基於語音與文本多模態特徵之條件式潛在擴散模型應用於阿茲海默症 PET影像生成

資訊系 黃禹璇 F74122014

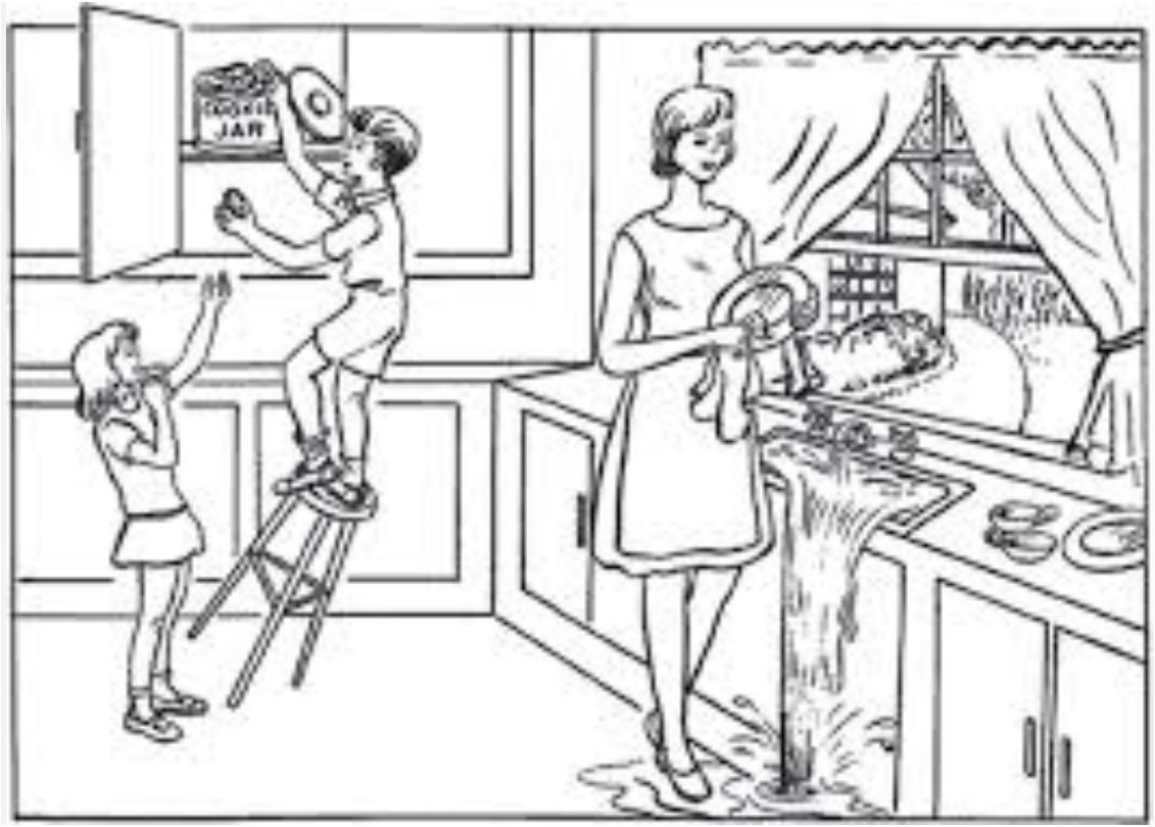
指導教授:蔡佩璇

背景與動機

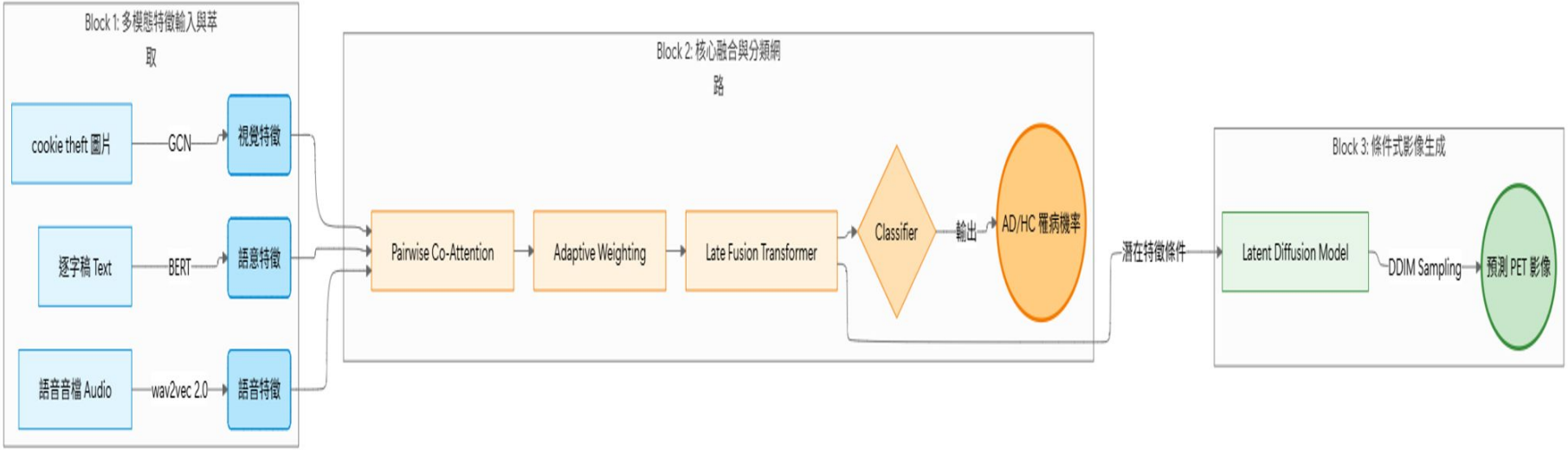
在當代阿茲海默症 (AD) 的診斷中，利用人工智慧對各種異質模態進行綜合評估的趨勢日益顯著。這些進展可支援早期疾病風險評估、個人化健康監測以及臨床決策輔助，為精準健康與智慧醫療應用奠定了基礎。在阿茲海默的診斷中，正子斷層造影(PET) 能提供關於大腦葡萄糖代謝異常或病理蛋白（如類澱粉蛋白）沉積的關鍵分子層級資訊，這對早期檢測至關重要，但高昂的檢查成本、設備限制以及放射性藥物製備的門檻，導致PET影像數據相對匱乏。

相對而言，語音訊號具備非侵入性、易於取得的優勢，且具有反映神經認知功能的特性，使其成為臨床環境中極具潛力的輔助診斷模態。因此，本研究旨在開發一項跨模態生成技術，利用來自「圖片描述任務」(偷餅乾任務 Cookie Theft task）的語音數據，結合年齡與性別等人口統計變數，來合成相應的腦部PET 影像，期盼此方法能實作醫療數據補全與擴充，解決高價值醫療模態短缺的問題。

cookie theft task



架構圖


Block 1: 多模態特徵輸入與萃取
cookie theft 圖片
GCN
視覺特徵
逐字稿 Text
BERT
語意特徵
語音音檔 Audio
wav2vec 2.0
語音特徵
Block 2: 核心融合與分類網路
Pairwise Co-Attention
Adaptive Weighting
Late Fusion Transformer
Classifier
輸出
AD/HC 罹病機率
潛在特徵條件
Block 3: 條件式影像生成
Latent Diffusion Model
DDIM Sampling
預測 PET 影像

多模態特徵輸入與萃取

- **語音模態 (Audio)**：使用預訓練的 wav2vec 2.0 提取聲學特徵 。為了處理語音與文本 Token 長度差異過大的問題（語音 Token 平均比文本多 25.8 倍），作者採用了降採樣策略（Downsampling），將每 10 個 Token 取平均，保留語音動態的同時提升對齊效率 。
- **文本模態 (Text)**：透過 Whisper 將語音轉錄為文本，並使用 BERT 進行編碼，捕捉語義與上下文資訊。
- **圖模態 (Graph / Image-Text)**：這也是本研究的一大亮點。作者不直接單獨輸入影像，而是利用 VLM（如 CLIP，BLIP，BLIP-2）建立影像與文本的**二分圖 (Bipartite Graph)** 。圖的節點分為影像節點與文本節點，邊的權重則由影像與文本的餘弦相似度決定 。接著，透過 3 層的圖卷積網路（GCN）進行處理。

多模態融合

為了讓不同模態之間能 夠進行細粒度的交互，模型採用了兩階段的注意力機制 ：

- **自注意力 (Self-Attention)**：語音與文本特徵先各自通過獨立的自注意力層，強化單一模態內的結構與上下文訊號。
- **成對協同注意力 (Pairwise Co-attention)**：讓語音、文本、圖特徵進行兩兩交叉的 Multi-Head Attention，使得模型可以動態學習跨模態的雙向依賴關係 。
- **自適應權重整合 (Adaptive Weighting)**：為了避免融合過程中喪失原本單一模態的獨特性，模型將「原始特徵」與「協同注意力特徵」進行加權總和。權重透過 Softmax 確保總和為 1 。
- **串接 (Concat)：** 將上述加權總合後的 128維特徵與年齡和性別特徵串接，通過幾層由 `Linear + ReLU` 組成的多層感知機（MLP），作為後面LDM的condition向量。

LDM條件式生成影像

模型並不是直接生成影像，它首先依賴一個預訓練的自編碼器 (Autoencoder, AE)：

- **編碼 (Encoder)：** 將真實的輸入影像壓縮成低維度的潛在特徵表示 (Latent Representation)，擴散模型的加噪與去噪過程都是在這個低維度空間中進行的 。
- **解碼 (Decoder)：** 等到 U-Net 在潛在空間完成去噪，生成了條件化的潛在變數 (Conditioned Latent) 後，再交由解碼器將其還原回像素空間的 PET 影像 。

融合「文本提示 (Text Prompt)」與「結構遮罩 (Structure Mask)」這兩種條件 ：

- **病患特徵路徑** ：前頁所提到的特徵會透過 **交叉注意力機制 (Cross-Attention)** 注入到主 U-Net 的各個層級中 。
- **結構路徑** ：為了確保生成的影像符合大腦的解剖結構，模型引入了一個輔助的 **引導網路 (Guidance Network)** 來處理二值化的邊緣遮罩 (Canny Edge)，這個引導網路會複製主 U-Net 凍結的編碼器與中間區塊的權重。

生成成果及未來改進方向

- 配合阿茲海默症專業的醫生對生成 PET的品質進行評估
- 嘗試其他不同的模型來生成患者的病理特徵