審 査 の 結 果 の 要 旨

氏　名　　孫 昱偉

本論文は，「Localized Learning and Generalization in Artificial Neural Networks with Properties of the Global Workspace（グローバルワークスペースの特性を持つ人工ニューラルネットワークにおける局所学習と汎化）」と題し，Artificial Intelligence（AI）における分布外問題に対して汎化性を向上させる局所学習を提案および評価するもので，英文で書かれ，全体で9章からなる．

本論文は，複数のニューラルモジュールで構成されるAIシステムにおいて知識の再利用性とモデル全体の汎化性を向上させることを目的とし，分散したニューラルモジュールを題材に，帰納バイアスと局所学習について，ニューロサイエンスにおけるグローバルワークスペース(GW)理論の機能群を参照しつつ，提案している．これは，大規模な実展開において知能は集中モデルとしての処理としては働かないこと，新しいタスクに対しては非同期かつローカルなシナプス更新という生理学的な原理に基づくこと，生涯にわたり学習を行うこと，からなる．本研究では連想トランスフォーマを提案し，帰納バイアスとの関係についても研究している他，Bidirectional Contrastive Split Learning（BiCSL）の提案では，そのようなニューラルネットワークにおける信頼性とセキュリティについても研究している．これらによって実世界を簡潔に表現できるAIシステムが構築され，新しいタスクへの効率的な適用を可能にし，数々の知識との干渉を最小化することが可能になる．

第 1 章の「Introduction（序論)」では，現代のニューラルネットワーク・アーキテクチャを取り上げ，数学的な定式化とその背景を整理している．局所学習を定義し，大域的最適化手法や大規模基盤モデルとの比較を行っている．そして本研究の基礎となるグルーバル・ワークスペース(GW)理論の重要な機能群を定義している．

第 2 章の「Central Hypotheses and Motivation（中心仮説と動機)」では，GW機能を備えた人口ニューラルネットワークにおける局所学習に関する重要な仮説を打ち出している．ニューロサイエンスにおける仮説と工学的実現可能性について言及している．分布シフト問題についても取り上げ，意味論における構成性と帰納バイアスがどのようにこれを解決するかについて述べている．AIを分散化させる場合に起こりうるセキュリティについて整理されている．

第 3 章の「Generalization and Transfer Learning（汎化と転移学習)」では，ニューラルモジュールがネットワーク的に分散された状況下での連合学習として，複数ソースドメイン適合を可能にするFederated Knowledge Alignment（FedKA）を提案している．複数のエキスパートモジュールによる連合学習において知識転移を行う仕組みを作成し，協調的にドメインを汎化させる手法となっている．これにより複数のエキスパートモジュールにより構成される局所学習と帰納バイアスが大局的なモデルの汎化に貢献することが示されている．

第 4 章の「Learning to Learn with Reusable Neural Modules（再利用ニューラルモジュール学習への学習)」では，完全分散化された状況下でのデータ多様性問題について取り上げ，複数のエキスパートモジュール間での Markov chain による Homogeneous Learning を提案している．強化学習による選択ポリシー最適化の手法を取り入れることで多様なデータ環境下で，局所学習と帰納バイアスにより大局的なモデルの汎化ができることを示している．

第 5 章の「Priors, Attractors, and Inductive Biases（事前確率，誘導，帰納バイアス)」では，局所学習の仕組みをトランスフォーマのような基盤モデルに適用させ，その汎化性向上を目指した連想

トランスフォーマを提案している．これは，GWの仕組みに触発された連想記憶が有効なスパーストランスフォーマで，事前に得た知識を再利用して，選択的にアテンションを誘導するように事前確率を学習するものとなっている．そして，ボトルネックすなわちスパースアテンションの仕組みをGWでのワーキングメモリと対比している．ニューラルモジュール間での共有ワークスペースが異なる知識エキスパート間での競争を誘発することを示し，大規模学習におけるスケーラビリティ，特に，分布外汎化性，計算と通信に関する効率性について評価を行っている．

第 6 章から第 8 章は，局所学習の安全性と堅牢性についての研究となっている．第 6 章の「Model Poisoning in Federated Learning（連合学習におけるモデル汚染)」では，最適化されたターゲットクラスを特徴量空間内で探索して悪性パラメータを流し込むことにより，連合学習におけるモデル汚染のリスクを明らかにする Attacking Distance-aware Attack（ADA）を提案している．画像分類タスクにおいてADAの攻撃インパクトと堅牢性を従来の連合学習での対策手法ともに評価している．

第 7 章の「Trojan Attack in Multi-Modal Learning（マルチモーダル学習におけるトロイ攻撃)」では，局所学習後のモデルに対して適用可能な，デュアルモダリティでの敵対的学習によるインスタンス単位のトロイ攻撃を提案している．この手法に関して，データ数に関する効率性，ステルス性，堅牢性を評価し，局所学習されたモデルに対しても確実に数少ない汚染データ数で攻撃が成立することを示している．

第 8 章の「Robust Learning with Local Supervision（局所監督による堅牢学習)」では，Bidirectional Contrastive Split Learning (BiCSL)を提案している．これは，分散したクライアントによる分散データ全体に対するマルチモーダルモデルを自己教師あり学習で訓練するものとなっている．デュアルキー・バックドア攻撃に対するBiCSLの堅牢性を従来の集中計算モデルと比較評価し，BICSLが分散マルチモーダル学習において敵対的攻撃を受けたときにレジリエンスを持つことを確認している．

第 9 章の「Conclusions（結論)」は，本研究のインパクトについてまとめ，生涯にわたる学習，局所更新，機械記憶などの将来的な研究課題について述べている．特に，GWにおける長期記憶と意識に関する研究をAIの研究に持ち込み，局所学習を通して行われる生涯にわたる学習の可能性について言及している．

以上を要するに，本論文は，ニューロサイエンスにおけるグローバルワークスペース理論をもとに着想を得て，実世界展開で必ず課題となる分布外問題に対して，分散ニューラルモジュールに基づく新しいAIシステムを提案し，事前知識の再利用と新規タスクへの効率的な適応が安全にできることを示したものであり，電子情報学の今後の発展に寄与・貢献するところが少なくない．

よって，本論文は，博士(情報理工学)の学位請求論文として合格と認められる．