

【招待講演】

自然言語処理における深層学習の省メモリ化と高速化

進藤 裕之

奈良先端科学技術大学院大学

深層学習を用いて言語解析や生成を行う場合には、大規模な語彙に起因するメモリ消費や計算速度の問題がある。

例えば、モデルの入力層では、単語埋め込みベクトルが GPU メモリに収まらないことや、出力層では、ソフトマックス関数の計算コストが非常に大きくなってしまふことがある。本講演では、これらの問題に対処し得る技術について紹介し、実際のテキスト解析による実験結果について報告する。