perplexity

Titans: Learning to Memorize at Test Time

簡報大綱

1. 研究背景與動機

- 序列建模的挑戰：Transformer雖精確建模依賴關係，但計算複雜度高，難以處理超長序列[1]。
- 線性遞歸模型（如RetNet、Mamba等）雖提升效率，但長序列記憶壓縮導致信息遺失[1]。
- 現有記憶模組的不足：多數僅考慮瞬時驚訝，缺乏全局信息流建模與有效遺忘機制[1]。
- 人腦啟發：短期、長期、持久記憶協同，促進有效學習[1]。

2. 相關研究回顧

- Transformer與Attention機制[1]
- 線性Transformer/線性遞歸模型與記憶壓縮[1]
- Hopfield網絡、LSTM、DeltaNet、Gated DeltaNet、Longhorn等記憶設計[1]
- 測試時學習（Test-Time Training）、快速權重程序（Fast Weight Programs）[1]

3. 論文目標

- 設計一種能於測試時動態學習、記憶與遺忘的神經長期記憶模組（LMM）[1]。
- 提出Titans架構，結合短期、長期、持久記憶，提升超長序列的建模能力與推理表現[1]。

4. Titans架構設計與方法

- **神經長期記憶模組（LMM）**：以驚訝度（surprise）為核心，動態更新記憶，並引入動量與遺忘門控[1]。
- **三分支架構**：
  - Core（短期記憶）：負責當前數據流處理
  - Long-term Memory（長期記憶）：存儲遠距離過去信息
  - Persistent Memory（持久記憶）：任務知識儲存
- **三種整合方式**：Context、Layer、Gated Branch
- **高效並行訓練**：張量化mini-batch梯度下降，動量與權重衰減[1]

5. 主要數學公式與直觀解釋

- **(1) Transformer Attention公式**
  - 會議隱喻：每個token根據關聯度分配注意力，彙總意見[1]
- **(2) 線性Attention公式**
  - 預先計算總和，每人根據特點加權獲取信息[1]
- **(3) 記憶更新（驚訝度）**
  - 筆記本隱喻：遇到驚訝事件就記下來，根據驚訝程度調整[1]
- **(4) 動量式驚訝累積**
  - 慣性隱喻：過去驚訝影響現在記憶更新，避免只記住一時事件[1]
- **(5) 遺忘機制**
  - 大腦選擇性遺忘不重要記憶，靈活調整保留/清除[1]
- **(6) 記憶檢索**
  - 查詢筆記本，找到對應記憶內容[1]

6. 實驗與結果

- 多領域評測：語言建模、常識推理、基因組學、時間序列等[1]
- Titans在超長序列下明顯優於主流Transformer與線性遞歸模型[1]
- Titans可擴展至2M以上context window，且效率與準確率兼具[1]

7. 結論與貢獻

- Titans架構有效結合短期、長期、持久記憶，突破超長序列建模瓶頸[1]
- 神經長期記憶模組具備動量、遺忘、深度結構，顯著提升記憶管理能力[1]
- 提供高效並行計算方案，為大規模序列建模提供新範式[1]

8. 未來展望

- 深化神經記憶模組結構設計
- 拓展至更多應用場景與任務
- 開源與社群合作推動技術落地[1]

附註：每個數學公式及變數解釋可於簡報中以圖解與生活化隱喻輔助說明，提升理解與互動效果。

⁂

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