5ZG-06

縮小画像による研究成果物分類手法

福本小夏† 佐野雅彦‡
徳島大学†

1. 背景

大学などの研究室では，日々研究活動を行っている．研究活動をする中で論文や発表スライド，研究データなどの様々な研究成果物が生成され続けている．研究成果は，学会における公開された成果物と研究室に蓄積された研究成果物は，新しい研究を進める際に活用され，研究内容の早期や共有のために再利用される．

しかし，ファイル名や日時などのメタ情報だけでは，膨大に蓄積されている研究成果物から参考にしたい研究成果物を探すことは困難である．

よって，研究に利用する研究成果物を再利用・分析を行いやすい蓄積方法が求められる．しかし，膨大な量の研究成果物を人力で分類する作業には研究者に対して時間と負担を要することになる．そこで本研究では，研究成果物を縮小画像に変換したものを学習して，自動的に分類する手法を提案する．

2. 関連研究

学術論文等の研究成果物に焦点を当てた分類手法に関する研究が存在する．柏木らは，アブストラクトを用いた論文分類システムを実装し，効率的に必要な論文を収集した[1]．榊らは，類似度による論文ネットワークを構築し，論文のカテゴリ分類を提案している[2]．これらの研究はいずれも研究に使用する調査論文の取得に期待できる．しかし，研究成果物の中でも論文に特化した分類手法であるため，発表スライドや実験データといった詳細な種類までは分類されておらず，分類する成果物の範囲を広げる必要がある．

3. 提案手法

本研究では，研究成果物である論文と発表スライドを PDF に変換し，画像を取得した後，32×32 ピクセルのグレースケール画像に縮小し機械学習して分類する手法を提案する．

3.1. 研究成果物分類手法

本手法では，研究成果物の分類において教師あり学習の機械学習を用いる．機械学習で利用する学習データには，研究室内の論文と発表スライドを使用する．学習データは画像に変換し，画像学習として進める．学習機器には，畳み込みニューラルネットワーク(CNN)で学習したモデルを利用する[3]．CNN は,画像認識分野を中心に利用されている深層学習の一つであり，図 1 に示すように入力層(Input layer)，畳み込み層(Convolution layer)，プーリング層(Pooling layer)，全結合層(Fully connected layer)，出力層(Output layer)から構成される．本手法では，畳み込み層とプーリング層の繰り返しを最小の層に構成することで画像データの局所性を捉えることを保ちながら冗長な情報を排除し,過学習を防ぐためにドロップアウト層(Dropout layer)を構成した．

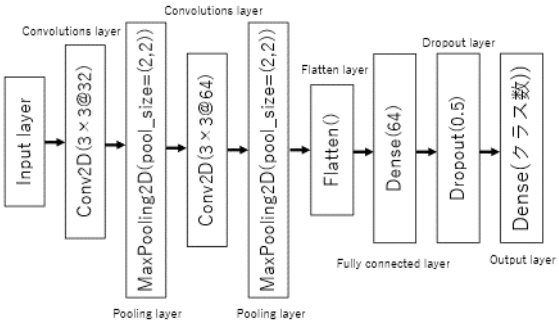

図 1 畳み込みニューラルネットワークの構成

3.2. 研究成果物分類における前処理

前節で述べた学習データをグレースケールに変換し，画像サイズを 32×32 のサイズに統一する．画像サイズを縮小することで文字や図形の情報量削減することにより，学習時間と推論時間を短縮する．この前処理により，文字情報としては失われるがレイアウト情報は保持される．

3.3. 学習

CNN の実装には Python 言語にて Google 製の機械学習ライブラリである TensorFlow を用いた．最適化アルゴリズムには AdaGrad[4]を採用し，学習率は,0.00921 で学習率の減衰率は,10^{-6} とした．

また，学習する際のフレームワークは，バッチサイズ 32，エポック数 55 とした．学習には NVIDIA GeForce RTX 3090 を用い，学習にかかる時間は約 11 秒である．

4. 評価実験

作成した画像分類器のテストデータに対する

Classification method of research artifacts using reduced images
Konatsu FUKUMOTO†, Masahiko SANO†
†Tokushima University

評価と各カテゴリのデータのモデル推論を行う．本実験では，画像分類器を学習データによって学習し，テストデータにおける分類器の評価値を算出する．分類器の評価値は，正解率，適合率，再現率，F 値，正解率として算出する．評価値は，プログラムの実行 5 回の平均を評価値とした．学習データ数とテストデータ数の内訳を表 1 に示す．

次に各カテゴリに当てはまるデータへのモデル推論を行う．推論には,学術論文や学内のスライドを使用した．

表 1 学習データ数とテストデータ数の内訳

カテゴリ名	学習データ数	テストデータ数
論文	198	66
スライド	183	61

4.1. 学習結果

5 回に分けて実行した評価値の平均を表 2 に示す．

表 2 学習した分類器の評価値平均

分類カテゴリ	適合率	再現率	F 値
スライド	0.98	0.68	0.79
論文	0.78	0.98	0.87
正解率	0.84		

4.2. 推論結果

画像分類器で各カテゴリに当てはまるデータの推論結果を表 3 に示す．

表 3 学習した分類器の推論結果

推論データ	データ数	正解数	不正解数	正解率
スライド	1613	1613	0	1.00
論文	328	288	40	0.88

5. 考察

表 2 に示した結果から，スライドの適合率は 0.98 と高く，モデルがスライドと正しく予測したもののうち，実際にスライドと認識したものがほとんどであることが分かる．再現率は 0.68 と低いことは，スライドの実際の割合に比べてモデルが検出できていないデータが多いことを示しているため，スライドが論文に分類され特定のパターンがモデルにとって難しいことが考えられる．

論文の適合率は，0.78 と論文と予測できたものが多いことを示している．再現率は,0.98 と高い結果となり，モデルにとって明確で区別しやすい特徴を持っているといえる．

F 値は，論文の評価が高く，全体的な正解率は,0.84 となった．モデルは全体的に分類できているがスライドの検出において改善の余地がある．

表 3 の結果から，スライドは正しく分類されており，正解率は 1.00 となった．スライド評価値では，スライドの再現率は低いが適合率が高いことによる結果だと考えられる．また，スライドは画像の特徴を捉えやすいため，高い性能を発揮できたと考える．論文は，一部不正解しており，正解率は 0.88 である．論文は，再現率は高いが適合率が低いことによる結果だと考えられる．また，論文には文章だけでなく図が含まれているものがあったため，改善の余地がある．

6. まとめと今後の課題

本研究では，研究成果物の分類による負担を低減するために CNN を用いた機械学習による研究成果物分類手法を提案した．機械学習で学習する画像データを 32×32 サイズに縮小することで学習時間と推論時間の短縮を試みた．提案手法の学習結果からスライドは誤分類があるが,論文は比較的高い評価となった．一方推論結果では，学習結果とは逆にスライドの方が正しく分類できている．この理由としては,論文の推論データには文章のほかに図形や写真なども載っていたことから，学習データ数の不足が考えられる．今後の課題としては，誤分類の具体的なパターンの分析によるモデルの弱点の理解と研究成果物と分類カテゴリを増やすことが挙げられる．また，ある成果物に関するページの異なる複数画像からの推論についても検討の余地がある．

謝辞

本研究は JSPS 科研費 JP18K11572 の助成を受けたものです．

参考文献

[1] 柏木裕恵，高田雅美，佐々木明，城 和貴，アブストラクトを用いた論文分類システムの設計と実装，情報処理学会研究報告数理モデル化と問題解決 (MPS)，2006(95 (2006-MPS-061))，pp33-36,2006.

[2] 榊剛志，松尾豊，石塚満，制限付きクラスタリングを用いた論文分類,人工知能学会全国大会論文集,第 20 回,2006.

[3] LeCun Y, Bottou L,Bengio Y,et al,Grandient-based learning applied to document recognition,Proc.IEEE 86,pp2278-2324,1998.

[4] Duch,John, Elad Hazan, Yoram Singer,Adaptive subgradient methods for online learning and stochastic optimization, Journal of machine learning research 12.7,2011.