文書分類タスクにおけるディープラーニングの学習プロセス可視化に関する研究

井上 達郎

近年, 深層学習と呼ばれる機械学習の手法が, 幅広い分野で活用されている. 例えば, 画像処理や自然言語処理の分野では，基盤となる分類タスクに機械学習を適用し，従来の手法よりも高い精度を達成している.

このように幅広い分野で活用が進む一方，深層学習の内部モデルは複雑であり，高パフォーマンスを達成できている理由や，性能改善のために何をすべきなのかなど，より良いモデルに改良するための手がかりを得ることが難しいという課題がある.

このような背景から, モデルの出力の根拠を説明しようという研究が盛んに行われている. これまでは, 学習済みのモデルが, 入力のどの部位に着目して出力しているか可視化する研究が多く行われている. これらの手法では, モデルが注目している入力部位を可視化することはできるが, 直接, 誤分類の原因を特定することはできない.

そこで本研究では, 畳み込みニューラルネットワークを用いた文書分類タスクにおいて, モデルの学習過程での入力に対する注目度変化を分析する手法を提案する. これにより, 学習による注目度が不安定な単語, すなわち, 誤分類の原因と考えられる単語を特定し, より少ない学習データセットで高い精度を達成することを目指す.

具体的には, まずデータセットを前半学習用, 後半学習用, 評価用の 3 つに分割する. 次に, 前半学習用データセットを用いて, モデルの学習を行う. この時, 誤差逆伝搬法により算出される, 入力の損失関数に対する勾配値を用いて, モデルが学習によって, 単語ごとの注目度をどのように変化させたか分析する. そして, 学習過程での注目度変化の激しい単語を, うまく学習できていない単語だと仮定し, そのような単語を多く含むデータと, 含まないデータに, 後半学習用のデータセットを分割して, 後半の学習を行う.

実験の結果, 注目度変化の激しい単語を多く含むデータを使って後半の学習を行った場合は, 精度 0.763 だったのに対し, 注目度変化の小さいデータを使った場合は, 精度 0.746 であった. このことから, 本研究の手法を用いることで, モデルの性能向上に寄与する, 質の高い学習データセットの特徴を特定できる可能性が示唆された.

（指導教員　佐藤 哲司）