Parametric and Non-Parametric Memory of Language Models

Name: Yuto Nishida

Laboratory's name: Natural Language Processing Laboratory

Supervisor's name: Taro Watanabe

Abstract (should be within 1st page)

Language models use previously observed linguistic patterns and factual associations to make predictions across tasks. In standard neural language models, such memory is represented implicitly in model parameters. Retrieval-augmented language models additionally use external data sources as non-parametric memory during inference. Analyzing these forms of memory is essential for understanding the reliability and limitations of language models, because the influence of memory on prediction can depend on conditions beyond the mere presence of stored patterns. In particular, it can depend on the properties of the data from which memory is formed or retrieved, and on the way relevant items are expressed in model inputs. This dissertation analyzes parametric and non-parametric memory in language models from these two viewpoints. It examines (1) how non-parametric memory behaves for low-frequency target token prediction in k-nearest-neighbor language models, and (2) how parametric memory behaves under surface-form variation in entity-based factual question answering.

First, we examine whether non-parametric memory actually supports low-frequency target token prediction, with the aim of reassessing the assumption that retrieval-based memory benefits long-tail phenomena. The results suggest that explicit memory does not necessarily improve predictions for rare targets, and that its benefits depend on how target tokens are distributed and retrieved. Second, we examine factual prediction under entity surface-form variation, with the aim of clarifying how evaluations based on a single canonical name reflect parametric memory. The results suggest that factual memorization is not fully invariant to input expressions. Together, these studies show that language model memory should be analyzed under the conditions that shape prediction.