perplexity

論文總結報告：《Titans: Learning to Memorize at Test Time》

一、論文主要目標

本論文旨在解決現有序列建模（如Transformer和線性遞歸模型）在長序列記憶與推理能力上的瓶頸，提出一種能於測試時學習並記憶的神經長期記憶模組（Neural Long-term Memory Module, LMM），並基於此設計全新架構Titans。該架構結合短期與長期記憶，能高效處理超長序列，並於多種任務（語言建模、常識推理、基因組學、時間序列等）中超越現有主流模型[1]。

二、相關研究背景

- **Transformer與Attention**：Transformer以attention為核心，能精確建模序列中各token間的依賴關係，但計算複雜度隨序列長度呈二次增長，難以擴展至超長序列[1]。
- **線性Transformer與線性遞歸模型**：如RetNet、Mamba、DeltaNet等，透過kernel替代softmax降低複雜度，雖提升效率，但因記憶壓縮至固定向量/矩陣，長序列信息易遺失[1]。
- **記憶模組**：從Hopfield網絡、LSTM到現代Transformer，記憶設計一直是核心議題。近期如Gated DeltaNet、Longhorn等模型，嘗試引入更複雜的記憶更新規則與遺忘機制，但多數仍僅考慮瞬時驚訝（surprise），缺乏對token流的全局建模，且遺忘機制有限[1]。
- **測試時學習與快速權重程序**：如TTT-layer、MNM等，強調模型於測試時根據新數據自適應調整，但記憶管理與表達能力仍有提升空間[1]。

三、論文採用的方法與架構

- **神經長期記憶模組（LMM）**：設計一個可於測試時動態學習、記憶與遺忘的神經網絡模組，能將過往序列抽象壓縮進參數中，並根據「驚訝度」動態更新記憶[1]。
- **三分支Titans架構**：
  - **Core（核心短期記憶）**：負責當前數據流處理，類似於有限窗口的attention。
  - **Long-term Memory（長期記憶）**：即LMM，負責存儲與檢索遠距離過去信息。
  - **Persistent Memory（持久記憶）**：一組可學習但與數據無關的參數，存儲任務知識[1]。
- **多種記憶整合方式**：提出三種Titans變體，分別將記憶模組作為context、layer或gated branch融入整體架構[1]。
- **高效並行訓練演算法**：將LMM的訓練過程張量化，利用mini-batch梯度下降、動量與權重衰減，實現高效並行運算[1]。

四、主要數學公式詳解與隱喻說明

1. **Transformer Attention公式**

$$y_i = \sum_{j=1}^{N} \frac{\exp(Q_i K_j^\top / \sqrt{d_{in}})}{\sum_{l=1}^{N} \exp(Q_i K_l^\top / \sqrt{d_{in}})} V_j$$

- **隱喻**：就像在一場會議中，每個人（token）根據與其他人的關聯度（query和key的相似度）分配注意力，然後彙總大家的意見（value）。
- **變數說明**：
  - Q, K, V：分別為query、key、value矩陣，均由輸入x經線性變換得到。
  - d_{in}：輸入維度，用於歸一化。
  - y_i：第i個token的最終輸出[1]。

2. **線性Attention公式**

$$y_i = \frac{\phi(Q_i) \sum_{j=1}^{N} \phi(K_j) V_j}{\phi(Q_i) \sum_{l=1}^{N} \phi(K_l)}$$

- **隱喻**：像是預先計算好所有人的意見總和，然後每個人根據自己的特點（經kernel變換後的query）加權獲取這些信息，減少重複計算。
- **變數說明**：
  - $\phi(\cdot)$：kernel函數，將原始特徵映射到新空間。
  - 其他同上[1]。

3. **神經長期記憶的驚訝度更新公式**

$$M_t = M_{t-1} - \theta_t \nabla l(M_{t-1}; x_t)$$

- **隱喻**：像是筆記本，遇到特別驚訝的事件時（梯度大），就會特別記下來，並根據這個驚訝程度調整記憶。
- **變數說明**：
  - M_t：當前記憶參數。
  - θ_t：學習率，控制更新幅度。
  - $l(\cdot)$：損失函數，衡量記憶與實際輸入的差距[1]。

4. **Momentum（動量）式驚訝度累積**

$$S_t = \eta_t S_{t-1} - \theta_t \nabla l(M_{t-1}; x_t)$$
$$M_t = M_{t-1} + S_t$$

- **隱喻**：像是記憶的慣性，過去的驚訝會影響現在的記憶更新，避免只記住一時的突發事件。
- **變數說明**：
  - S_t：累積的驚訝度（動量）。
  - η_t：控制過去驚訝的衰減程度[1]。

5. **遺忘機制（weight decay/gating）**

$$M_t = (1 - \alpha_t) M_{t-1} + S_t$$

- **隱喻**：像是大腦會有選擇地遺忘不重要的記憶，α_t決定遺忘多少過去的信息。
- **變數說明**：
  - α_t：遺忘門控，範圍[1]，0代表完全保留，1代表完全清除[1]。

6. **記憶檢索公式**

$$y_t = M^*(q_t)$$

- **隱喻**：像是用查詢（query）去翻閱筆記本，找到對應的記憶內容。

- **變數說明**：
  - M^*：記憶模組的前向傳播（不更新權重）。
  - q_t：查詢向量，由輸入經線性變換得到[1]。

五、結論與貢獻

- Titans架構通過結合短期、長期與持久記憶，能在超長序列下高效訓練與推理，並於多項任務上超越現有Transformer及線性遞歸模型。
- 神經長期記憶模組引入動量式驚訝度、遺忘機制與深度記憶結構，顯著提升記憶表達能力與信息管理效率[1]。
- 所有設計均可高效並行實現，具備良好可擴展性，為未來大規模序列建模提供新範式[1]。

參考來源[1] 論文全文《Titans: Learning to Memorize at Test Time》

⁂

1. https://ppl-ai-file-upload.s3.amazonaws.com/web/direct-files/31472917/ed4236e7-cbce-469e-874c-91d5051f6e3e/Titans_Architecture_v1.pdf