

1. はじめに

ニューラルネットワークにおける枝刈りは、重要度の低いニューロンを削除することで、モデルサイズを軽量化する手法である。ニューラルネットワークの 1 つである Recurrent Neural Network (RNN) は、再帰的構造によって系列性を学習する。従来の枝刈り手法の多くは非系列データを対象としており、系列データに適した手法は少ない。そこで、本研究では系列性を考慮した枝刈り手法を提案する。

2. 提案手法

本研究では、RNN の系列性を考慮する枝刈り手法として、移動エントロピーによる枝刈りを提案する。また、移動エントロピーによる枝刈りの認識精度を向上する手法として、L2 正則化を併用することを提案する。

2.1. 移動エントロピーを用いた枝刈り手法

移動エントロピー [1] は情報量の一種であり、情報の流れを抽出することができる。RNN における時刻 t の出力を x_t 、時刻 $t+1$ の出力を x_{t+1} 、学習データの時刻 t における正解ラベルの認識精度を y_t とすると、移動エントロピー $T_{Y \rightarrow X}$ は式 (1) により算出できる。

$$T_{Y \rightarrow X} = \sum_{x_{t+1}, x_t \in X} \sum_{y_t \in Y} p(x_{t+1}, x_t, y_t) \log \frac{p(x_{t+1}|x_t, y_t)}{p(x_{t+1}|x_t)} \quad (1)$$

$Y \rightarrow X$ は、 Y から X への因果的な影響の情報理論的な定量化を示す。つまり、2 つの系列間の情報伝播の大きさと、正解ラベルの認識精度が与える因果性を測定することで、RNN における各ユニットの系列間の前後関係を考慮した評価が可能である。図 1 に、RNN と移動エントロピーの対応関係を示す。そして、算出した移動エントロピーを基に昇順、または降順でパラメータを削除する。

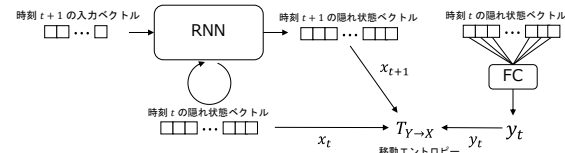


図 1: RNN と移動エントロピーの対応関係

2.2. L2 正則化の併用

従来の枝刈り手法の多くは、L1 正則化を併用することで冗長な説明変数を 0 付近に誘導し、枝刈りによる認識精度への影響を抑制する。しかし、0 付近に誘導されなかったパラメータは認識精度への寄与率が高く、枝刈りされた場合に大きな精度劣化を起こしやすい。本研究ではユニット単位で枝刈りをするため、0 付近に誘導されたパラメータと、それ以外のパラメータが混同する。そのため、L1 正則化によって十分な認識精度を維持できないと考える。そこで、L2 正則化を併用する枝刈りを提案する。

L2 正則化は、パラメータ全体の大きさを抑制することで過学習を防ぐ正則化手法である。L2 正則化は式 (2) で表される。ここで、 L はネットワークの損失、 f は交差エントロピーなどの損失関数、 w はパラメータを表す。 λ はハイパーパラメータである。

$$L = f(x) + \lambda \|w\|_2^2 \quad (2)$$

L2 正則化を適用したモデルは説明変数を多く持つことから、過剰パラメータ表現を引き起こしやすい。その結果、ユニット単位で枝刈りをする場合、L1 正則化よりも高い認識精度を維持することが期待できる。

3. 評価実験

本実験では、枝刈りしたモデルの認識精度を比較する。認識精度の比較は、(I) 移動エントロピーの有効性、(II) L2 正則化を併用する効果の検証を目的とする。(I) は系列性

を考慮する移動エントロピーと、系列性を考慮しない指標 (エントロピーと L1 ノルム) を用いて枝刈りを行い、認識精度を比較する。(II) は、L1 正則化を用いてパラメータをスパースにした場合と、L2 正則化を用いてパラメータ全体の大きさを抑制した場合で認識精度を比較する。

3.1. 実験条件

本実験では、ImageNet で事前学習された ResNet-18 を特徴抽出器として利用し、1024 次元のユニットを持つ RNN 層が 1 層、全結合層が 2 層のモデルを構成する。RNN 層は Long Short Term Memory (LSTM) を用いる。データセットは、51 種類の人の動作について収録した HMDB-51 と、101 種類の人の動作について収録した UCF101 を使用する。画像サイズを 240×320 ピクセル、動画のフレーム数を 25 フレームとする。また、学習データは全体の 8 割とし、残りを評価用データとする。

3.2. 実験結果

移動エントロピーの有効性: 図 2 に、LSTM に対して枝刈りを行ったときの認識精度の推移を示す。横軸はユニットの削除率、縦軸は認識精度である。実線は指標における降順、破線は昇順による枝刈りを表す。HMDB51 と UCF101 における、枝刈り前の認識精度は 89.4% と 98.3% であった。そこから、移動エントロピーの降順による枝刈りを用いてユニット数を 50% 削除したとき、HMDB51 と UCF101 の認識精度は 89.6% と 98.3% であり、比較指標の中で最も高い認識精度を確認できた。

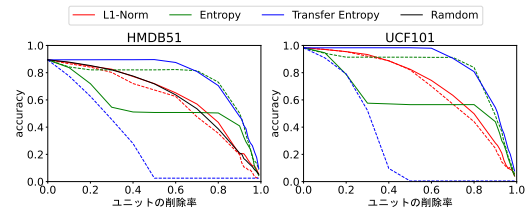


図 2: 認識精度の推移

L2 正則化を併用する効果: 図 3 に、L1 正則化と L2 正則化を適用した LSTM で枝刈りを行ったときの認識精度の推移を示す。枝刈り前の認識精度は 90.5% と 89.8% であり、L1 正則化を適用した場合の方が枝刈り前の認識精度は高い。移動エントロピーの降順による枝刈りを用いてユニット数を 60% 削除したとき、それぞれの認識精度は 87.8% と 89.9% であった。L1 正則化の併用で 2.7pt 認識精度が低下し、L2 正則化を併用で 0.1pt 認識精度が向上した。これは、L2 正則化によって過剰パラメータ表現が強くなり、移動エントロピーで特徴量の変動を抑えやすくなったためである。一方、移動エントロピーを除く指標は特徴量やパラメータの大きさのみを考慮するため、認識精度が低下した。

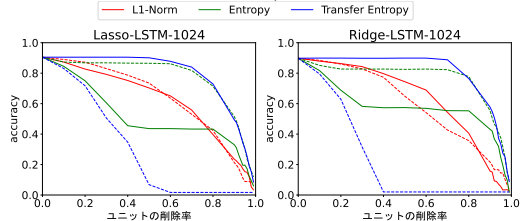


図 3: 正則化認識精度の推移

4. おわりに

本研究では、系列性を考慮する移動エントロピーによる枝刈り手法と、移動エントロピーによる枝刈りの認識精度を向上する手法として L2 正則化の併用を提案した。これらにより、従来の枝刈り手法よりも高い認識精度の維持を実現した。今後は、ユニット単位ではなく、ゲート毎等のより細かな単位で枝刈りを行い、認識精度の改善を図る。

参考文献

- [1] T. Schreiber, “Measuring information transfer”, Phys. Rev. Lett., 2000.