

高次元空間での外れ値検出と教師なし分類による ワードグラウンディング技術の開発

Developing Word Grounding Technology via Unsupervised Learning and Outlier Detection in a High- Dimensional Space

2171019



研究代表者

京都大学教育学部

助 教

橋 本 敦 史

[研究の目的]

本研究では、実社会で展開するサービスにより収集されたテキストと画像（映像）の組から人手による正解データの作成なしに自動で機械学習を行うワードグラウンディング技術の開発を目指した。深層学習技術によって、機械学習の技術は飛躍的に発展し、人と機械の調和の基盤となる対話技術も大きく向上している。この発展は、ImageNet など大量の学習データの存在にも大きく支えられており、機械学習技術のさらなる発展のためには、ImageNet よりもさらに充実したデータベースの構築が求められる。しかし、人手で学習用の正解データを作成するには限界があり、実社会に存在するアプリケーションから自動で学習を行う技術への期待は大きい。

本研究では、人のモノづくりを対象とした技術開発の一環として、ユーザ投稿型レシピサイトを対象として、料理の教示テキストを元に、画像中の物体を自動的に学習する技術の構築を目指した。クックパッドなどのレシピサイトには100万単位の大量のテキストと画像が、それぞれ対応付けられて投稿されている。このデータは深層学習に十分な量であるだけでなく、ブログなどと違ってフォーマットが決まっている。一方で、料理には様々な種類があり、データの

多様性も非常に大きい。

また、これらのデータから学習を行うことができれば、人のモノづくりの作業状態について、機械に人と共通の認識を持たせることができる可能性が高い。人と機械が円滑にコミュニケーションするためには、互いの眼（機械の場合はカメラ）で観測される現在の状況に対するコンセンサスは必要不可欠なものであり、この問題を解決するための技術の一つとなり得る。

[研究の内容、成果]

テキストの中には動詞や形容詞など多様な概念が含まれるが、まずは第一歩として画像中の領域との対応関係が明確に取れる材料に対応する名詞を対象とした学習を行った。

教示テキスト中の名詞は画像中の何らかの物体領域に対応する材料ラベルの候補であると考えられる（図1）。

また画像中の物体領域の候補も、近年の一般物体検出手法などを用いることで推定すること



図1 ラベルと候補領域の取得

は可能であると考えられる。一方で、それらラベル候補と物体領域候補の対応関係は未知である。このように考えた場合、本研究の課題は、ラベルと画像領域の候補を、その対応関係を推定しながら学習する問題として定義できる。

このような問題は、従来から弱教師付き学習として取り扱われてきた。従来対象としていた典型的な応用では、ラベル候補はユーザが画像を説明するためにつけられたものであり、画像と全く関係のないラベルが存在することは想定されてこなかった。これに対して、本研究では自然言語から抽出した名詞（材料名）をラベル候補とする。特に教示文の場合、他の手順で既に述べられた食材が文中に現れない【ラベル不足】の事例が、従来よりも頻繁に生じる。また、反対に画像に全く関係のない材料名が文中に現れる【ラベル余り】は人がつけたタグを候補とする従来の弱教師付き学習の枠組みではこれまで考えられてこなかった。

このうち【ラベル不足】は多くの場合、初期の手順で名前が呼ばれた後、省略され続けることになる。これは、作業後半の加工中の材料の外見を学習する上で致命的な情報不足を引き起こす可能性が高い。このため、【ラベル不足】に対しては、自然言語処理によって、手順を遡り、過去に取り扱われた材料名を全てラベル候補とすることで対処した。

この対処方法では、明らかに【ラベル余り】が増加してしまう。この【ラベル余り】に対処するための方法として、ある材料が絶対に含まれない領域、つまり【ネガティブサンプル】の利用に着眼した。ある材料が絶対に含まれない画像は、その材料が使われない作業の画像から容易に取得することができる。従って、そのような画像からえられる物体領域候補は【ネガティブサンプル】として利用が可能となる。

ある材料が【ラベル余り】となっている状況とは、その材料が画像中の候補領域と対応付かない状況である。消去法的に考えれば、ネガティブサンプルとよく似た候補領域は、その材

料ではないと考えられる。

このような発想の元に、材料ごとにポジティブ/ネガティブの2値を判定する問題に対して、以下の2通りの学習方法を提案し、比較・評価を行った。

[手法1] 第一の手法 [2] では、ネガティブサンプルは見た目が多様で学習を進めても容易にその特徴量は類似せず、ポジティブサンプルは相対的に早く学習が進むと仮定した。この仮定のもとでは、一定量の学習を行ったネットワークモデルから得られた特徴は、ポジティブクラス同士は類似し、ネガティブクラス同士は多様な特徴を取る。従って、このモデルから得られた特徴量を互いに類似する集合とそれ以外の2つに分割することで、個別の特徴量に対する判定結果とは別に、群としての特性からポジティブ/ネガティブの判定をできると考えた。

実装としては、VGG16のネットワークに対して一定量の学習を行ったモデルにより特徴量を抽出し、外れ値をクラスタとして検出する従来研究 [1] の理論に基づいて Spectral Clustering により2クラスタに分類を行い、それぞれのクラスタをモデルにおける判定結果の平均尤度に従ってポジティブ/ネガティブに割り振った。すなわち、Spectral Clustering で得られるクラスタの一つが外れ値のみから構成される場合があることを利用し、ネガティブサンプルを外れ値クラスタとして検出した。

[手法2] 手法1では、候補領域のポジティブ/ネガティブ判定のための学習用データのうち、ネガティブサンプルをランダムに抽出した。しかし、本来は候補領域に類似するネガティブサンプルのみを利用することが望ましい。このようなネガティブサンプルを画像以外のメタ情報に基づいて収集するために、作業に用いられる材料集合の類似性に着目した。つまり、ある材料に対して、その材料が含まれる作業とよく似た材料集合からなる作業の画像からネガティブサ

ンプルを収集する、という処理を行った。また、手法1ではポジティブサンプルが、その加工状態によらず一つの大きなクラスタを形成すると仮定した。

これに対して、手法2[3]では、ポジティブサンプルがその状態に合わせて複数のクラスタを形成している可能性を考慮し、このようなクラスタの形成とポジティブ/ネガティブ判定を同時に学習できないかを考えた。具体的には、学習中に適宜クラスタリングを適用し、クラスタを構成する要素に含まれるネガティブサンプルの割合を元に、クラスタごとにラベルの信頼度を計算し、信頼度を候補領域がポジティブであるという仮説の重みとした学習を行った(図2)。

なお、各クラスタに含まれるサンプルが互いに類似するよう、Triplet Lossなどを組み合わせてより状態の違いを強調することで、仮説の通り状態ごとに材料の特徴が分かれた構造を持つ特徴空間を得ることも可能であると考えられるが、研究期間の都合上、そこまでの実装・評価には残念ながら至らなかった。

表1、表2([3]より引用)に人参・ジャガイモの2種類に対して、手法1(表中の既存手法(学習量10 epoch, 20 epochの場合)、手法2(表中の提案手法)、および、【ラベル余り】が含まれるラベルをそのまま学習に使った手法(表中のベースライン)を適用した結果を、それぞれ示す。この表において「手順番号」とは

表1 人参を対象としたポジティブ/ネガティブ判定の平均適合率

評価データの手順番号	[1-2]	[3-4]	[5-6]	[7-]	[all]
ベースライン手法	0.6850	0.6661	0.6377	0.6839	0.6718
既存手法 (epoch=10)	0.7391	0.7227	0.7288	0.7553	0.7365
既存手法 (epoch=25)	0.7242	0.7198	0.7422	0.7267	0.7267
提案手法	0.7391	0.7097	0.7311	0.7089	0.7236

表2 ジャガイモを対象としたポジティブ/ネガティブ判定の平均適合率

評価データの手順番号	[1-2]	[3-4]	[5-6]	[7-]	[all]
ベースライン手法	0.6520	0.6338	0.5704	0.5020	0.5985
既存手法 (epoch=10)	0.5558	0.6161	0.5909	0.6428	0.5981
既存手法 (epoch=25)	0.5219	0.5580	0.5863	0.6591	0.5742
提案手法	0.5247	0.5661	0.5704	0.6081	0.5631

教示文が作業の何手順目かを示しており、数字が小さいほど序盤の手順となる。

特に作業の序盤においては画像中の材料の名前が教示文中に記述されている場合が多く、ベースラインの手法が高い精度を達成している。一方で、ベースライン手法は作業後半の識別精度が低い。これに対して、手法1、手法2は作業後半の方が高くなった。教示文が既知である状況下では、このような手順番号も既知であることが大きいいため、調理の進捗に合わせてベースライン手法と提案手法を統合するなどの工夫が考えられる。しかし、そもそもポジティブ/ネガティブの二値判定問題であることを考えると、識別精度はかなり低く、残念ながら改善の余地が多く残る結果となった。

[今後の研究の方向、課題]

本研究では弱教師付き学習でこれまで扱われてこなかったラベル余りの問題に対して検討を行った。ラベル余りに対処することができれば、自然言語処理表現と画像の視覚的要素の学習を膨大なWeb上の資源を利用して学習できるようになる可能性がある。一方で、今回の手法では2クラス識別問題としては非常に低い精度しか得られなかったことは大きな問題であると考えている。これは、そもそも識別機が今回扱っ

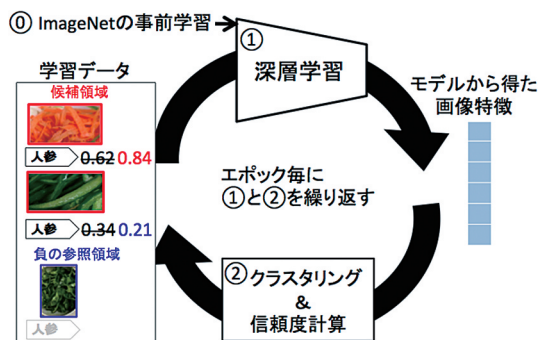


図2 逐次的な信頼度更新つき学習

た調理作業中の材料という識別対象を上手く特徴空間に射影できていないことが原因であると考えている。良い特徴空間を得るためには今回の弱教師信号以外にも，オートエンコーダの獲得や文章生成そのものを同時に学習するなど，様々なタスクを同時に学習する方法が必要であろうと考えられる。

[参考文献]

[1] Takuro Ina, Atsushi Hashimoto, Masaaki Iiyama,

Hidekazu Kasahara, Mikihiro Mori, Michihiko Minoh: Outlier Cluster Formation in Spectral Clustering, CoRR, Vol. abs/1703.01028, 2017

[成果の発表, 論文等]

- [2] Atsushi Hashimoto, Takumi Fujino, Jun Harashima, Masaaki Iiyama, Michihiko Minoh: Learning Food Appearance by a Supervision with Recipe Text, Proc. of CEA2017, pp. 39-44, 2017
- [3] 伊奈拓郎: 対象材料を含まないレシピを用いたラベルの信頼度計算による画像からの材料推定, 京都大学情報学研究科知能情報学専攻 修士論文, 2018