商業智慧與營運策略

Business Intelligence and Operations Strategy

多模態(Multimodal)

Week7

授課老師: 林冠成 教授

Kuan-Cheng Lin

5501.mis.nchu.edu.tw

課程大綱

- **多模態基礎與應用**
  - 多模態資訊處理（結構化資料、多媒體資料）
  - 傳統 BI 系統限制 vs. AI 驅動型 BI 的優勢
  - 多模態 vs 單模態的差異與挑戰
- **多模態核心技術與實務應用**
  - 多模態流程的五個階段介紹(表示、對齊、融合、翻譯、協同學習)
  - 多模態融合在金融中的實際應用
  - 預訓練與遷移學習（ CLIP、BLIP）
  - 多模態可解釋性分析（決策樹、SHAP、LIME、Grad-CAM）

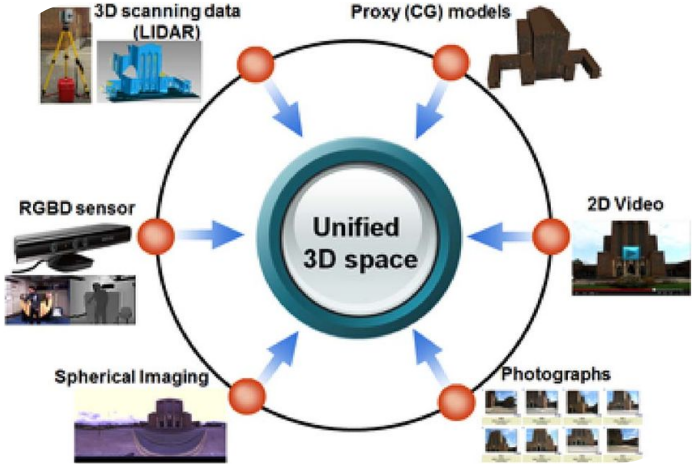
一、 多模態基礎與應用

二、多模態核心技術與實務應用

什麼是多模態資訊處理？

現今資料驅動決策的時代，企業所擁有的資料早已不再僅限於結構化表格（如 Excel、報表、銷售紀錄），還包括：

- 商品圖片（影像）
- 顧客評論（文字）
- 客服錄音（語音）
- 使用者操作紀錄（時間序列）



➢ 這些來自不同「模態」的資料，構成了企業營運與顧客互動的「全貌」

➢ 若能融合這些模態資訊，就能更完整地認識顧客、預測行為、提升決策品質

什麼是多模態資訊處理？

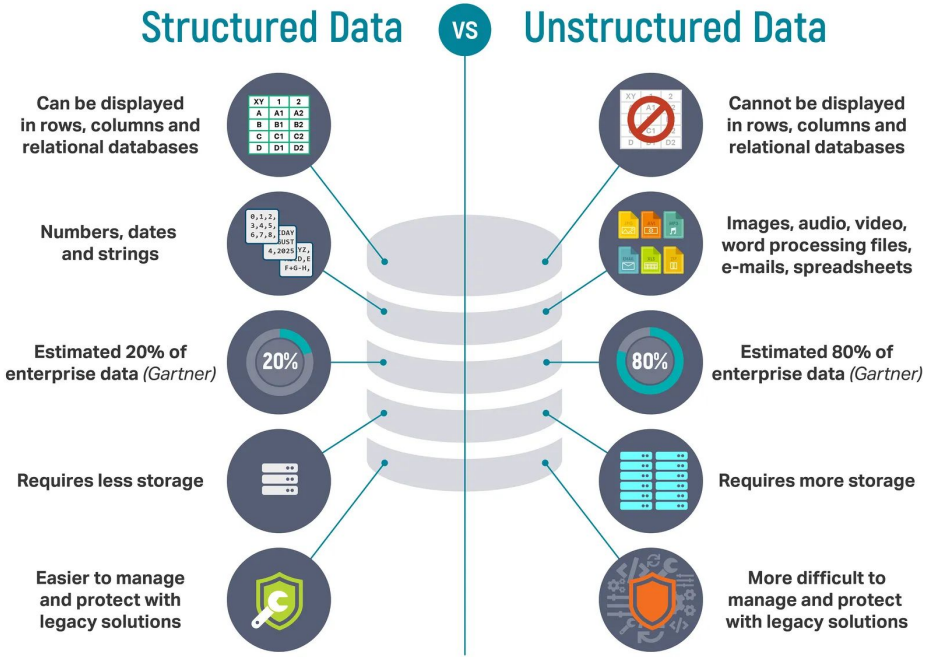
模態類型	資料格式	商業應用範例
結構化資料	Excel / SQL 表格	顧客消費紀錄、庫存、營收報表
非結構化資料	文字	顧客留言、社群評論、客服對話紀錄
視覺資料	影像 / 視訊	商品圖、監控攝影機畫面、實體門市陳列
聲音資料	語音 / 音訊	客服通話內容、語音指令
時間序列資料	Log / Sensor	瀏覽紀錄、溫度／機台數值、點擊事件紀錄

為何無法只靠表格就做好資料分析？

- **結構化資料** 是企業進行商業分析與報表的基礎，可直接透過 SQL 查詢、BI 工具視覺化，易於自動化與標準化。

- **非結構化資料** 雖然佔據大多數比例，但因無固定格式，難以直接分析與利用。然而其中潛藏豐富資訊，例如：
  - 消費者意見（社群貼文）
  - 法律風險（合約文字）
  - 客戶關係（客服對話）

- **為什麼要處理非結構化資料？**
  - 顧客留言中藏著真實滿意度
  - 客服錄音能分析情緒與抱怨重點
  - 商品圖片能辨識品牌一致性與異常
  - AI 可以幫忙將這些資料轉為結構化資訊（如 NLP、CV、OCR）



Retrieved from https://lawtomated.com/

什麼是多模態資訊處理？

▼ 各產業的資料型態與應用範例

產業	結構化資料	非結構化資料	應用任務
電商	訂單表、價格、庫存	商品照片、使用者評論、影片	商品推薦、退貨預測
金融	帳戶紀錄、交易明細	財報 PDF、新聞、客服語音	股價預測、信用風險預測
醫療	病歷紀錄、檢驗資料	X光影像、MRI、醫師語音紀錄	疾病診斷、異常偵測
法律	案號、開庭時間、案件類別	合約 PDF、掃描件、e-mail	條款辨識、盡職調查（Due Diligence）

- **結構化資料**：結構清晰、格式穩定、可直接統計分析。
- **非結構化資料**：無固定格式、數量龐大，需 AI 或人工作業才能挖掘價值。

➢ 目前企業中高達 **80% 的資料為非結構化**，處理能力會成為企業競爭力。

從傳統 BI 到 AI -BI

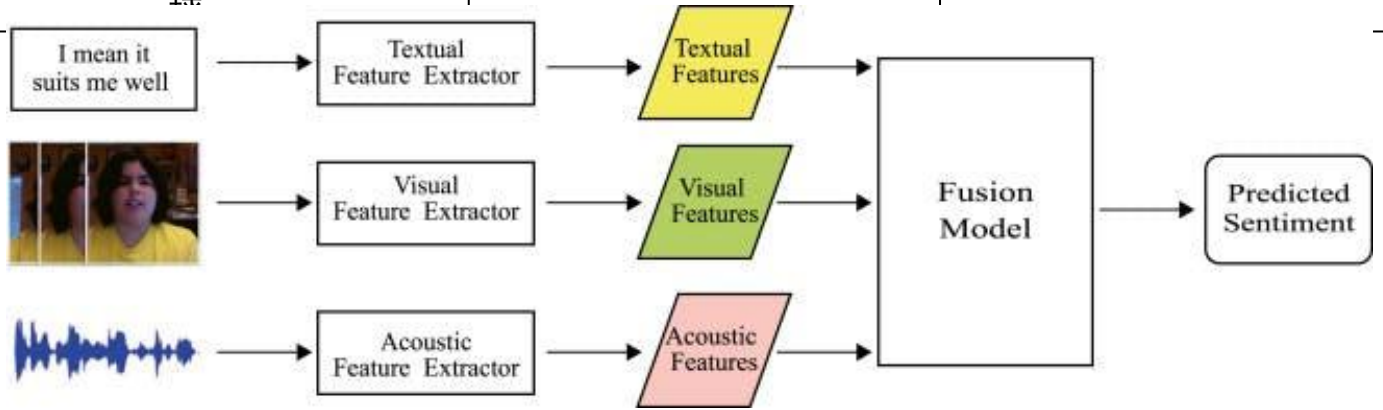
- **過去企業多依賴「靜態報表」或「 OLAP 多維分析」進行營運決策，例如** ：
  - 從 SQL 資料庫拉出銷售資料生成 Excel 圖表
  - 使用 OLAP 工具查看特定區域、產品的銷售趨勢
- 但這些方法多為「回顧式分析」，反應不夠即時，且無法處理如顧客評論、圖片等非結構化資料。
- 隨著深度學習技術的普及，我們能 夠讓機器「看圖說話」、「聽懂文字」、「讀懂語氣」，實現真正的 **AI-BI（人工智慧商業智慧）**，具有以下幾大優勢：

項目	傳統 BI	AI-BI（多模態）
資料型態	僅結構化資料（表格）	可整合文字、圖像、語音等多模態
分析方式	人工查詢／拖拉欄位	自動學習、模式挖掘
決策時間	月／週報表	實時或預測式決策
應用場景	銷售分析、庫存報表	退貨預測、情緒辨識、圖文推薦

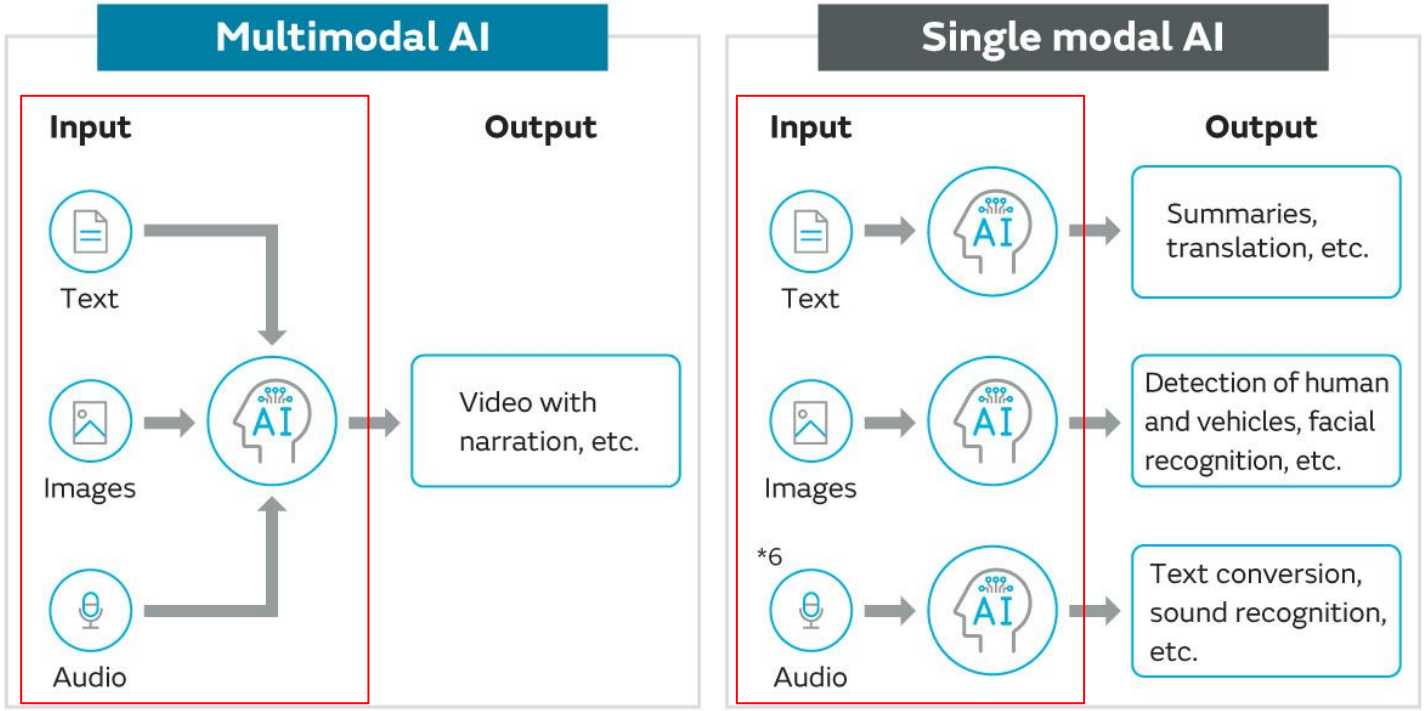
商業智慧中的多模態應用場景

產業	多模態資料組合	任務範例	商業價值
電商	商品圖 + 文字評論 + 點擊紀錄	精準推薦、退貨預測	提升轉換率、客戶滿意度
金融	財報 + 新聞文字 + 高管語音	股價預測、情緒監測	改善投資策略、風險控管
製造	Sensor 資料 + 維修紀錄 + 圖片	異常檢測、預測保養	降低停機損失、維修成本
客服	聲音通話 + 對話文本 + 點擊紀錄	客訴預測、對話品質分析	提升客戶體驗、服務效率

➢ ***Multimodal Sentiment Analysis***



多模態 vs 單模態的差異



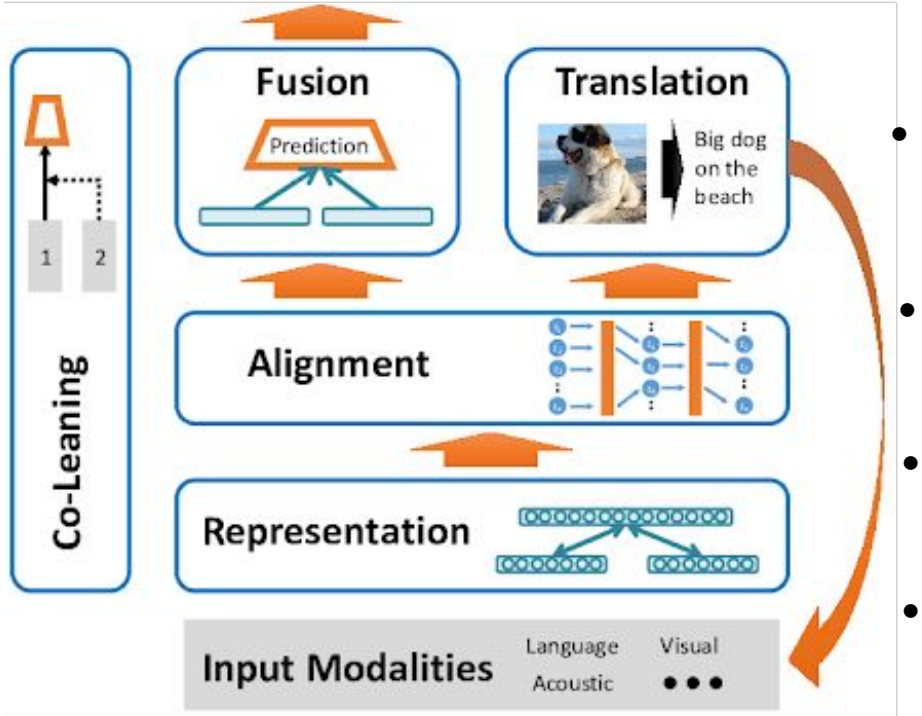
多模態 vs 單模態的差異

面向	單模態（Single-modal）	多模態（Multi-modal）
資料來源	**單一來源**（如文字、影像、時間序列）	來自**不同來源與型態**（文字＋影像＋聲音＋結構化資料等）
特徵空間	特徵一致，容易進行建模	各模態特徵空間不同（維度、分佈、時間尺度等），**需要對齊與融合**
模型設計	結構相對簡單，通常針對單一資料形式設計	需處理跨模態對齊、注意力分配、表示學習與融合策略
資料標註與蒐集	相對容易（如 NLP 的文字標註）	多模態資料蒐集與同步成本高，標註困難（例如影像與文字需對應）
資訊表達能力	受限於單一模態提供的訊息	多模態互補，能更完整捕捉語義、情境與事件特徵
應用場景	單一維度的任務（如文字分類、影像辨識）	跨領域任務（如圖文匹配、影音理解、金融結構＋新聞融合分析）
主要挑戰	資料品質與模型能力	特徵對齊困難- 融合策略設計- 模態間權重分配與資訊不平衡- 缺失模態處理
訓練與推論成本	較低，資源需求相對可控	較高，需要更多計算資源與訓練資料

一、 多模態基礎與應用

二、多模態核心技術與實務應用

多模態流程的五個階段



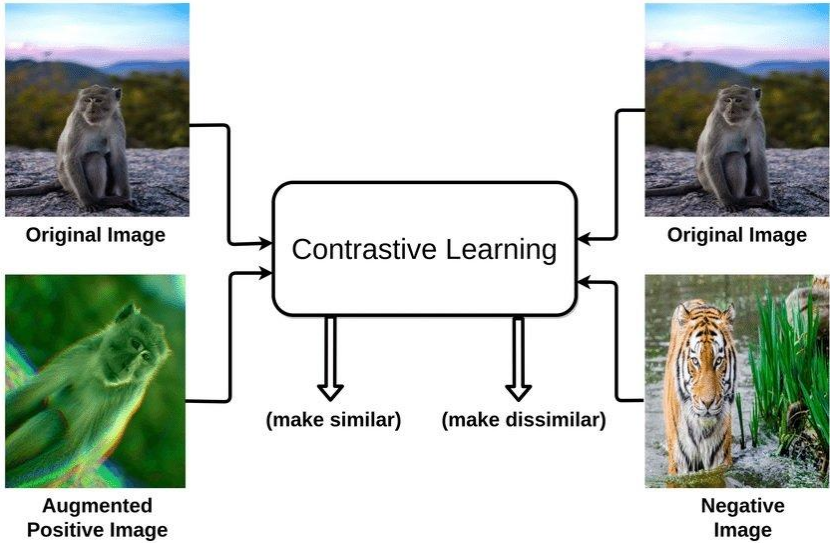
- **Representation(表示)**
  - 將不同模態轉換為向量表示
  - CNN 抽圖像特徵、BERT 抽文字特徵；可分 Joint / Coordinated 兩類
- **Alignment(對齊)**
  - 建立不同模態元素的對應關係
  - 文字與圖像區域對齊（Cross-Attention、對比學習）
- **Fusion(融合)**
  - 整合不同模態資訊
  - Early / Intermediate / Late / Bilinear Fusion
- **Translation(翻譯)**
  - 在不同模態間轉換訊息
  - Image Captioning（圖➡文）、Text-to-Image 生成
- **Co-learning(協同學習)**
  - 模態間互相增強知識
  - 少模態學習時，用有標註的模態幫助另一模態訓練

多模態表示學習Multimodal Representation

多模態共享／對應表示（Multimodal Shared / Corresponding Representations）

- 在多模態學習中，不同模態（如文字、影像、音訊）具有各自的特徵表示。為了有效整合這些資訊，需將不同模態的資料映射到一個共享的表示空間，使模型能夠理解並比較來自不同模態的資訊。

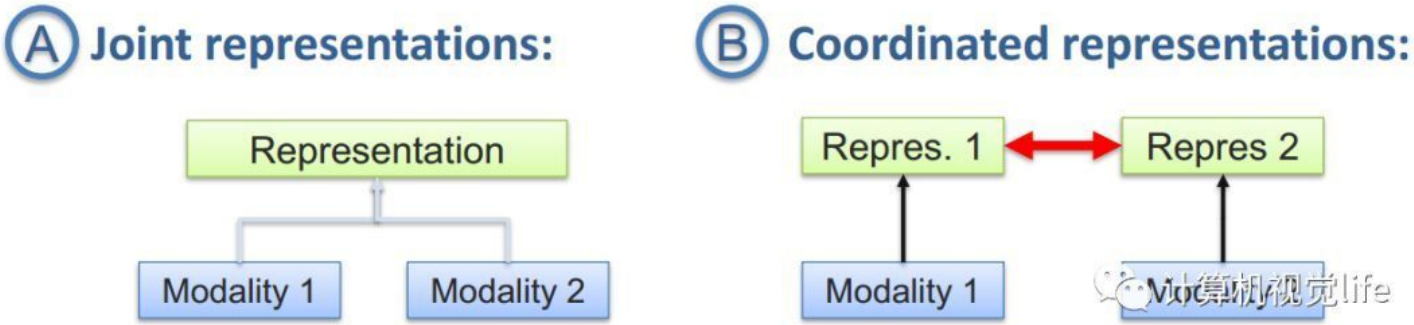
- **技術方法**
  - **Canonical Correlation Analysis (CCA)**：尋找兩個模態之間的線性關係，使其在投影後的表示具有最大相關性。
  - **Deep CCA**：利用深度神經網路學習非線性的投影，捕捉更複雜的模態關係。
  - **Contrastive Learning**：透過對比學習方法，如 CLIP，將文字與影像對應起來，學習到共享的表示空間。



CLIP: Contrastive Language-Image Pre-Training (2025)

多模態表示的兩大類型

- 聯合表示（Joint Representations）
  將多個模態的資訊一起映射到一個統一的多模態向量空間
  例如把文字＋圖片一起變成一個向量。
- 協同表示（Coordinated Representations）
  將多模態中的每個模態分別映射到各自的表示空間，但映射後的向量之間滿足一定的相關性約束（例如線性相關）
  不需要完全融合，但要讓不同模態在各自空間中「對齊」。

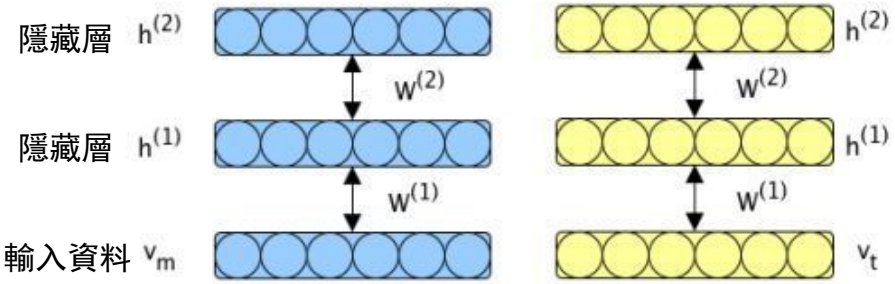


多模態表示-聯合表示

影像專屬模型　　文字專屬模型

Image-specific DBM　　**Text-specific DBM**

逐步抽取影像的高階特徵表示



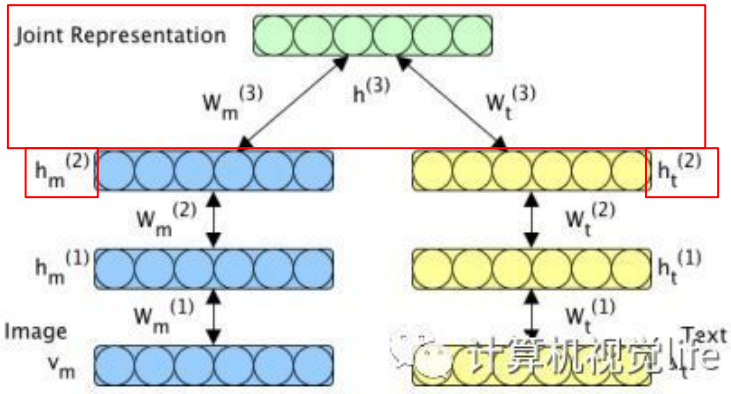
對單一模態（圖片）做獨立的表示學習。

另一個模態（文字）做獨立表示。

多模態模型

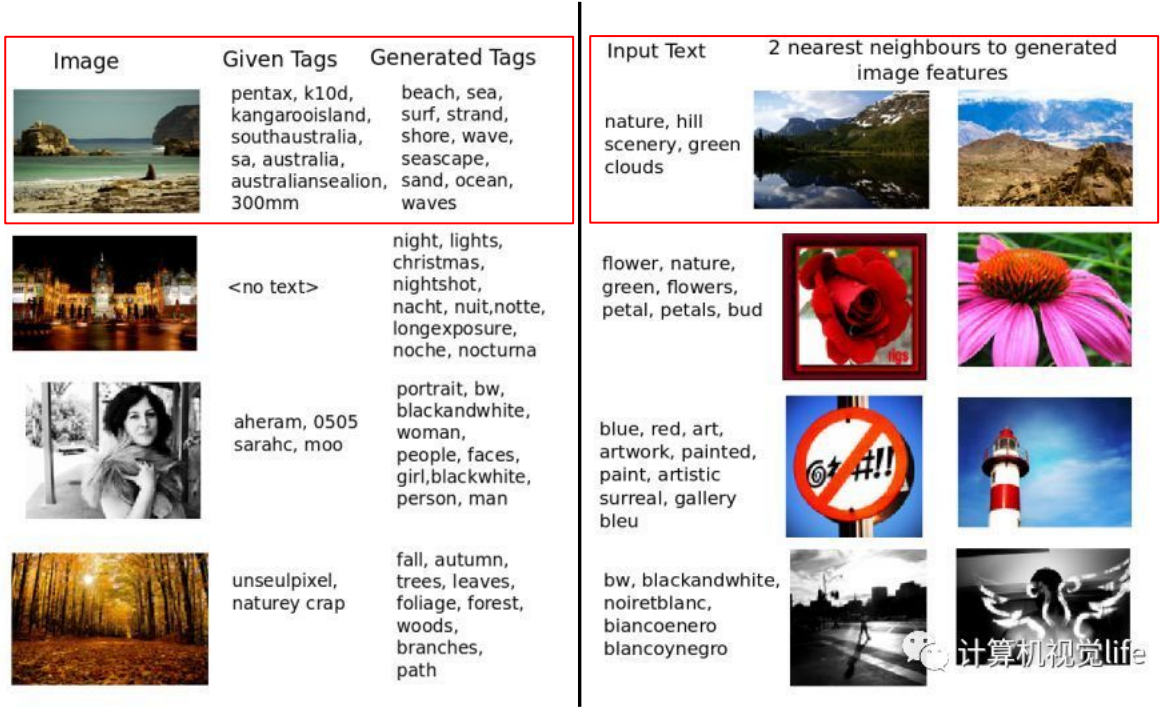
Multimodal DBM

前面影像與文字的高階特徵結合到一個 Joint Representation



這樣模型就能同時考慮影像與文字的資訊，在這個共享空間中進行學習。

多模態表示–聯合表示

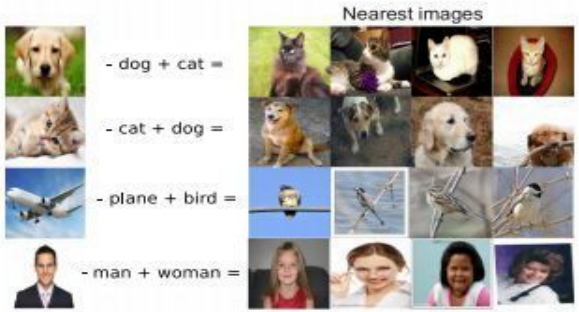


輸入圖片 → 模型能自動產生「beach, sea, wave…」等文字標籤

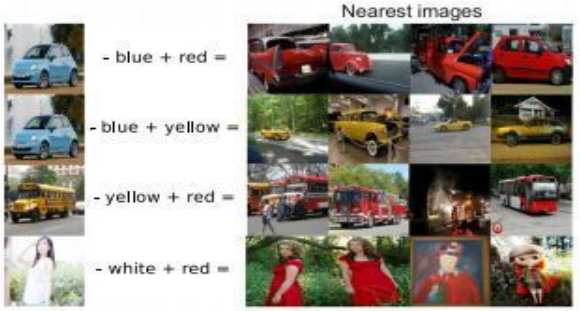
輸入文字標籤 → 模型反過來找到兩張與描述接近的風景圖片（山與湖）

多模態表示-協同表示

- 利用協同學習到的特徵向量之間滿足加減算數運算這一特性，可以搜索出與給定圖片滿足“指定的轉換語義”的圖片。



(a) Simple cases



(b) Colors

- **狗的圖片特徵向量**
  - **狗的文字特徵向量**
  + **貓的文字特徵向量**
  = **貓的圖片特徵向量**

  -> 在特徵向量空間，根據最近鄰距離，檢索得到貓的圖片。

(c) Image structure

(d) Sanity check

计算机视觉life

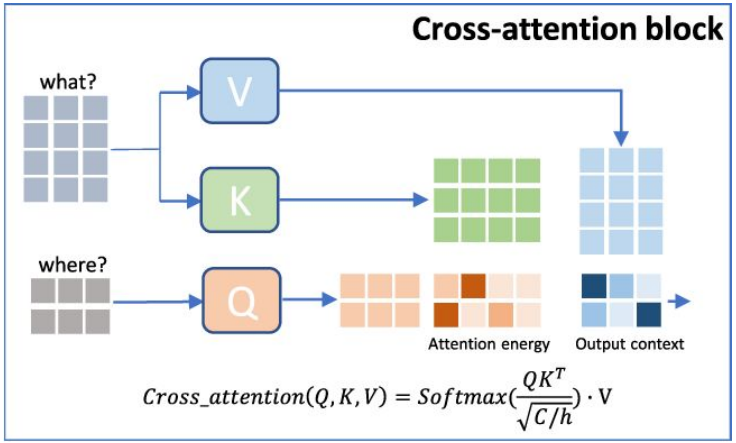
多模態對齊(Alignment)

Cross-Attention

- Cross-Attention 是一種注意力機制，允許模型在處理一個模態的同時，參考另一個模態的資訊。例如，在影像問答中，模型在生成答案時，會根據問題的文字資訊，關注影像中的相關區域。

- **技術方法**

  - **Transformer 架構中的 Cross-Attention 模塊**：
    將一個模態的表示作為 查詢(Query)，另一個模態的表示作為鍵(Key)和值(Value)，計算注意力權重，融合資訊。
  - **應用於多模態模型，如 ViLBERT、LXMERT**：
    這些模型在處理影像與文字的任務中，廣泛使用 Cross-Attention 機制。



➢ In-Depth Guide to Visual Language Models

多模態對齊-視覺問答（Visual Question Answering, VQA）

在 VQA 任務中，模型輸入一張圖片與一個問題，例如：

> 🖼️ 圖片：一位女生在餵長頸鹿
> ❓ 問題：「這位女生在餵什麼？」

模型需要根據圖片與文字問題，輸出正確答案（例如：「胡蘿蔔」）。

Cross-Attention 的作用：

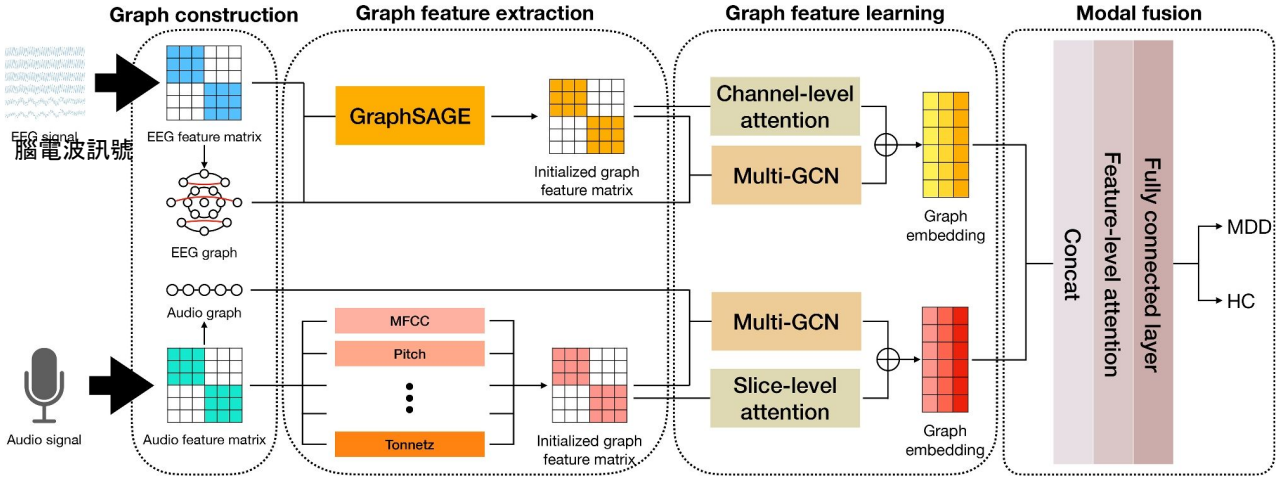
- **Q（Query）** ← 來自文字問題的向量（例如「What is the woman feeding the giraffe?」）
- **K、V（Key/Value）** ← 來自 CNN 提取的圖像特徵（例如各區域的 ResNet 向量）
- 模型透過 Cross-Attention 計算：
  - 問題字詞與圖像各區域特徵的關聯性
  - 對應到圖像中「女生餵長頸鹿」的區域
- 最終聚合關聯區域資訊，生成答案「carrot」。

多模態對齊(Alignment)

圖神經網路對齊技術（Graph Neural Network Alignment）

- 圖神經網路（GNN）能夠有效地建模資料中的結構關係。在多模態學習中，GNN 可用於對齊不同模態的資料，特別是在存在明確結構的情況下，如知識圖譜或社交網路。

- **技術方法**
  - **Cross-modal Graph Neural Networks (CM-GNN)**：構建跨模態的圖結構，學習節點之間的對應關係
  - **Graph Matching Networks**：專注於在兩個圖之間找到最佳的對齊方式，應用於模態間的對應學 習

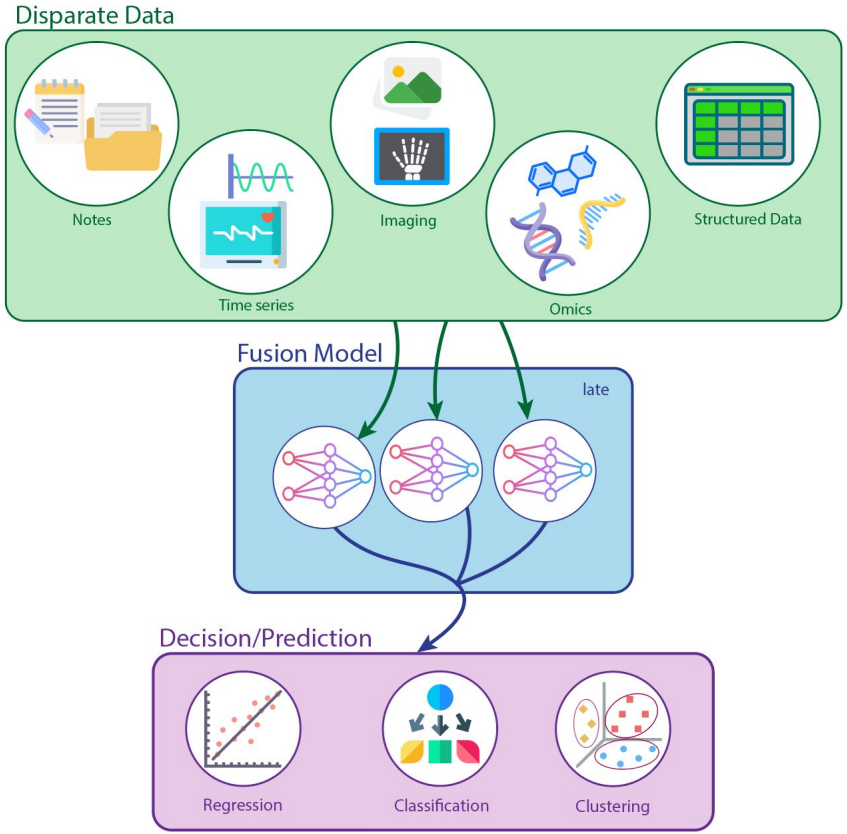


➢ An adaptive multi-graph neural network with multimodal feature fusion learning for MDD detection

多模態階段融合（Multimodal Fusion ）

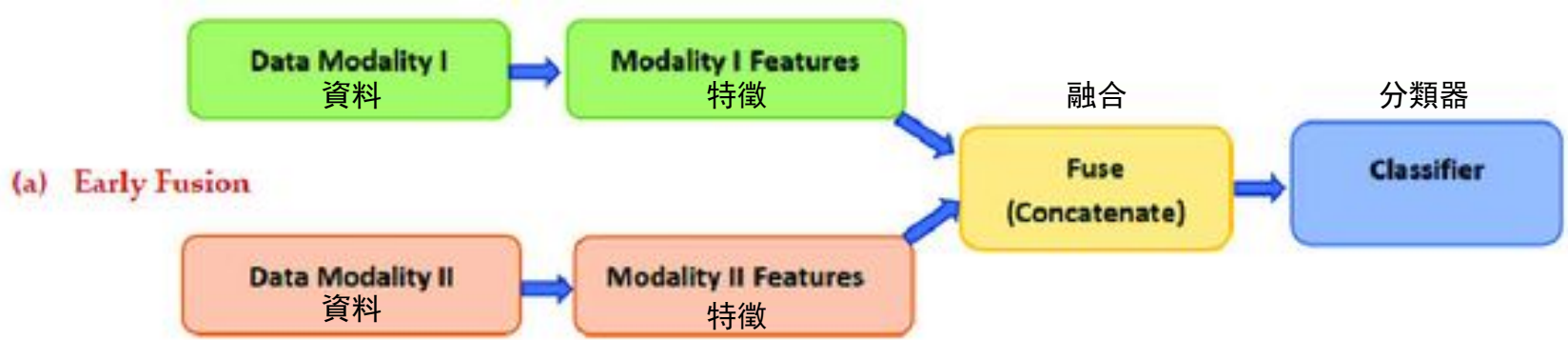
- 是指結合多種不同類型或來源的資料，如文字、圖像、音頻、數值等，進行綜合分析。在金融領域，這種方法有助於從多角度獲取資訊，提升模型的全面性和準確性。

- **多模態階段融合的方法：**
  - 早期融合（Early Fusion）
  - 中期融合（Intermediate Fusion）
  - 晚期融合（Late Fusion）
  - 混合融合（Hybrid Fusion）



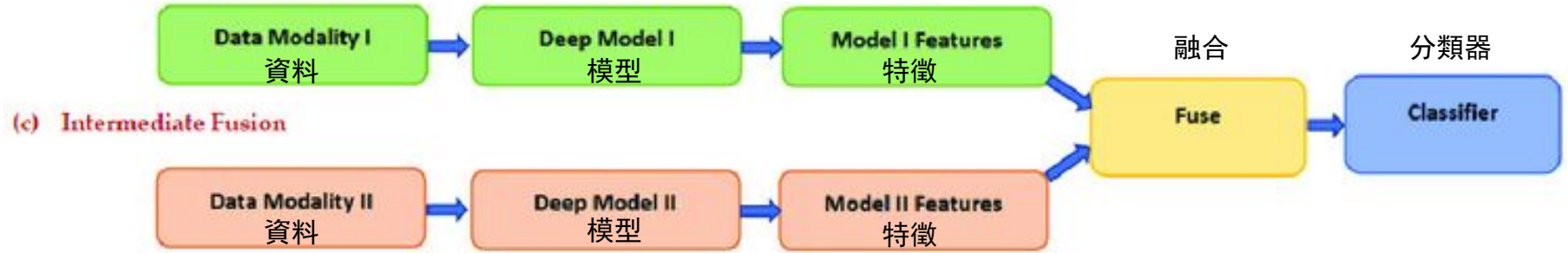
早期融合（Early Fusion）(資料層融合)

- **定義**：在模型輸入階段，將不同模態的原始資料（如文本、圖像、數值等）直接合併成一個統一的表示，然後輸入模型。
- **優點**：1. 融合方法簡單直接，適合資料來源格式一致的場景。
  2. 能夠在初期捕捉模態之間的潛在關係。
- **缺點**：1. 原始資料可能缺乏語義資訊，模型難以捕捉高階關係。
  2. 維度高，計算資源需求大。
- **應用案例**：股票市場中，將技術指標（如移動平均線、RSI）和財務資料（如EPS、市盈率）直接拼接成一個特徵向量，輸入模型預測股價漲跌。



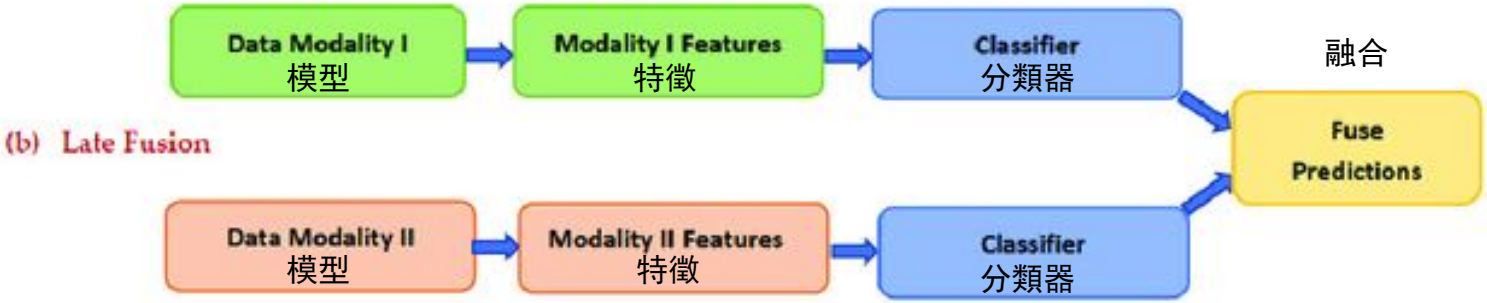
中期融合（Intermediate Fusion） （特徵層融合）

- **定義**：各模態資料先經過專屬處理提取特徵，形成嵌入向量（Embeddings），然後將這些嵌入向量進行融合，再輸入後續的深度學習層處理。
- **優點**：1. 能夠捕捉模態之間的深層交互資訊。
  2. 利用嵌入向量統一表示不同資料模態，提高模型理解能力。
- **缺點**：1. 增加了資料處理的步驟，對計算資源需求更高。
- **應用案例**：將股票的技術面資料（如歷史K線圖）通過CNN提取圖像特徵，新聞情緒（如相關財經報導）通過Transformer模型提取文本特徵，然後將兩種模態的特徵向量進行拼接，最終輸出預測股價走勢。



晚期融合（Late Fusion）（決策層融合）

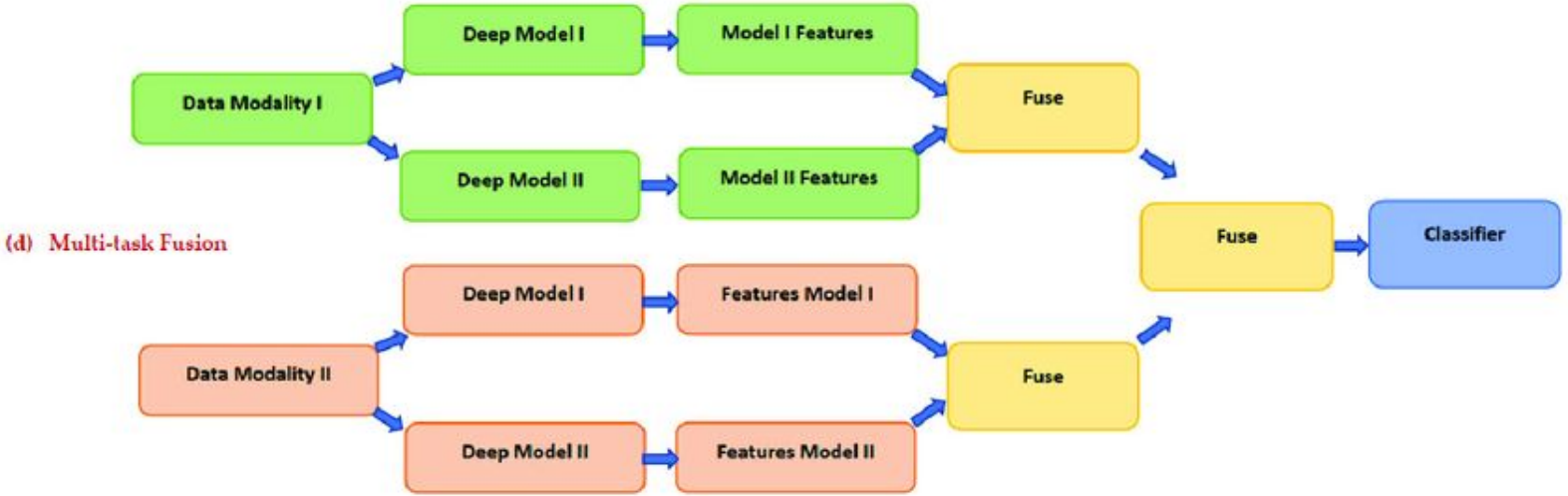
- **定義**：各模態資料獨立輸入各自的模型進行預測，然後將預測結果（如分數或概率）在決策層進行融合。
- **優點**：1. 每個模態模型可以獨立設計與訓練，靈活性高。
  2. 避免模態間資料衝突，模型易於解釋。
- **缺點**：1. 無法充分挖掘模態間的潛在交互關係。
- **應用案例**：使用獨立模型分別處理股票的技術面資料（如交易量、價格波動）和社交媒體資料（如投資者情緒），最後在決策層進行加權平均以生成綜合預測結果。



混合融合（Hybrid Fusion）

- **定義**：結合早期融合、中期融合和晚期融合的優點，根據需求靈活設計融合流程。
- **優點**：兼具多種融合方法的特性，適合複雜的多模態資料分析場景。
- **缺點**：設計和實現的複雜性高。

- **應用案例**：在股票分析中，先使用早期融合將技術面和基本面資料進行拼接，隨後在特徵層加入文本情緒分析，最終在決策層對多模態模型的輸出進行加權整合，實現更精確的股價預測。



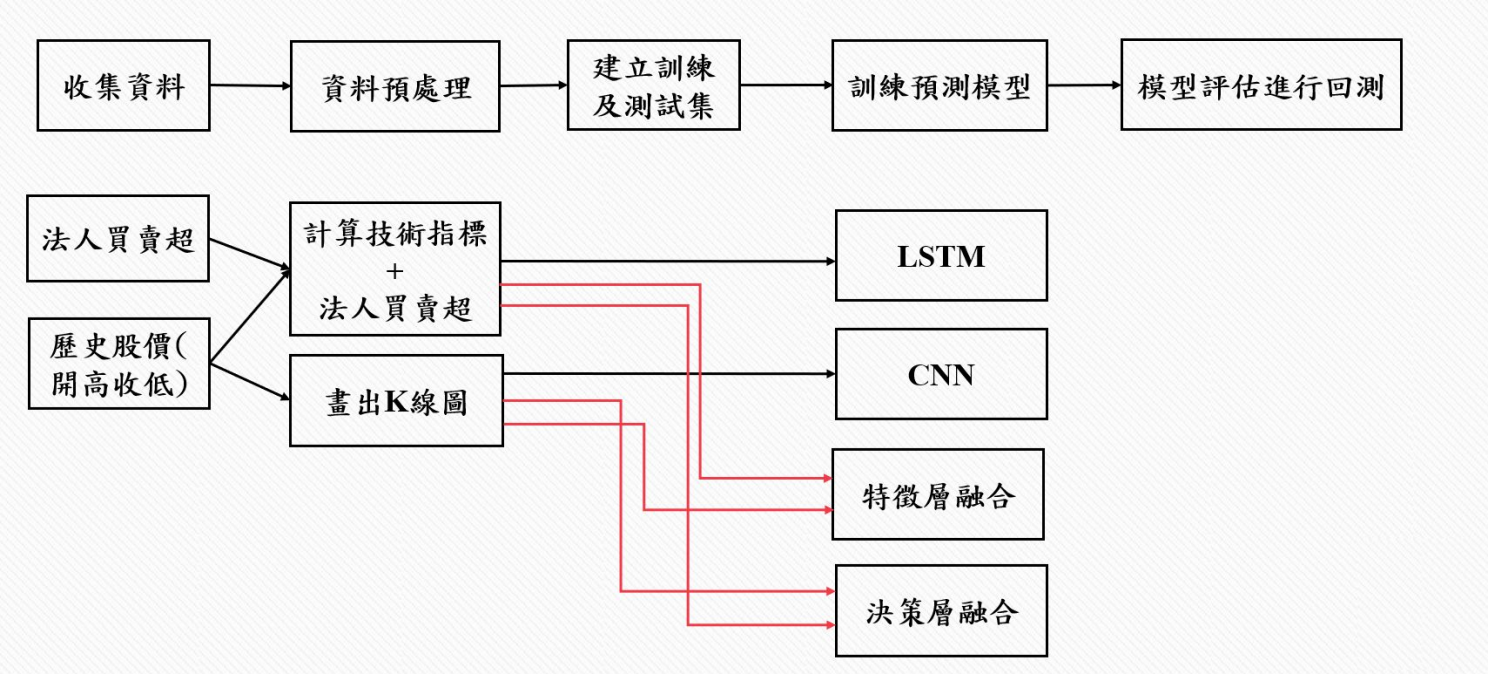
多模態融合-實際案例1

《基於多模態深度學習之一個月股價預測實證》

（ 羅煜凱，2022 ）

指導教授：林冠成 、許志義 教授

實驗流程



多模態融合-實際案例1

《基於多模態深度學習之一個月股價預測實證》

- **LSTM模型結構** 參考自(Mehtab et al., 2020)

  輸入為過去20天的時間序列數值資料(開高收低價、成交量、技術指標、三大人買賣超)

  兩層LSTM層 和 兩層Dropout層

  接下來分兩邊連接三層Dense層預測20天內最高價和最低價

- **CNN模型結構** 參考自(Kusuma et al., 2019)

  輸入為過去20天的日K線圖

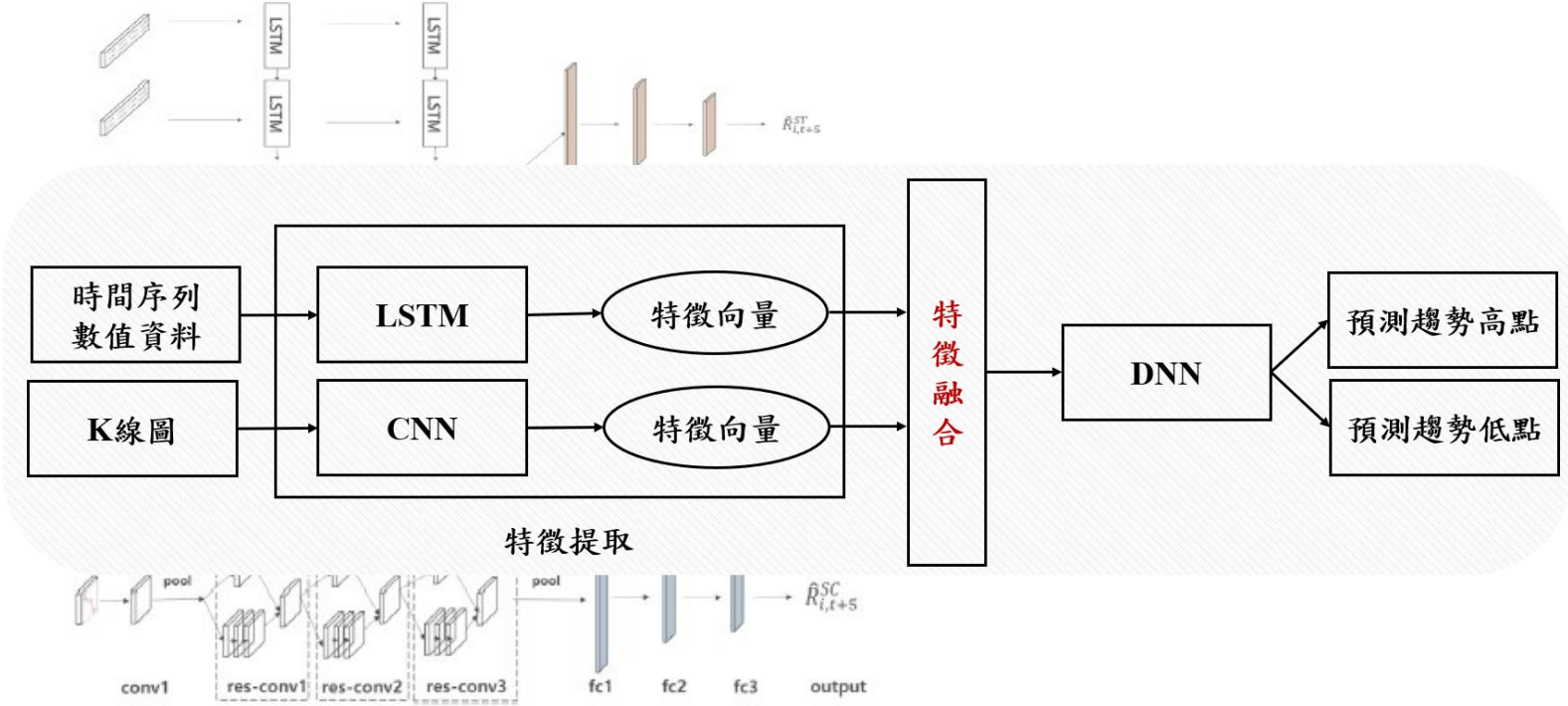
  四層卷積層 和 **四層最大池化層** 連接 flatten 層

  接下來分兩邊連接三層Dense層預測20天內最高價和最低價

多模態融合-實際案例1

《基於多模態深度學習之一個月股價預測實證》

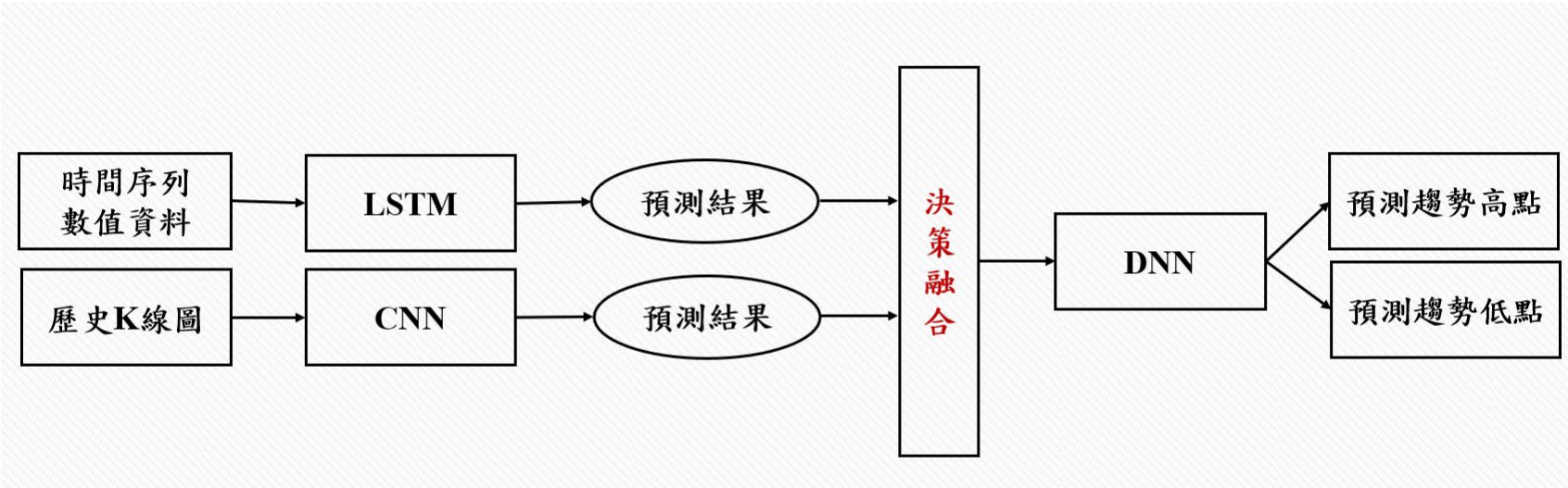
- 特徵層融合模型(Feature Fusion)



多模態融合-實際案例1

《基於多模態深度學習之一個月股價預測實證》

- 決策層融合模型(Decision Fusion)



多模態融合-實際案例2

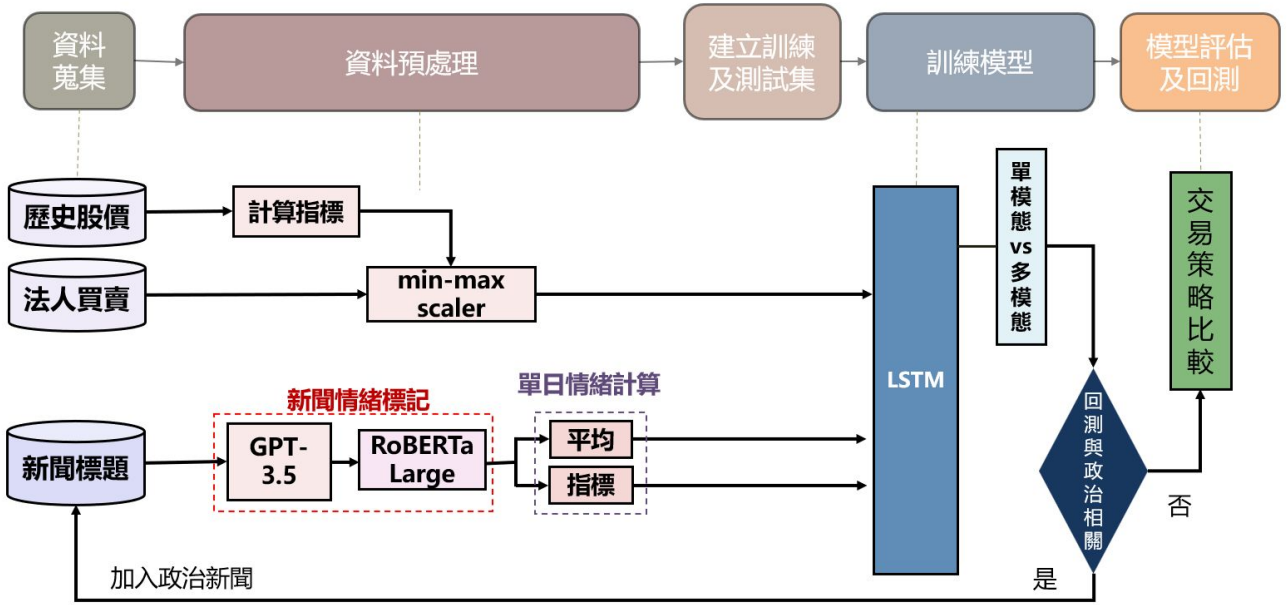
職

《結合新聞情緒標記計算與多模態深度學習應用於股價趨勢預測》

（ 許新湲，2024 ）

指導教授：林冠成 教授

實驗流程



多模態融合-實際案例2

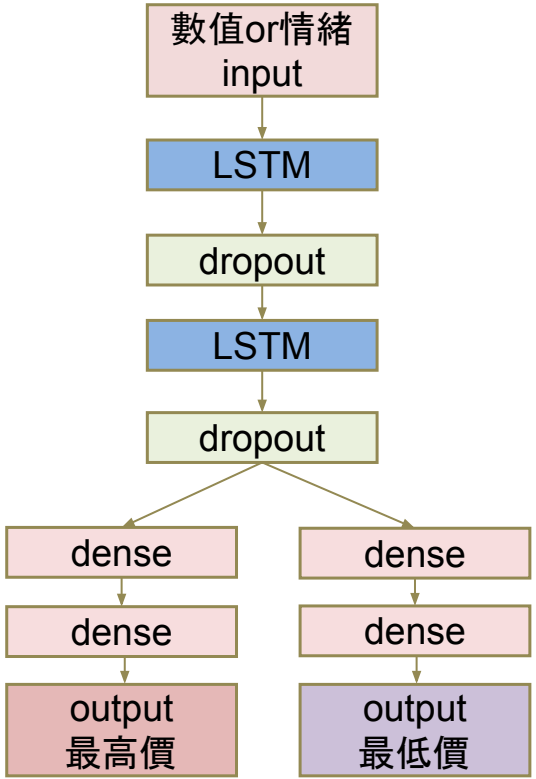
《結合新聞情緒標記計算與多模態深度學習應用於股價趨勢預測》

- LSTM模型結構 參考自Edmure Windsor & Wei Cao, 2022

  輸入為過去10天的時間序列資料(數值、情緒)

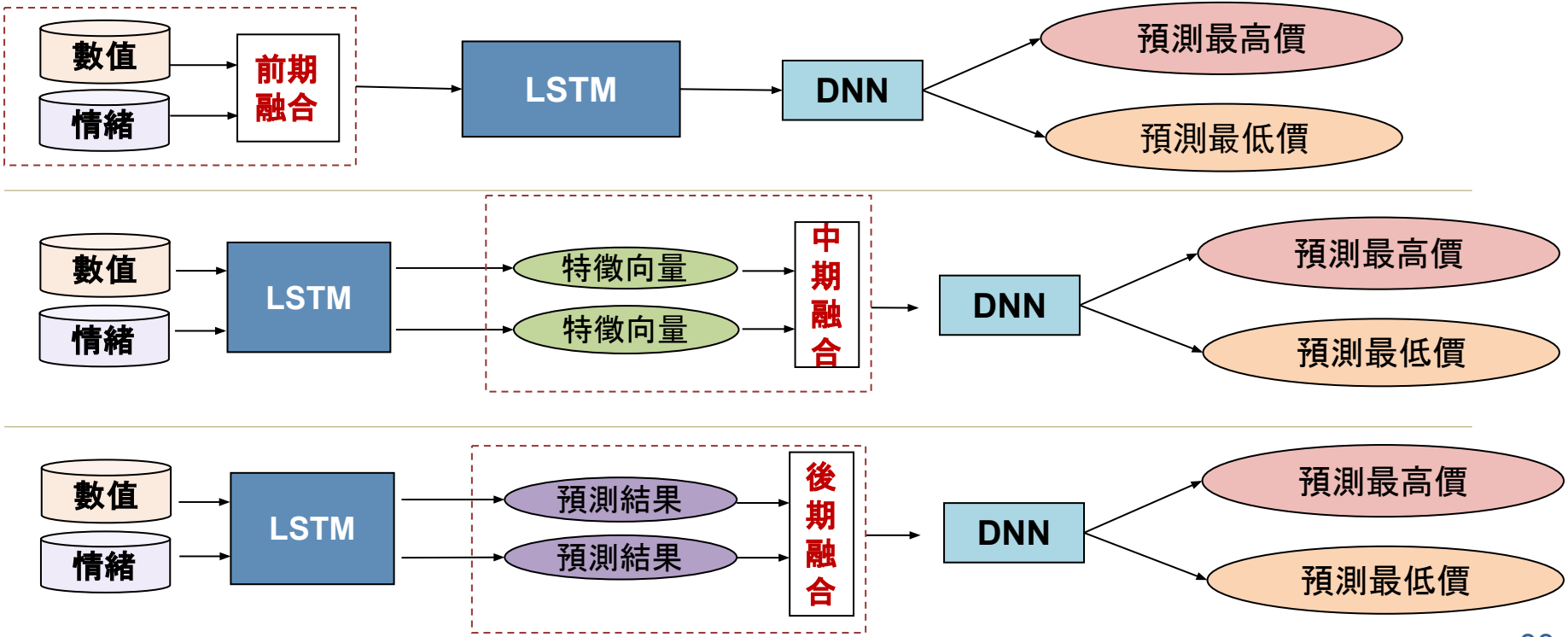
  兩層LSTM層、兩層Dropout層

  分兩邊連接兩層Dense層預測10天內最高價和最低價



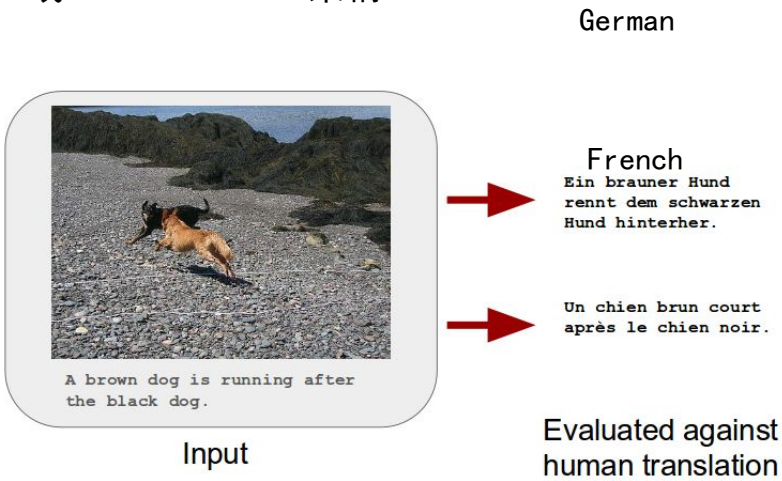
多模態融合-實際案例2

《結合新聞情緒標記計算與多模態深度學習應用於股價趨勢預測》



多模態-Translation(翻譯)

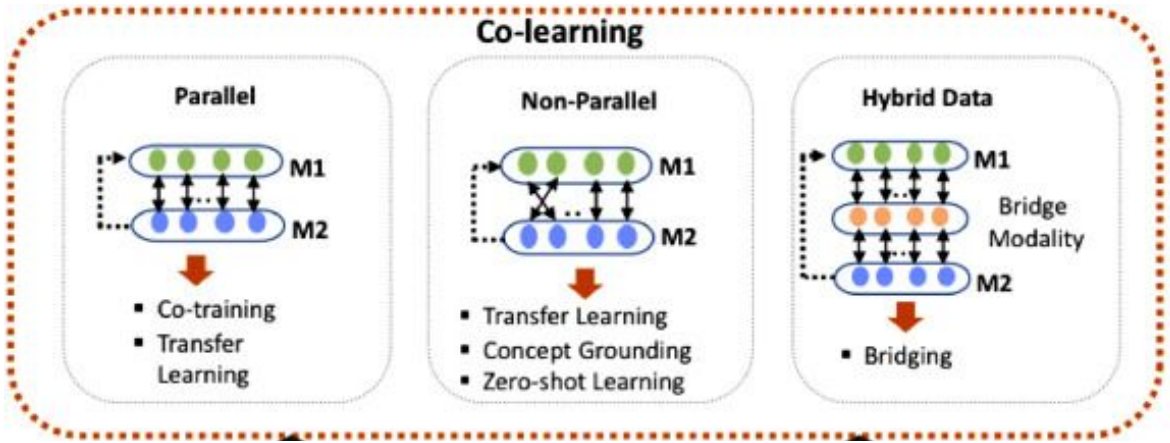
- 結合多種模態信息（例如文本、圖像、語音等）來提升機器翻譯質量的技術。
- 傳統機器翻譯多依賴純文字，而多模態機器翻譯則利**用與文本相關的視覺、語音信息，增強語義理解和上下文判斷。**
- 圖 → 文、文 → 圖、圖+文 → 文
- 常用 Encoder－Decoder 或 Transformer 架構



將一段英文描述（描述圖片內容）翻譯成德文（German）或法文（French），並且在翻譯的過程中，**同時利用圖片提供的輔助資訊(上下文資訊），讓翻譯結果更準確(避免因一詞多義導致翻譯錯誤)。**

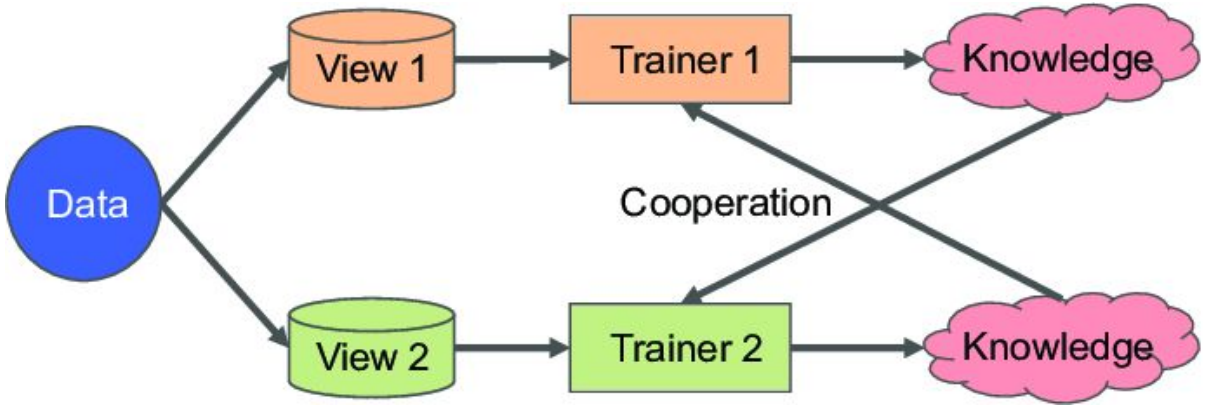
多模態-Colearning(協同學習)

- 結合利用一種模態的知識來幫助或增強另一種模態的學習，達成跨模態的互補與共同提升。
- **Parallel Data（平行資料）**
  - 兩種模態之間有明確的 一對一對應（例如圖像-文字配對資料）
  - Co-training：雙方模型交替提供標籤，互相強化。
  - Transfer Learning：將一種模態學到的知識遷移到另一模態，提升弱模態的效果。
  - 適用於標註齊全、對應關係明確的資料集



Co-training（協同訓練）

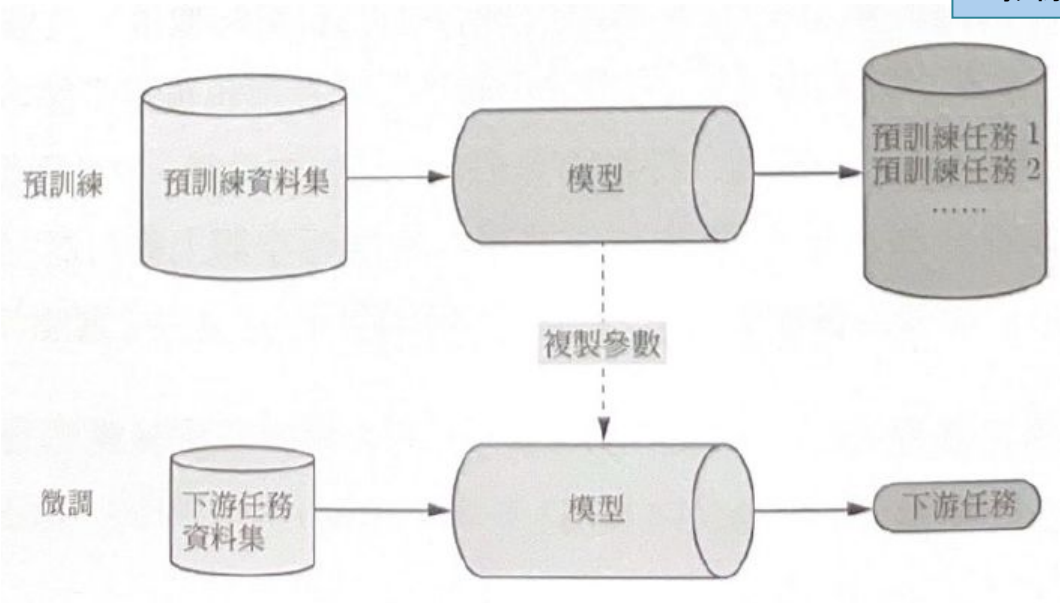
- 透過分別訓練多個模型來互相補充資訊，從而提高對未標記數據的分類性能
- Trainer 1 與 Trainer 2 分別從 View 1、View 2 的資料中學習，形成各自的知識表示。
- 接著，兩個訓練器會進行 Cooperation（合作）：
  - Trainer 1 將自己從 View 1 學到的知識，用於協助 Trainer 2 的學習（例如產生 pseudo-label 或提供輔助特徵）。
  - Trainer 2 同樣將從 View 2 學到的知識，回饋給 Trainer 1，補充另一個模態的訊息。
- **透過這樣的互相教學與知識共享，雙方在迭代過程中不斷提升彼此的準確率與泛化能力。**



多模態預訓練與遷移學習

➢ **整體框架**

遮罩語言模型
遮罩視覺模型
影像文字匹配



跨模態檢索
影像描述
視覺問答
文字生成影像
指代表達
視覺常識推理
視覺語言推理
視覺蘊含

多模態預訓練與遷移學習

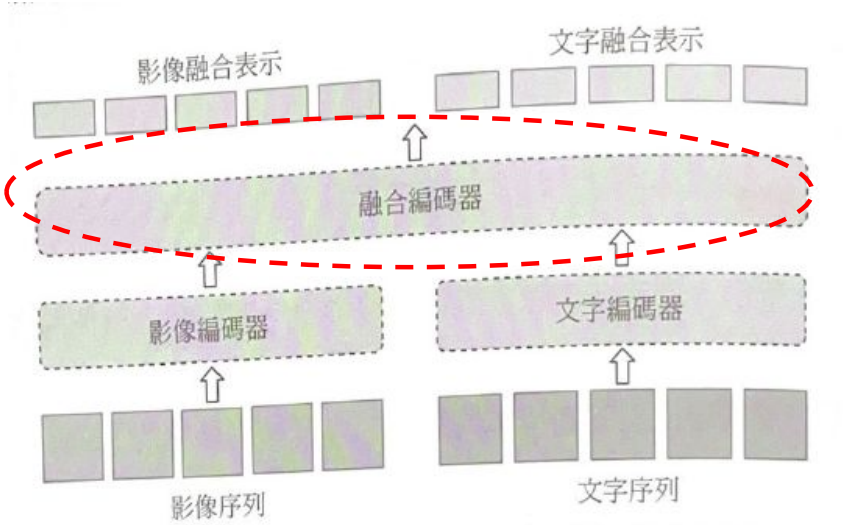
➢ 下游任務

- **視覺常識推理：**
  - 給定一張以隻很多物件區域的圖片
  - 任務要求完成:根據問題選擇答案、解釋選擇該答案的原因
- **視覺語言推理**
  - 給定兩張圖片和一句文字描述
  - 任務要求判斷文字描述是否正確地描述了兩張圖片的內容
- **視覺蘊含：**
  - 給定一張圖片和一句文字描述
  - 任務要求判斷由圖片(前提)推斷句子(假設)是否合適(蘊含/中立/矛盾)

多模態預訓練與遷移學習

➢ 模型結構 -基於融合編碼器 (encoder)

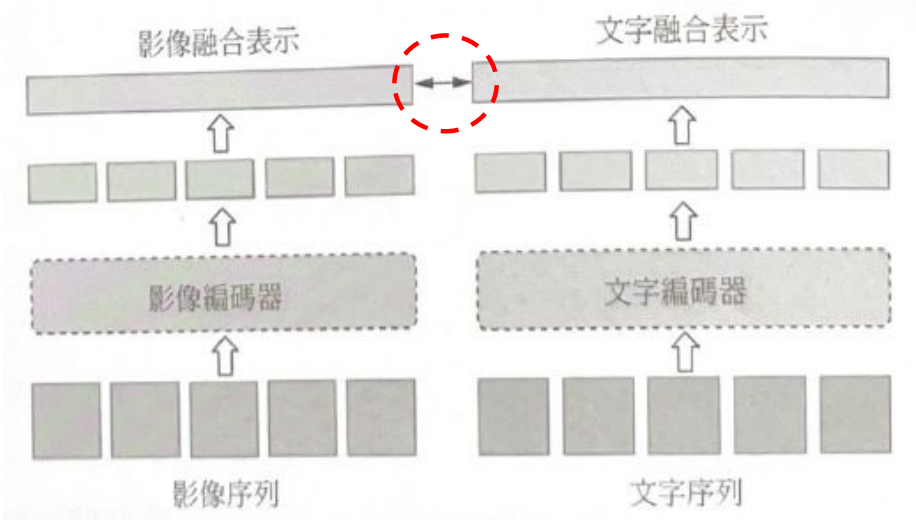
- 使用兩個單模態模型分別編碼影像資訊和文字資訊，之後再利用多模態融合模型對圖文單模態編碼進行建模，獲得圖文融合表示。
- 其中單模態模型往往採用標準的 transformer 模型，而最常用的多模態融合模型則採用交叉 transformer 模型。



多模態預訓練與遷移學習

➢ 模型結構 -基於雙編碼器

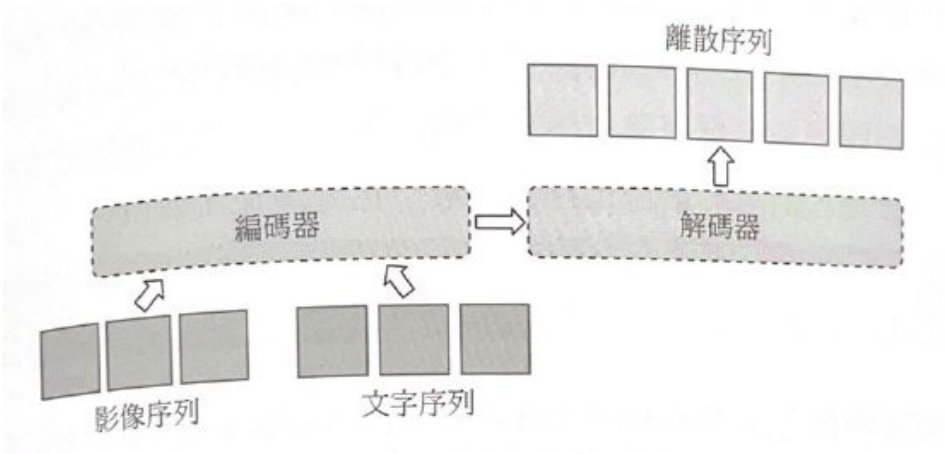
- 使用兩個單獨的編碼器分別學習影像和文字的對應表示
- 通常在對應表示空間中增加圖文相似性連結約束以建立圖文連結



多模態預訓練與遷移學習

➢ 模型結構 -基於編解碼框架的模型

- 將影像序列和文字序列拼接成一個輸入序列，並將多個任務的輸出轉化成共用詞表的離散序列，最終使用編解碼模型學習輸入和輸出的連結。
- 此類模型不再關注通用多模態表示的學習，而是將不同任務的輸入/輸出轉化成統一的形式，以達到使用一個單一的模型同時建模多種任務的目標。



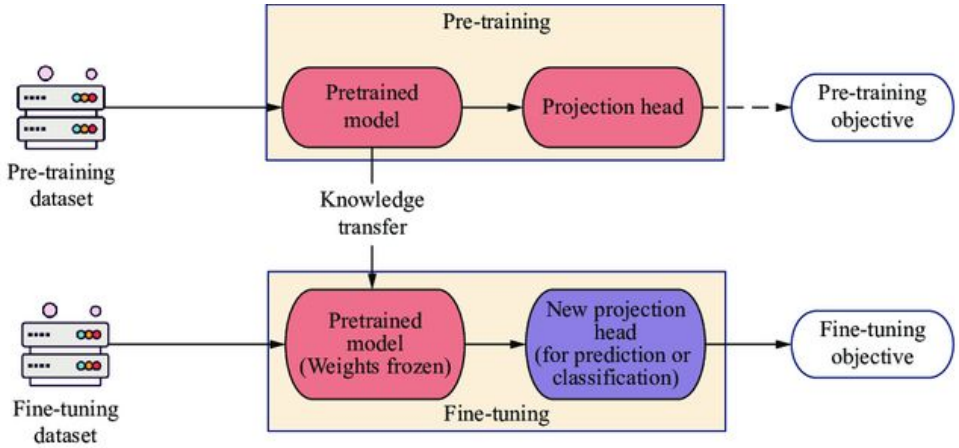
多模態預訓練與遷移學習

➢ 多模態預訓練與遷移學習的價 值

多模態預訓練模型透過大量的圖文對資料進行訓練，學習到圖像與文字之間的對應關係。
這些模型在訓練後，能 夠將不同模態的資料映射到同一個語意空間，實現跨模態的理解與生成。遷移學習則允許這些預訓練模型在特定任務上進行微調，快速適應新的應用場景。

在商業智慧應用中，這種技術能夠：

- 提升資料處理效率，減少人工標註成本。
- 加強模型的泛化能力，適應多變的商業環境。
- 實現跨模態的資料分析與決策支持。



Pre-training_in_Medical_Data_A_Survey

多模態預訓練與遷移學習－CLIP

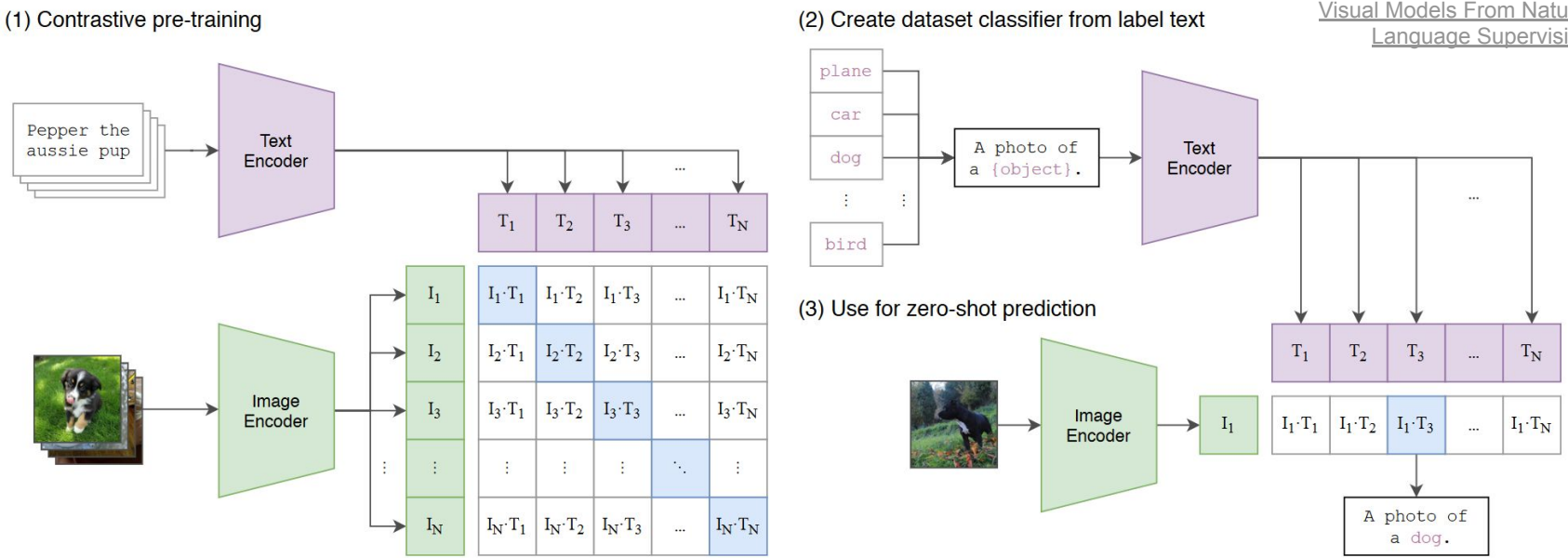
職

➢ CLIP：對比式語言 -圖像預訓練模型

OpenAI

CLIP（Contrastive Language–Image Pre-training）由 OpenAI 開發，透過對比學習的方法，將圖像與其對應的文字描述映射到同一個嵌入空間 。在訓練過程中，CLIP 同時處理大量的圖像與文字對，學習到圖像與文字之間的語意對應關係。

CLIP Learning Transferable Visual Models From Natural Language Supervision



多模態預訓練與遷移學習－CLIP

➢ CLIP：對比式語言 -圖像預訓練模型

CLIP（Contrastive Language–Image Pre-training）由 OpenAI 開發，透過對比學習的方法，將圖像與其對應的文字描述映射到同一個嵌入空間 。在訓練過程中，CLIP 同時處理大量的圖像與文字對，學習到圖像與文字之間的語意對應關係。

實際應用案例：電商平台的商品推薦系統

假設某電商平台希望提升商品推薦的準確性，考慮結合用 戶的瀏覽行為（文字描述）與商品圖片（影像資料）。

應用方式：

- 使用 CLIP 模型，將商品的文字描述與圖片映射到同一個表示空間。
- 在推薦模型中，利用用 戶的瀏覽歷史作為 查詢，商品的多模態表示作為鍵和 值，計算注意力權重，生成個性化的推薦列表。

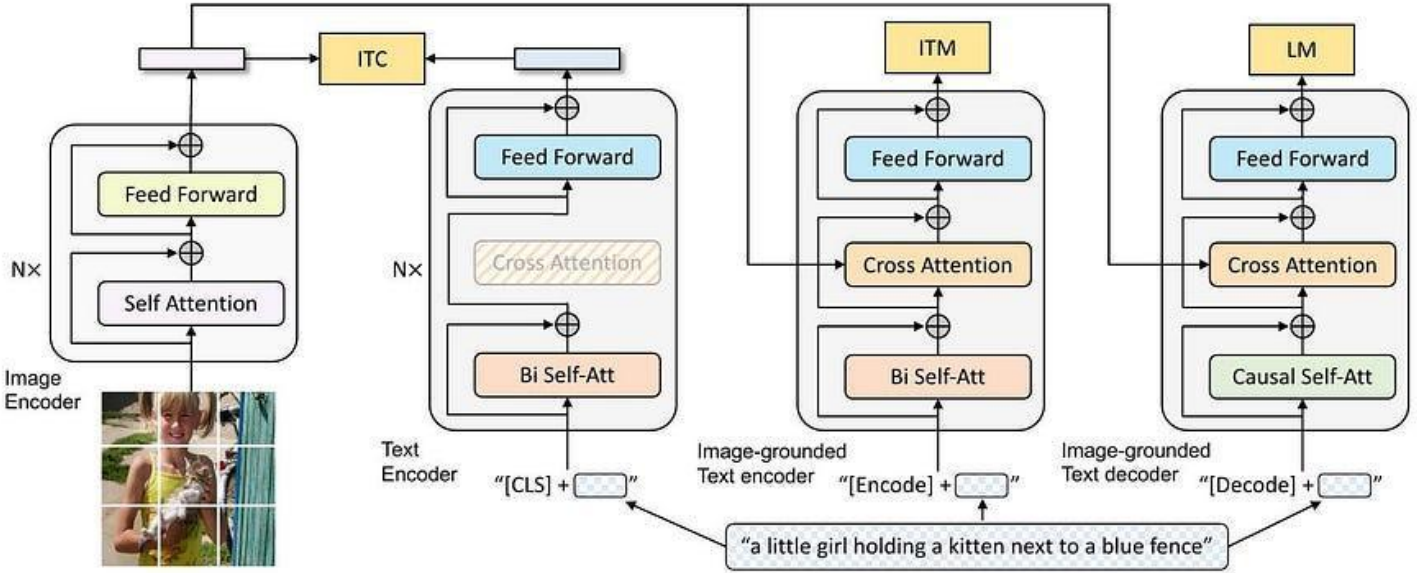
預期成效：

- 推薦點擊率提升 15%
- 用戶停留時間增加 10%

多模態預訓練與遷移學習－BLIP

➢ BLIP：語言-圖像預訓練的統一框架

BLIP（Bootstrapping Language-Image Pre-training）由 Salesforce AI Research 開發，旨在統一視覺-語言的理解與生成任務。BLIP 引入了多模態的編碼器-解碼器架構，能夠同時處理圖像與文字資料，並進行跨模態的生成任務。



BLIP for Unified Vision-Language Understanding and Generation

多模態預訓練與遷移學習－BLIP

➢ BLIP：語言-圖像預訓練的統一框架

BLIP（Bootstrapping Language-Image Pre-training）由 Salesforce AI Research 開發，旨在統一視覺-語言的理解與生成任務。BLIP 引入了多模態的編碼器-解碼器架構，能夠同時處理圖像與文字資料，並進行跨模態的生成任務。

實際應用案例：金融風險評估系統

金融機構需要對大量的財務報告與新聞文本進行情緒分析，以評估潛在的風險。

應用方式：

- 使用 BLIP 模型，將財務報告中的圖表與文字描述進行整合，生成統一的語意表示。
- 對新聞文本進行情緒分析，識別可能影響市場的負面訊息。
- 結合上述資訊，建立風險評估模型，提供決策支持。

預期成效：

- 風險預測準確率提升 20%
- 風險識別時間縮短 30%

多模態預訓練與遷移學習

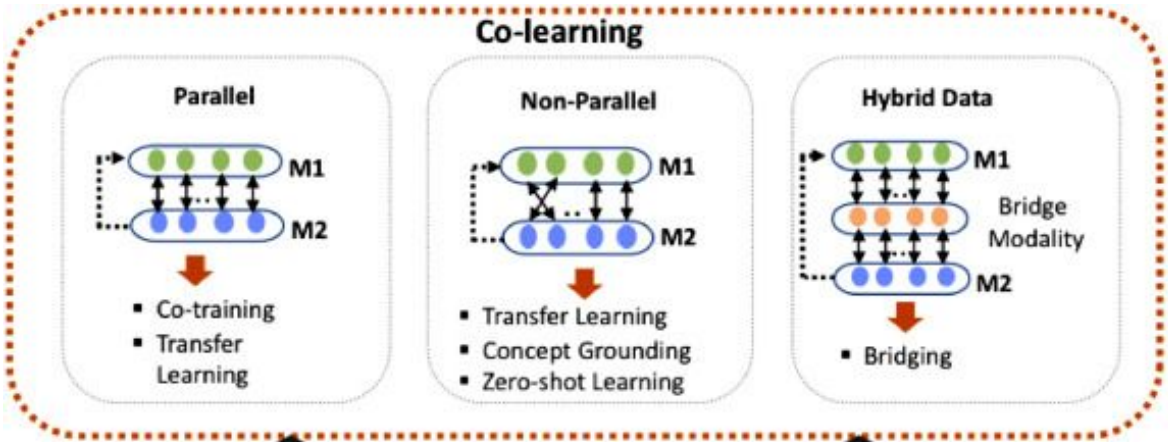
職

➢ CLIP VS. BLIP

模型	開發者	架構	主要特點	適用場景
CLIP	OpenAI	雙編碼器	對比學習、零樣本學習	圖像分類、文本-圖像檢索
BLIP	Salesforce	編碼器-解碼器	統一理解與生成、多任務學習	圖像描述、視覺問答、風險評估

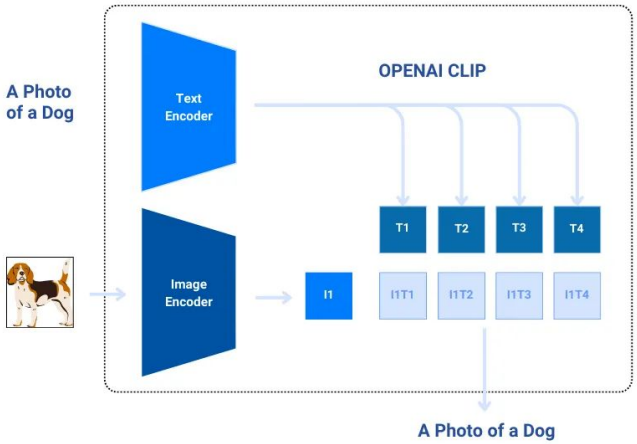
多模態-Colearning(協同學習)

- **Non-Parallel Data（非平行資料）**
  - 不同模態之間沒有直接對應關係（例如圖像資料集與文字資料集分開）
  - Concept Grounding：將兩種模態映射到共同的語意空間，以建立間接對應。
  - Zero-shot Learning：即使未見過配對資料，也能跨模態進行任務（如 CLIP 能以文字檢索圖片）。
  - 適用於資料來源異質、缺乏對應的情境。
  - 模型如CLIP、BLIP為代表



Concept Grounding(概念對應)-CLIP

- 將一種模態中的抽象語意對應到另一種模態的具體訊息
- 不需要平行資料，也能在共享概念空間中完成跨模態對齊
- 常用於 Non-parallel Co-learning，例如：
  - 文字概念「dog」對應到影像中的狗的特徵
  - 利用語意空間連結文字與視覺，使模型跨模態共享知識



輸入：左邊有一張「狗的照片」，以及一段對應的文字描述「A photo of a dog」

兩個編碼器（Encoders）：

CLIP 使用 **Image Encoder**（圖像編碼器）將圖片轉換成一個向量表示（I1）。

同時，使用 **Text Encoder**（文字編碼器）把文字轉成文字向量（T1～T4）。

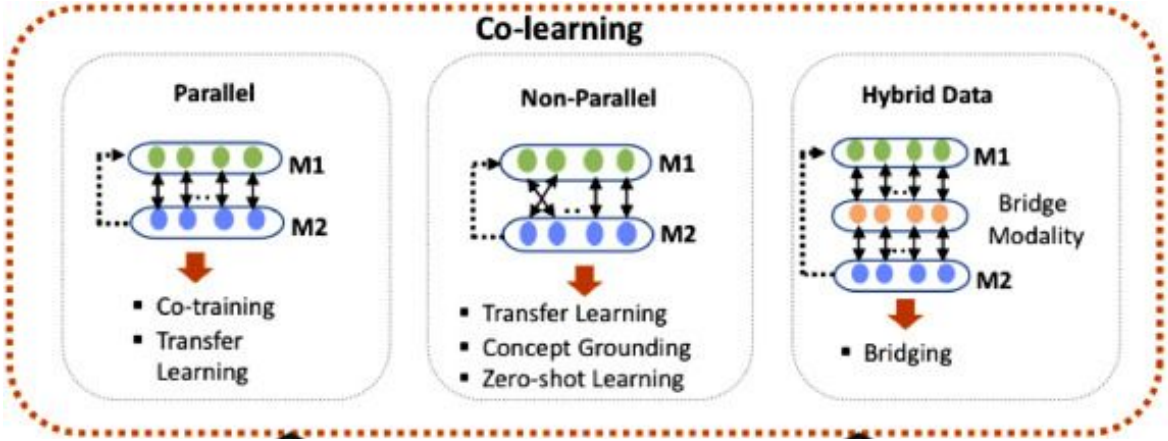
投影到共同語意空間：

這兩種向量會被映射到 **同一個語意空間** 中。

目標是：相對應的圖文（例如「狗的照片」與「a photo of a dog」）要在這個空間中距離很近；不相符的圖文則距離遠。

多模態-Colearning(協同學習)

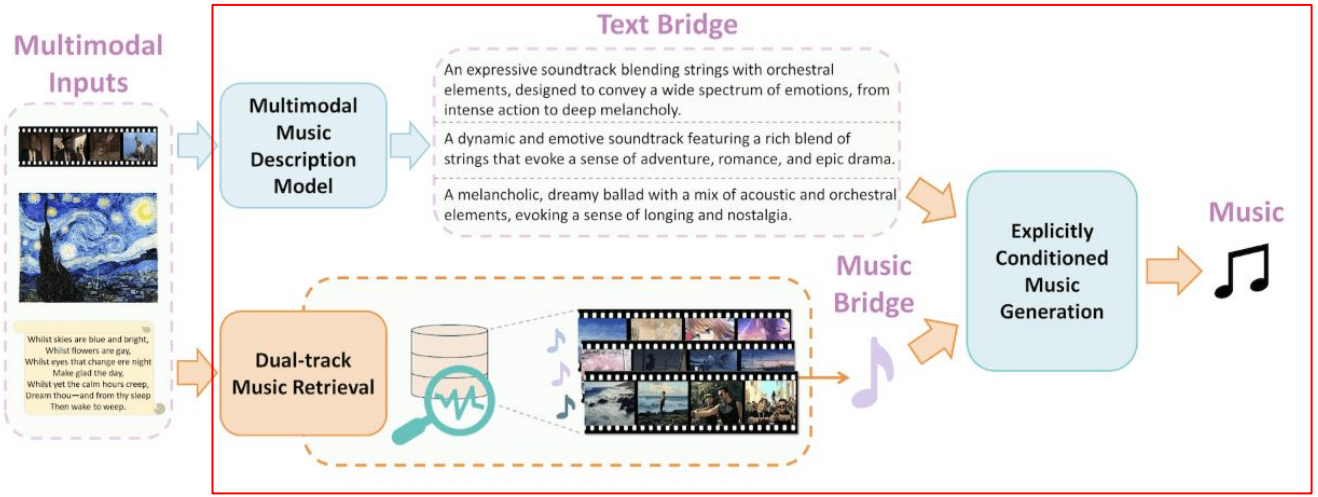
- **Hybrid Data（混合資料）**
  - 結合部分平行資料與大量非平行資料，是實務中最常見的情況
  - **Bridging（橋接模態）** ：引入第三種「橋接模態」（如共同語意標籤、語音、meta feature），作為中介，幫助不同模態建立聯繫。
  - 適用於標註資料有限，但有大量單一模態資料的應用場景（例如醫療影像 + 病歷文字）。



Bridging（橋接模態）-Visuals Music Bridge (VMB)

圖像與音樂之間 沒有直接配對資料集，引入「第三方模態」作為橋樑
透過文字描述（Text Bridge）或音樂檢索（Music Bridge）間接建立跨模態對應

1. 輸入圖像 → 透過 **Multimodal Music Description Model** 產生描述性文字（如「一首交響弦樂與史詩冒險氛圍的曲子」）
2. 這些文字描述再被用來指導音樂生成模型 → 形成音樂
3. 相當於先把視覺內容轉成語意描述，再讓音樂模型依據文字生成對應音樂



圖像（電影畫面藝術圖像）
文字（詩句）

1. 文字輸入（如詩） → 使用 Dual-track Music Retrieval（雙軌音樂檢索）
2. 從資料庫中找到對應音樂，進而引導音樂生成模型
3. 當於利用現有音樂片段作為「中間媒介」，實現跨模態對應

可解釋AI （XAI）

- 什麼是模型可視化？

模型可視化是指以圖形或可視化方式來呈現和理解機器學習模型的內部結構、運作方式和學習到的特徵。

這些視覺化工具和技術幫助我們更好地理解模型在處理資料時的行為，並有助於調試、優化和改進機器學習模型。

可解釋AI （XAI）

•不同類型的模型可視化

- **資料探索** -資料探索是使用探索性資料分析(EDA) 完成的。應用 T 分佈隨機鄰域嵌入 (t-SNE) 或主成分分析 (PCA) 技術來理解該特徵。
- **構建模型** -用於衡量分類和回歸模型的各種指標。分類中使用準確率、精確率和召回率、混淆指標、對數損失和 F1 分數，回歸中使用均方誤差 (MSE)、均方對數誤差、均方根誤差 (RMSE)。構建模型後的所有這些指標都用於理解和衡量性能。
- **決策樹模型** -靜態特徵摘要，例如從模型中檢索的特徵重要性。它僅存在於基於決策樹的算法中，例如隨機森林和XGBoost。
- **評估模型** -評估模型的錯誤預測。

可解釋AI （XAI）

職

- **訓練期間的模型可視化**
  - **標量（損失和準確度）** -標量可用於顯示訓練過程中誤差的趨勢。除了定期將損失和準確性記錄到標準輸出之外，我們還記錄並繪製它們以分析其長期趨勢。
  - **直方圖** -可視化模型圖中張量的分佈如何隨時間變化。顯示不同時間點張量的許多直方圖可視化。
  - **權重和偏差** -通過在直方圖上可視化來監控訓練期間的權重和偏差。
  - **激活** -為了使梯度下降發揮最佳性能，節點通常在激活函數分發之前輸出。
  - **梯度** -每層的梯度都可以可視化，以識別深度學習問題，例如梯度遞減或梯度爆炸問題。
  - **圖表** -圖表可視化模型的內部結構或體系結構。
  - **圖像** -訓練每一步的圖像意味著生成的中間圖像可以可視化並可視化張量。

可解釋AI （XAI）

- **為什麼模型可視化很重要？**
  - 有必要了解所有這些算法如何做出決策。
  - 認識模型的基本特徵可以讓我們深入了解其內部運作方式，並為消除偏差和提高其性能提供方向。
  - 有助於調試模型
  - 提供預測解釋的主要原因是可解釋的機器學習模型對於獲得最終用戶的信任是必要的。

可解釋AI （XAI）

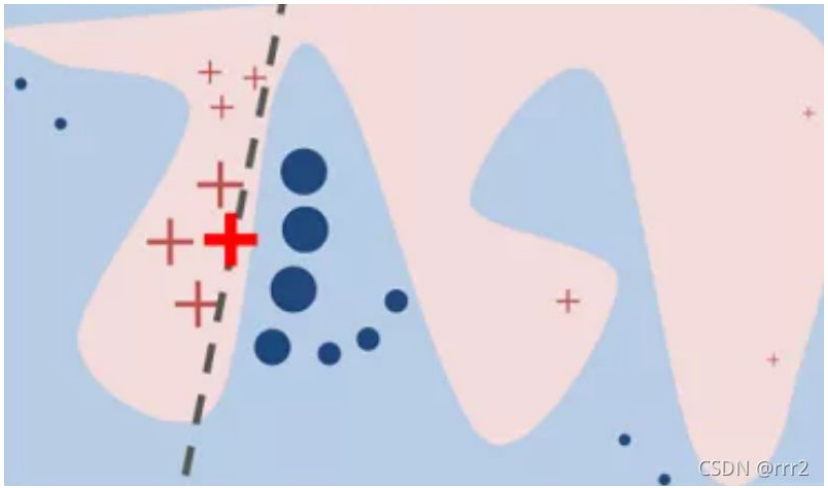
- **為什麼模型可視化很重要？**

  - 我們需要在大多數情況下進行解釋:
    - **可信度** －如果使用分類或預測結果，則需要像股票交易者一樣了解一些領域知識，以提供購買或出售特定股票的決策。
    - **透明度**－機器學習不能成為黑匣子，而應該為客戶、消費者和管理層提供模型結構和清晰度。就像開源一樣，模型理解也應該是開源的。
    - **責任**－模型應該有責任向消費者提供正確的答案。作為模型所有者，我們應該驗證模型特徵以保證其對決策的幫助。

LIME
(Local Interpretable Model Agnostic Explanation)

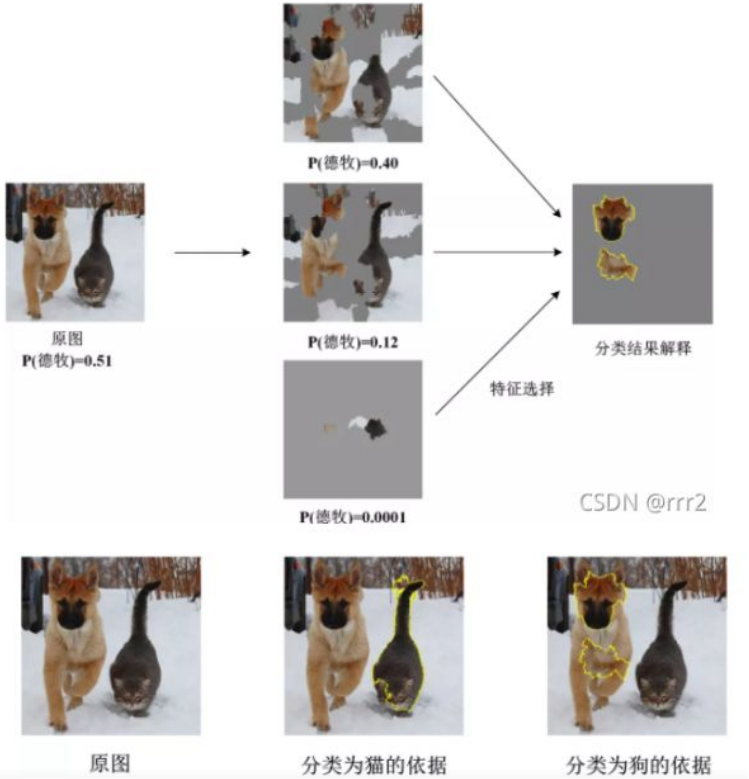
定義:對一個複雜的分類模型(黑盒),在局部擬合出一個簡單的可解釋模型。

需要解釋的樣本



紅色和藍色區域表示一個複雜的分類模型(黑盒),從加粗的紅色十字樣本周圍採樣,
將採樣出的樣本用分類模型分類並得到結果,同時根據採樣樣本與加粗紅十字的距離賦予權重。
虛線表示通過這些採樣樣本學到的局部可解釋模型

LIME
(Local Interpretable Model Agnostic Explanation)



從特徵的角度考慮，不再以單個像素為特徵，而是以超像素為特徵，整個圖片的特徵空間就小了很多

LIME找出對分類結果影響最大的幾個超像素，也就是說模型僅通過這幾個像素塊就已經能 夠做出預測。

右圖為貓狗辨識舉例
將圖片利用超像素切割取得特徵，將各為狗的特徵組合分析出結果

多模態可解釋分析 實際案例-LIME

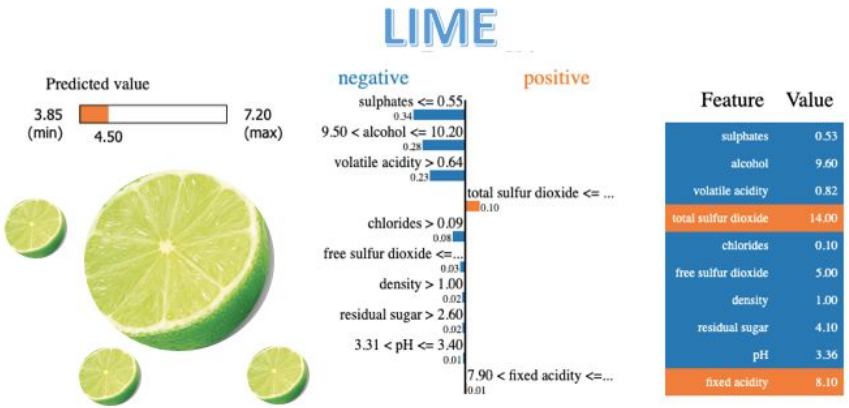
➢ **結構化資料: LIME 在信貸風險評估的應用**

在信用風險管理中，銀行常使用機器學習模型（如 XGBoost）來預測借款人的違約風險。然而，這些模型的黑箱特性使得解釋預測結果變得困難。

- **LIME（Local Interpretable Model-agnostic Explanations）**

  透過在特定資料點附近生成擾動樣本，建立簡單模型解釋原始模型的預測，適用於各種模型。

- LIME 通過在申請人資料附近生成擾動樣本，建立簡單模型解釋原始模型的預測，幫助銀行理解特定預測的原因。



Explain Your Model with LIME. Compare SHAP and LIME | by Chris Kuo/Dr. Dataman

多模態可解釋分析 實際案例-SHAP

➢ **結構化資料：金融風控中的 SHAP 解釋**

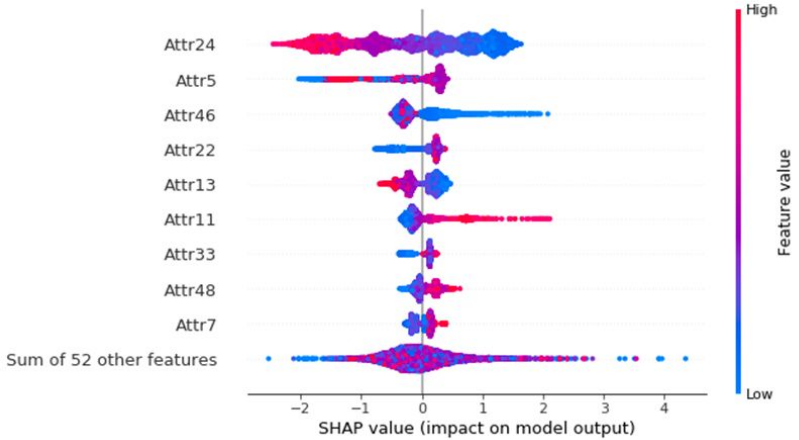
在信用風險管理中，銀行常使用機器學習模型（如 XGBoost）來預測借款人的違約風險。然而，這些模型的黑箱特性使得解釋預測結果變得困難。

- **可解釋性方法：SHAP（SHapley Additive exPlanations）**

  SHAP 基於合作博弈論中的 Shapley 值，計算每個特徵對模型預測的貢獻度 。

- SHAP 提供了整體特徵重要性排序，幫助銀行了解哪些變數（如收入、負債比、信用分數）對模型預測影響最大。

- 對於特定申請人，SHAP 可以指出該申請人哪些特徵值（如高負債比）導致模型預測其為高風險。



[1705.07874] A Unified Approach to Interpreting Model Predictions

多模態可解釋分析 實際案例-Grad CAM

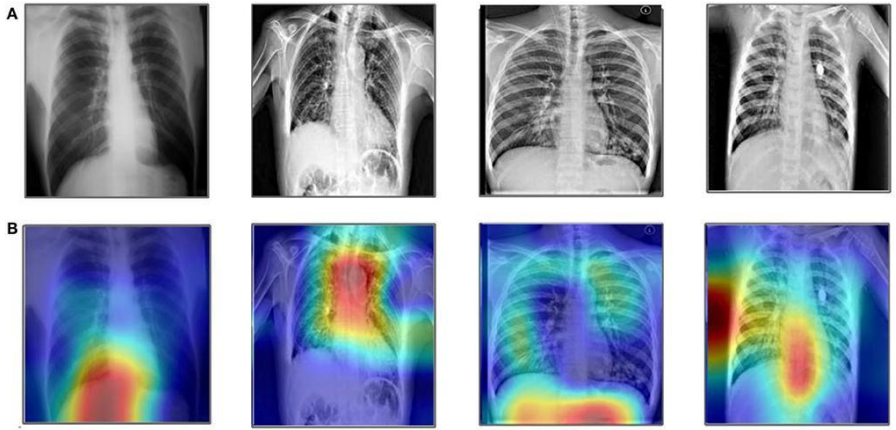
➢ **實際案例:肺炎 X 光影像診斷**

醫院使用卷積神經網絡(CNN)模型分析胸部 X 光影像，以診斷患者是否患有肺炎。為了提高診斷的透明度和可信度，醫院引入了 Grad-CAM 方法來解釋模型的決策依據。

- **可解釋性方法：Grad-CAM(Gradient-weighted Class Activation Mapping)**

Grad-CAM 是一種視覺化技術，通過計算模型對特定類別的梯度，生成對應的熱力圖，顯示模型在影像中關注的區域。

- Grad-CAM 生成的熱力圖可以顯示模型在影像中關注的區域，幫助醫生確認模型是否關注了正確的肺部區域。
- 如果熱力圖顯示模型關注了非肺部區域，可能表明模型學習到了資料中的偏差，需進一步調整模型。



COVID-19 classification using chest X-ray images

多模態可解釋分析 實際案例-Grad CAM

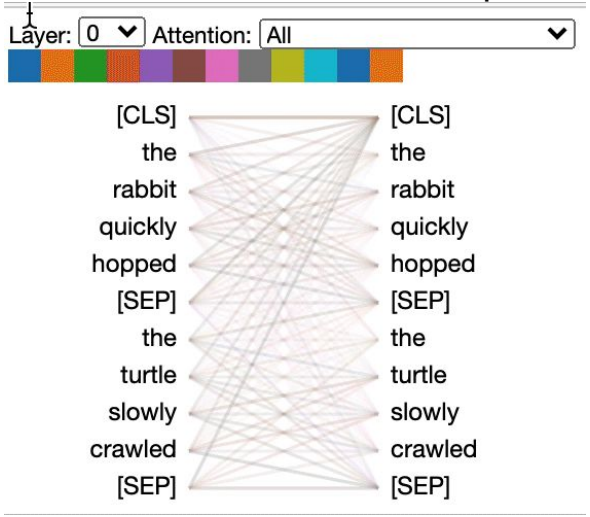
➢ **實際案例：產品評論的情感分析**

電商平台使用 BERT 模型對產品評論進行情感分析，以了解顧客對 產品的滿意度。為了提高模型的解釋性，平台引入了 Attention Heatmap 方法來可視化模型對評論中各詞語的關注程度。

- **可解釋性方法：Attention Heatmap**

  在自然語言處理中，注意力機制（Attention Mechanism）允許模型在處理每個詞語時，根據其與其他詞語的關係分配不同的權重 。注意力熱力圖（Attention Heatmap）可視化這些權重，幫助理解模型的決策過程。

- 可視化模型對評論中各詞語的注意力權重，幫助理解模型是如何做出情感分類的。
- 例如，模型可能對「優秀」、「糟糕」等情感詞給予較高的注意力權重。



 BertViz 工具

https://github.com/jessevig/bertviz

Week7 程式作業　程式碼說明

實作一個多模態分析模型,
結合金融資料與
新聞文本進行市場預測

參考程式碼連結