# 深度學習在文字上的應用

組員：謝奇容、唐祥勝

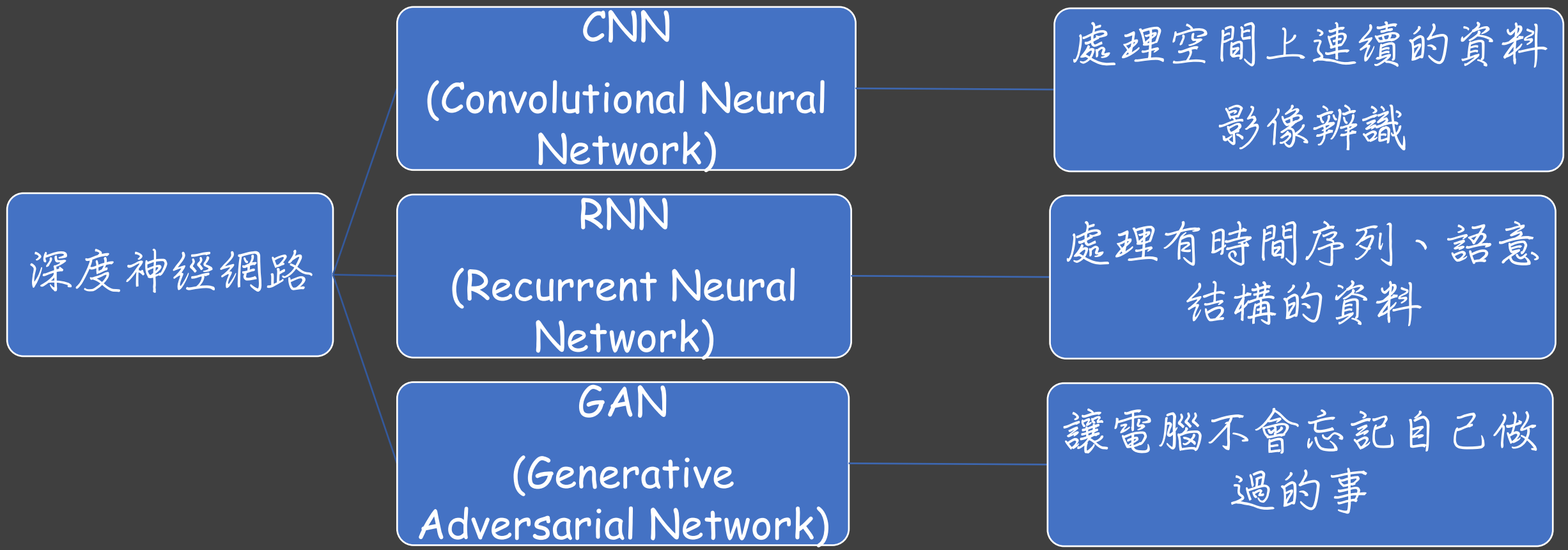
# 目錄

# 什麼是深度學習？

# 深度學習

- 機器學習的一種、現今人工智慧的主流
- 學習：經過大量的訓練過程，找出一個最佳函數，得出最佳解
- 針對特定需求設計
- 用“規則”及“大量學習資料”，告訴機器什麼是對的
- 使用深度神經網路

# 深度學習


深度神經網路
CNN
(Convolutional Neural Network)
RNN
(Recurrent Neural Network)
GAN
(Generative Adversarial Network)
處理空間上連續的資料
影像辨識
處理有時間序列、語意結構的資料
讓電腦不會忘記自己做過的事

# 深度學習遇到文字的困難

# 深度學習 VS 文字

- 電腦無法讀取文字，僅能讀取數字
- 語文不像數學，需要根據前後文才能理解意義
- 根據地域性會有當地的方言、用法

# 自然語言處理NLP

# 自然語言處理

- 程式語言：人類為了與電腦溝通而設計的語言
- 自然語言：人類與人類溝通所發展、創造出的語言


The Neural History of
Natural Language Processing
2001
Neural language models
2008
Multi-task learning
2013
Word embeddings
2013
Neural networks for NLP
2014
Sequence-to-sequence models
2015
Attention
2015
Memory-based networks
2018
Pretrained language models

自然語言處理

# 自然語言處理

- 文本分詞
- 將文字轉成數字序列
- 建立模型
- 定義衡量指標
- 訓練模型
- 用已訓練模型對新數據做預測

# 文本分詞

- 將非結構化的文本數據，轉成結構化數據
- 詞是表達含意的最小單位
- 中英文分詞區別：
  - 分詞方式不同
  - 詞的型態分別不同
- 中文分詞的難點：
  - 分詞沒有統一標準
  - 歧義詞切分
  - 新詞識別

# 文本分詞

- 中文分詞工具：
  - Hanlp
  - Stanford 分詞
  - ansj 分詞器
  - 哈工大 LTP
  - KCWS分詞器
  - jieba
  - IK
  - 清華大學THULAC
  - ICTCLAS

# Jieba

分詞模式：

精確模式

全模式

搜尋引擎模式

自定義字典

# 深度學習在文字上的應用

# 深度學習模型

- TensorFlow
- Keras
- PyTorch
- CRF
- Kashgari

# Kashgari

- 集中多種模型　　、　　、　　、
- 擁有多種預訓練模型　　、　　、
- 方便快速訓練一個序列標註或文字分類模型
- 擁有統一的　和說明文件

# Kashgari

- 
- 
- 無監督式訓練
- 開發的語言代表模型
- 先以 的方式預先訓練出一個對自然語言有一定「理解」的通用模型
- 再將該模型拿來做特徵擷取或是下游的（監督式）任務

# 應用實例

# Kashgari 應用

- 玉山人工智慧公開挑戰賽　　夏季賽
  - 比賽目的：透過　，抓出　　焦點人物，提升銀行效率

```
bert_embed = BERTEmbedding('code/chinese_L-12_H-768_A-12',
                           task=kashgari.LABELING,
                           sequence_length=100)

model = LSTM_CNN_Model(bert_embed)
# This step will build token dict, label dict and model structure
model.build_model(train_x, train_y, valid_x, valid_y)
# Compile model with custom optimizer, you can also customize loss and metrics.
# optimizer = RAdam()
# model.compile_model(optimizer=optimizer, loss=categorical_focal_loss(gamma=2.0, alpha=0.25))

eval_callback = EvalCallBack(kash_model=model,
                             valid_x=valid_x,
                             valid_y=valid_y,
                             step=1)

model.compile_model()

# Train model
model.fit(train_x, train_y, valid_x, valid_y, batch_size=128, epochs=2, callbacks=[eval_callback])
```

# Kashgari實作

# 資料來源

- https://kknews.cc/zh-tw/code/rlyeezn.html
- https://blog.kennycoder.io/2020/02/12/Python-%E7%9F%A5%E5%90%8DJieba%E4%B8%AD%E6%96%87%E6%96%B7%E8%A9%9E%E5%B7%A5%E5%85%B7%E6%95%99%E5%AD%B8/
- https://research.sinica.edu.tw/nlp-natural-language-processing-chinese-knowledge-information/
- https://medium.com/@leemeng/%E9%80%B2%E5%85%A5-nlp-%E4%B8%96%E7%95%8C%E7%9A%84%E6%9C%80%E4%BD%B3%E6%A9%8B%E6%A8%91-1e90e21f3838
- https://medium.com/%E5%AD%B8%E4%BB%A5%E5%BB%A3%E6%89%8D/nlp%E7%99%BC%E5%B1%95%E5%8F%B2%E6%91%98%E8%A6%81-1-a30af62cbcec
- https://www.itread01.com/content/1565680922.html
- https://panx.asia/archives/53209
- https://research.sinica.edu.tw/deep-learning-2017-ai-month/

# 分工

| 姓名 | 謝奇容 | 唐祥勝 |
| --- | --- | --- |
| 工作 | 資料整理製作 | 文本分詞 |

# THANK YOU