Transformers with memory

Mikhail Burtsev

The talk will cover development of memory-augmented Transformer architectures started from Memory Transformer which introduced special [mem] tokens to store global context. Building upon this, the Recurrent Memory Transformer (RMT) introduced segment-level recurrence, enabling linear scaling of inference operations and constant memory requirements for inputs up to two million tokens. The most recent advancement is the Associative Recurrent Memory Transformer (ARMT), which enhances RMT with an associative memory mechanism for improved capacity. ARMT has set a record by demonstrating reasonable question answering performance on over 50 million tokens.