智慧運算學院
College of Intelligent Computing
人工智慧學系
Department of Artificial Intelligence

# 智慧運算技術導論

# 自然語言處理篇 - Transformer

林英嘉 (Ying-Jia Lin)
長庚大學人工智慧學系

2026/01/27

Slido
# AIMD (Link)

# Outline

- Contextual Embedding 的概念
- RNN
- Transformer
- BERT

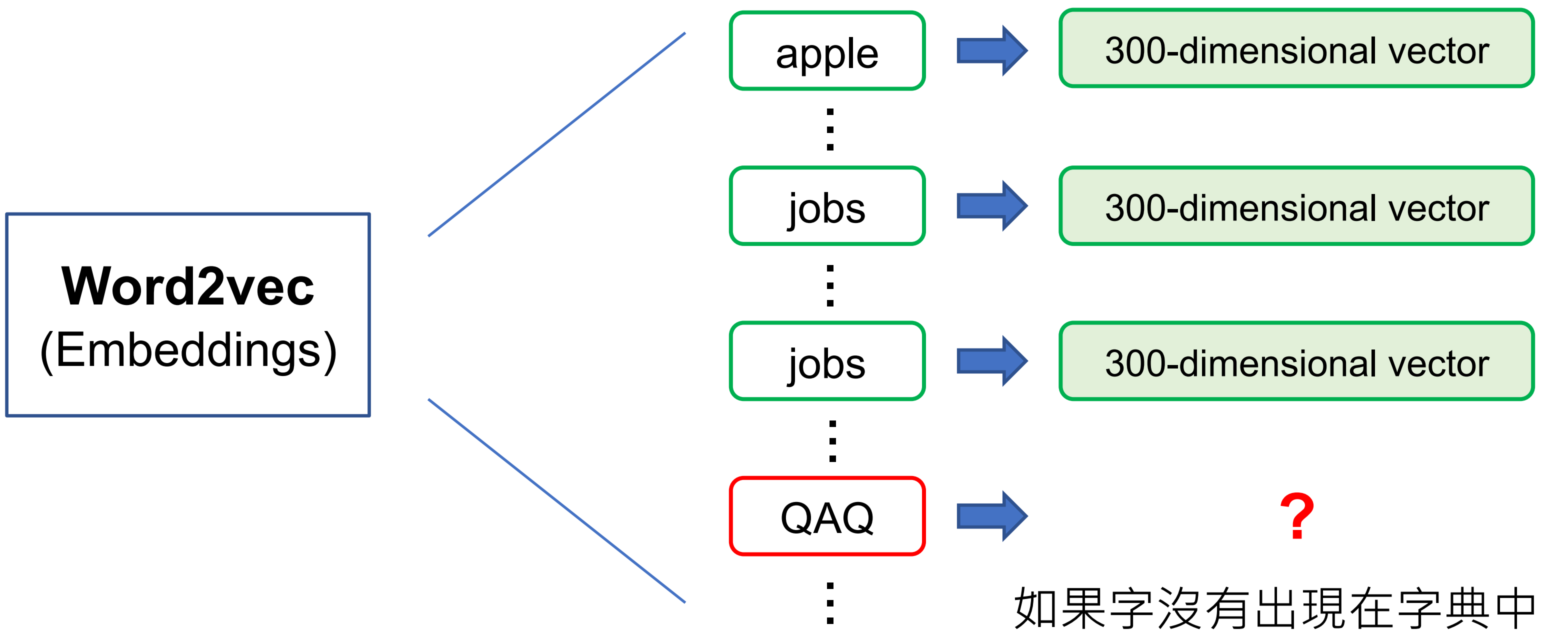
# Problems of Traditional Word Embeddings

1. Out-of-vocabulary problem
2. Model architecture for down-stream applications
    - Context-dependent representations

# The Out-of-vocabulary Problem

- Before training word embeddings, words are split by whitespaces for English corpora

**Vocabulary**
(假設字典的內容由a排到z)




如果字沒有出現在字典中 (可能此字不存在於訓練資料)，那它就不會有對應的字向量

# Sub-word Tokenization (Byte Pair Encoding)

- Byte Pair Encoding (BPE)[1] is a sub-word (子詞) tokenization technique used to handle rare words by breaking them into more frequent sub-word units.

Example sentence: I printed Hello world.

**Traditional word tokenization**

I | printed | Hello | world

**Traditional with BPE**

I | prin | ted | Hell | o | world

分詞方法的決定是無監督式的



[1] Sennrich et al. "Neural Machine Translation of Rare Words with Subword Units." ACL 2016.

# Sub-word Tokenization 的好處

- With sub-word tokenization algorithms like BPE, we can handle representations for unknown words (or mis-spelled words).
- In machine translation , the compound word issues between source and target languages can be alleviated.
- State-of-the-art pre-trained language models (e.g., GPT-3, BERT) adopt sub-word tokenization algorithms before pre-training.

# Problems of Traditional Word Embeddings

1. Out-of-vocabulary problem
2. Model architecture for down-stream applications
   - Context-dependent representations

# 單純使用 Word Embeddings 並無上下文知識

- [Recap] Word2Vec 的範例長相:

Dimension size (e.g., 300)

Vocabulary size (e.g., 30,000)

apple -0.110960 0.016115 -0.004809 0.033589 0.121455 …
banana -0.027713 -0.015676 0.003314 0.077602 0.159718 …
…
…
…

open a bank account

on the river bank

the same word vector

# 語言的模式

⭐ 上下文非常重要

句子1：

小明昨天遇到小美，他對她說：「下次一起去看電影吧！」，請問小明下次想去跟誰看電影？

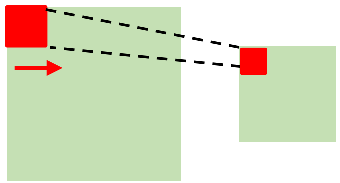
句子2：

「庭院深深深幾許」，請問有幾個深？

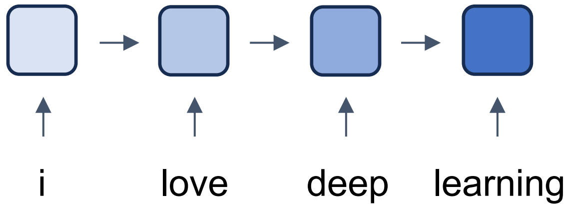
# Contextual Embedding

# 深度學習常見模型架構

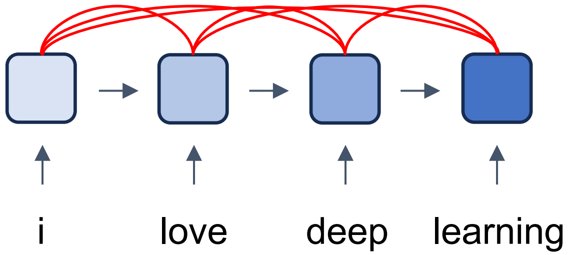
**CNN**
**(著重局部資訊)**



**RNN**
**(著重記憶力)**

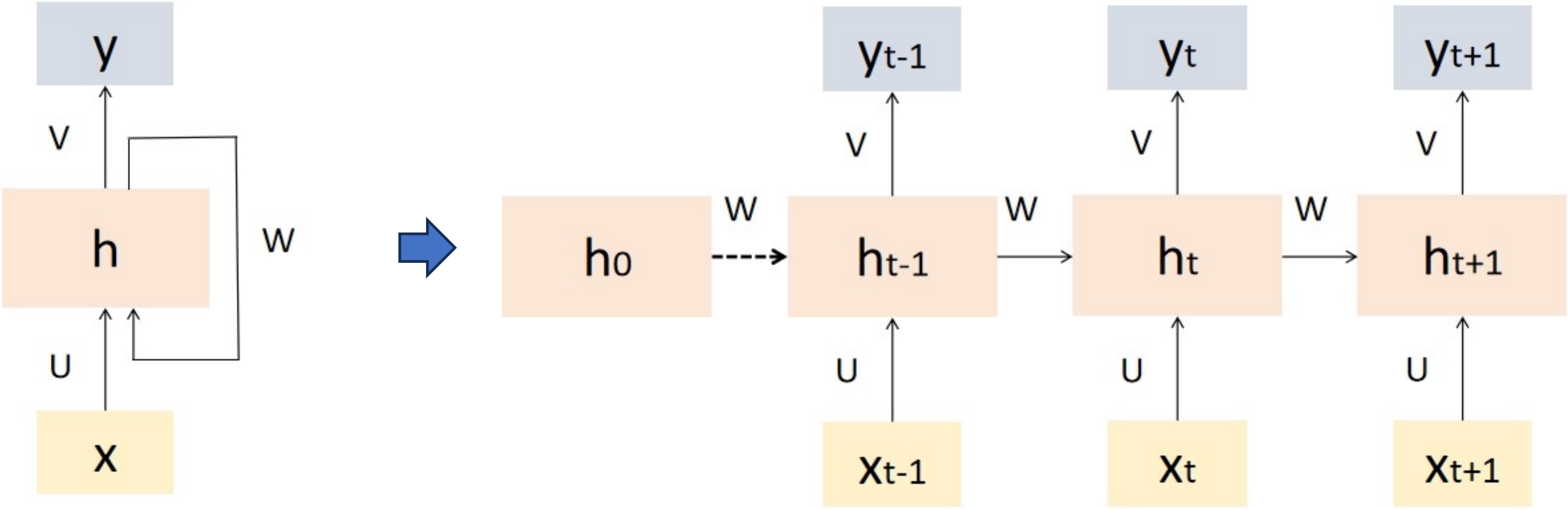



**Transformer**
**(著重全局資訊與記憶力)**

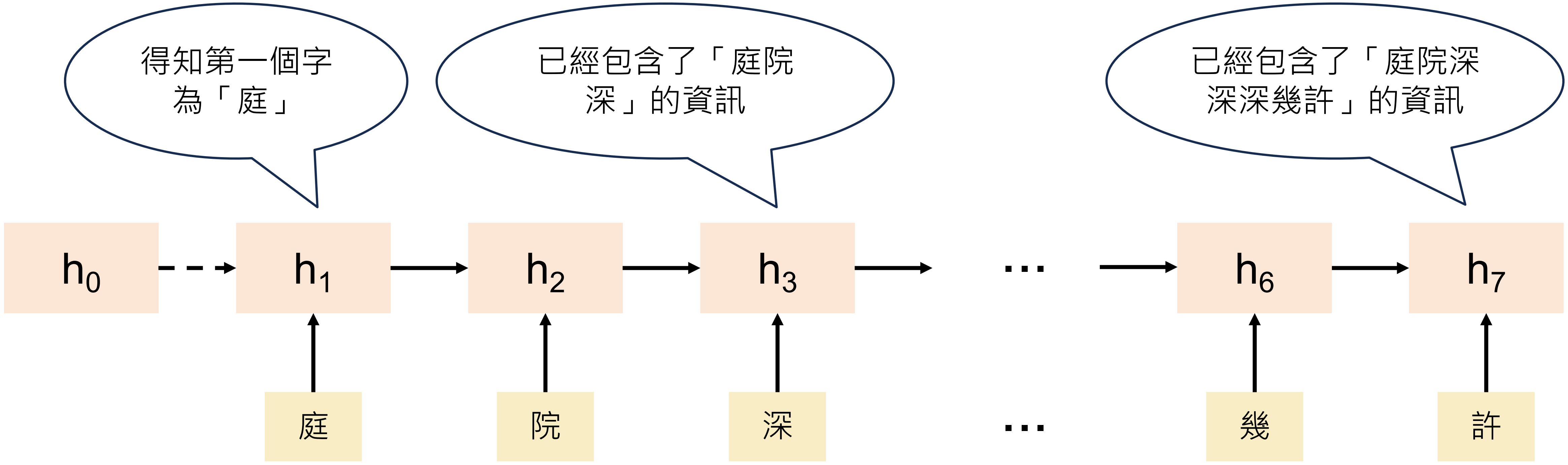


# Recurrent Neural Networks (RNN)

- Recurrent Neural Networks (RNNs) are a type of artificial neural network designed to handle sequential data by capturing temporal dependencies.
  - RNN 的“遞迴”是指它在每一個時間步中重複使用相同的參數（同一組權重），並把先前的隱藏狀態傳到下一個時間步



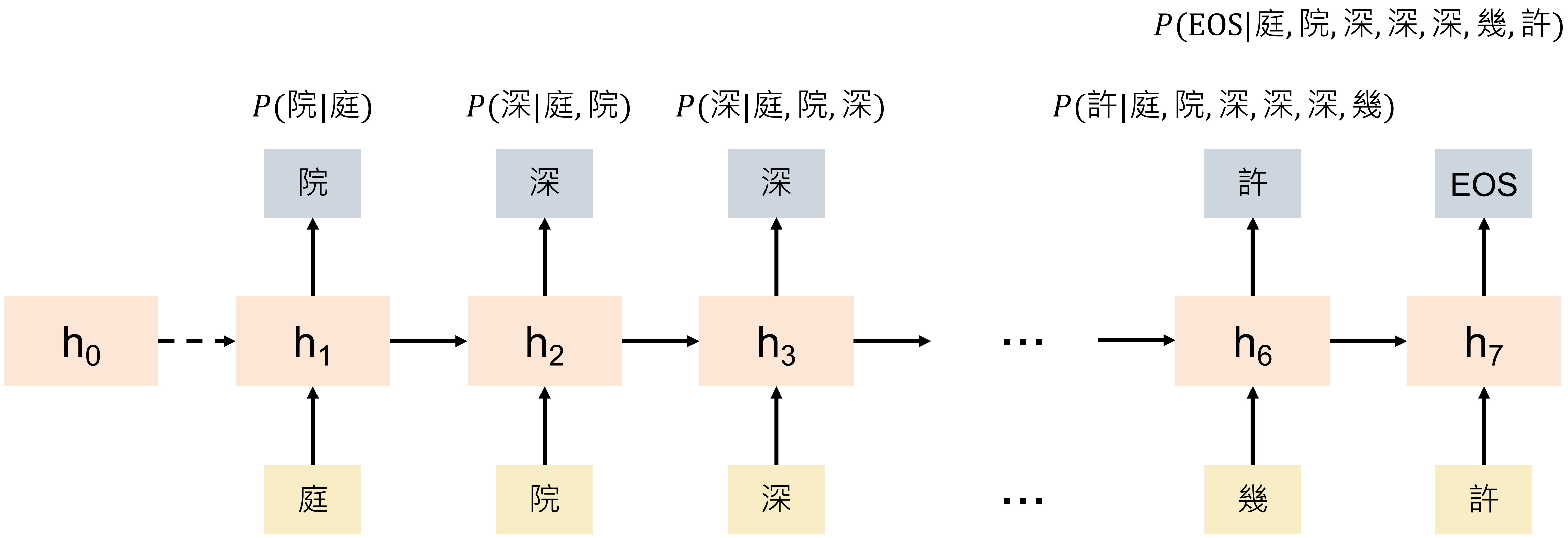
# RNN 的記憶力意義



得知第一個字
為「庭」
已經包含了「庭院
深」的資訊
已經包含了「庭院深
深深幾許」的資訊
$h_0$
$h_1$
$h_2$
$h_3$
…
$h_6$
$h_7$
庭
院
深
…
幾
許

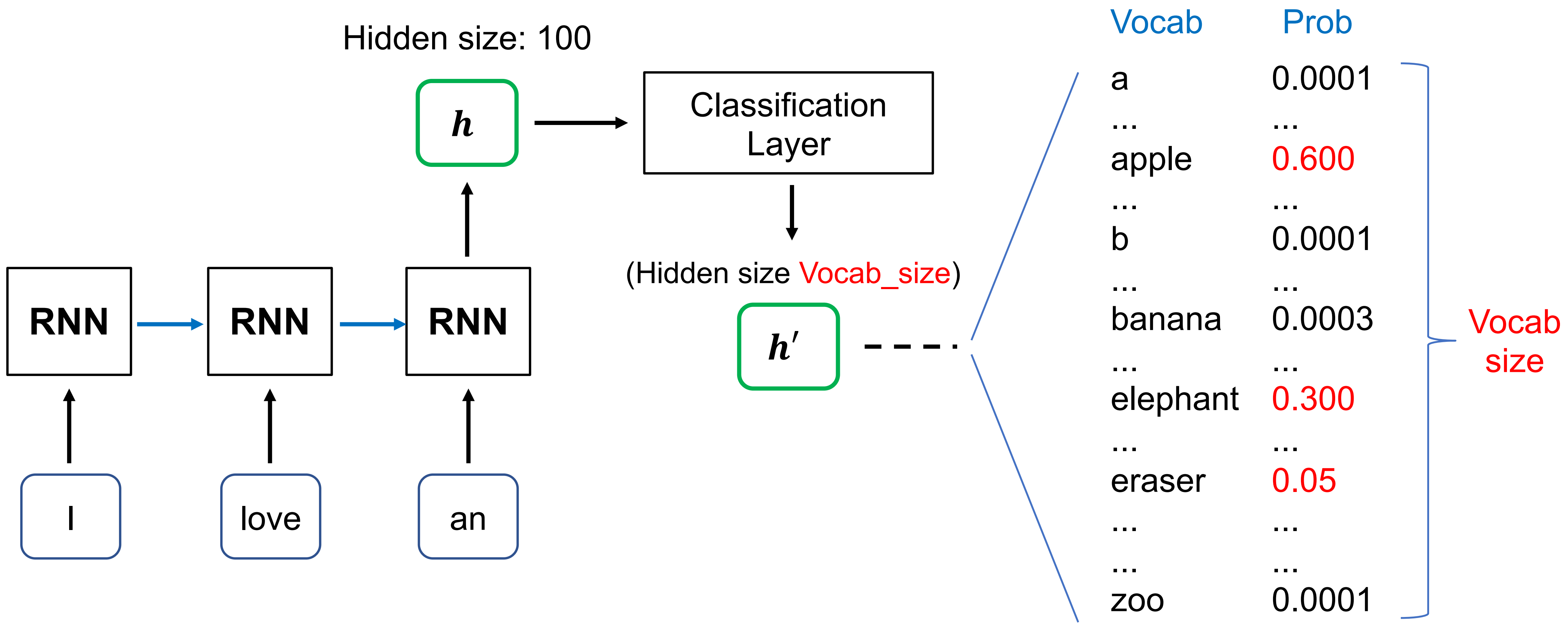
# RNN 以文字接龍進行訓練

EOS: End of Sentence

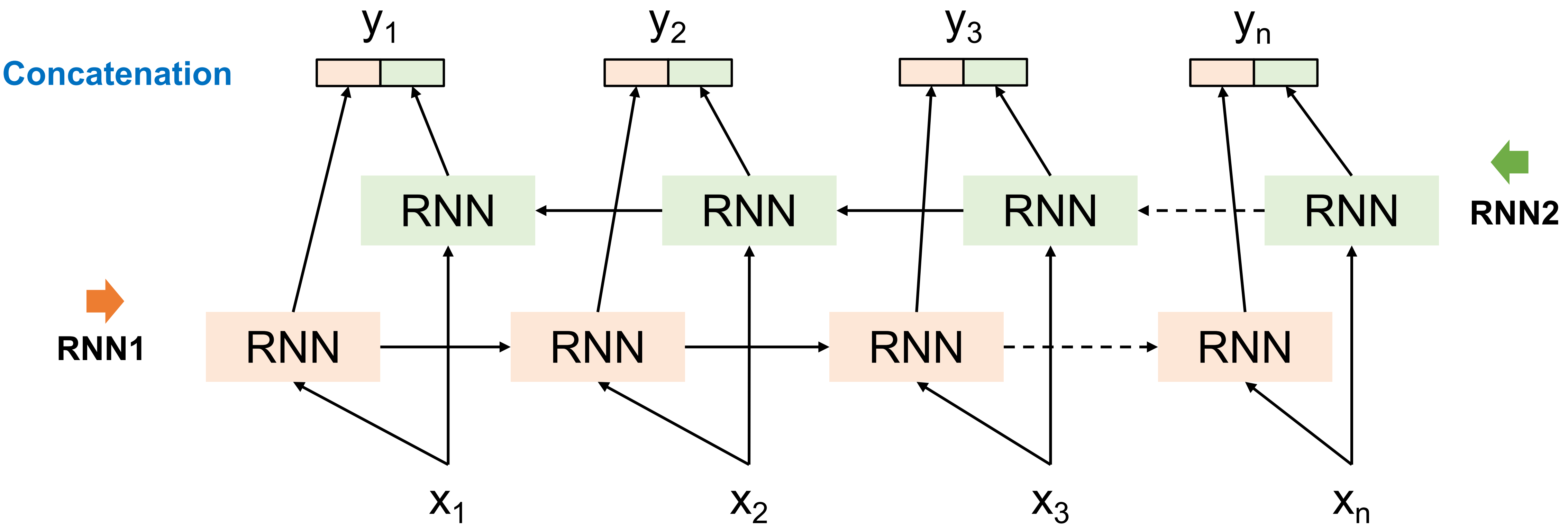



目標函數：$\log(P(\text{院}|\text{庭}) \times P(\text{深}|\text{庭},\text{院}) \times P(\text{深}|\text{庭},\text{院},\text{深}) \times \ldots \times P(\text{EOS}|\text{庭},\text{院},\text{深},\text{深},\text{深},\text{幾},\text{許}))$

# Text Generation with RNN

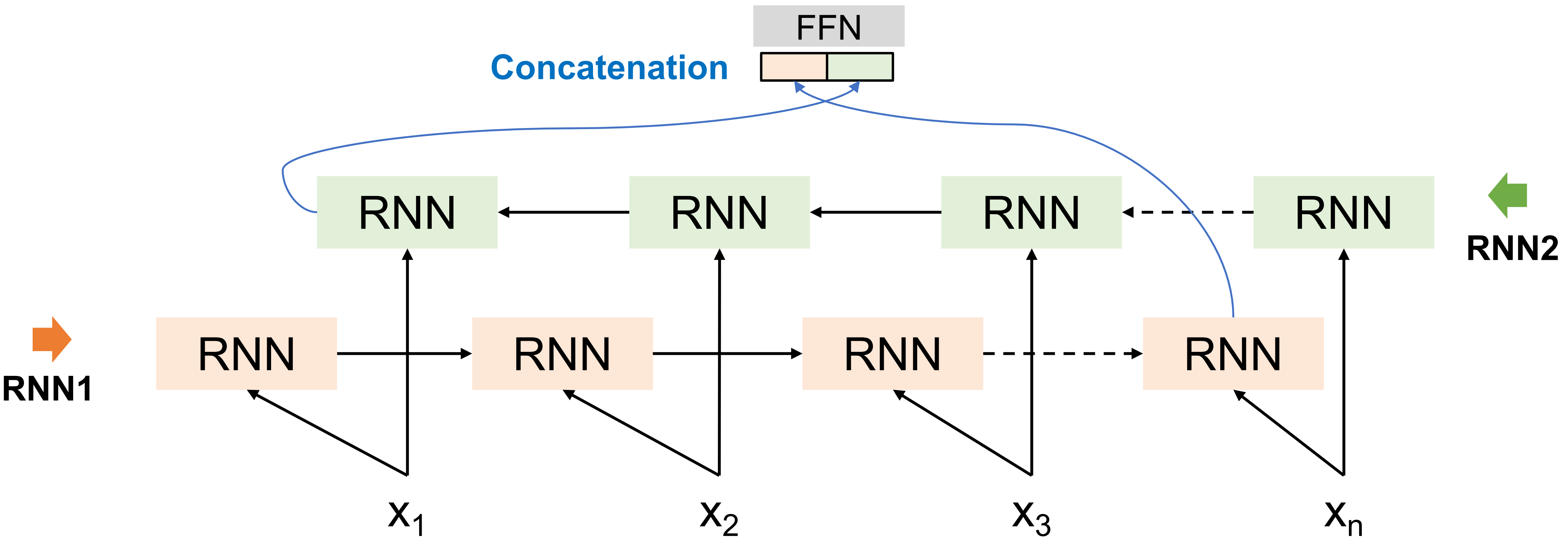


# Bidirectional RNNs

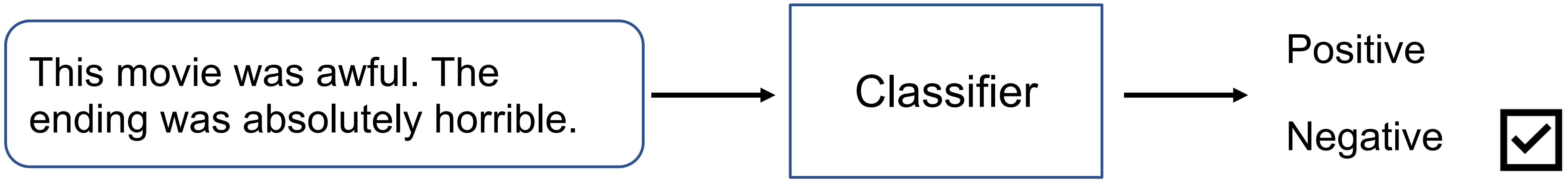


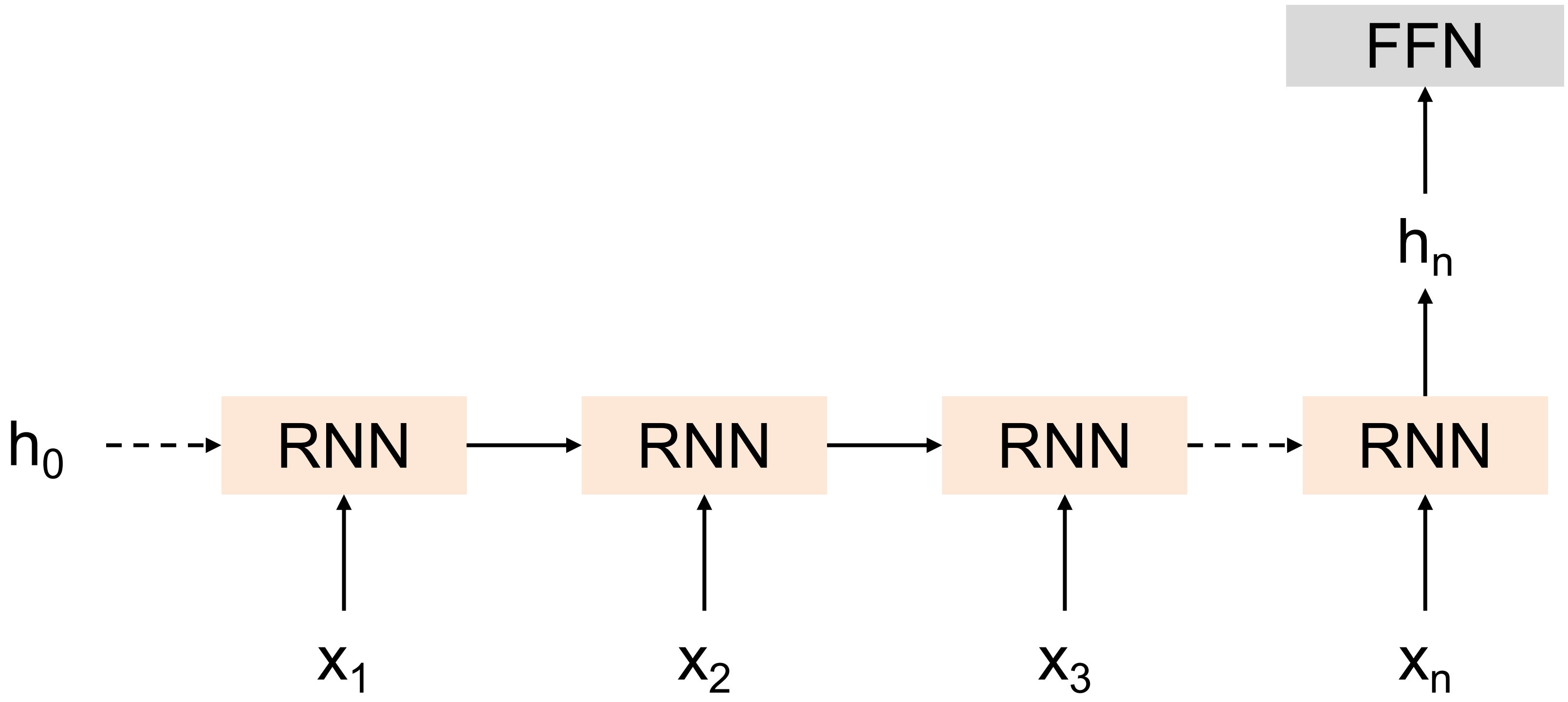
# Bidirectional RNNs for Sequence Classification

- The final hidden units from the forward and backward passes are combined to represent the entire sequence.
- This combined representation serves as input to the subsequent classifier.

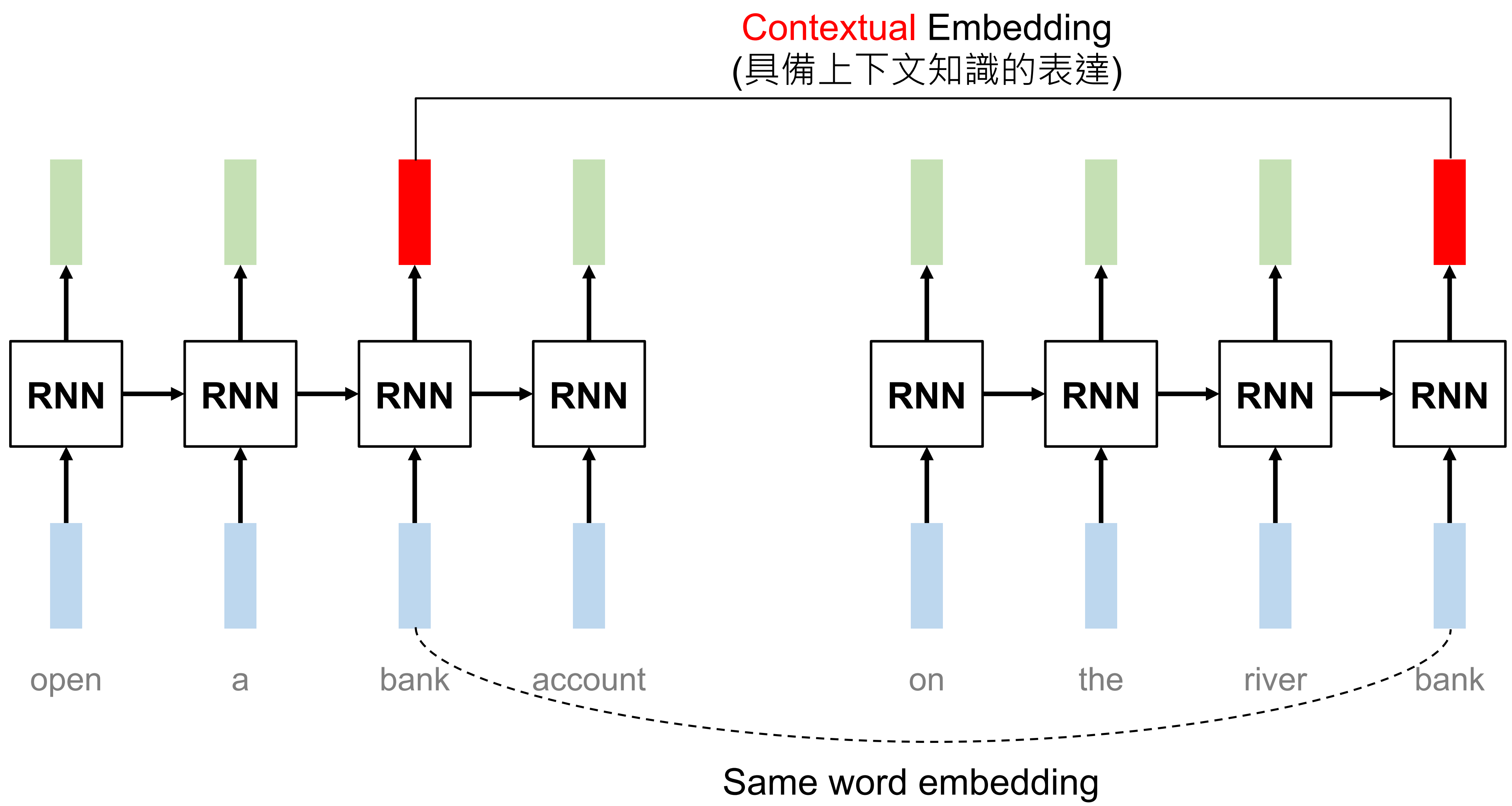


# RNNs for Sequence Classification



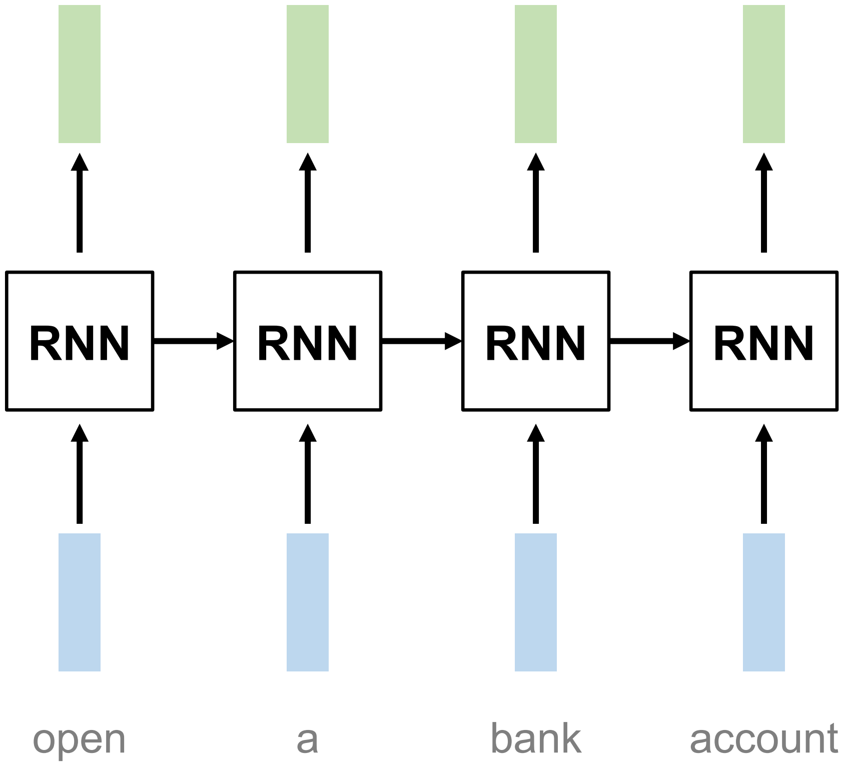



# 為什麼我們需要序列模型?

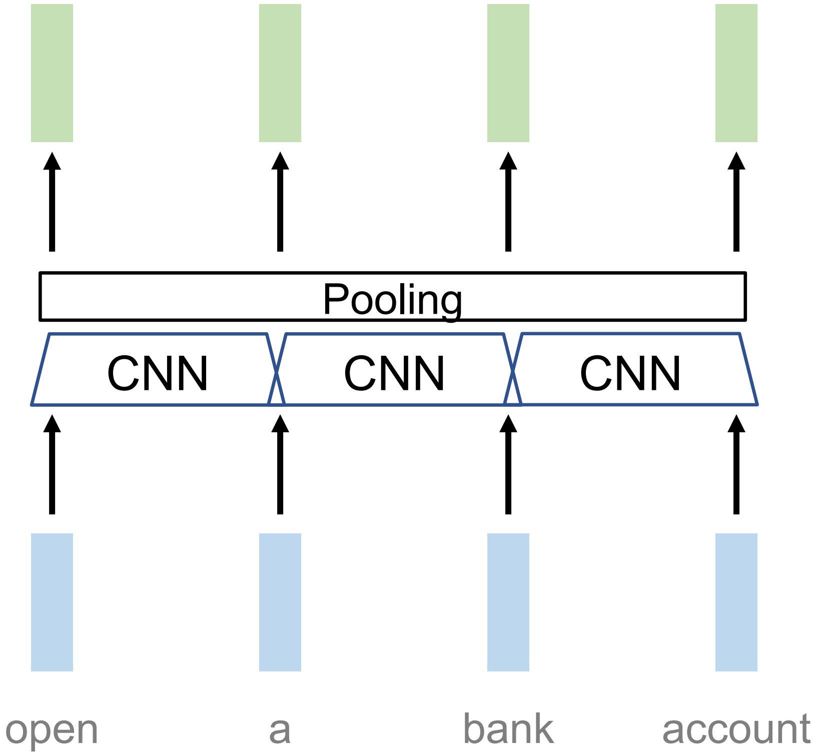


# RNN 與 CNN 作為序列模型的問題

問題：RNN 難以平行化




問題：CNN 著重局部資訊




NLP

# Transformer

# Tokens vs. words

- token 是 (語言) 模型在每個時間點處理的單位
- word 是語言本身的單位
  - token 可以是 word，也可以是 sub-word
  - 一個 word 可以是一個 token，但單純講 token 不一定指的是 word

**Traditional word tokenization**

| I | printed | Hello | world |
|---|---|---|---|

**Sub-word Tokenization**

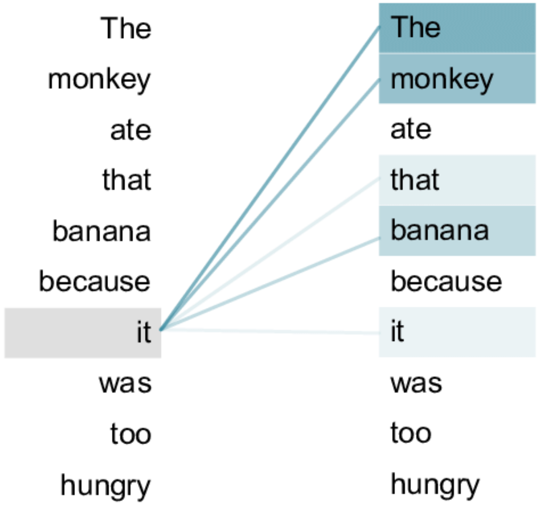
| I | prin | ted | Hell | o | world |
|---|---|---|---|---|---|

# Transformer

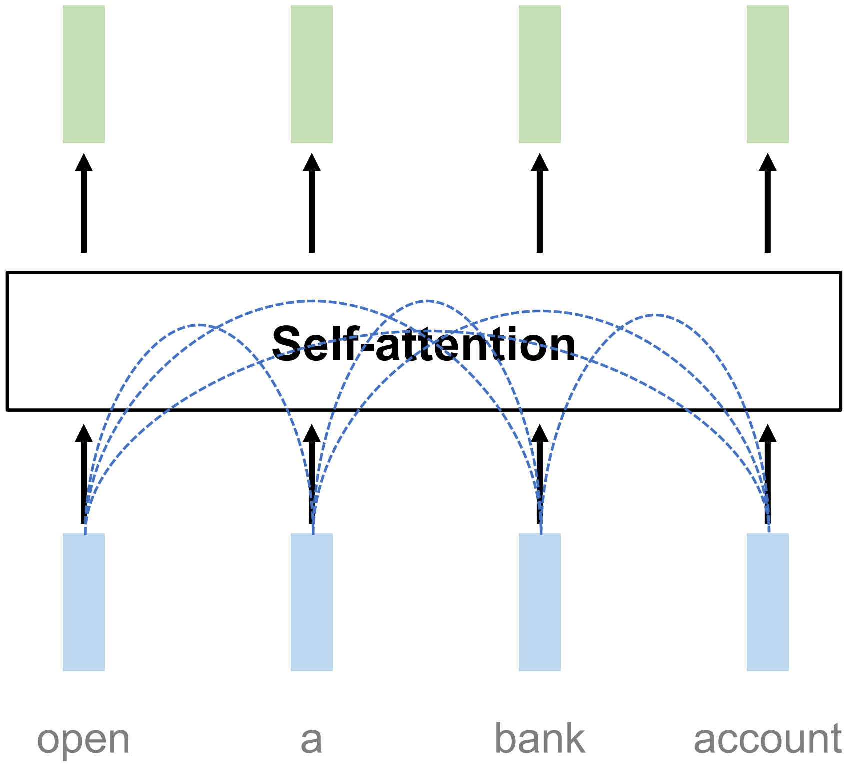
Source: An example of the self-attention mechanism following long-distance... | Download Scientific Diagram (researchgate.net)

- Transformer 來自論文：Attention Is All You Need (Vaswani et al., 2017)
- Transformer 內部的主要機制為 Self-attention

Self-attention: 句子內的每個 token 自己跟自己做 attention




attention 進行的過程需要主動與被動

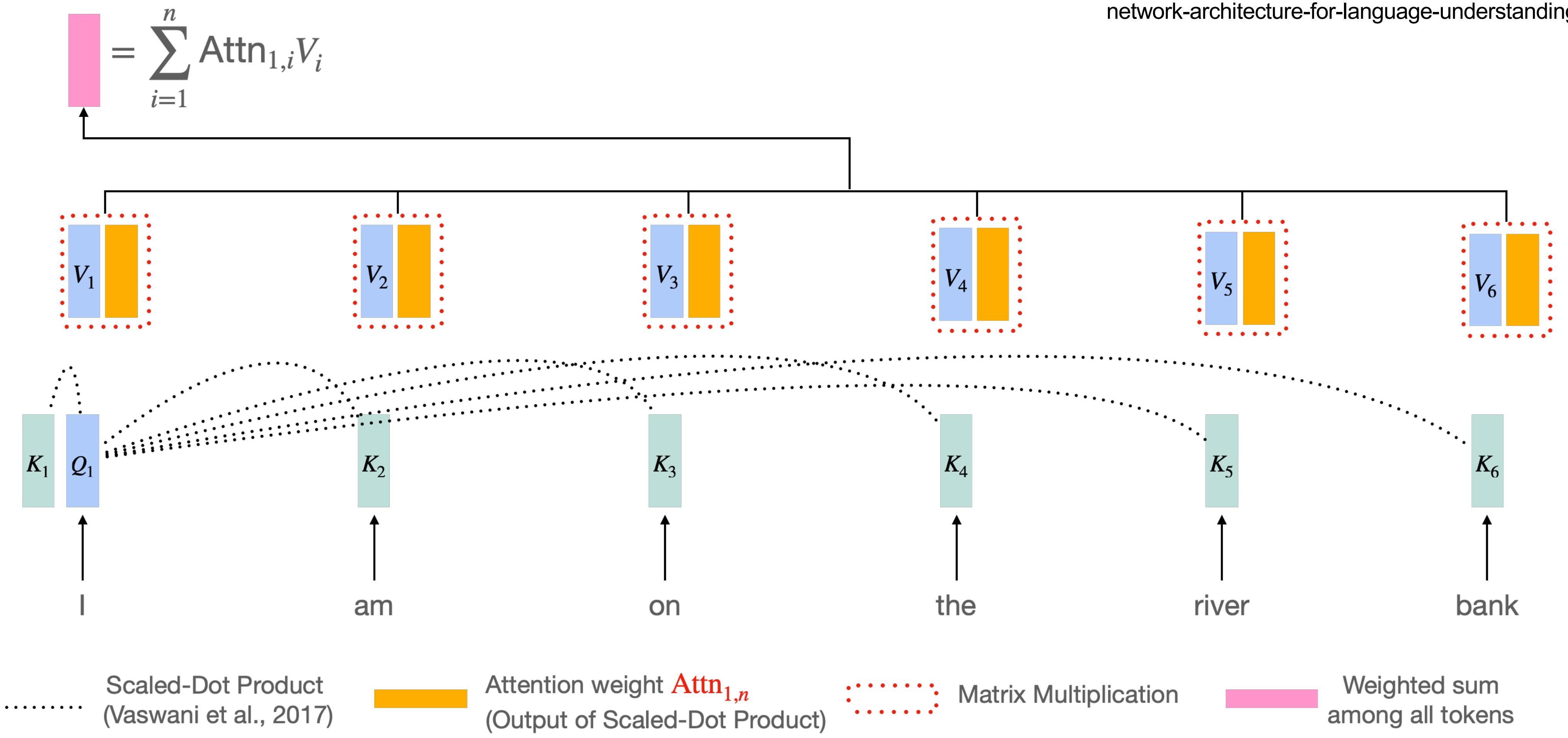


# Self-attention 動畫

Figure source: https://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/

# Self-Attention (Transformer) 過程

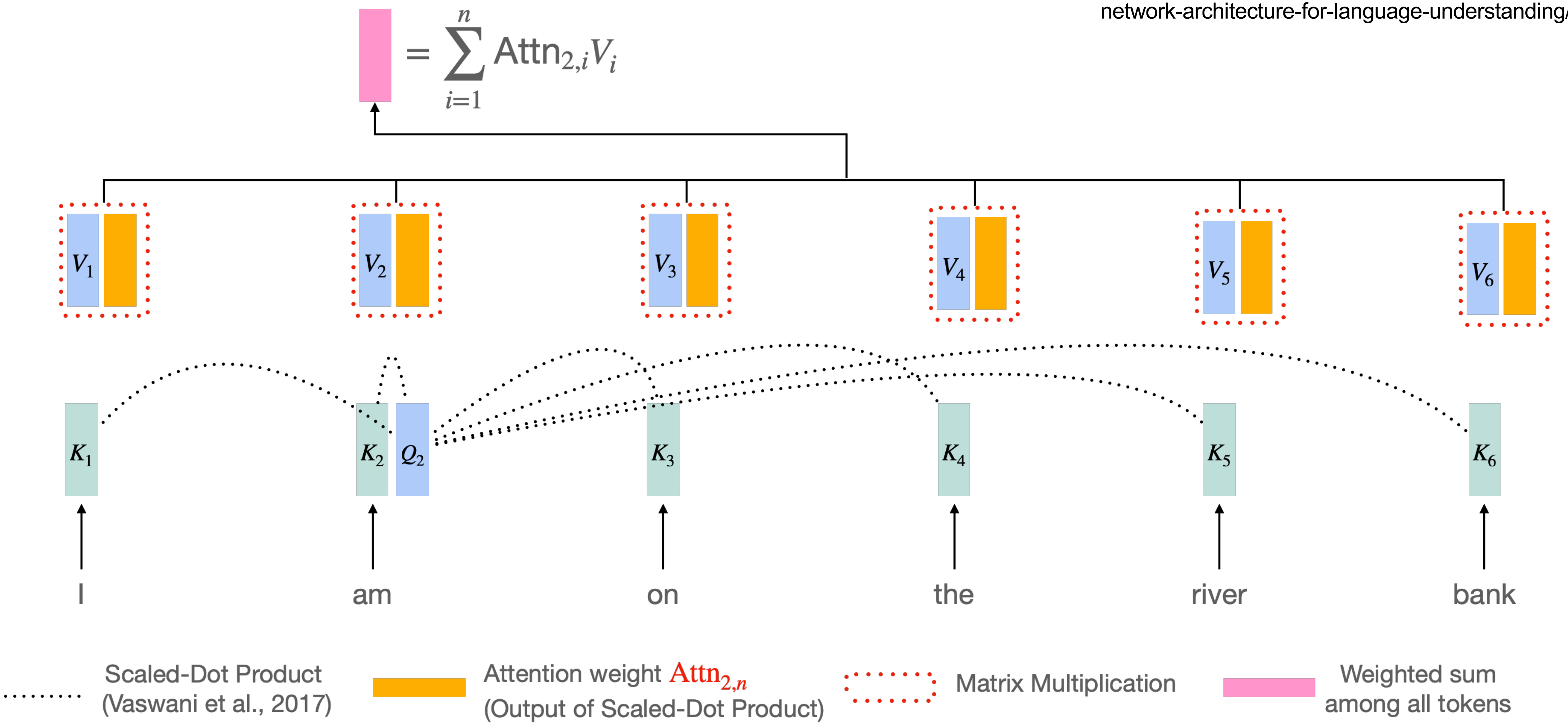
Vaswani et al. "Attention is all you need." NeurIPS 2017.
https://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/




(For simplicity, I only draw the self-attention process for the first token.)

# Self-Attention (Transformer) 過程

Vaswani et al. "Attention is all you need." NeurIPS 2017. https://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/

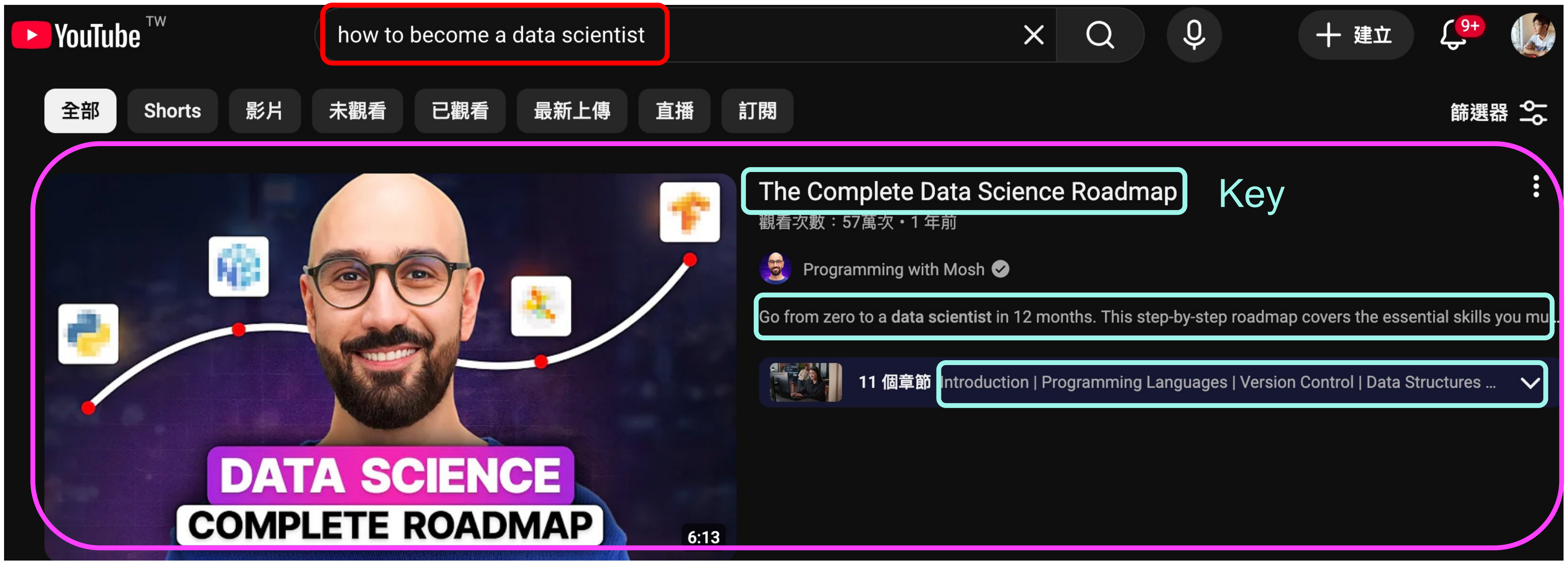



(For simplicity, I only draw the self-attention process for the second token.)

# QKV 的比喻

Vaswani, Ashish, et al. "Attention is all you need." NeurIPS (2017).

Query



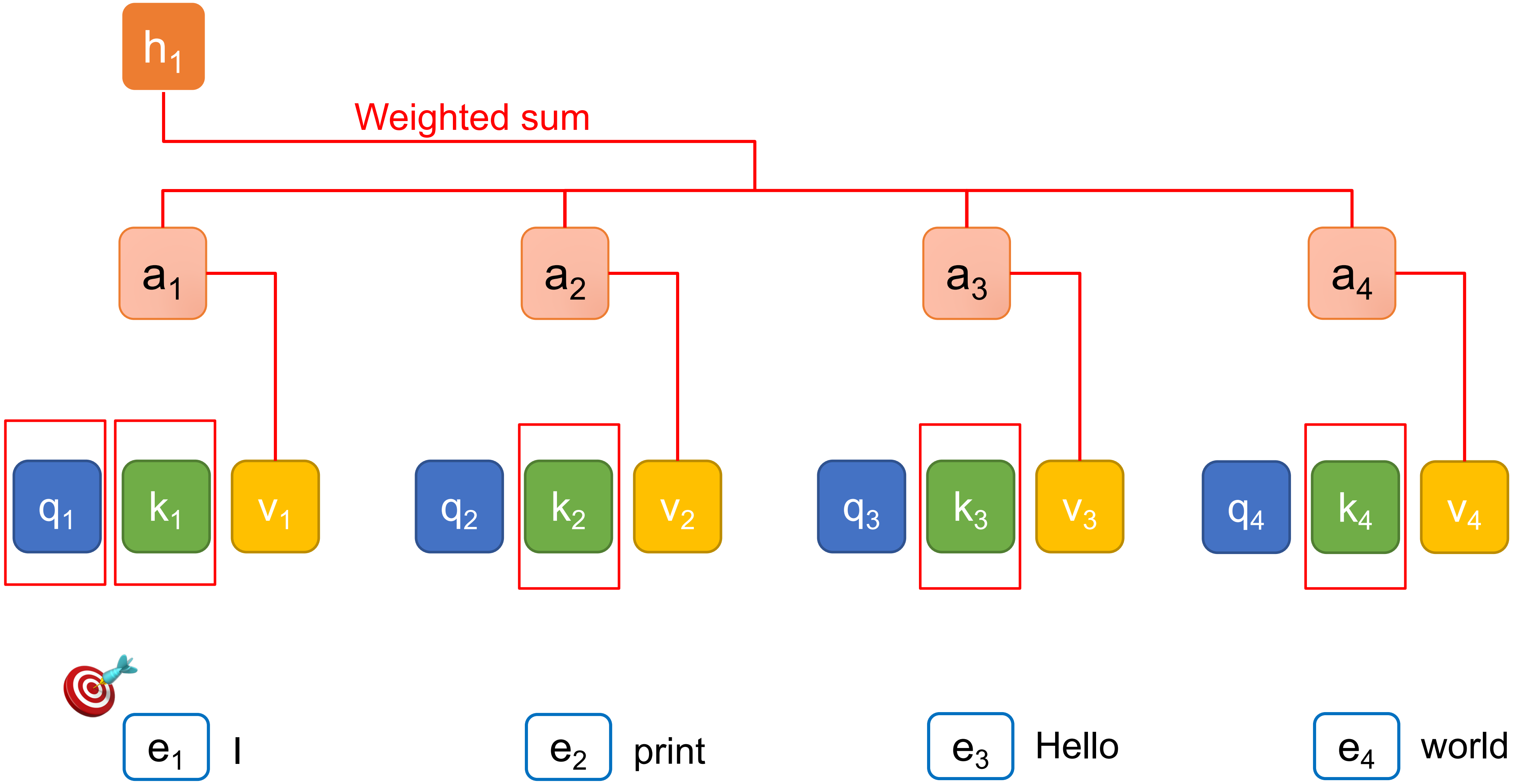

Value

- Query: (主動) 查詢的關鍵字；Key: (被動) 查詢的對象
- Value: 值，關鍵字與對象的匹配程度

# Query (Q), Key (K), and Value (V)

t = 1

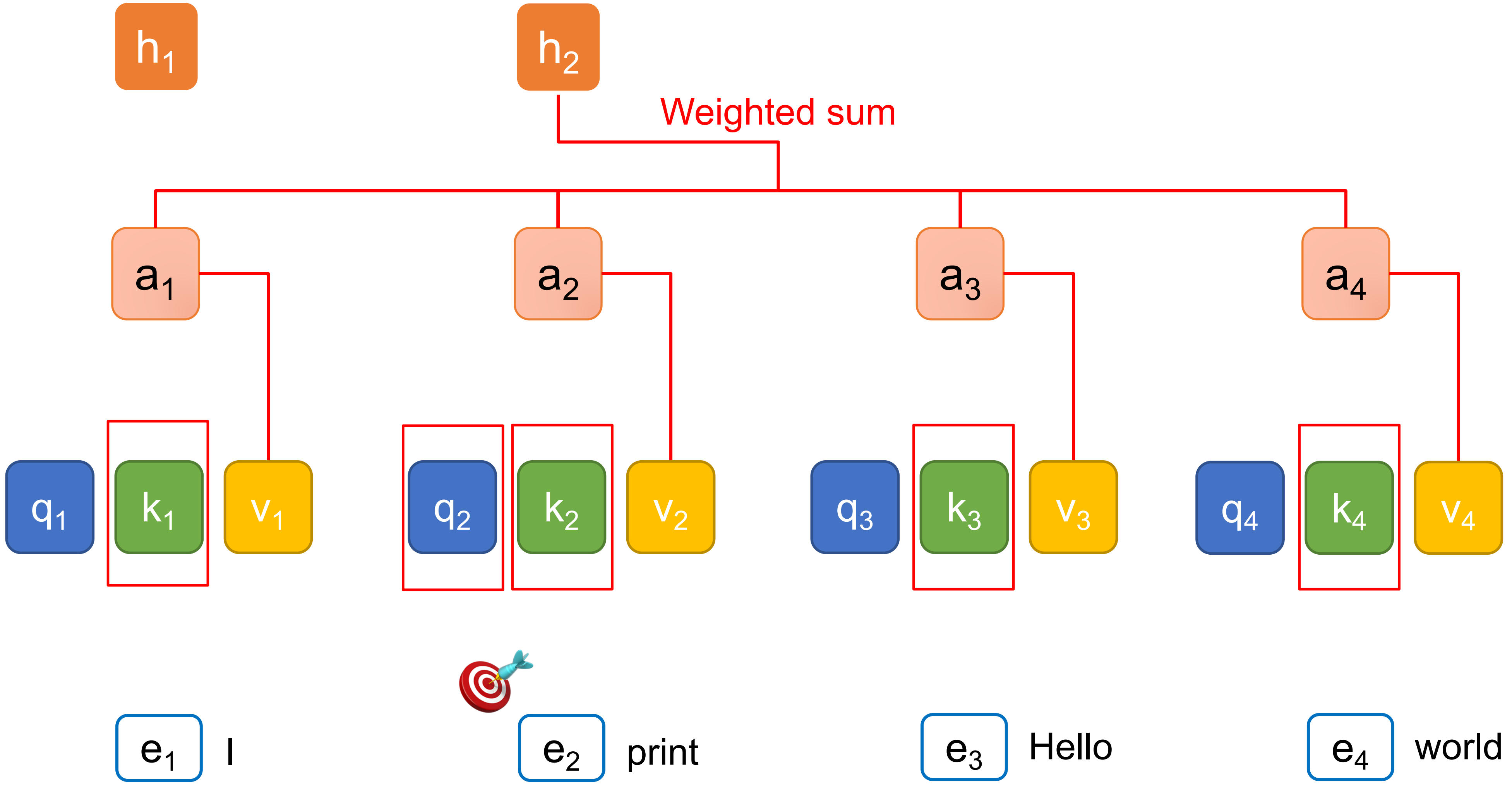
Self-attention 的過程



# Query (Q), Key (K), and Value (V)

t = 2

Self-attention 的過程

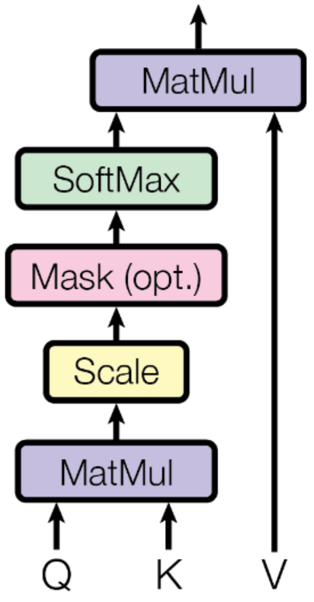


# 為什麼 Transformers 可以平行化？

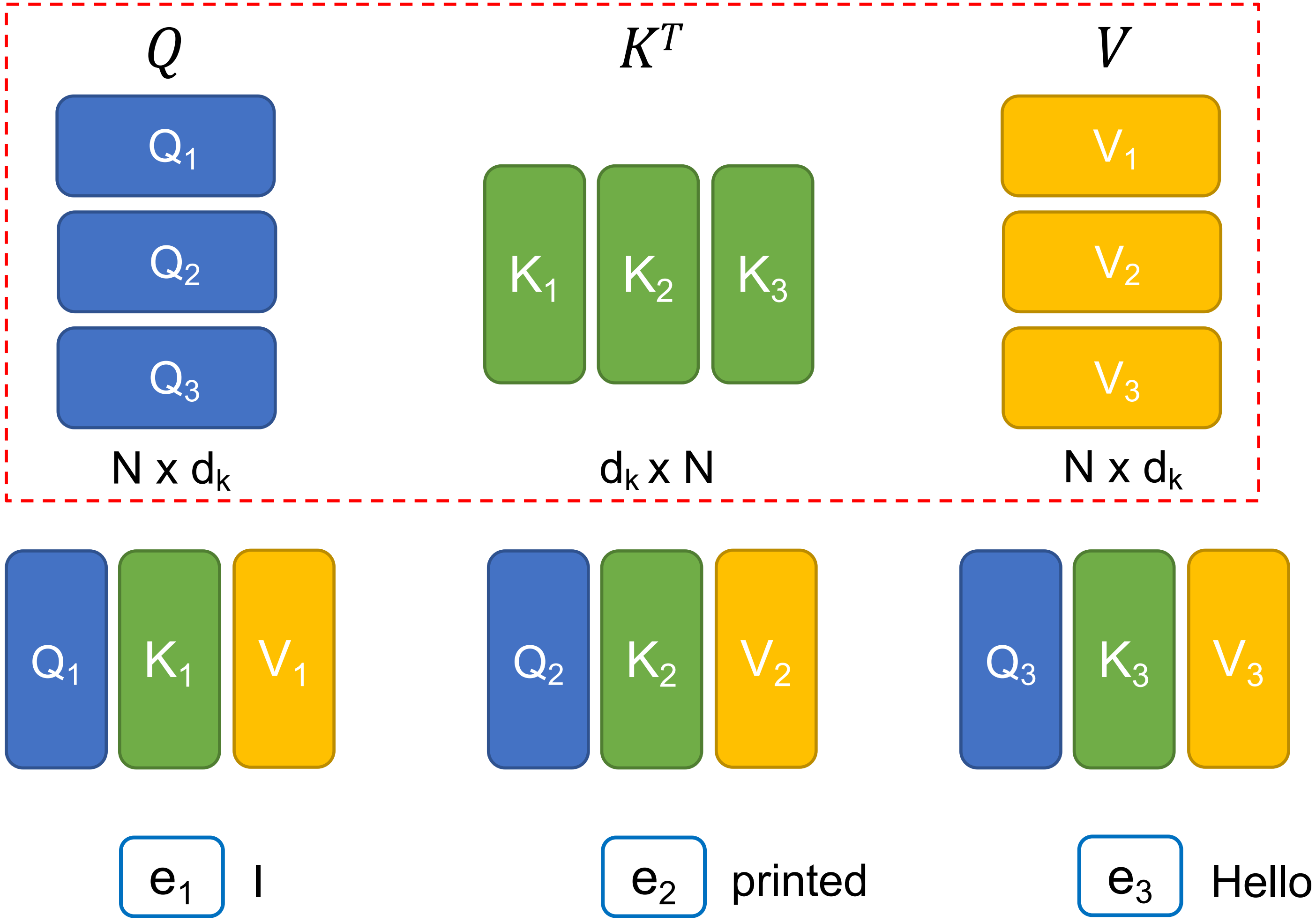
$d_q = d_k = d_v$

(自注意力機制可以利用矩陣乘積來進行平行化計算)

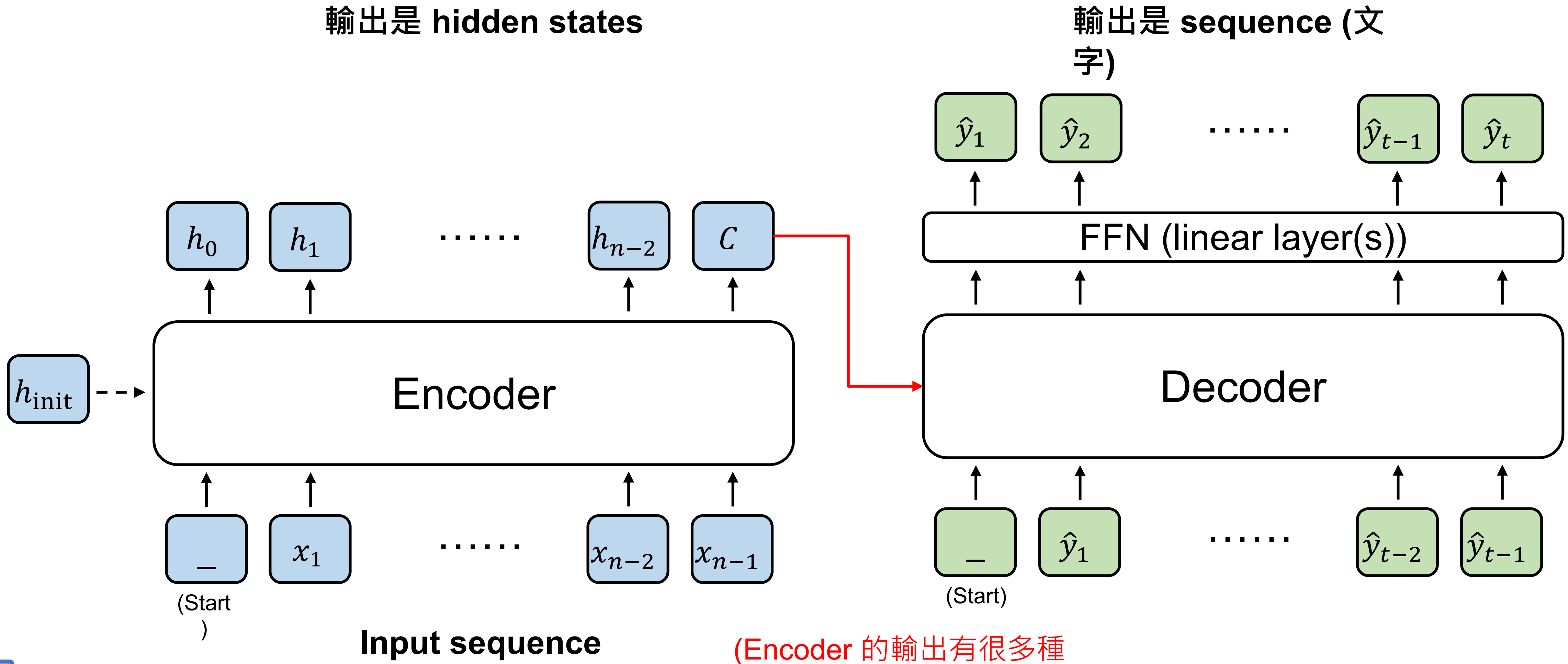
Scaled Dot-Product Attention




$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



# Encoder and Decoder




(Encoder 的輸出有很多種被Decoder運用的方式)

# Machine Translation

**Translation, EN-ZH**
**(WMT-19 dataset [1])**

Yet not nearly enough has been invested in this effort. → Machine Translation Model → 但目前這方面的投入還遠遠不夠。

EMNLP 2024

**NINTH CONFERENCE ON**
**MACHINE TRANSLATION (WMT24)**

November 15-16, 2024
Miami, Florida, USA

[HOME] [PROGRAM] [PAPERS] [AUTHORS]
TRANSLATION TASKS: [GENERAL MT (NEWS)] [LOW-RESOURCE LANGUAGES OF SPAIN] [INDIC MT] [CHAT TASK] [BIOMEDICAL] [MULTIINDIC22MT TASK] [ENGLISH-TO-LOWRES MULTIMODAL MT TASK] [NON-REPETITIVE] [PATENT] [LITERARY]
EVALUATION TASKS: [METRICS TASK] [MT TEST SUITES] [QUALITY ESTIMATION]
OTHER TASKS: [OPEN LANGUAGE DATA INITIATIVE]

Figure source: https://www2.statmt.org/wmt24/mtdata/
[1] https://huggingface.co/datasets/wmt/wmt19/viewer/zh-en

# Attention Is All You Need

Essential AI

**Ashish Vaswani***
Google Brain
~~avaswani@google.com~~

**Noam Shazeer***
Google Brain
~~noam@google.com~~

Anthropic

**Niki Parmar***
Google Research
~~nikip@google.com~~

Inceptive

**Jakob Uszkoreit***
Google Research
~~usz@google.com~~

Sakana AI

**Llion Jones***
Google Research
~~llion@google.com~~

Cohere

**Aidan N. Gomez*** †
University of Toronto
~~aidan@cs.toronto.edu~~

**Łukasz Kaiser***
Google Brain
~~lukaszkaiser@google.com~~

OpenAI

NEAR

**Illia Polosukhin*** ‡
~~illia.polosukhin@gmail.com~~

https://arxiv.org/abs/1706.03762v6

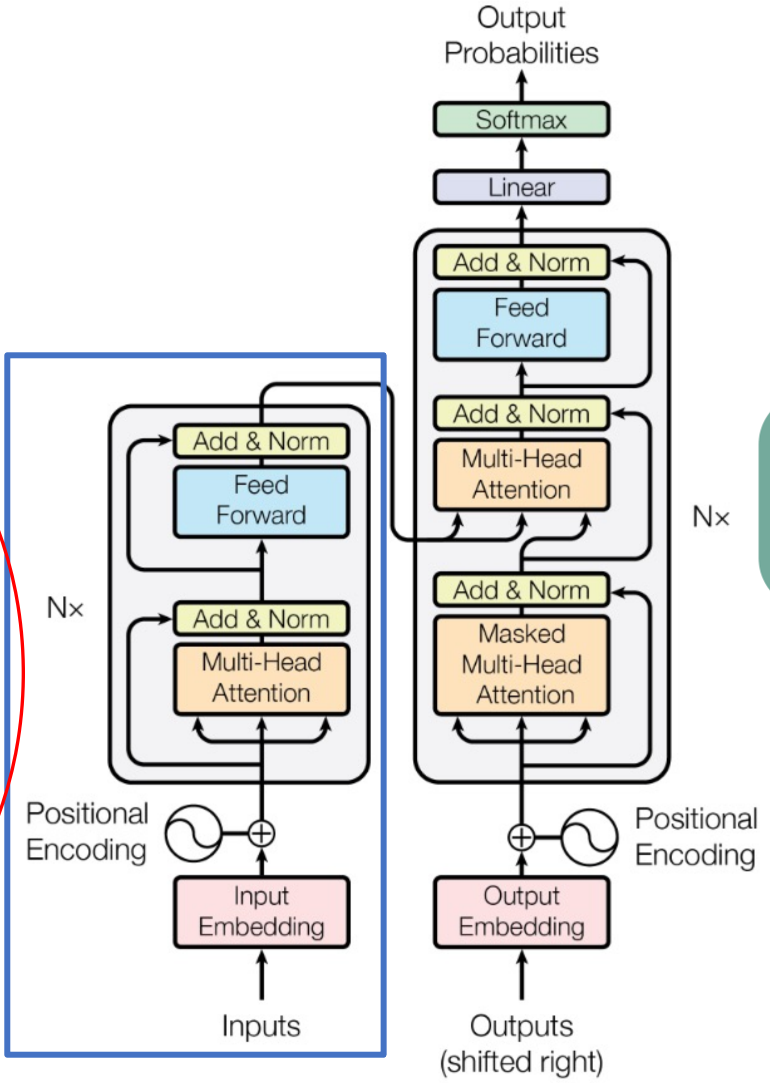
# Transformer 架構的後續應用

- Generative Pre-training (GPT) series
- BERT (Bidirectional encoder representations from transformers)
    - [Devlin et al., 2018](#)

N=12

**Encoder**

(BERT)




(GPT)

# 先 Pre-training，再 Fine-tuning

**Pre-training** → **Fine-tuning**

在大量資料上進行訓練，通常是自監督式 (Self-Supervised Training, SSL)

在**目標**資料上 (Down-stream tasks, 下游任務) 進行訓練，通常是監督式 (Supervised Training)，也就是需要有標註的資料才能進行模型訓練

# Model Training: Pre-training and Fine-Tuning

## Step1: Pre-training (use large-scale corpora)

Open a [MASK] account → Model (e.g., Word2vec, BERT) → bank

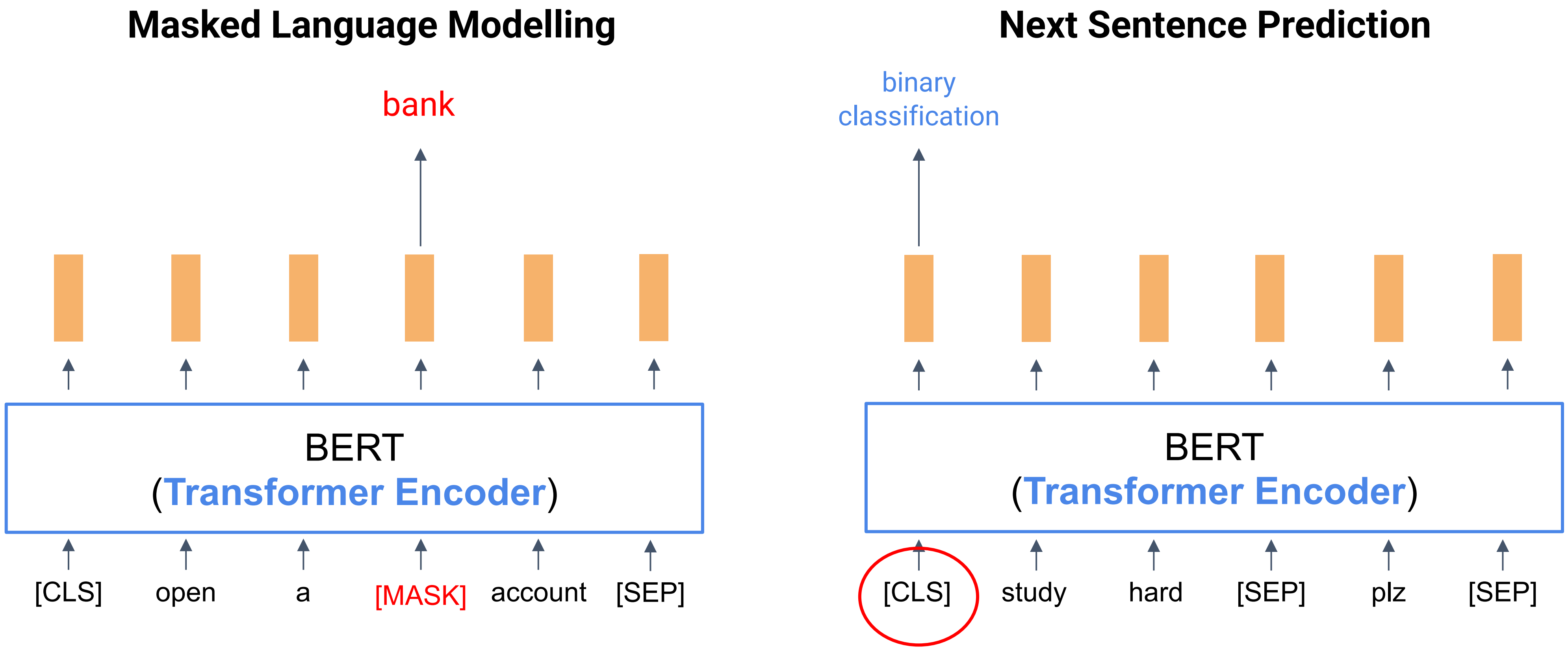
## Step2: Fine-Tuning (use datasets from target tasks)

It was a good movie! → Model (e.g., Word2vec + RNN; BERT + Classifier) → Positive

Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." NAACL 2019.

# BERT (Bidirectional Encoder Representations from Transformers)

- BERT 有兩種預訓練任務：




Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." NAACL 2019.

# Next Sentence Prediction (NSP)

- 二元分類任務
- 目標：使模型能夠理解不同語句之間的關聯性

Input = `[CLS] the man went to [MASK] store [SEP]`
`he bought a gallon [MASK] milk [SEP]`

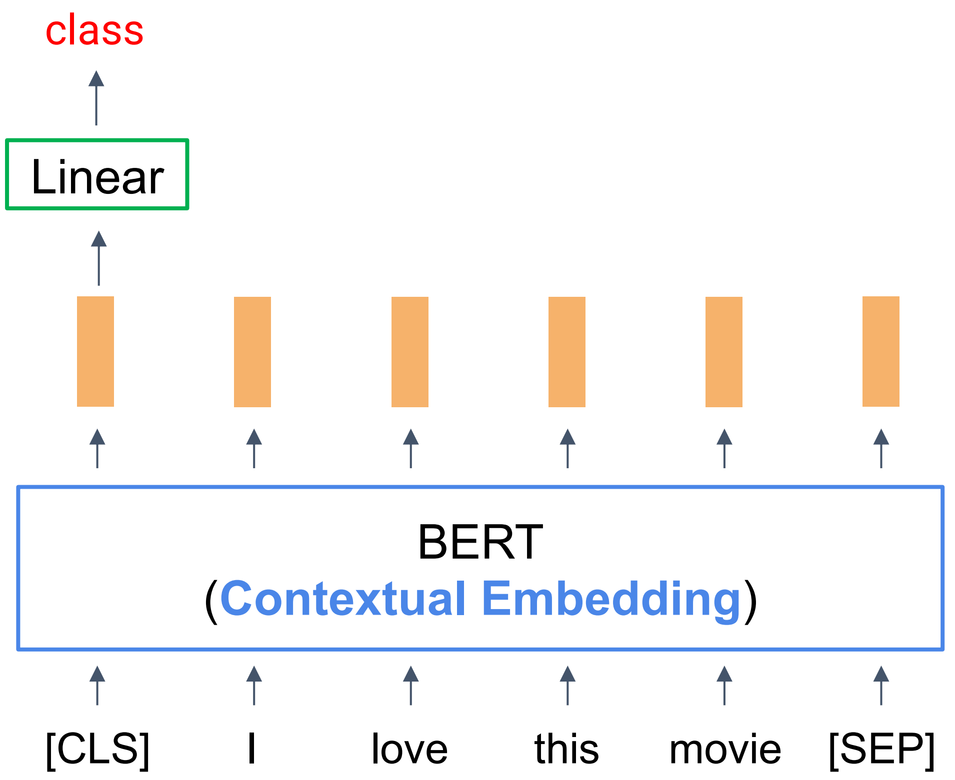
Label = `IsNext`

Input = `[CLS] the man [MASK] to the store [SEP]`
`penguin [MASK] are flight ##less birds [SEP]`

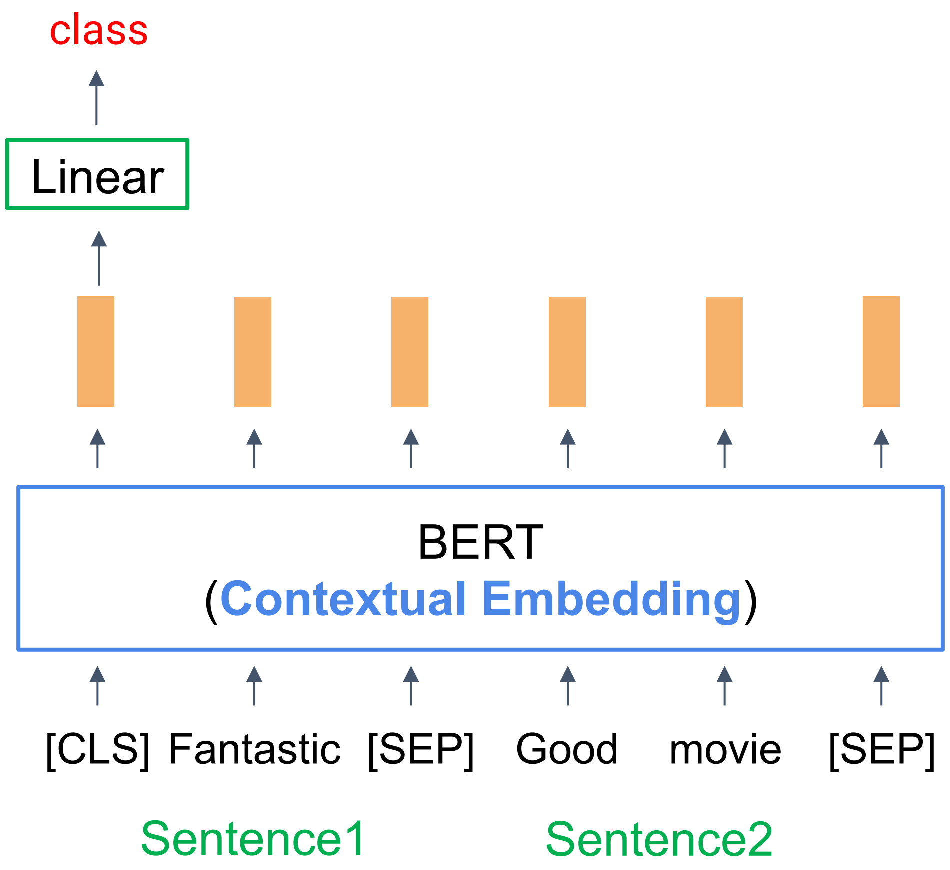
Label = `NotNext`

# Fine-tuning BERT (Sentence Classification)

**Single-sentence classification**




**Sentence pair classification**



# Summary - Key information

- Self-supervised Learning 的訓練方式是從資料中取得「答案」
  - 節省大量標註成本
  - 適用於大規模未標資料
- 自然語言處理領域以預測文章或句子中的下一個字為 SSL 的主要方法

# Thank you!

長庚大學人工智慧學系 林英嘉

✉️ yjlin@cgu.edu.tw