


# Beyond Word Embedding : Keyword and BERT

Google機器學習開發專家  
Jerry Wu  
國立台灣大學 工程所碩士生  
Charlie Li

# Google機器學習開發專家(GDE) 暨 歐奔頭殼(OpenTalk)創辦人兼CTO




Products ▾ Events Developer Programs ▾ Blog

Search

LANGUAGE ▾ ALL PRODUCTS SIGN IN

ABOUT DIRECTORY



**Jerry Wu**  
Taipei, Taiwan

Social links



- [LinkedIn](#)
- [GitHub](#)
- [Personal Website](#)

Technologies

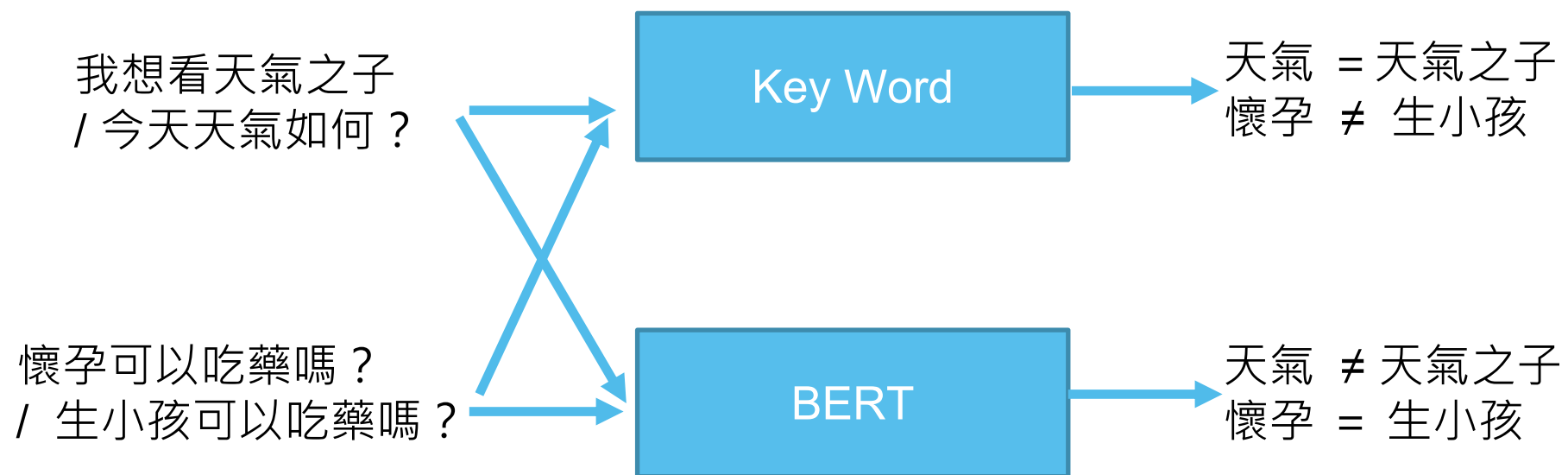
Machine Learning

Biography

JerryWu is a Data Scientist, currently teaching of Information Management at National Taiwan University of Science and Technology (NTUST). He is also a Founder & Chief Technology Officer (CTO) in the Asia Pacific Machine Intelligence Company (APMIC). Jerry Wu's teaching and research interests include Machine Intelligence, Computer Vision (CV) and Natural Language Understanding (NLU). He eager to do research related to data technology, share them with people and solve problems. He also has many years of experience in predictive analysis with traditional industry.

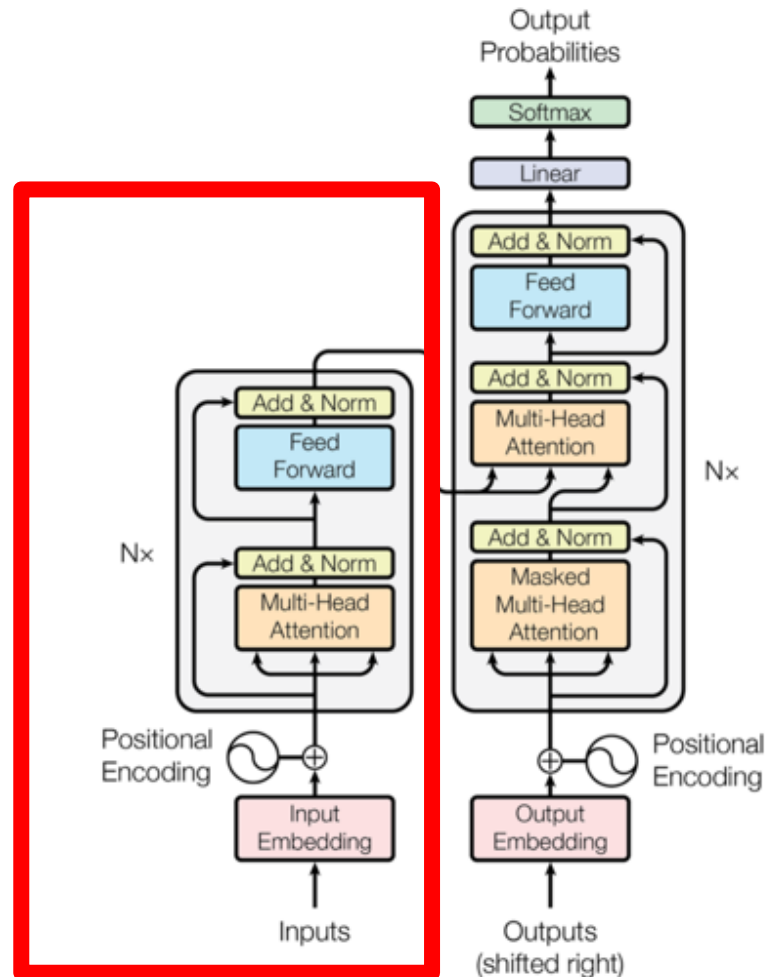


# Keyword and BERT



# BERT

## Model Architecture - Encoder from Transformer



圖片來源：  
<https://arxiv.org/pdf/1706.03762.pdf>

# BERT

## Pretraining

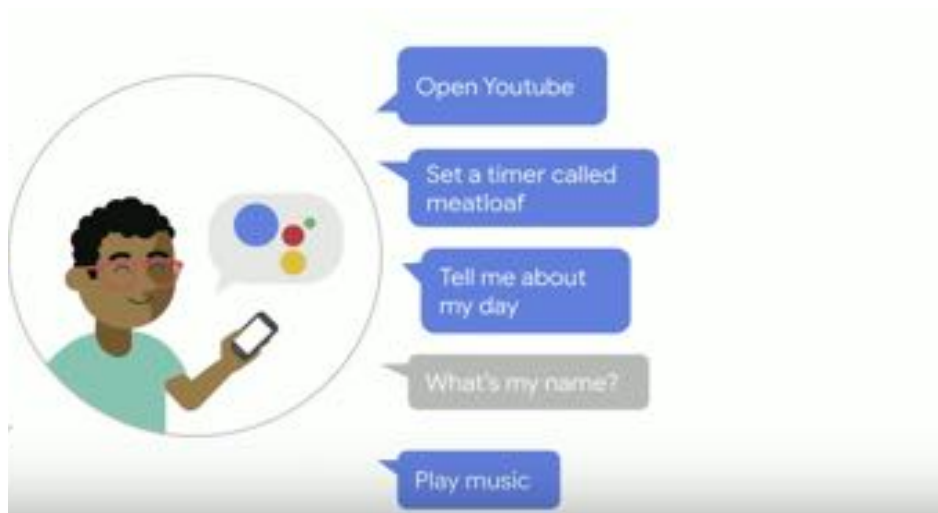
- Masked Language Model (MLM)

Mask about 15% of tokens in the input sequence and predict the original tokens

- Next Sentence Prediction (NSP)

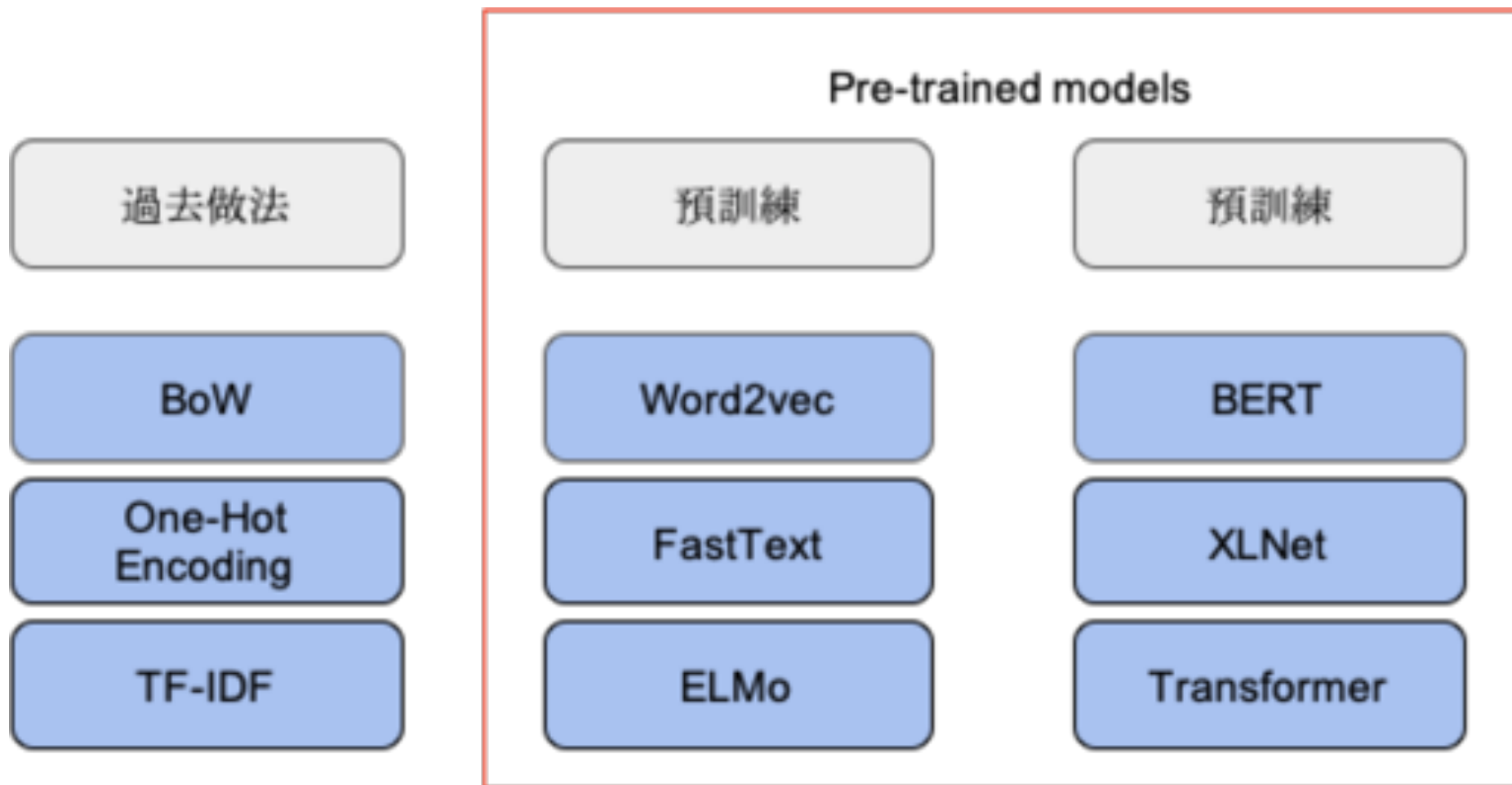
Feed the two sequences and predict whether they are consecutive

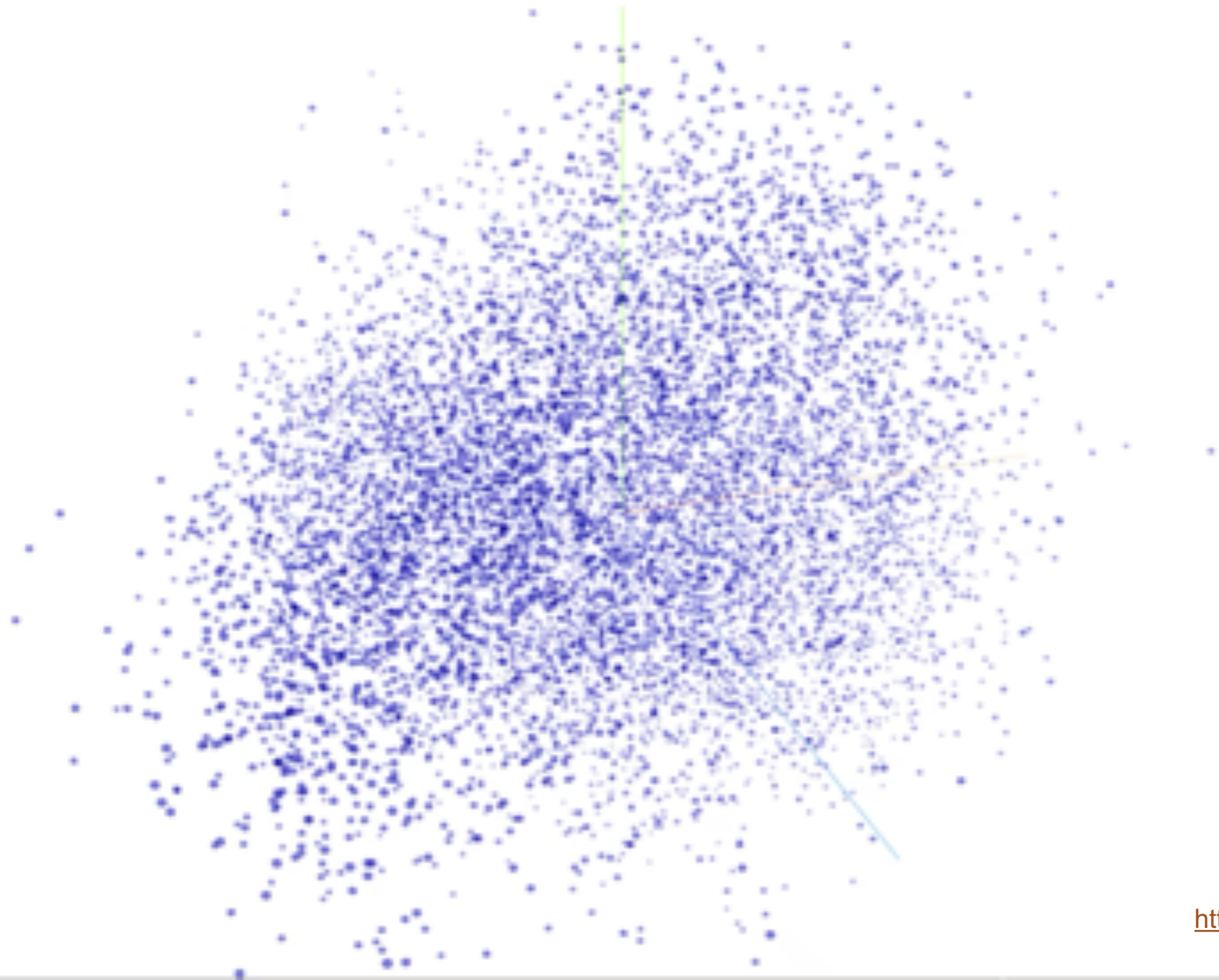
# KeyWords -> Word Embeddings -> Sentence Embeddings



Bow [0 1 0 0 3 0 0 0 2 0 0 0 0 0 1]  
One-Hot [0 0 0 0 0 0 0 0 0 0 0 0 0 0 1]  
Word Embeddings [1.2 3.2 1.2 3.5 5.1 8.1 1.2 1.1]

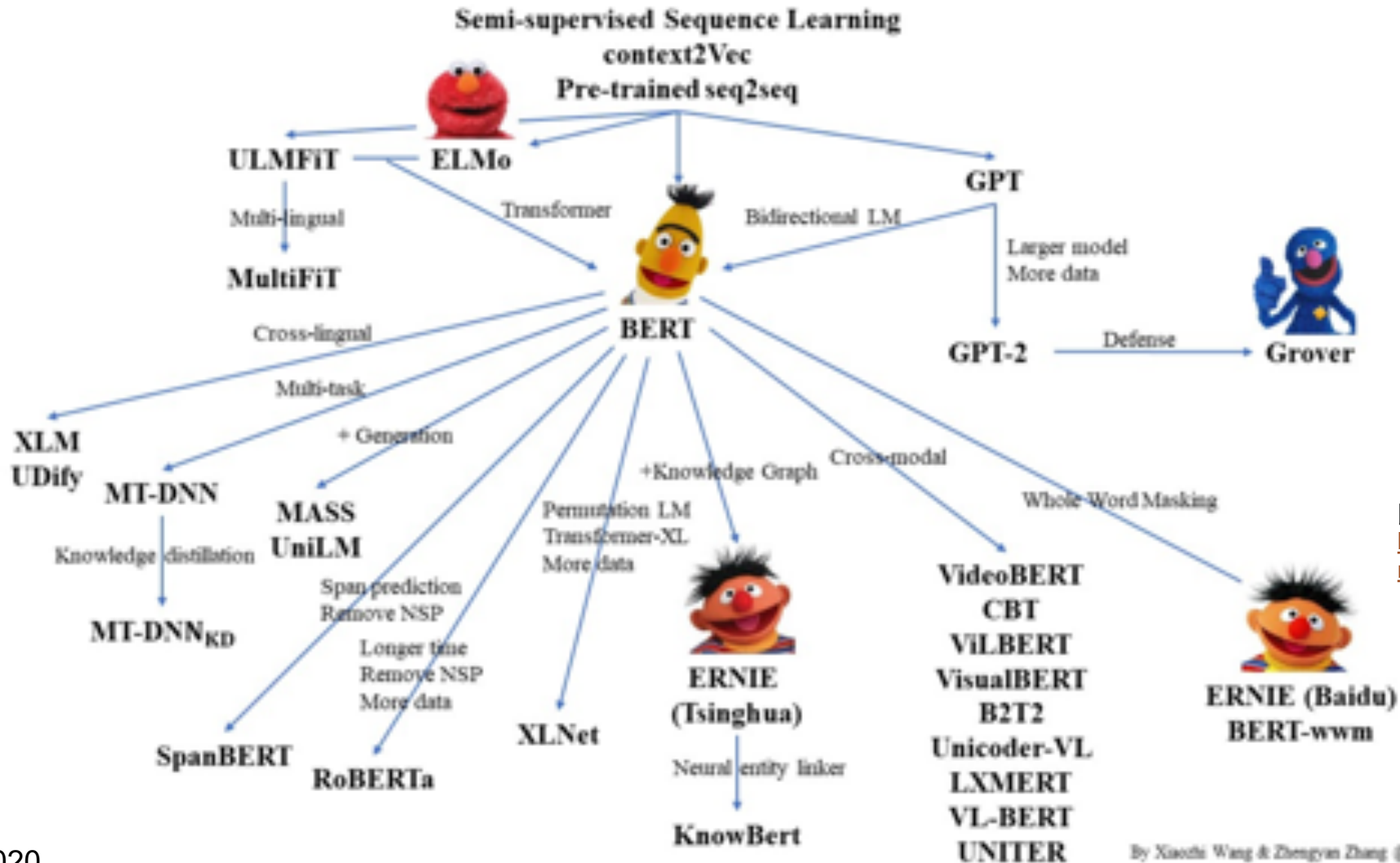
# KeyWords -> Word Embeddings -> Sentence Embeddings







# NLP Model



圖片來源：  
<https://github.com/thunlp/PLMpapers/blob/master/PLMfamily.jpg>

Title	BERT
Organization	Google
author	Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova
year	2019
Advantage	讓 pre-training 對後來NLP研究有重大影響
Disadvantage	模型太大，一般人不容易訓練，且BERT 不擅長 Generation task

Title	ERNIE
Organization	Tsinghua University, Beijing, China
author	Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun and Qun Liu
year	2019
Advantage	超越原生BERT，使模型具有知識結構，有助於 Few-shot 訓練
Disadvantage	需要良好的 NER model，且除了 Few-shot 資料的任務外，提昇不多

Title	ERNIE-百度
Organization	Baidu
author	Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian and Hua Wu
year	2019
Advantage	使模型具有知識結構，透過實驗找到較佳的mask 範圍
Disadvantage	需要良好的 NER model，且提昇不多，2-3%

Title	XLNet
Organization	Facebook AI
author	Guillaume Lample, Alexis Conneau
year	2019
Advantage	採用多種遮蔽語言模型，比其他跨語言模型優
Disadvantage	無

Title	MT-DNN
Organization	Microsoft
author	Xiaodong Liu, Pengcheng He, Weizhu Chen, Jianfeng Gao
year	2019
Advantage	實驗證明全部任務一起處理比較好
Disadvantage	一次訓練多任務，相當耗資源

Title	roBERTa
Organization	Facebook AI
author	Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov
year	2019
Advantage	dynamic masking 可以避免mask掉的字沒有被訓練到
Disadvantage	模型變大了，運算資源要求也更多了

Title	Grover
Organization	University of Washington
author	Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, Yejin Choi
year	2019
Advantage	用 GPT-2 生成假新聞，也分辨假新聞
Disadvantage	和GPT-2一樣，需要大量資源訓練

Title	XLNet
Organization	Google
author	Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le
year	2019
Advantage	算是最強模型之一
Disadvantage	因為排列問題，時間複雜度會增加約 $O(n!)$ ，沒有幾家公司能訓練

Title	VideoBERT
Organization	Google
author	Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, Cordelia Schmid
year	2019
Advantage	讓 attention同時處理不同domain的資料
Disadvantage	無

Title	VisualBERT
Organization	University of California, Allen Institute of Artificial Intelligence and Peking University
author	Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, Kai-Wei Chang
year	2019
Advantage	讓 transformer 能理解圖片，可以實現圖片問答等應用
Disadvantage	無

Title	K-BERT
Organization	Peking University, Beijing, China, Beijing Normal University, Beijing, China
author	Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, Ping Wang
year	2019
Advantage	可辨別專業詞彙
Disadvantage	需要建立知識圖譜，且在一班領域問答上和原本的 BERT 差不多

Title	ELECTRA
Organization	Google
author	Kevin Clark, Minh-Thang Luong, Quoc V. Le, Christopher D. Manning
year	2020
Advantage	比 masked language model 更快，消耗資源更少
Disadvantage	無

# 來談幾個有趣的自然語言理解模型

# VideoBERT

(A BERT model learned over sequence of visual and linguistic tokens)

特色：  
用ASR+CNN（S2D pretrained model）取 video features  
作為 BERT large input

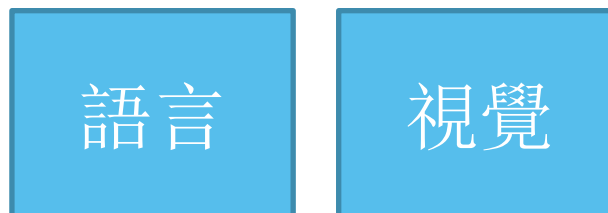
An illustration featuring four blue puzzle pieces arranged in a 2x2 square. A hand with orange skin and a grey sleeve is shown from the bottom right, reaching towards the pieces. The background is white with light blue wavy shapes in the top left and bottom right corners.

Source	<a href="https://arxiv.org/pdf/1904.01766.pdf">https://arxiv.org/pdf/1904.01766.pdf</a>
Organization	Google
author	Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, Cordelia Schmid
year	2019
Advantage	讓 attention同時處理不同domain的資料
Disadvantage	無

# VideoBERT

Combining two domain

- Visual domain
- Linguistic domain



Three methods

- Automatic speech recognition (ASR): Used to convert speech to text
- Vector quantization (VQ): Quantize visual feature from pre-train video classification model
- BERT: Learning joint distributions over sequences of discrete tokens.

# VideoBERT

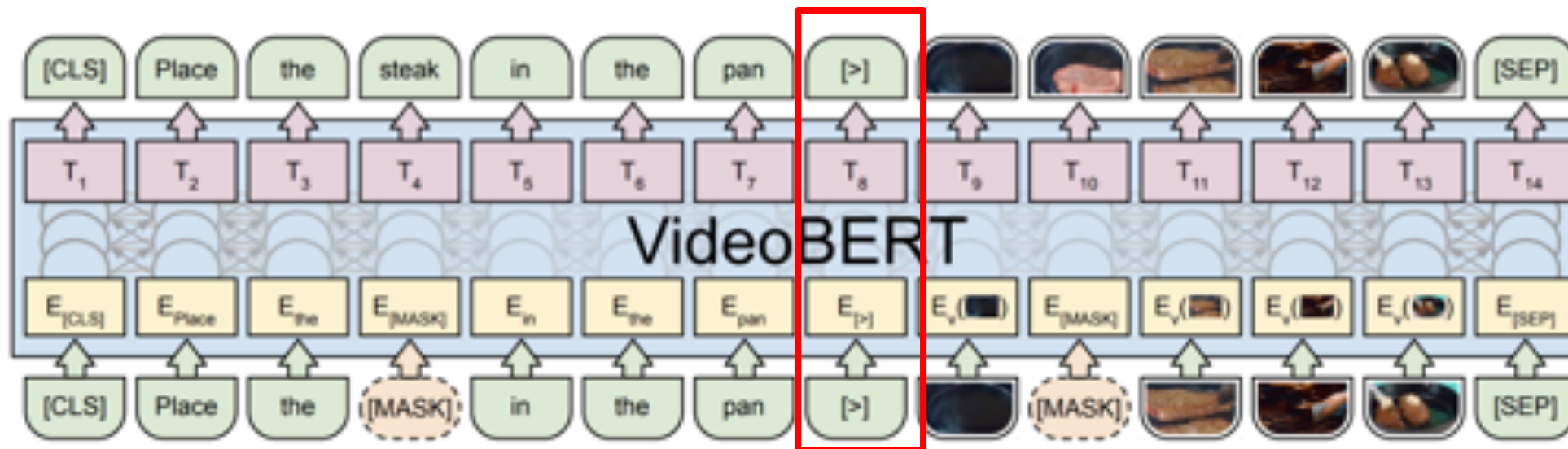
## Dataset

從Youtube爬大量影片

- Text：用ASR 抓文字，再轉 token
- Video：每 30 frame（1.5 seconds）用 Pretrained Convolutional Neural Network（S3D model）抓 Feature

# VideoBERT

- Text and Video input 用「>」分隔
- Mask pretraining



The start of video features



# Application

Text to video: illustrate a set of instructions(such as a recipe)



Video to text: Caption generation



chop the basil and add to the bowl

cut the bread into thin slices

# K-BERT

Enabling Language Representation with Knowledge Graph

特色：  
在 BERT 加入 **knowledge graph**，以解決 BERT 在專業詞彙中表現不佳的問題

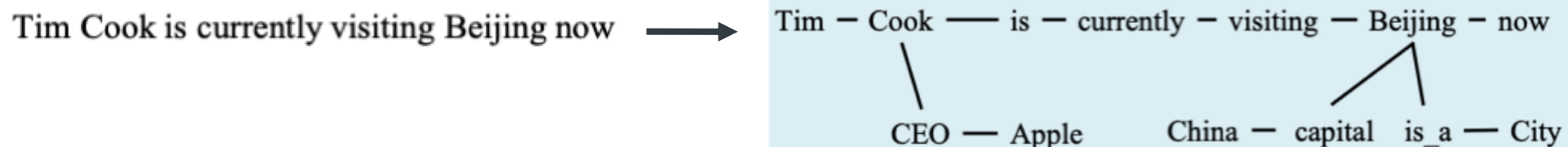


Source	<a href="https://arxiv.org/pdf/1909.07606.pdf">https://arxiv.org/pdf/1909.07606.pdf</a>
Organization	Peking University, Beijing, China, Beijing Normal University, Beijing, China
author	Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, Ping Wang
year	2019
Advantage	可辨別專業詞彙
Disadvantage	需要建立知識圖譜，且在一班領域問答上和原本的 BERT 差不多

# K-BERT

## Architecture

- **Knowledge layer:** convert input sentence and knowledge graph to sentence tree



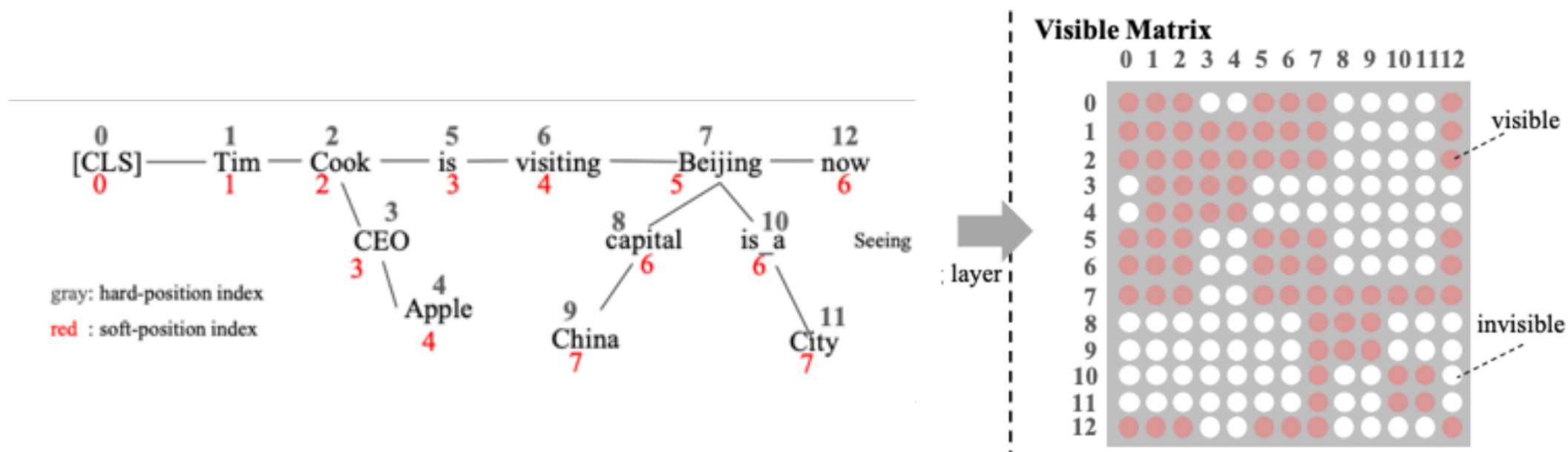
- **Embedding layer:** convert sentence tree to embedding vector

Token embedding	[CLS]	Tim	Cook	CEO	Apple	is	visiting	Beijing	capital	China	is_a	City	now
	+	+	+	+	+	+	+	+	+	+	+	+	+
Soft-position embedding	0	1	2	3	4	3	4	5	6	7	6	7	6
	+	+	+	+	+	+	+	+	+	+	+	+	+
Segment embedding	A	A	A	A	A	A	A	A	A	A	A	A	A

# K-BERT

## Architecture

- **Seeing layer:** convert sentence tree into **visible matrix**



$$M_{ij} = \begin{cases} 0 & w_i \ominus w_j \text{ red point 兩者在 branch 上} \\ -\infty & w_i \oslash w_j \text{ white point} \end{cases}$$

# K-BERT

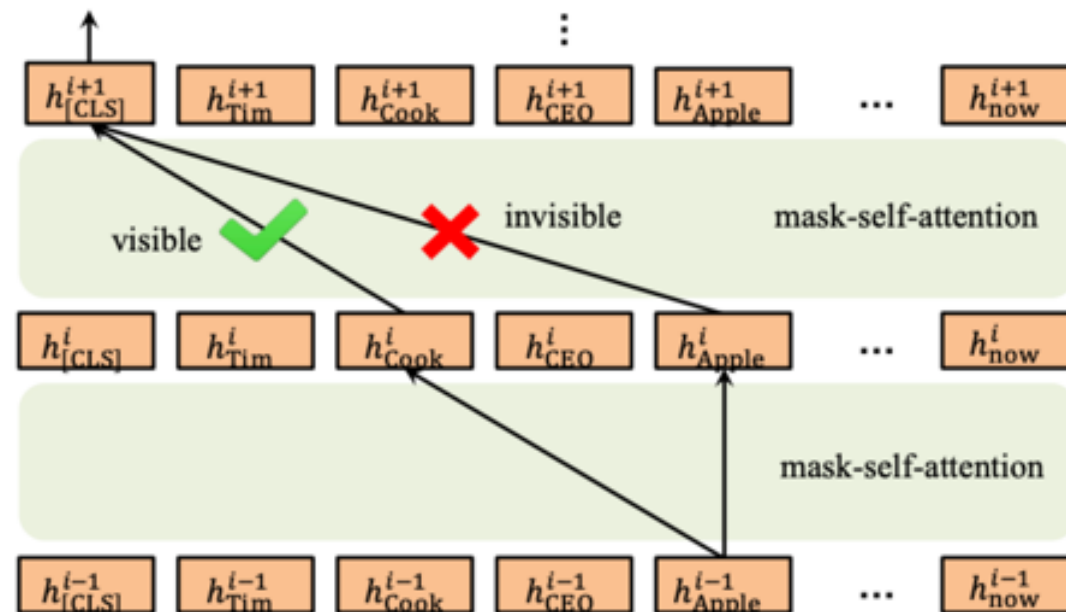
## Architecture

- **Mask-transformer:** just like BERT, but use **mask-self-attention**

$$S^{i+1} = \text{softmax}\left(\frac{Q^{i+1}K^{i+1\top} + M}{\sqrt{d_k}}\right),$$
$$h^{i+1} = S^{i+1}V^{i+1},$$

和原本的 **self-attention** 差別只在這個  $M$ ，是 **visible matrix**

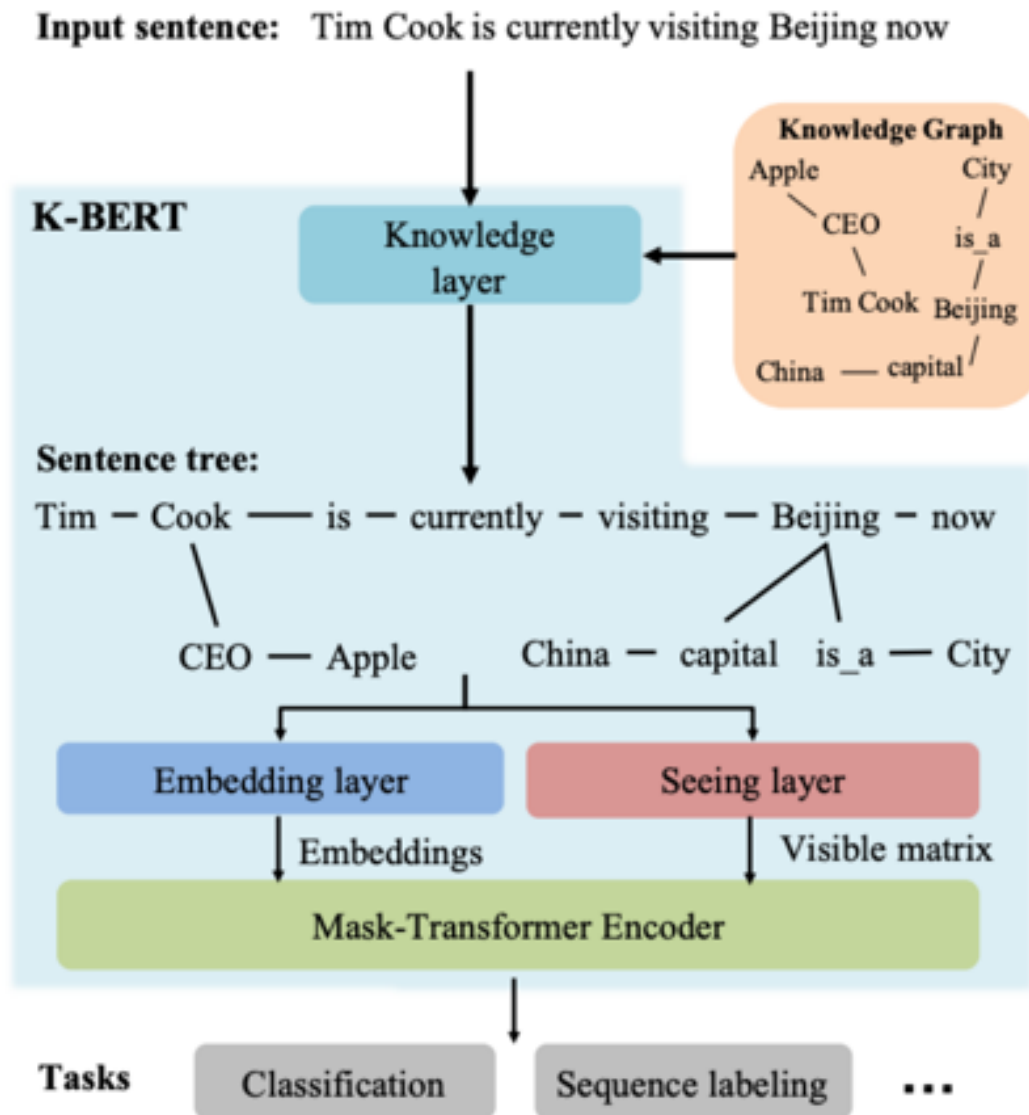
目標是讓不在分支的 token 無法互相 attend



圖片資料取自原始論文

# K-BERT

## Whole architecture



# ELECTRA

Pre-training Text Encoders as Discriminators  
Rather Than Generators

特色：

small BERT 後面接一個 discriminator

replace token detection, 比 masked language model 更快，消耗資源更少



Source	<a href="https://openreview.net/pdf?id=r1xMH1BtvB">https://openreview.net/pdf?id=r1xMH1BtvB</a>
Organization	Google
author	Kevin Clark, Minh-Thang Luong, Quoc V. Le, Christopher D. Manning
year	2020
Advantage	比 masked language model 更快，消耗資源更少
Disadvantage	無

# ELECTRA

## Method

### Pretraining

- **Generator**

Using a “BERT like” **Masked Language Model** to generate token to replace the words that are masked

- **Discriminator**

Detect which tokens are replaced

### Fine-tuning

Fine-tune **Discriminator** according to your task

## Benefit

**Faster, more efficient and lower resource requirement !**



# ELECTRA

## Generator - A BERT like model

- Masked Language Model (MLM)

Mask about 15% of tokens in the input sequence and predict the original tokens

- Next Sentence Prediction (NSP)

Feed the first token to the model and expect the next

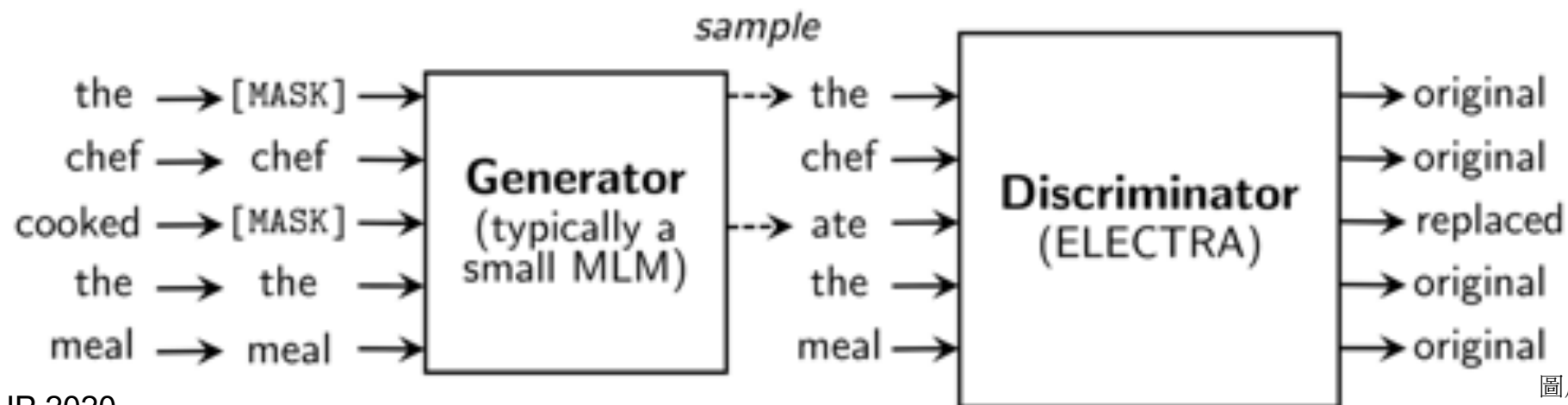
# ELECTRA

## Discriminator - Replaced Token Detection

1. Feed the model with the output sequence of the generator
2. Detect which tokens are original token

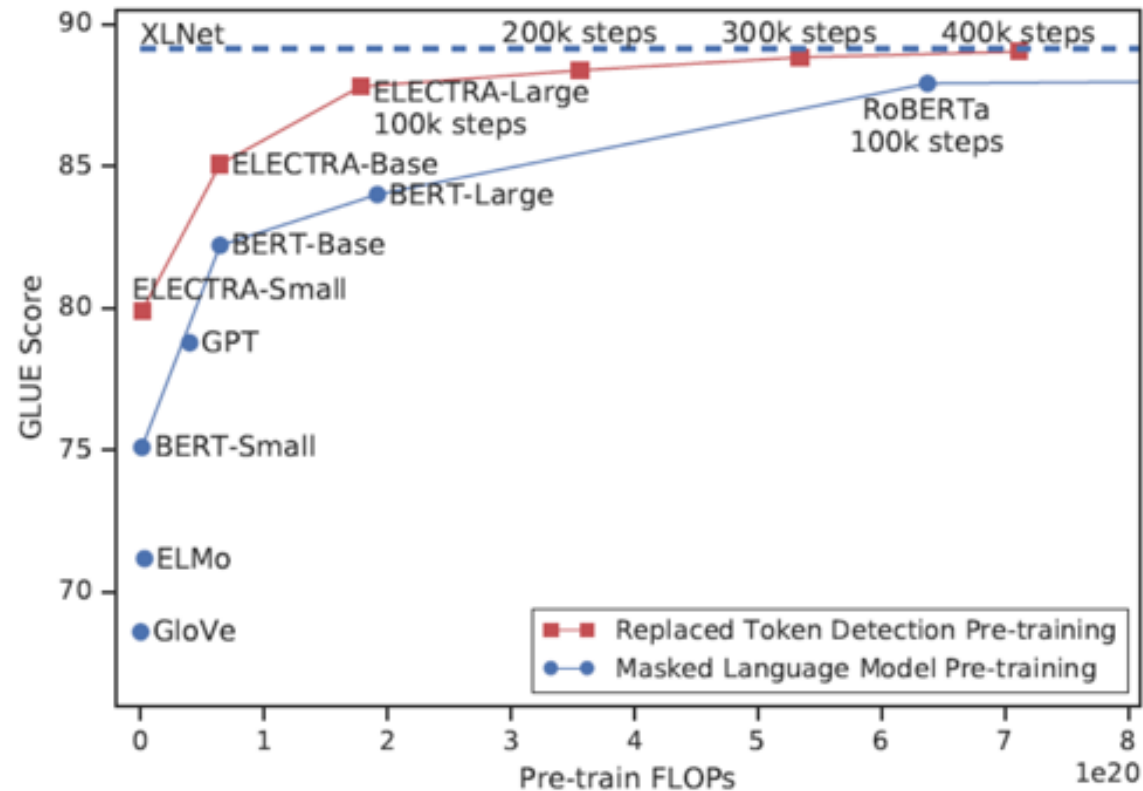
## Fine-tuning

Use **only Discriminator** to your downstream tasks



# ELECTRA

Better result on GLUE benchmark than BERT Series



我不想寫程式，想直接體驗BERT

可以直接到OpenTalk平台體驗90天

使用教學文章

<https://medium.com/apmic-opentalk/apmic-opentalk-%E4%BD%BF%E7%94%A8%E6%95%99%E5%AD%B8-%E8%A8%BB%E5%86%8A%E8%B3%BC%E8%B2%B7%E7%AF%87-adf779af402e>

Coupon: COSCUPBERT0801

如有使用問題請寫信到  
客服信箱 [arthur@ap-mic.com](mailto:arthur@ap-mic.com)

我想訓練BERT只想寫一點程式

<https://github.com/google-research/bert>

<https://github.com/hanxiao/bert-as-service>

我想徹底研究BERT，開發不同的方法

Attention Is All You Need :

<https://arxiv.org/pdf/1706.03762.pdf>

Pre-training of Deep Bidirectional Transformers for  
Language Understanding :

<https://arxiv.org/pdf/1810.04805.pdf>

A Span-Extraction Dataset for Chinese Machine Reading  
Comprehension

<https://www.aclweb.org/anthology/D19-1600/>

# THANK YOU

Jerry老師的學習群組



jerry@ap-mic.com