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

平成　２９　年　　６　月　１５　日現在

機関番号：　１３６０１
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３０２４９

研究課題名（和文）異種情報を統合する階層型特徴表現モデルの学習とその応用に関する研究

研究課題名（英文）learning of multimodal representation and its application

研究代表者
丸山　稔（MARUYAMA, Minoru）

信州大学・学術研究院工学系・教授

研究者番号：８０２８３２３２

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：本研究の目的は異種情報を統合する情報表現手法を確立し、それを画像検索結果の言語による修正などに応用することである。本研究では主として画像とそれを記述したテキストに関する情報統合を対象とした。まず、DBMを用いた情報統合を行い、情報統合層を特徴に用いることにより、CNNによる特徴抽出よりも画像識別精度を向上できることを示した。またLSTMとCNNによる統合モデルを用いた場合の画像検索結果のテキストによる修正のための類似度尺度を考案し、その有効性を示した。これらの処理の効率化のためにはCNNの処理時間の短縮が必要になる。そこで、CNNの圧縮方法を考案し、手法の有効性を示した。

研究成果の概要（英文）：Research on multimodal learning and its application to image search has been carried out. We have studied the DBM to jointly represent the image and corresponding text. We showed feature vector obtained from the joint layer could give rise to better classification results than CNN-based features. We also studied image query method based on multimodal representation which is enabled by using visual-semantic embedding model based on CNN and LSTM. It allows us to perform analogical reasoning over images by specifying properties to be added and subtracted by words. We introduced a novel similarity measure based on the difference between additive and subtractive query. Our methods strongly depend upon the CNNs for image processing. To reduce the computational cost of the image processing by the CNN, we examined a method for compressing the given CNN. We proposed the compression method based on the block-wise distillation and examined its effectiveness.

研究分野：　知能情報学

キーワード：　画像認識　機械学習　深層学習　画像検索　ニューラルネットワーク

１．研究開始当初の背景

我々人間は、画像、文書、音楽などの多様な情報を理解する高度な認識能力を有すると共に、画像とそれを表現する文章などのように全く異なる情報形態であってもこれらの内容に基づいて対応付け、認識することができる柔軟性を有している。各種の認識を行う際には、それぞれの情報から特徴量を抽出して認識を行うことが通常であるが、これらの特徴量は個別の情報形態に即したものに留まっており、共通の内容を表現するための統合表現などの検討は十分とは言えない。現在各種パターン認識問題に対して深層学習などの機械学習技術が応用され大きな成果を挙げているが、これらは機械学習による情報表現方式の獲得によるものであると考えられる。単一種類の情報表現にとどまらず、異種情報を統合する情報表現手法を機械学習に基づいて確立できれば、例えば画像や音楽などの検索結果を言語（テキスト）を用いて修正したり、画像や音楽などを記述する文章を合成するなどの応用が考えられ、実用上の意義は大きいものと考えられる。

２．研究の目的

本研究においては異種情報を統合するための基本方式の確立と、統合表現の例題からの学習方法とその高度化、また得られた統合表現の応用に関する研究を目的とする。本研究における情報種類としては、画像とそれを記述する文章（テキスト）を用い、これらの情報統合を対象として研究を実施した。

３．研究の方法

画像とテキストという異種情報を統合する特徴表現モデルとしては、画像とテキストを与えて統合表現を得るだけではなく、単一の情報を与えたときにそれに対応する別表現の特徴を生成できる能力を持つ生成モデル（generative model）型の DBM（Deep Boltzmann Machine）を考え、異種特徴を上位の隠れ層で統合する方式を検討し、画像識別を対象として情報統合モデルの有効性を検証した。また情報統合モデルの応用例として、画像とテキストを統合したベクトル空間上で表現することにより、画像検索結果をテキストにより修正・変動させる方式を検討した。これらの研究においては画像はまず CNN を用いて特徴ベクトルに変換して用いている。全体の処理時間の短縮・実時間性の実現のためには CNN による処理速度の向上も重要であることから、CNN の構造圧縮に関する手法の検討も行い、提案手法の有効性を検証した。

４．研究成果

（１）　DBM に基づく画像情報とテキスト情報の統合学習とその効果の検証

画像情報とそれを記述したキャプション（テキスト）情報を対象として、情報統合方式とその効果に関する検討を行った。画像とテキスト情報の統合のための構造としては DBM（deep Boltzmann machine）を用いた。情報統合用のネットワークは、画像特徴を処理する３層構造の Gaussian-RBM（restricted Boltzmann machine）部、テキスト情報を処理する Replicated softmax RBM 部及びこれらを統合する最終層から構成されている。各情報の処理を行う２つの隠れ層はそれぞれ 2048、1024 ユニットから成り、最終層は 2048 ユニットから構成されている。図１に DBM の構造を示す。

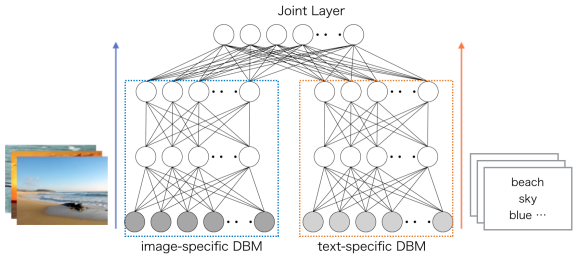

図１　画像とテキスト統合のための DBM 構造

画像部の入力としては CNN（convolutional neural network）により抽出された特徴量（4096 次元）を用いた。CNN は画像識別のために ImageNet データベースを用いて学習済の CaffeNet を用いた。テキスト部の入力としては BoW 表現（bag of words、4096 次元）を用いた。stop words の除去や stemming 処理などの前処理は行っている。このような統合ネットワークの学習には MIR-FLICKR データセットを用いた。学習には 975K の画像及び画像に付随するタグとメタデータを用いた。

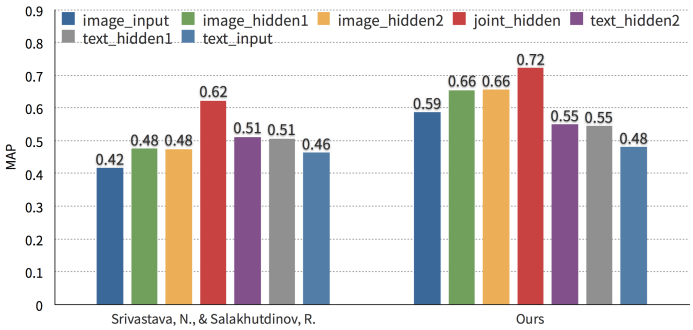

図２　情報統合ネットワークの隠れ層出力を用いた識別実験結果

画像-テキスト統合表現獲得の効果を検証するために画像のカテゴリ識別を対象として実験を行った。情報統合ネットワークを構成する各層の状態を特徴量とした場合のカテゴリ識別性能の比較を行った結果を図２に示す。識別に際しては logistic regression モデルを用いた。画像入力部（image-input）の識別性能が CNN の識別能力に相当すること

になる。実験結果の比較図より、情報統合層の結果は通常の CNN の識別能力を大きく上回ることが分かり、情報統合の有効性が検証できた。

（２）　画像情報とテキスト情報の統合ベクトル空間を用いた画像検索

異種情報統合による応用の一つとして画像検索を言語（テキスト）を用いて修正することが考えられる。検索結果がユーザの望む画像とは若干異なっている場合、画像に含まれているどのような内容をどのように変化させたいかという意図を言語（テキスト）により伝達することができればより柔軟な検索が可能になると考えられる。本研究ではこのような追加したい内容と削減・減少させたい内容を単語により与え、元画像にこのような変動を反映させたものに最も近い画像をデータベース中から探索するための手法について検討を行い実験によりその有効性の検証を行った。

本研究においては Kiros らが画像のキャプション自動生成に用いた画像とテキスト情報の統合モデルを用いた。これは言語処理部分では、単に単語頻度（BoW 表現）ではなく文章（テキスト）を処理できるようにするために系列データを扱う RNN（recurrent neural netwok）の一種である LSTM（long short-term memory）ネットワークを用い、最終状態における中間層の状態をテキスト情報のベクトル表現として用いるものである。画像情報についてはCNNにより得られるベクトル表現を用い、これら２つのベクトルが統合ベクトル空間中に写像されて統合表現が得られるモデルとなっている。このとき、対応する画像・文書対を学習データとして用い、統合ベクトル空間において同一の表現となるようにパラメータの学習が実行される。このモデルを用いることにより、画像または文書単体であっても統合ベクトル空間の表現を得ることができる。

図３に Kiros らのモデル概略を示す。

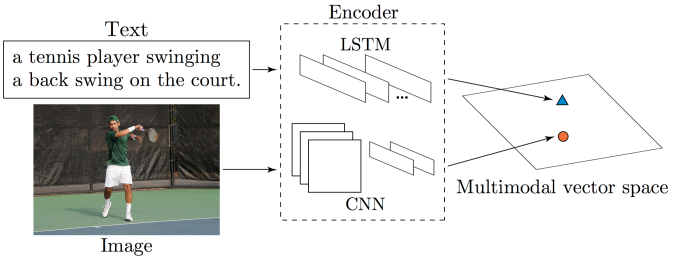

図３　CNN と LSTM の統合ベクトル空間

与えられた画像に対して追加する特性と削除する特性をテキストで与え、そのような特質に合致する画像をデータベース中から検索するタスクを考える。元となる画像、追加特性を記述する単語（文章）、削除特性を記述する単語はそれぞれ統合ベクトル空間中のベクトルとして表すことができる。これらのベクトルをそれぞれ qimg, qadd, qsub とおく。Kiros らは属性に変更を加えた画像を以下のように検索する手法を提案している

X* = arg min S(X, qimg-qsub+qadd)

ここに S(u,v)はベクトル u,v 間の cosine 類似度である。これは qimg-qsub+qadd と類似するベクトルにマップされる画像を検索する手法である。この手法の場合、望ましいベクトルは元のベクトルに変化分を加えた形で表現されているが 、従来手法では変動ベクトルの方向だけでなく長さも全て指定していることになる。しかしながら通常ユーザが与える変化方向は定性的なものであり、ユーザが与えているものは変化の方向性だけであると考えることができる。そこで本手法においては従来手法とは異なる以下の手法により類似度計算を実行する。

X*=arg min S(X-qimg, qadd-qsub)

従来手法と本手法による類似度尺度の違いを図４に示す。

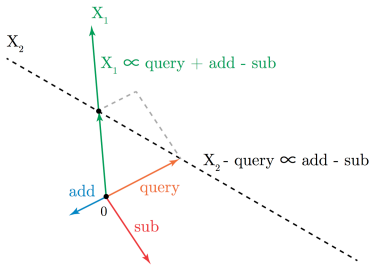

図４　追加・削除クエリによる検索概念図

図５　画像検索事例１

本手法の効果を検証するために画像検索実験を行った。実験においては Microsoft COCO

データベースを使用して統合ネットワークの学習を行った。得られた統合ベクトル空間を用いて検索の修正実験を行った結果を図５に示す。図５(a)に示すように本手法により変化が必ずしも画像特徴として顕著に表れないケースでも良好な結果が得られていることが分かる。

但し、常に従来手法より良好な結果が得られる訳ではなく、場合によっては悪化する場合も見られた。これらの例を図６に示す。

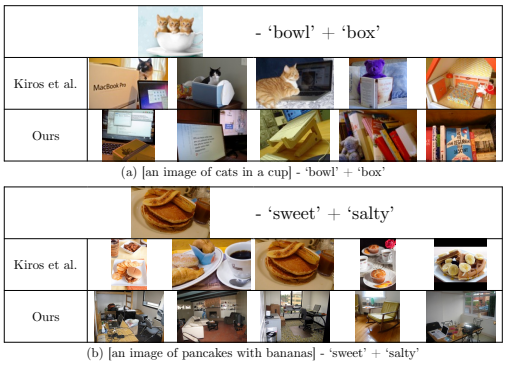

図６　画像検索事例２

さらなる精度の向上のための類似度尺度の改良やネットワークそのものの改良は今後の課題である。

（３）　画像特徴抽出用ニューラルネットワークの圧縮

本研究において、画像情報とテキスト情報を統合する際には、まず元の情報からニューラルネットワークによる特徴抽出を行い、得られた特徴ベクトルに基づいて情報統合を行っている。このため、全体の処理時間は特徴抽出に要する計算時間に大きく依存することになる。特に画像特徴抽出に用いている CNN は多数の層から構成されており計算時間を要するため、全体の計算時間を短縮するためには CNN の処理速度向上が必要となる。CNN の処理速度向上のための方策としては、CNN の精度を維持しつつ層、素子、結合重みなどを削減するネットワーク圧縮が研究されている。それらの中でも Network Prune 法は非常に高い情報圧縮率を実現できることが示されている。しかしながら、この手法は結合重み単位の圧縮手法であり、得られたネットワークモデルは多くの場合規則性が損なわれており、パラメータ数の削減がそのまま処理速度の向上につながらない問題点を有している。そこで、本研究では結合単位ではなく層単位での圧縮を行う手法について検討を行った。

本手法においてはまず画像識別を行う CNN を学習した後で、識別能力をできるだけ低下させることなく層単位の削減によりネットワークの圧縮を行うことを考える。このためにまずネットワーク全体を幾つかのブロックに分割し、各ブロック単位で既存ネットワークの入出力関係を模倣する圧縮された部分ネットワークを学習する。このためのブロック単位での模倣・学習には distillation と呼ばれる既存のモデル圧縮手法を用いる。distillation においては最終層において softmax 関数適用前の値（logit）を使用して学習を行う。本来 distillation はネットワーク全体の圧縮のために提案された手法であるが、本手法においては、distillation をネットワーク全体ではなく、ブロック化された部分ネットワークの模倣学習に適用する。distillation 適用によって圧縮された各ブロック構造を組み合わせたネットワークに対して、さらに学習を適用することにより、圧縮後のネットワークの識別精度向上を図った。

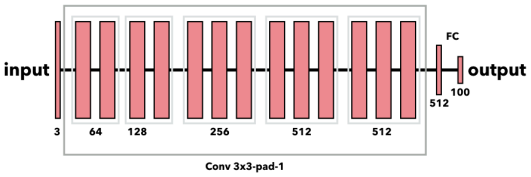

図７　VGG15 の構造概略図

本手法の有効性を検証するために画像データベースとしては 100 カテゴリの画像を有する CIFAR100、圧縮前の CNN としては VGG15（図７）を用いて実験を行った。VGG15 は VGG16 から全結合層を１つ減少させたものであり、13 個の convolution 層と２個の全結合層の合計 13 層からなるネットワークである。VGG15 を CIFAR100 を用いて学習を行った結果識別精度 63.0%を得た。VGG は５個の pooling 層を持ち、これを境界として５個のブロックに分割することができる。ブロック 1,2 は２個の層を含み、残りのブロックは３個の層を含む。これらのブロック構造に対して本手法を適用した。なお、実験においては全結合層に対する圧縮操作は行っていない。

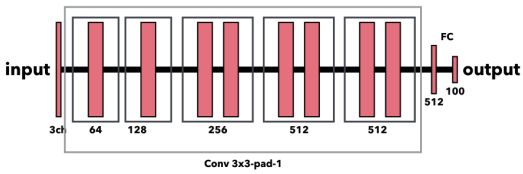

図８　層削減によるネットワーク１

表１　圧縮実験１の結果

	既存モデル	Network Prune	一から学習	Distillation	提案法
ブロック1の層数	2	2	1	1	1
ブロック2の層数	2	2	1	1	1
ブロック3の層数	3	3	2	2	2
ブロック4の層数	3	3	2	2	2
ブロック5の層数	3	3	2	2	2
Conv層の総数	13	13	8	8	8
モデルサイズ(%)	100	**10.0**	63.4	63.4	63.4
識別精度(%)	63.0	62.8	42.8	59.0	**63.1**

まず、各ブロックに対して１つずつ層を減少させる実験（圧縮実験１）を行った。表１に圧縮実験１の結果を、圧縮の結果得られる構造を図８に示す。圧縮実験の結果（層の数、識別精度等）を表１に示す。また実験の結果得られるネットワークの識別能力を学習進行過程（epoch）に関してプロットした結果を図９に示す。図において baseline は元のネットワーク、(1)モデル全体に distillation を適用した結果、(2)本手法、(3)圧縮アーキテクチャを最初から学習した結果である。

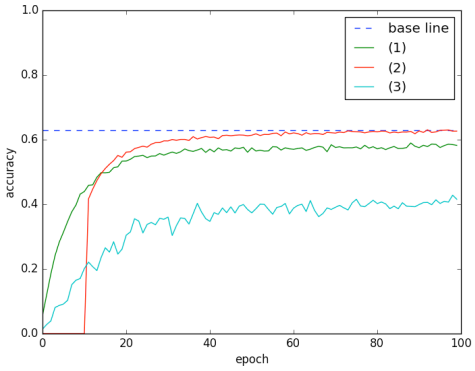

図９　学習進行過程とネットワーク性能１

以上より本手法により既存モデルとほぼ同等の制度を持ち、層数を削減した構造が得られたことが分かる。

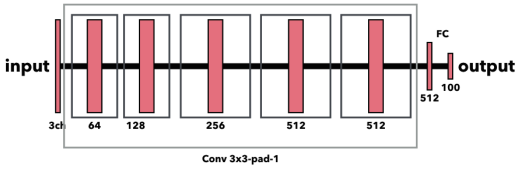

図 10 層削減によるネットワーク２

次いで、既存 CNN（baseline）の各ブロック内の層数を１にまで減少させ（図１０）同様の実験（圧縮実験２）を行った結果を表２、学習進行過程の模様を図１１に示す。前の実験と同様、ネットワーク全体をブロック化し、各部分ネットワークについて distillation によって模倣学習を行った上で全体の学習を行うことで、識別精度を維持しつつ層数を減少させることが可能であることが分かる。

表２　圧縮実験２の結果

	既存モデル	一から学習	Distillation	提案法
ブロック1の層数	2	1	1	1
ブロック2の層数	2	1	1	1
ブロック3の層数	3	1	1	1
ブロック4の層数	3	1	1	1
ブロック5の層数	3	1	1	1
Conv層の総数	13	5	5	5
モデルサイズ(%)	100	28.1	28.1	28.1
識別精度(%)	63.0	40.7	47.1	**63.0**

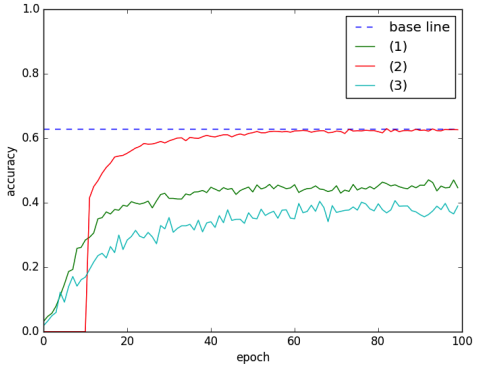

図 11 学習進行過程とネットワーク性能２

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計　８　件）

① Kosuke Ota, Keiichiro Shirai, Hidetoshi Miyao, Minoru Maruyama, Interactive image search system based on multimodal analogy, Proc. 19th International Conference on Human-Computer Interaction, 12-14 July, Vancouver Convention Centre, Vancouver, Canada (2017)（査読有）

② Kosuke Ota, Keiichiro Shirai, Hidetoshi Miyao, Minoru Maruyama, Multimodal representation learning by utilizing deep convolutional features and preprocessed texts、IEEE Shin-etsu session, 9C-4, pp.155, 10 月 8 日, 長岡技術科学大学（2016）（査読無）

③ 横島府、白井啓一郎、宮尾秀俊、丸山稔、既存深層構造に対する圧縮方法の検討、電子情報通信学会信越支部大会講演論文集、7A-2、pp. 88、10 月 8 日、長岡技術科学大学（2016）（査読無）

④ 横島府、白井啓一郎、宮尾秀俊、丸山稔、既存深層構造を用いた新規カテゴリ学習方式に関する検討、電子情報通信学会信越支部大会講演論文集、2C-2、pp. 25、10 月 3 日、新潟工科大学（2015）（査読無）

⑤ 宇都宮気伸、白井啓一郎、宮尾秀俊、丸山稔、画像の属性識別のための Convolutional Neural Networks の学習、電子情報通信学会信越支部大会講演論文集,、8D-2, pp. 116、10 月 3 日、新潟工科大学（2015）（査読無）

⑥ 宇都宮気伸、丸山稔、宮尾秀俊、画像の早期性識別のための特徴量選択、電子情報通信学会信越支部大会講演論文集、

6C-3, pp. 82、10 月 4 日、信州大学(2014)（査読無）

⑦ 福島崚平、宮尾秀俊、丸山稔、識別関数の学習に基づくピアノ音楽の自動採譜、電子情報通信学会信越支部大会講演論文集、6C-4、 pp. 83、10 月 4 日、信州大学（2014）（査読無）

⑧ 桜井翔、宮尾秀俊、丸山稔、確率的トピックモデルを用いた論文データの経年変化解析、電子情報通信学会信越支部大会講演論文集、7C-1、 pp. 96、10 月 4 日、信州大学（2014）（査読無）

〔その他〕

http://soar-rd.shinshu-u.ac.jp/profile/ja.WCnCbpkh.html

６．研究組織

(1)研究代表者

丸山　稔　（MARUYAMA, Minoru）

信州大学・学術研究院工学系・教授

研究者番号：80283232