邊緣共情與全生命週期記憶：

打造真正「懂你」的個人化 AI 好友

111652042 顏友君

一、 AI 的未來能力：從「模擬情感」到「真正理解」的深度共情

當前的 AI 雖然能流暢對話，但本質上僅是基於通用數據模擬情感的「陌生人」，缺乏對個體的深層理解。我認為未來 20 年最具意義的突破，是 AI 發展出具備「全生命週期記憶」與「個人化心智理論」的伴侶型智能。它不再是無狀態的機器，而是能紀錄並理解使用者數十年的生活軌跡，讀懂基於過往共同經歷的「潛台詞」，像是出現負面情緒時，它能根據長期記憶判斷這是求救訊號，採取主動且策略性的關懷。這種從「被動回應」進化到「主動理解」的能力，將成為高齡化與原子化社會中，維護人類心理健康最重要的防線。

二、 所需的成分與資源：小數據深度與邊緣運算

1. **資料 (Data)**：縱向多模態日誌

   核心在於蒐集「小而深」的高度隱私個人數據，涵蓋數十年的對話紀錄與聲學特徵。標註方式採用「隱式回饋」，即追蹤使用者在互動後的焦慮緩解程度或作息變化，作為訓練模型的 Ground Truth。

2. **工具 (Tools)**：視覺化記憶壓縮技術

   為突破邊緣裝置處理 20 年長記憶的算力瓶頸，能夠引入「文字即圖像 (Text-as-Image)」技術[1]，將長篇歷史文本渲染為圖像輸入多模態模型，能將解碼器所需的 Token 數量減少近一半。此方法在不犧牲理解準確率的前提下，大幅提升了本地端處理長上下文的效率，確保隱私數據無需上傳雲端。

3. **學習架構 (Learning Setup)**：個人化 RLHF

   建立「個人化獎勵模型」，透過長期互動讓 AI 動態演化，學會針對特定使用者的最佳策略(例如發現對該用戶而言，安靜陪伴的獎勵值高於言語安慰)，而非套用通用的社交腳本。

三、涉及的機器學習類型：混合架構

1. **監督式學習 (建立認知)**：這是 AI 理解世界的地基。正如 Toy Model 所示，模型需透過標註資料，學習將「指派作業」映射為「焦慮」、將「動漫更新」映射為「興奮」，建立 AI 對事件與情緒關聯的基礎判斷能力。
2. **人類回饋強化學習 (RLHF) (演化策略)**： 這是教導 AI「如何應對」的關鍵。監督式學習僅能識別情緒，而 RLHF 利用使用者的「隱性回饋」(如打字速度回穩、語氣放鬆) 作為獎勵訊號 (Reward)。透過持續的試錯與權重修正，AI 能動態調整策略，從單純的分類器演化為真正能「讀懂」並安撫你的好友。

四、Toy Model：個人化情緒情境檢索 (PCER) 模型

1. 問題設計：

我們將目標降維成一個「基於歷史事件的情緒分類任務」。為凸顯個人化，設定使用者為「討厭作業、熱愛動漫的社恐數學系學生」，模型接收生活事件描述(Input)，預測該使用者的特定情緒反應(Output)。利用 Gemini 生成 600 筆符合此人設的英文資料集，涵蓋 10 種情緒類別，並採 8:2 比例切分訓練與測試集，以驗證個人化數據在情感預測中的關鍵作用。

2. 模型選擇與實作：

考量到資料量稀疏，深度學習模型容易 Overfitting，因此我們選擇了多項式樸素貝氏分類器 (Multinomial Naive Bayes)：

為了從文字輸入中推斷使用者的情緒，我們將此問題形式化為一個機率分類問題。給定一個歷史事件描述 E(Event)，目標是找出最可能的情緒類別 C_{map}(Maximum A Posteriori Emotion)。

根據**貝氏定理 (Bayes' Theorem)**，後驗機率可表示為：

$$P(C|E) = \frac{P(E|C) \cdot P(C)}{P(E)}$$

其中：

- $P(C|E)$：**後驗機率 (Posterior)**，即在看到事件 E 發生的情況下，使用者產生情緒 C 的機率。

- $P(C)$：**先驗機率 (Prior)**，即該情緒在使用者生活中出現的基礎頻率。
- $P(E|C)$：**似然性 (Likelihood)**，即如果使用者處於情緒 C，描述該事件的文字 E 出現的機率。
- $P(E)$：**證據機率 (Evidence)**，對所有類別皆為常數，在比較時可忽略。

由於事件文字 E 是由一連串特徵 $w_1, w_2, \ldots, w_n$ 組成，為了簡化計算，我們引入「樸素 (Naive)」假設，即假設各個特徵之間相互獨立。因此，似然性可以展開為各個特徵機率的乘積：

$$P(E|C) \approx \prod_{i=1}^{n} P(w_i|C)$$

最終，模型的分類決策函數為：

$$\hat{y} = \text{argmax}_{k\in\{1,\ldots,K\}} \left(\log P(C_k) + \sum_{i=1}^{n} \log P(w_i|C_k) \right)$$

這個公式精確地解釋了為何模型能捕捉「個人化特徵」。

- $\boldsymbol{P(w_i|C_k)}$ **的學習**：模型透過訓練資料學會了特定關鍵字與情緒的連結。例如，對於我們設定的數學系學生，$P(proof|\text{Anxious})$ 的數值會極高，而對於一般人，$P(proof|\text{Neutral})$ 可能較高。
- **決策過程**：當輸入一句話時，模型會累加句子中每個字的「情緒權重」。即使句子裡有雜訊，只要關鍵特徵(如 "deadline", "cat")的條件機率夠顯著，就能主導最終的分類結果 $\hat{y}$。

3. 結果分析與困難探討：

經過訓練與測試，模型展現了以下性能：

這個結果雖然顯示出模型抓住了大方向(Top-3 很高)，但準確度仍有待加強(Top-1 偏低)。利用這個結果分析我們距離終極目標的真實差距：

成功的部份：模型成功捕捉到了強烈的個人化特徵。例如，只要出現 "cat" 或 "anime"，模型能準確預測為正面情緒；出現 "deadline" 或 "social"，則準確預測為負面。Top-3 高達 84.2% 的準確率，證明模型已經掌握了使用者的情緒光譜，只是在細微差別上還是容易混淆。

困難與瓶頸：Top-1 只有 59.2%，我認為主要受限於兩個問題，

(1). 語意斷層：樸素貝氏將單字視為獨立特徵，無法理解語境。例如輸入 "I finished the difficult proof easily"，模型看到 "difficult proof" 這個通常連結到焦慮的關鍵字，就直接判斷為負面，而忽略了 "finished" 與 "easily"。

(2). 類別重疊：在我們設定的 10 種情緒中，`Anxious`(焦慮)、`Annoyed`(煩躁)與 `Meltdown`(崩潰)的界線對模型來說過於模糊，它們共享類似的觸發詞(如 "Professor", "Code")，導致模型在決策時容易搖擺。

4. 未來解決方案：

這個 Toy Model 的侷限性，正好指出了未來 20 年技術發展的關鍵，

(1). 引入語意理解：必須從統計關鍵字進化到 Transformer 架構，利用預訓練的詞嵌入來理解上下文與否定語氣。

(2). 檢索增強生成 (RAG)：為了解決資料稀疏問題，未來的系統不應只是分類器，而應結合向量資料庫。當遇到新事件時，AI 應去檢索「過去發生類似事件時，用戶的反應是什麼?」，這種基於記憶的檢索機制將能大幅提升預測的準確度與個人化程度。

五、結論

透過這次的 Toy Model 的實作，我們發現了要打造共情 AI 的關鍵，不在於模型有多大，而在於它對使用者個人歷史理解有多深，透過結合全生命週期記憶、邊緣運算與強化學習，我相信在未來一定能夠研發出每個人專屬的 AI 好友、讓無處傾訴的情緒都能夠被接住。

參考資料:

[1] Text or Pixels? It Takes Half: On the Token Efficiency of Visual Text Inputs in Multimodal LLMs(Yanhong Li, Zixuan Lan, Jiawei Zhou, 2025, cs.CL)

這篇報告以及程式碼皆有使用 Gemini 輔助生成