学際大規模情報基盤共同利用・共同研究拠点公募型共同研究2025年度採択課題
jh250029

マルチモーダル基盤モデルのための選択的忘却に関する研究

入江　豪（東京理科大学）　　相澤　清晴（東京大学／東京理科大学）

概要

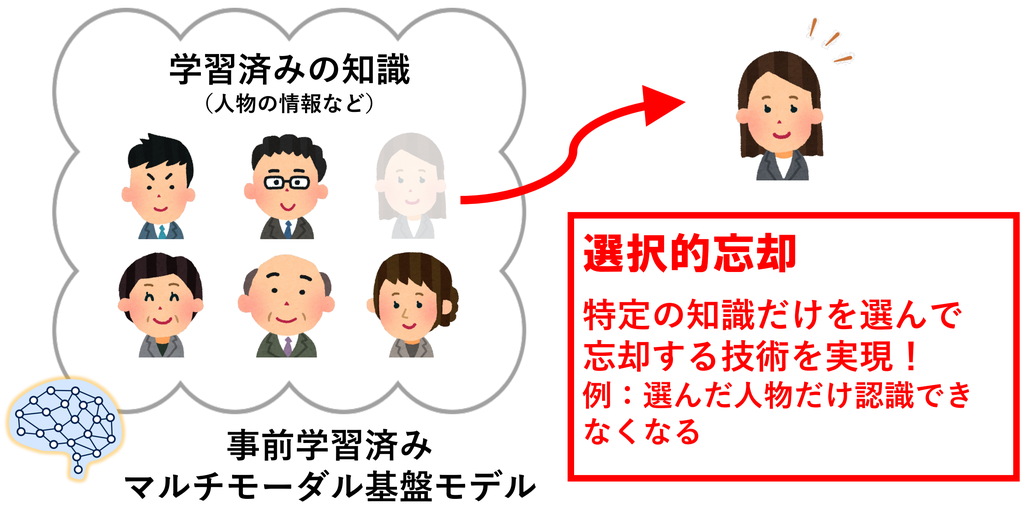
マルチモーダル基盤モデルの選択的忘却技術を実現する

マルチモーダル基盤モデル

- 言語・画像・音声など、複数のモダリティの情報を統合的に学習することによって、高い汎用性・タスク解決能力を獲得した汎用AIモデル（例：GPT-4）
- **現代AIの中核**：多くの研究機関や企業などで**研究開発・社会実装が活発化**

選択的忘却

- 学習済みのモデルから**不要な知識だけを選択的に削除**（必要な知識だけを維持）する技術
- 広義のマシンアンラーニングの一種



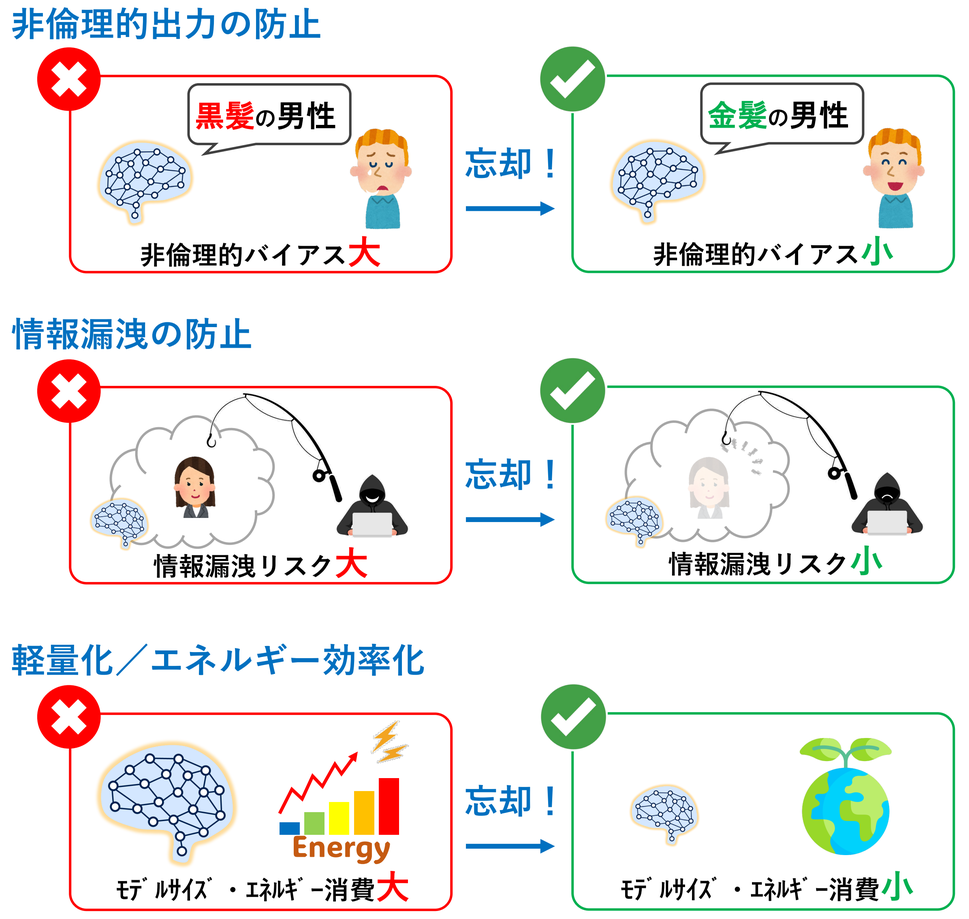
背景と狙い

背景：マルチモーダル基盤モデルの社会的リスクの顕在化

非倫理的バイアス
- 学習データに潜むバイアスに起因して、モデルが倫理的に望ましくない結果を出力してしまう事例が頻発（例：「金髪」の男性の髪色を「黒髪」と誤認する）
- 利用者に不利益を招き、不適切な倫理観を助長する可能性

情報漏洩
- 学習済みモデルから学習データを復元・窃取する技術が存在（例：モデル反転攻撃）
- 膨大・多様なデータで学習されているがゆえに影響はより深刻

膨大な運用コスト
- モデル規模の大きさによる計算資源／エネルギー消費量が甚大
- 運用性・適用性・持続可能性を阻害



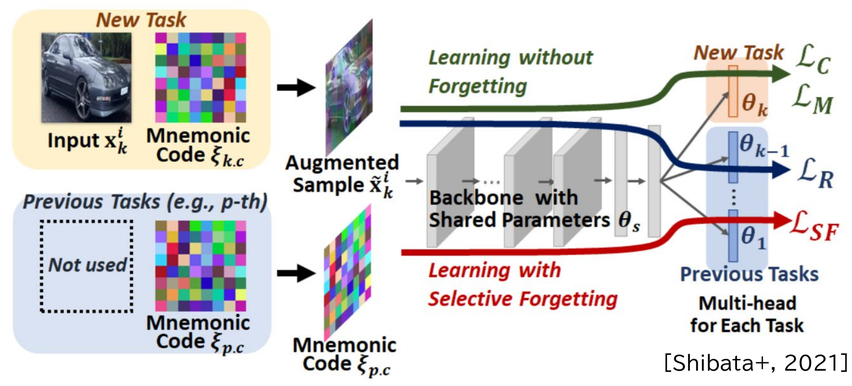
本研究の狙い：選択的忘却による解決

上記リスクは**モデルが不要な知識を過剰に保持している**ことに由来
⇒ **包括的な解決**を目指して選択的忘却の創出に挑戦

実現目標

① 深いブラックボックスモデルに対する選択的忘却の実現

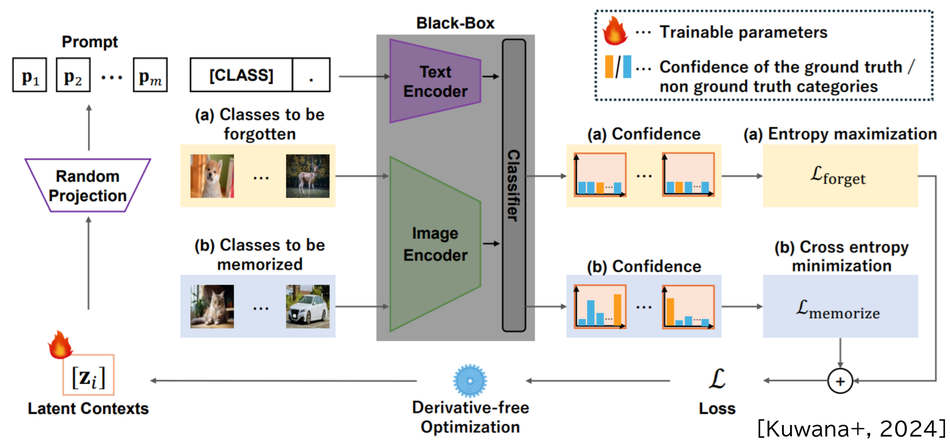
- マルチモーダル基盤モデルはブラックボックス（詳細が非公開）であるものも多い
  ⇒ **ブラックボックスモデルにも適用可能**な選択的忘却法が必要
- 従来法はトークン埋め込みの最適化により忘却を実現
  ⇒ トークナイザにもアクセスできない**深いブラックボックスモデルには適用不可能**
- テキストプロンプトそのものを離散最適化する新たな選択的忘却法を創出
  特に、精度と効率を両立する定式化・解法を探求



[Shibata+, 2021]

② マルチモーダル画像生成モデルに対する選択的忘却の実現

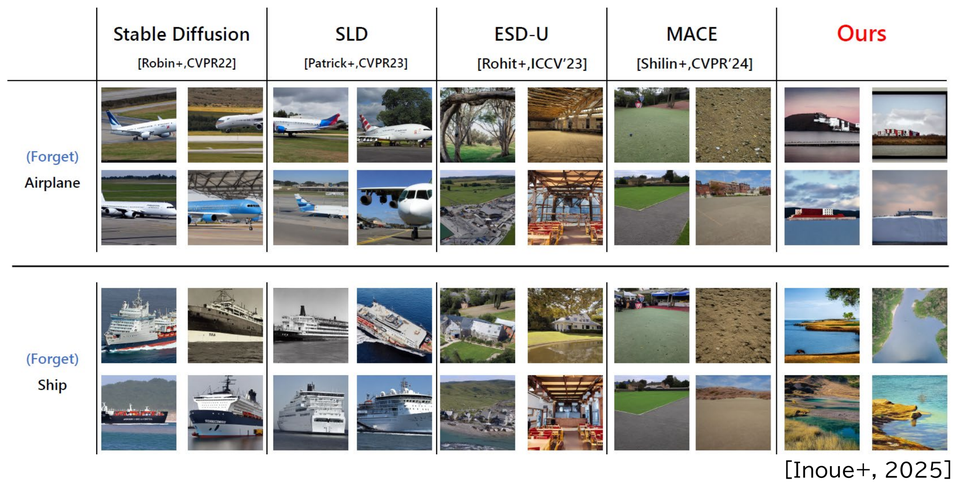
- 画像生成モデルの普及とともに著作権侵害や不適切なコンテンツの生成が問題に
  ⇒ **画像生成モデルに適用可能**な選択的忘却法が必要
- 従来法は生成モデル自体のチューニングにより忘却を実現
  ⇒ **コストが高い＆生成モデルにアクセスできない場合には利用不可能**
- コンディショナー（入力テキストエンコーダ）のみをチューニングする軽量な方法を創出
  特に、画像の空間的特徴を捉えることのできるチューニング方式を探求



[Kuwana+, 2024]

③ マルチモーダル大規模言語モデルに対する評価ベンチマークの整備

- マルチモーダル大規模言語モデルの選択的忘却については**評価プロトコル自体が未確立**
- 汎用性や広範な応用シナリオをカバーした評価ベンチマークを設計
  特に、広範な応用シナリオにおいて求められる**基本性能要件を規定**し、
  現状の技術の選択的忘却性能と課題を明らかにする評価ベンチマークを構築



[Inoue+, 2025]