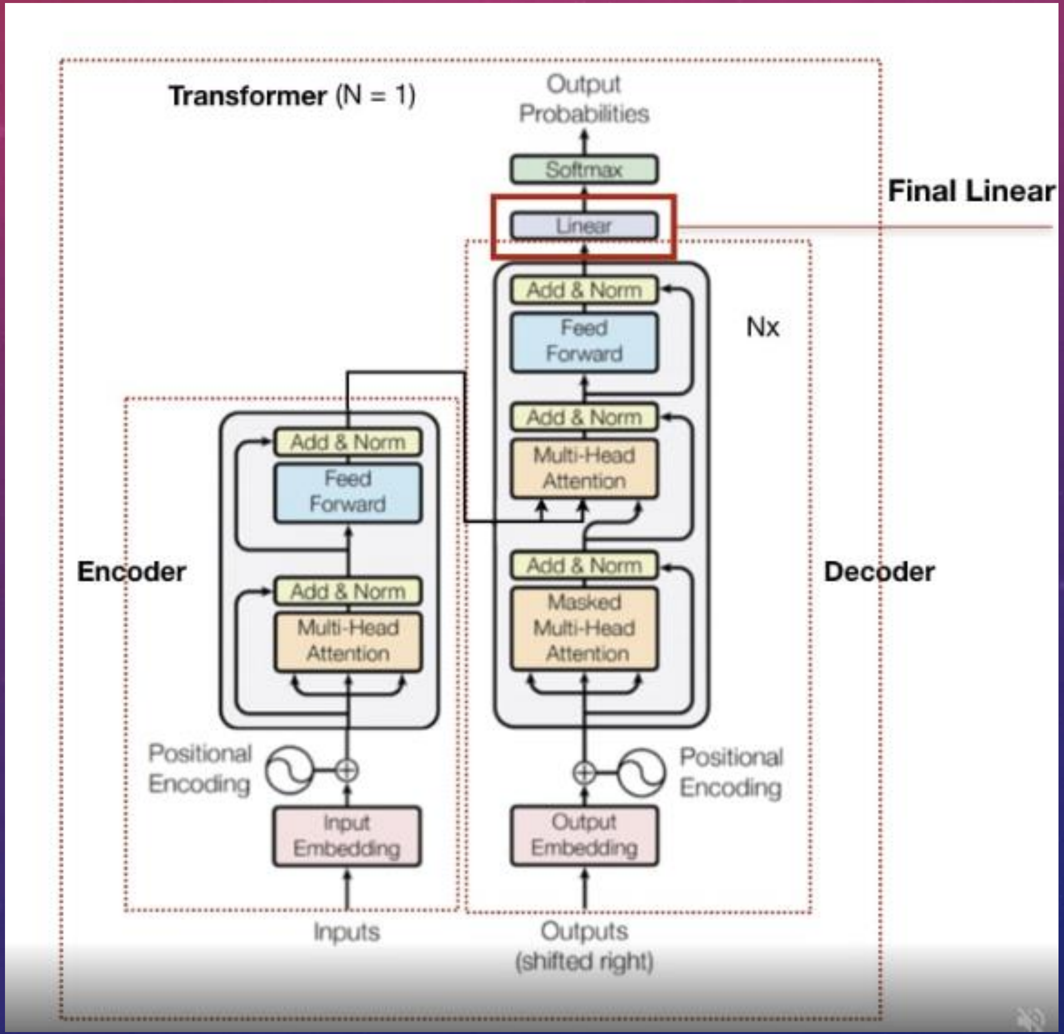


無法有效地平行運算

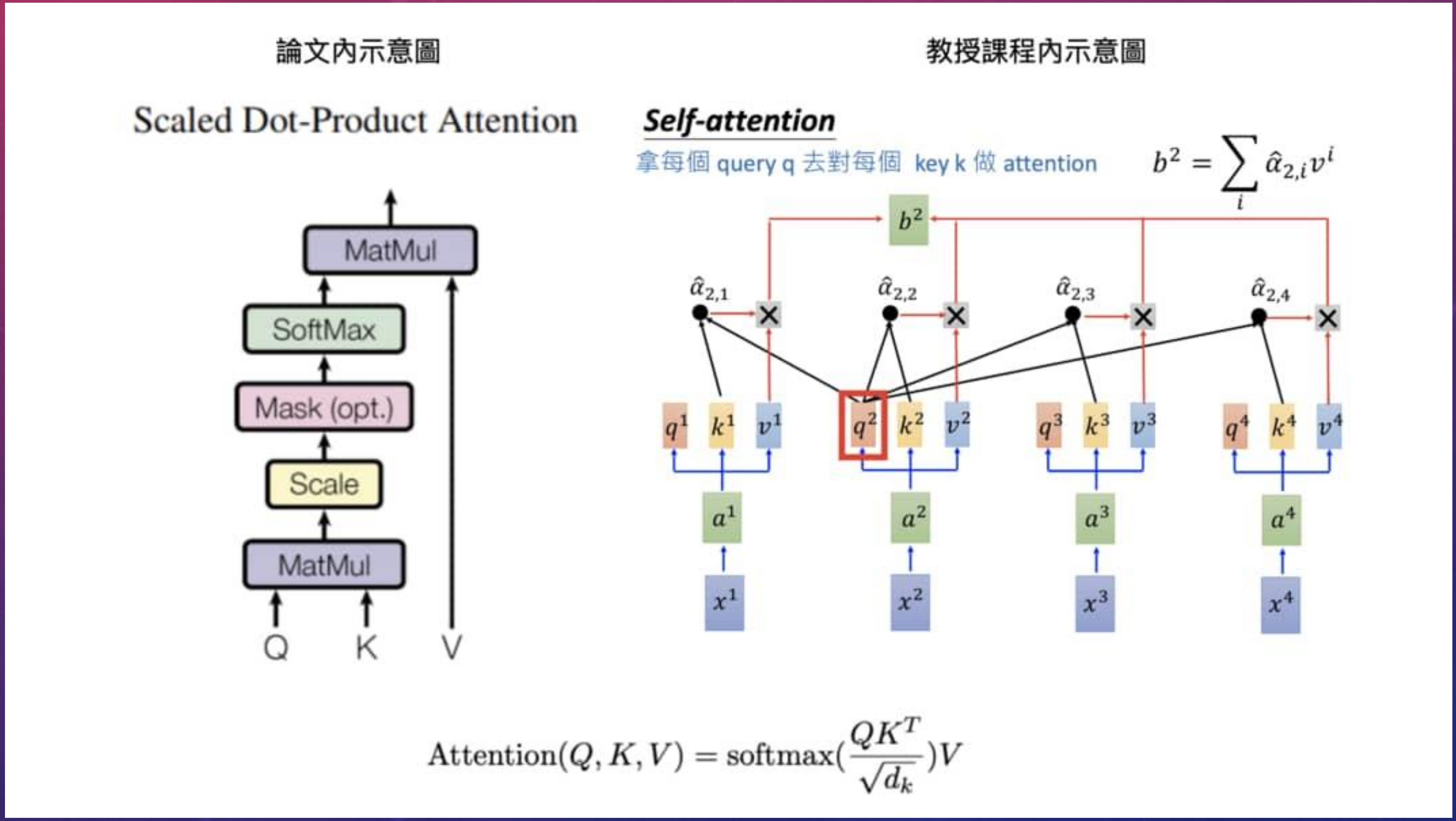
TRANSFORMER STRUCTURE



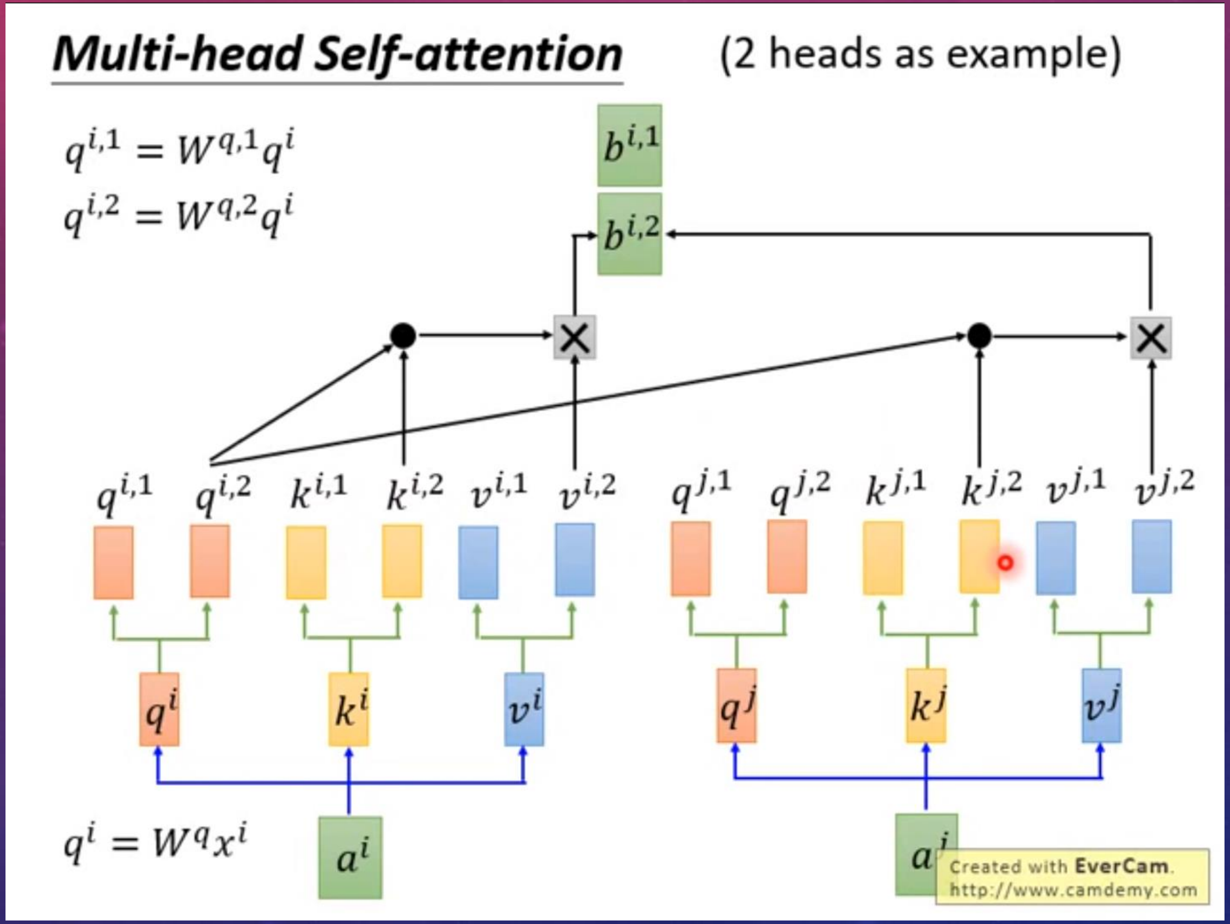
TRANSFORMER STRUCTURE



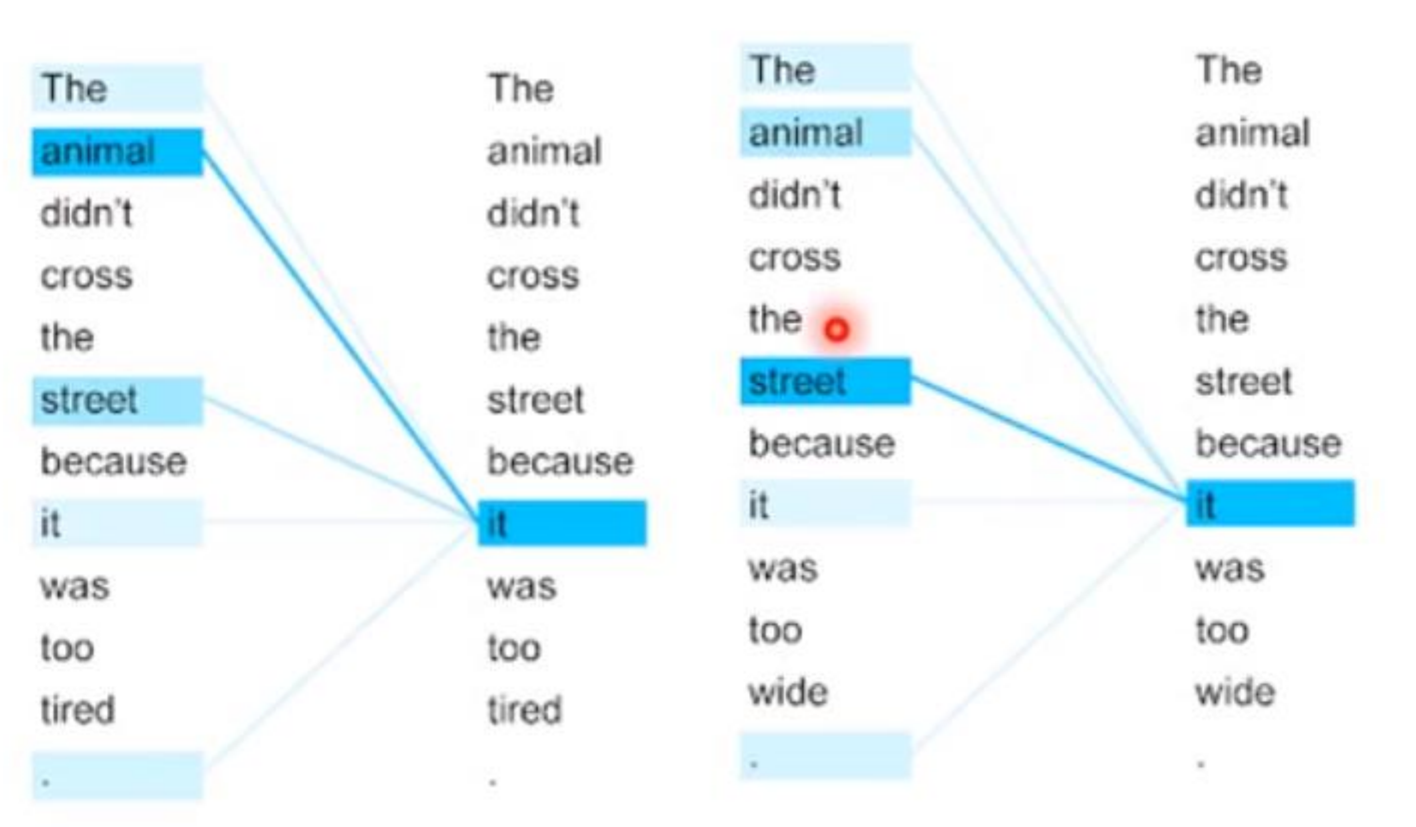
TRANSFORMER STRUCTURE



TRANSFORMER STRUCTURE



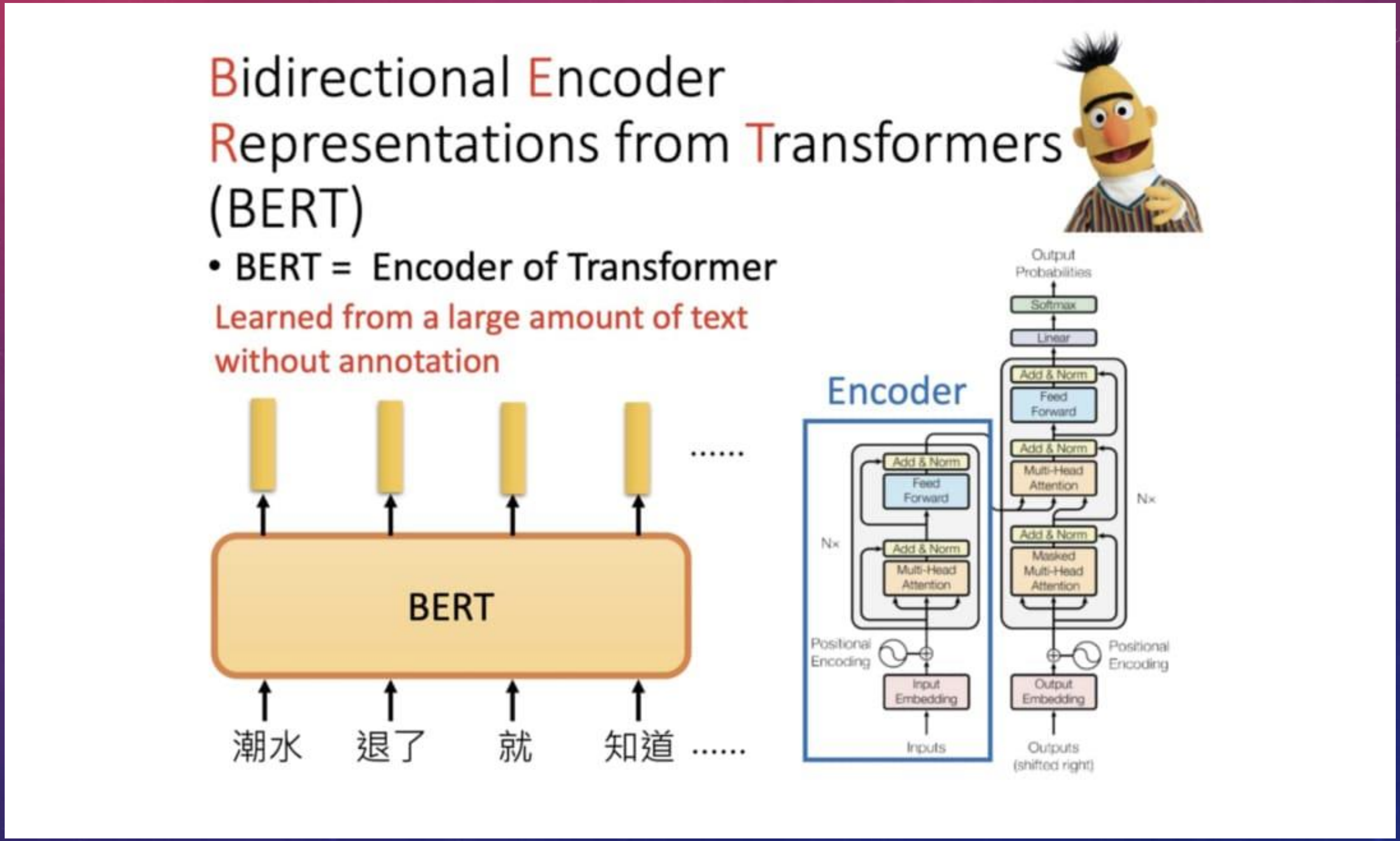
RESULTS



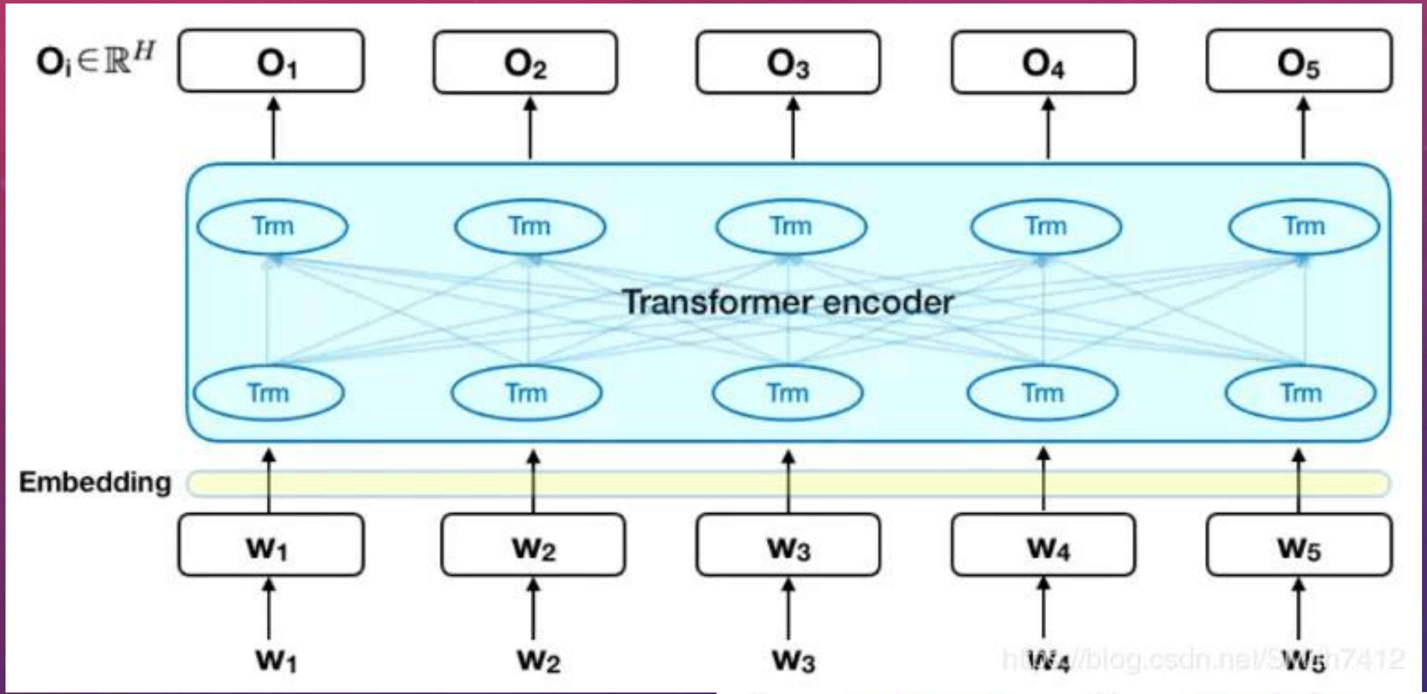
The
animal
didn't
cross
the
street
because
it
was
too
tired
.
The
animal
didn't
cross
the
street
because
it
was
too
tired
.
The
animal
didn't
cross
the
street
because
it
was
too
wide
.
The
animal
didn't
cross
the
street
because
it
was
too
wide
.

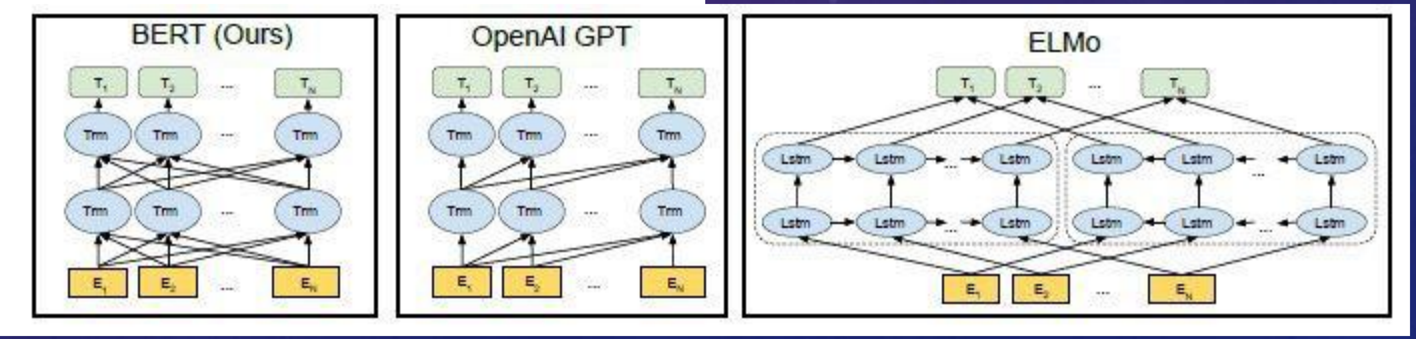
BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS, BERT

BERT= ENCODER OF TRANSFORMER

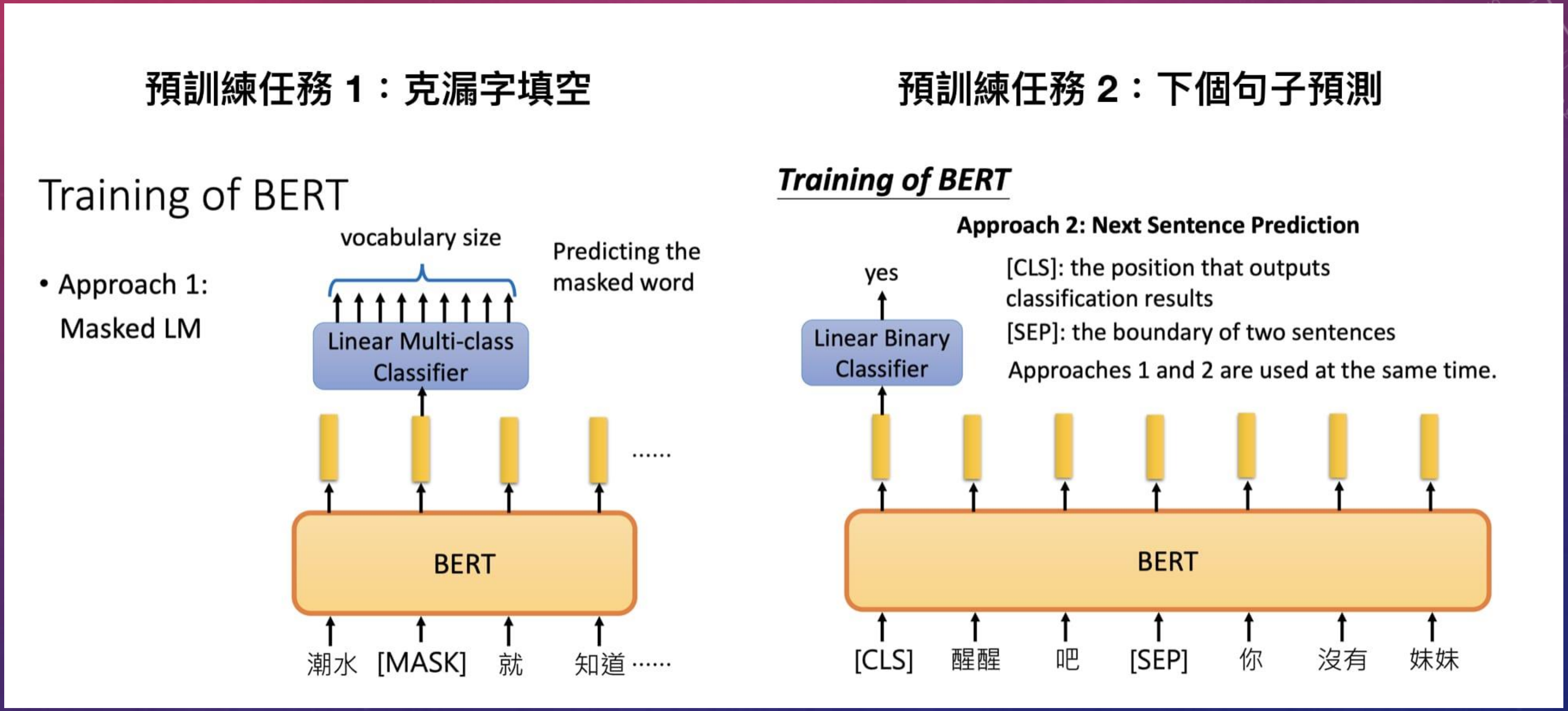


BERT STRUCTURE

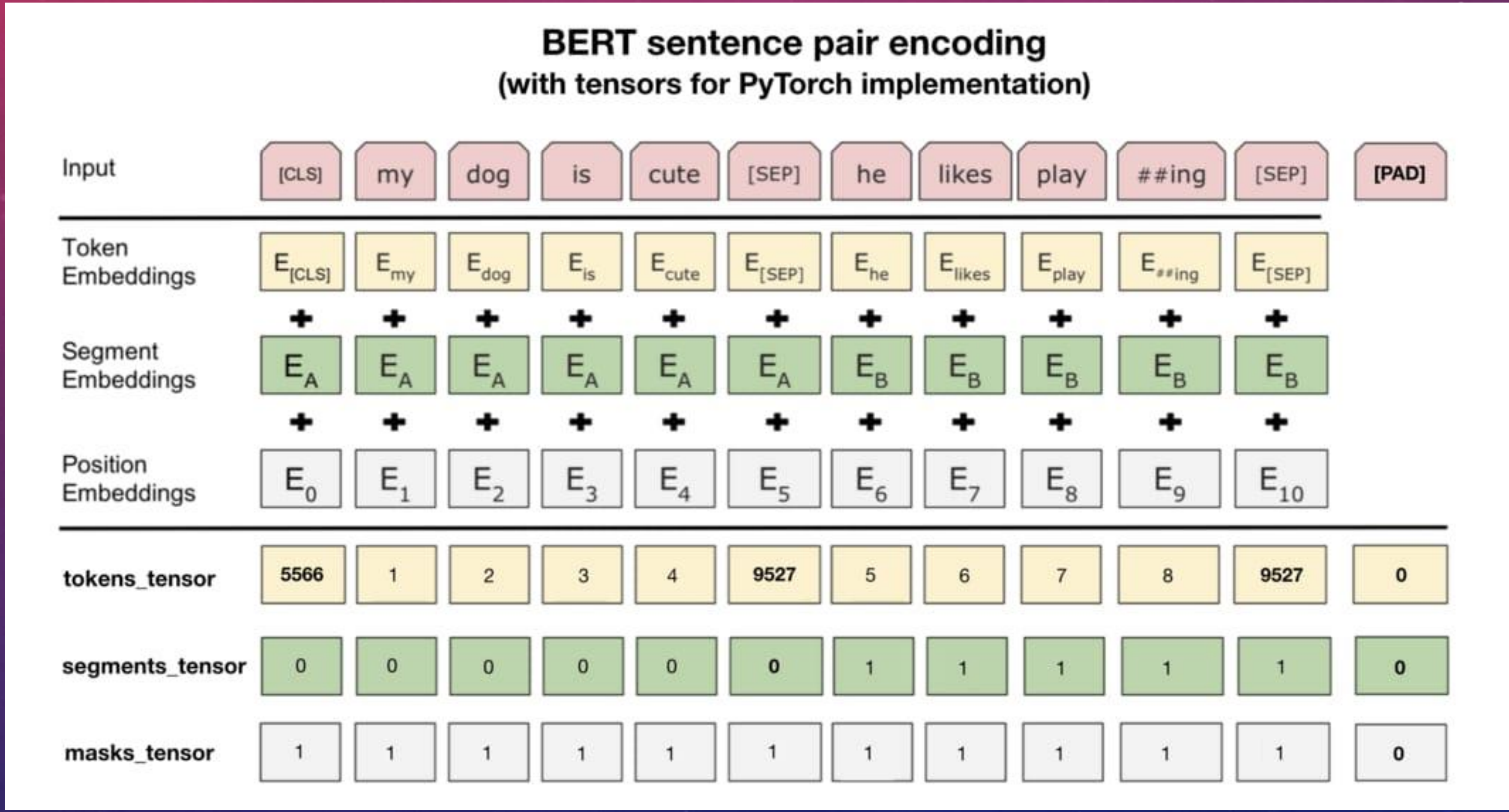




TRAINING OF BERT



KEY POINTS OF BERT



THANK YOU