ショートノート

物体認識結果を利用したLDAによるシーン分類の試み[†]

池田 祐輔*・高野 敏明*

近年、ビッグデータを対象とした深層学習による手法の研究が盛んに行われている。それに伴いデータセットが必要となるが、大量のデータに対して手作業での処理となるため時間と手間がかかってしまう。この問題は、画像データセット作成におけるシーン分類においても同様に存在する。本論文では、シーン分類における問題の解決のために、人間の行うシーン分類での根拠となる画像内の物体に着目した手法を提案する。今回は試みとして、提案手法に基づき物体認識結果を利用したLDAによるシーン分類についての実験とその考察を行った。

キーワード：物体認識, LDA, シーン分類

1. はじめに

近年、ビッグデータを対象とした深層学習による手法の研究が盛んに行われている。それに伴い、機械が行う学習のためのデータセット作成が必要となる。しかしデータセット作成のための処理の多くは、手作業で行われるため多大な時間がかかってしまう。この問題は、画像データセット作成におけるシーン分類という処理においても同様である。さらにシーン分類には、この手作業の処理時間の膨大さに加え分類結果が処理を行う人間によって異なってしまう恐れがある等の複数の問題が存在する。これらの問題を解決し処理時間の削減、画一化した分類手法の確立の手段として機械による画像の自動振り分け手法が必要とされている。

本論文では、この課題に対して人間がシーン分類を行う際に利用する画像内の物体に着目した手法を提案する。ただし、本提案におけるシーン分類とは画像におけるシーンの体系をつくることを意味している。今回はその第一歩として、既存手法による物体認識結果を用いて、LDA（Latent Dirichlet Allocation）によるシーンの体系化を行った。実験では、その学習結果が人間の行ったシーン分類結果に類似しているのかについて確認を行った。

周辺研究として、画素情報に基づく局所的な特徴量を利用したニューラルネットワーク構築による画像の分類手法が存在する[1]。既存の手法における分類処理は、画像に規則的な処理を行うことで特徴量を算出しその結果を元に分類を行っていくというものである。しかしこの特徴量を根拠とした分類は人間にとって理解が困難であり、この分類処理は機械独自の根拠を元にした分類処理であると考えられる。本研究では、既存手法とは異なる視点として、人間が理解可能な根拠として物体に着目した手法を提案する。

図1 提案する分類手法の概要図

2. 物体認識結果を利用したLDAによるシーン分類の試み

人間がシーン分類を行うとき、生活における経験や知識に基づき画像内の物体の有無などによって情景を分類していると考えられる。つまり、人間は画像内の物体を認識し、それら物体の組合わせにより場所や情景を分類しているといえる。これを模倣することにより、画像のシーン分類が行えるのではないかと考えた。そこで今回提案する手法の概要図を図1に示す。具体的には、人間が画像内の物体を認識する行為については物体認識の既存手法により物体名称を抽出する行為と置き換える。また、物体の組合わせによる場所や情景の分類については、LDA（Latent Dirichlet Allocation）を利用することで、物体の有無によるシーンクラス作成のための分類が実現できると考えた。LDAは、文書の潜在的なトピックをボトムアップに生成するモデルである。本稿では、物体認識による結果（物体名称）を単語として捉え、画像1枚は1つの文書として、シーンはLDAにより生成されるトピックと考える。

本章の構成として、2.1節では既存手法であるYOLOv3[2]を用いた物体認識について、2.2節にシーン分類のためのLDAについて、それぞれ説明を行う。

2.1 物体認識

画像内の物体認識として、近年物体認識分野で有名なYOLOv3を既存の物体認識手法として利用した。YOLOv3は、物体認識における分類問題を回帰問題として扱うという手法であり、認識までの処理速度が早くその精度がそれまでの手法よりも比較的に良いという特徴を持っている既存の物体認識の手法である。YOLOv3による物体認識を行う際には、画

[†] Scene Clustering by LDA Using Object Recognition Result
Yusuke IKEDA and Toshiaki TAKANO

* 静岡理工科大学大学院理工学研究科
Graduate of Science and Engineering, Shizuoka Institute of Science and Technology

図2 LDAのグラフィカルモデル

像データセットによる学習が必要になるが、今回は、学習済みのモデルとして公開されている COCO [3] を学習したモデルを用いた。

2.2 LDA

まとめられたテキストデータを基に LDA による潜在トピックの学習を行う。LDA とは、自然言語処理分野における文書生成過程の確率的モデルの 1 つである。このモデルでは、各文書には複数の潜在トピックがあると仮定し、その潜在トピックを統計的に共起しやすい単語の集合が生成される要因と考え、確率変数で定式化されている。LDA のグラフィカルモデルを図 2 に示す。M を文書数、K をトピック数、 n_d を文書 d における総単語数、 $z_{d,i}$ を文書 d における潜在トピック、 $w_{d,i}$ を文書 d における生成された単語とする。また文書 d におけるトピック k の出現確率分布を $\theta_d = (\theta_{d,1} \cdots \theta_{d,K})$ とし、トピック k における単語 v の出現確率分布を $\phi_k = (\phi_{k,1} \cdots \phi_{k,V})$ とする。この時それぞれの確率分布の生成過程をディリクレ分布にしたがうものと仮定し α, β はそのパラメータとする。 θ_d, ϕ_k のそれぞれから多項分布にしたがい $z_{d,i}, w_{d,i}$ が生成されるというモデルになっている。

グラフィカルモデルを元に LDA では次のような各文書の生成がされる。まず、各文書 d についてディリクレ分布 $Dir(\alpha)$ にしたがった各文書におけるトピックの構成比率の多項分布 θ_d が選択される。今度はこの多項分布 θ_d にしたがった各単語ごとに潜在トピック $z_{d,i}$ が選択される。そしてこのトピック $z_{d,i}$ に対応する単語 $w_{d,i}$ が多項分布にしたがい選択される。これを文書 d に含まれる単語の数だけ行う。これは単語選択において毎回トピックの選び直しを行うため、単語毎に語彙の違いを想定した 1 文書における複数トピックの選択が可能となる。ただし、トピック $z_{d,i}$ からの単語生成については、ディリクレ分布 $Dir\beta$ にしたがった各トピックにおける単語の出現分布の多項分布 $\phi_{z_{d,i}}$ にしたがって生成されるものとする。以上のことから、LDA における結合分布の式 (1) を以下に示す。

$$p(w, z, \theta, \phi | \alpha, \beta) = p(w | z, \phi) p(z | \theta) p(\theta | \alpha) p(\phi | \beta) \quad (1)$$

文書集合全体 w におけるパラメータ推定は、以下の式 (2)

$$\log p(w | \alpha, \beta) = \log \int \sum_z p(w, z, \theta, \phi | \alpha, \beta) d\phi d\theta \quad (2)$$

を最大化することで行う。しかしこの推定は解析的には解けないため以下の式のように、

$$\begin{aligned} \log p(w | \alpha, \beta) &= \log \int \sum_z q(z, \theta, \phi) \frac{p(w, z, \theta, \phi | \alpha, \beta)}{q(z, \theta, \phi)} d\phi d\theta \\ &\geq \int \sum_z q(z, \theta, \phi) \log \frac{p(w, z, \theta, \phi | \alpha, \beta)}{q(z, \theta, \phi)} d\phi d\theta \end{aligned} \quad (3)$$

変分下限の導出を行い、変分法による $q(z, \theta, \phi)$ の最適化によるパラメータ推定を行う。ただし、近似事後分布として以下の因子仮定をおく。

$$q(z, \theta, \phi) = \left[\prod_{d=1}^M \prod_{i=1}^{n_d} q(z_{d,i}) \right] \left[\prod_{d=1}^M q(\theta_d) \right] \left[\prod_{k=1}^K q(\phi_k) \right] \quad (4)$$

LDA は、文書生成過程の確率モデルであり多量の文書とその文書内の単語から学習を行い潜在トピックの推定を行う。今回は LDA における文書 d を画像とし、単語 $w_{d,i}$ を各画像内の物体名称と置き換えたのち、LDA によるトピック推定を行う。つまり、多量の画像内の物体名称を基にした潜在的なトピックの推定を行う。ここで、LDA における潜在トピック $z_{d,i}$ を画像のシーンとして捉え、LDA によるシーンの体系化を行う。

3. 実験と考察

実験では、LDA によって推定された潜在トピックと人間によって作成されたクラスとでの Jaccard 係数による類似度計算を行い、その評価を行った。

3.1 実験方法

今回は、LDA における潜在トピックの学習と人間によって作成されたクラスを定めるために画像データセットとして ADE20K [4] の訓練データを利用した。この訓練データに対する物体認識結果をリスト化し LDA のモデルにしたがい潜在トピックの学習を行った。また、人間によって作成されたクラスについては、ADE20K でのシーン分類結果を用いた。ただし、画像の分類が明確ではないものやクラス内の画像数が少ないものは、潜在トピックの学習や類似度の計算の際に結果が偏る可能性があるため対象外とした。またこのデータセットで取り扱っているオブジェクトが学習済みの YOLOv3 の認識可能なオブジェクトとは異なるため、今回はオブジェクトを YOLOv3 の認識可能なものに統一し実験を行った。この統一は、Jaccard 係数による類似度計算における物体名の表記ぶれの影響防止のために行った。

LDA のプログラム実装には、生成する潜在トピックの設定と確率分布として出力されるトピック内の単語数を設定する必要がある。今回は、人間によって作成されたクラス数 (= 715) を参考にトピック数による変化を見るため大まかではあるが生成トピック数 K を $K = 50, 100, 350, 700$ とした。

トピック内の単語数については、YOLOv3 の認識可能なオブジェクト数を参考にした。しかし類似度を計算する際には潜在トピックにおいてこの YOLOv3 の認識可能なオブジェクト数をすべて考慮するのは、不必要な情報を含む可能性がある。そのため各トピック内での単語の出現頻度を高いものから累積をして、その合計が 95% を超えないように、各トピック内の単語数を定めた。

前処理については、通常の LDA 内部では学習の際に除外されるべき記号や頻出しやすい単語を除外する処理がある。今回は、全画像における物体認識結果での物体名称の出現頻度が 80% 以上となる物体名称を除外し潜在トピックの学習を行った。

図3 Jaccard 係数の分布

図4 人間によって作成されたクラスにおけるオブジェクト数が3未満となるクラスを実験結果から除外した結果による上位5つの比較

以上のような設定を行った上で、LDA によって推定した潜在トピックにおける単語集合 W_1 と定義した人間によって作成されたクラスの単語集合 W_2 による Jaccard 係数を以下の式 (5) により算出する。

$$Jaccard(W_{1i}, W_{2j}) = \frac{|W_{1i} \cap W_{2j}|}{|W_{1i} \cup W_{2j}|} \quad (5)$$

ただし、LDA は確率的モデルであるので潜在トピックの偏りを防ぐために 30 回の生成の試行を行い、それぞれの Jaccard 係数の算出結果の最大値の平均を用いた。

3.2 結果と考察

LDA によって作成されたシーンと ADE20K でのシーンにより算出された Jaccard 係数の分布をまとめたものを図 3 に示す。図 3 より、大半のトピックが 0.2 以上の値になっていることがわかる。また、同じ条件で無作為にクラスを生成した場合の Jaccard 係数を算出すると 0.1 未満であった。このことから、LDA によってシーン分類することで、人間の分類に近づくことがいえる。なお、本実験において人間によって作成されたクラスにおけるオブジェクト数は、0~3 個が 43%、4~10 個が 41%、11 個以上が 16% という分布であった。

各トピックにおける上位 5 つの Jaccard 係数をまとめたものを図 4 に示す。ただしここでは、オブジェクト数が少ないクラス (0~3 個のクラス) を除外した。その理由として、Jaccard 係数はその性質上、クラス内のオブジェクトが少ない場合、係数値が極端になりやすいためである。図 4 において、Jaccard 係数が最も高くなったものは「トピック数_350」であった。特

表 1 推定されたトピックと人間による分類結果でのシーン (fire_station) におけるオブジェクトの比較

	obj.1	obj.2	obj.3	obj.4
ADE20K (YOLOv3)	bus	car	truck	person
トピック数_350	car	truck	person	—

に値の高かった 1 は 0.776 であり、2 も同じで 0.776 であった。このことから、提案手法により生成されたシーンでも、人間が分類するシーンにある程度近い分類が可能であるといえる。

表 1 に LDA によって推定されたトピック (トピック数_350) の一例と人間によるシーン分類結果における fire_station のクラスのそれぞれのオブジェクトを示す。ここでのオブジェクトは、どちらも YOLOv3 によるオブジェクトの認識結果である。ちなみに、表 1 における Jaccard 係数は 0.75 となる。表 1 からわかるように、YOLOv3 で認識できるオブジェクトから fire_station というシーンの連想は難しいといえる。

4. まとめ

人間の行うシーン分類は、画像内における物体の有無がシーン分類の根拠となっている。そこで本稿では、物体の有無について着目したシーン分類手法を提案した。具体的には、YOLOv3 による物体認識結果を LDA に入力することで、シーンを分類することを図った。実験では、LDA により生成されたトピックと人間によるシーン分類の結果とで Jaccard 係数による評価を行った。結果として、生成されたトピックが無作為に作成したトピックよりも人間によるシーン分類結果に類似していることを確認した。

今回使用した画像データセット ADE20K の分類は階層構造を持っている。そのため今後としては、この階層構造に着目した手法の提案と検証を行っていく予定である。

参考文献

- [1] K. Simonyan and A. Zisserman: “Very deep convolutional networks for large-scale image recognition,” arXiv preprint, arXiv:1409.1556, 2014.
- [2] J. Redmon and A. Farhadi: “YOLOv3: An incremental improvement,” arXiv preprint, arXiv:1804.02767, 2018.
- [3] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, “Microsoft COCO: Common objects in context,” arXiv preprint, arXiv:1405.0312, 2014.
- [4] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba: “Scene parsing through ADE20K dataset,” *Proc. of*

the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Jul. 21-26, Honolulu, pp. 633-641, 2017.

(2020 年 4 月 11 日 受付)

(2020 年 6 月 24 日 採録)

[問い合わせ先]

〒437-0032 静岡県袋井市豊沢 2200-2

静岡理科大学大学院理工学研究科システム工学専攻

池田 祐輔

E-mail: 1921003.iy@sist.ac.jp

著者紹介

いけだ ゆうすけ
池田 祐輔 [非会員]

2019 年静岡理科大学情報学部コンピュータシステム学科卒業。現在、同大学院理工学研究科修士課程システム工学専攻在学中。

たかの としあき
高野 敏明 [正会員]

2008 年三重大学工学部電気電子工学科卒業。2010 年同大学院工学研究科博士前期課程電気電子工学専攻修了。修士（工学）。2013 年同大学院博士後期課程システム工学専攻修了。博士（工学，三重大学）。2013 年立命館大学情報理工学部特任助教。2016 年静岡理科大学情報学部講師を経て，2020 年同学部准教授として現在に至る。機械学習，特に強化学習の研究に従事。ファジィ学会，人工知能学会，日本ロボット学会，IEEE 各会員。

Scene Clustering by LDA Using Object Recognition Result

by

Yusuke IKEDA and Toshiaki TAKANO

Abstract:

The purpose of this study is to enable machines to cluster images like humans do. We assume that human image clustering is based on the co-occurrence of objects. In this paper, we propose a scene clustering method based on LDA using the result of object recognition. Through the experimental result, we confirmed whether this clustering method is similar to scene clustering by human, or not.

Keywords: object recognition, LDA, scene clustering

Contact Address: **Yusuke IKEDA**

Graduate of Science and Engineering, Shizuoka Institute of Science and Technology

2200-2 Toyosawa, Fukuroi, Shizuoka 437-0032, Japan

E-mail: 1921003.iy@sist.ac.jp