様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月　 ８ 日現在

機関番号：８２６２６
研究種目：若手研究
研究期間：2020～2022
課題番号：２０Ｋ１９８８８

研究課題名（和文）深層学習におけるデータ拡張の戦略的利用法の開発

研究課題名（英文）Development of Strategic Utilization Methods for Data Augmentation in Deep Learning

研究代表者

高瀬　朝海（TAKASE, Tomoumi）

国立研究開発法人産業技術総合研究所・情報・人間工学領域・研究員

研究者番号：３０８４４１６２

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究は、データ拡張が不適切に利用されることを避け、深層学習にとって有益となるようなデータ拡張の適用法を開発した。提案法であるSelf-paced augmentation法は、訓練中の損失関数の値をもとに、データ拡張を適用するサンプルおよび適用しないサンプルを動的に決定する手法である。多数のデータセットおよびニューラルネットワークを用いた実験において、提案法は汎化性能の向上を達成することができた。また、データ拡張の指標をもとに、データ拡張の手法とハイパーパラメータを探索する方法についても考案し、探索時間の短縮を実現することができた。

研究成果の学術的意義や社会的意義
データ拡張は経験や直感に基づいて利用されることが多いのが現状であるが、不適切なデータ拡張の利用はモデルの汎化性能を落とすことになる。本研究でデータ拡張をデータに応じて適切に適用する手法を開発したことは、深層学習の性能や安定性を高めることに対して大きく貢献すると期待される。研究成果は、IF付き国際誌Neurocomputingで発表された。

研究成果の概要（英文）：In this study, we developed a method that avoids the inappropriate use of data augmentation and is beneficial for deep learning. The proposed method, Self-paced augmentation, dynamically determines which samples to apply data augmentation based on the value of the loss function during training. Through experiments using numerous datasets and neural networks, the proposed method achieved improvement in generalization performance. Furthermore, we devised a method to explore data augmentation techniques and hyperparameters based on metrics for data augmentation, leading to a reduction in exploration time.

研究分野：深層学習、ニューラルネットワーク、機械学習

キーワード：データ拡張　Data augmentation　深層学習　ディープラーニング　ニューラルネットワーク　カリキュラム学習　教師あり学習

１．研究開始当初の背景

深層学習は、高い性能を発揮するために十分な数のデータを必要とするが、多くのデータを入手することのできないケースや、クラスごとに不均衡な数のデータが得られるというケースも多く存在する。そのため、深層学習において、データ数を増やすということが必要とされており、手元のデータに変形を加えて新たなデータを生成するデータ拡張技術が、数多く提案されてきている。ランダムノイズやアフィン変換に加え、様々なバリエーションをもつマスク処理や、複数のサンプルを合成して新しいサンプルを作る手法などがある。

データ拡張は古くから利用されている技術にも関わらず、一般に非線形な処理であり解析が困難であるため、その体系的な利用法の確立に取り組んだ研究は少ない。そのため、経験や直観に基づいたデータ拡張の利用が主流となっているが、不適切な利用が行なわれると、学習性能が悪化してしまう。この問題を踏まえ、深層学習モデルの汎化性能を向上させるために、データ拡張の適切な利用法を発見することが重要となってくる。

２．研究の目的

本研究の目的は、深層学習モデルの汎化性能に対するデータ拡張の影響を明らかにし、その影響に基づき、データ拡張の戦略的な利用法を考案することである。データ拡張は用いるデータに依存するので、不適切なデータが生成されると、学習に悪影響があり、汎化性能が低下してしまう。そこで本研究では、汎化性能に対するデータ拡張の影響を実験的に調べる。そして、データ拡張の影響を考慮し、データ拡張を戦略的に適用する手法を考案・検証する。

３．研究の方法

まず、深層学習に対するデータ拡張の影響の解明に取り組む。データ拡張の戦略的な適用法を開発するために、データ拡張により生成されるデータの適切性を解明する必要がある。そこで、データの適切性をニューラルネットワークの損失関数から判断する。損失関数はミニバッチ内の入力データに依存するので、得られる解もデータによって異なる。解の性質に関しては、損失関数上のシャープな解にパラメータが収束すると汎化性能が悪化するため、フラットな解に収束することが望ましいということが議論されている。そこで本課題では、フラットな解にパラメータを収束させられるように、データ拡張を最適化することによってミニバッチ内のデータ分布を調整するための方法を発見する。

これにより得られたデータ拡張と損失関数の関連性に関する知見と、深層学習の動的な学習に関する先行研究をもとに、データ拡張の最適化対象に対し、データ拡張を戦略的に適用する手法を考案し、実験により検証する。学習中にハイパーパラメータを動的に変化させる先行研究として、カリキュラム学習がある。これらは、学習の序盤では易しいデータのみを学習に利用し、徐々に難しいデータを利用するように設計されており、学習を効率的・効果的に行うことができる。この考えを応用して、本研究では、学習の序盤は易しいデータを生成し、徐々に難しいデータを生成するように、データ拡張を調整することを試みる。学習の進捗やデータの難易度は、損失関数の値を用いて判断できる。データ拡張の最適化対象として、①データ拡張を適用するデータ、②各データセットに応じたデータ拡張手法の種類、③各データ拡張手法内のパラメータの値が挙げられる。本課題ではこれらの全てを扱うが、一つ一つが大きなテーマであり、まとめて取り扱うことは難しいため、一つずつ順に取り組む。

実験では、機械学習用のベンチマークデータセットを用い、提案手法と従来手法のテスト精度と学習時間を比較し、効果を検証する。対象とする問題として、一般物体画像および数値データを用いた多クラス分類問題を扱う。訓練データ数は、ImageNet データセットなどの 1,000 万を超えるものから、100～1 万程度の少数データまで幅広く用いて検証を行う。モデルには、数値データであれば多層パーセプトロンを用い、画像データであれば ResNet などの典型的な畳み込みニューラルネットワークを用いる。

４．研究成果

本研究では、深層学習モデルの汎化性能に対するデータ拡張の影響を明らかにし、その影響に基づき、データ拡張の戦略的な利用法を考案するという目的を、概ね達成することができた。データ拡張の戦略的な利用法として、データ拡張を適用するサンプルの選択に着目し、手法開発を行った。従来の方法は、すべての訓練サンプルにデータ拡張を適用するが、本研究では、データ拡張を適用すべきサンプルつまりデータ拡張が学習に効果あると考えられるサンプルと、そうでないサンプルに分かれていると考え、前者のサンプルに対してのみ、データ拡張を適用することを考えた。また、データ拡張を適用すべきかどうかの判断は、学習中に変化するものであると考え、データ拡張の機械的かつ動的な調整に着目した。本研究では、これを実現するために Self-paced Augmentation 法を考案した。この手法は訓練中の損失関数の値をもとに、データ拡張を適用するサンプルおよび適用しないサンプルを決定する手法、サンプル難易度に応じてデータ拡張を適用することで学習を効果的に進めることができる。提案手法のアルゴリズム設計およ

びプログラム実装を行い、CIFAR-10 などの機械学習の基本的な画像データのベンチマークデータセットを用いた幅広い実験を通して、提案手法の性能を検証した。提案手法は、すべてのサンプルにデータ拡張を適用するという従来法を超える汎化性能を示す傾向がみられた。本研究の研究成果は、査読付き国際ジャーナルである Neurocomputing で発表した。

また、他の最適化対象として、データ拡張手法と各手法のハイパーパラメータの最適化について取り組んだ。従来の、訓練後に計算されるバリデーションデータの精度をもとに最適なデータ拡張を探索する方法は、多くの計算コストを必要とする。そこで、データ拡張の探索に関する関連研究を調査し、Affinity および Diversity というデータ拡張の指標が、汎化性能をうまく表すことができたという研究結果に着目した。本研究では、これらの指標を考慮に入れた新しい指標を提案し、これを利用することで、探索に必要な訓練ステップ数を大幅に減らすことができることがわかった。機械学習ベンチマークデータセットを用いた実験を行い、Affinity および Diversity を考慮した指標に基づいてデータ拡張の手法およびハイパーパラメータを探索することで、バリデーションデータの精度を用いる方法よりも、最適化の時間を大きく短縮することができ、さらに最適なデータ拡張をより精度よく選び出すことができるという結果が得られた。

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Tomoumi Takase, Ryo Karakida, Hideki Asoh	４．巻 442
２．論文標題 Self-paced data augmentation for training neural networks	５．発行年 2021年
３．雑誌名 Neurocomputing	６．最初と最後の頁 296~306
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.neucom.2021.02.080	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関