論 文 の 内 容 の 要 旨

論文題目　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）

氏　名　孫昱偉

This thesis investigates localized and contextual learning to tackle the out-of-distribution (OOD) problem, which is central to a model's capability to adapt to previously unseen samples for lifelong learning. This thesis aims to understand the knowledge reuse capability and the emergent generality within a collection of neural modules. This study is closely related to brain mechanisms of attention and memory, by incorporating the properties of Global Workspace (GW) in the prefrontal cortex and long-term memory (LTM) in the Hippocampus. The objective is to construct intelligent systems that possess concise representations of the world, enabling them to acquire new tasks more efficiently, with reduced processing time and minimized interference among various areas of knowledge.

To this end, I first study knowledge transfer among a collection of expert models that can only observe partial environments in a federated learning setting. I demonstrate that generalization can be achieved through the coordination of localized models without global objectives. Building upon this observation, I propose a novel federated domain generalization method for learning a global model by distilling domain-invariant knowledge from various localized models.

Another approach to localized learning I proposed is the Markov chain-based Homogeneous Learning, where a meta-observer aims to learn an efficient communication policy of individual models. To determine whether such localized learning can also enhance the generality of foundational models like Transformers, I introduce a novel Associative Transformer that learns distinct priors to guide selective attentions and reuse knowledge from previous observations based on associative memory-based replay. Importantly, the sparse attention with a bottleneck and the memory replay can find resonance with the working memory in the GW and the LTM in the Hippocampus, respectively. The consolidated implementation of the GW and LTM based on neural networks has demonstrated improved model performance and interpretability across a wide spectrum of tasks, compared to various existing

Transformer architectures. Finally, I investigate and reveal potential risks associated with the localized learning methods that achieve emergent behavior of generalization through inter-module representation learning. Overall, this thesis proposes a novel approach using modular and reusable neural knowledge to tackle the OOD problem based on sparse attention and memory study in cognitive neuroscience. I deeply believe this study will make a substantial contribution to our comprehension of general intelligence in humans and mammals, as well as the development of intelligent machines with the potential for lifelong learning and long-term memory in the near future.