1A4-OS-27b-1

Deep Learning技術をベースとした異常画像検出

Inappropriate image detection based on Deep Learning

菊田　遥平 *1　野村　眞平 *2　李　石映雪 *2　小林　秀 *1　神津　友武 *1
Yohei Kikuta　Shimpei Nomura　Shiyingxue Li　Shu Kobayashi　Tomotake Kozu

*1有限責任監査法人トーマツ　デロイトアナリティクス **
Deloitte Analytics, Deloitte Touche Tohmatsu LLC.

*2株式会社リクルート住まいカンパニー
Recruit Sumai Company Ltd.

Deep Learning is widely used in many applications such as image/speech recognitions, natural language processing, robot/game controllers and so on. Among them, an image analysis has a strong affinity with business issues from the view point of the availability of huge data sets and the understandability of analysed results. In this paper, using human detections by Faster R-CNN and anomaly class detections by clustering methods with extracted feature maps, we perform the detection of inappropriate images that are undesirable in terms of a web site policy. As an experiment, we use the image data in SUUMO, the largest housing information site. The proposed approach acheives the high recall rate for human detections and the sucessive detection of images that include property before completion and objects unrelated to property, which turns out it is useful for business purposes.

1.　はじめに

本稿では,Region-based Convolutional Neural Network(R-CNN) とその中間層から抽出した feature map を用いて大量の画像の中に混入する不適切な画像を検出する手法を提案する. 不動産ポータルサイト SUUMO において実際にサービスで使用されている画像データを対象に, 人が写っている画像の検出や画像をクラスタリングして適切に物件が写っている画像とそうでない画像の選別を行うことで, 提案手法の有用性を検討する.

Deep Learning 技術は日進月歩であり, 現在では特定タスクにおける識別性能は人間と比肩するレベルに達し, 強化学習との組み合わせや feature map を利用した演算などで扱えるタスクの領域も更なる広がりを見せている. 本稿で注目する画像データに対する Deep Learning の応用は, 最も研究が盛んな分野の一つである. 分類問題や姿勢推測などの画像判定に始まり, テキストと合わせた captioning や同時に分類と年齢・性別推測などを学習する heterogeneous learning, decoding 部分を活用した画像生成など, 実に幅広い研究が行われている. 画像を用いる利点としてデータが大量に存在することが挙げられるが, それに加えて人間が目で見てその結果を判断しやすいというのも特筆すべき性質である. この性質はビジネスにおけるサービス化という観点からも重要であり, 難解な理論を介さずとも技術サイドとビジネスサイドをつなぐことを可能とする. Deep Learning により性能と表現力が高まっている画像分析サービスは今後も更なる広がりを見せていくだろう.

本稿では, ウェブサイト上に掲載する（人手ではチェックしきれない）大量の画像に対して, 掲載にふさわしくない異常画像を自動検出してリスク回避やユーザビリティ向上の支援とすることを目的とする. 本稿で扱う異常画像として, 人が写りこんでいるためにプライバシーの侵害が生じうる恐れのある画像と, 遮蔽物やそもそも物件が写っていないなどの理由でユーザにとって有益でない可能性のある画像, を主たる対象とする.

本稿では Faster R-CNN[Ren 15] をベースとした検出手法を構築しこの問題に取り組む. R-CNN を用いて人検出を実施した結果の例が図 1 であり, 自動的に人部分の矩形領域を検出してモザイク処理を施している. この人検出により, 実際の掲

図 1: 上図が元画像で下図が分析処理を施したもの. 緑の枠が検出された人の矩形領域で数字は予測確率である. 画像はウェブ (http://www.photo-ac.com) から収集した著作権フリーのものだが, 実際のサービスでも現れるものに近い画像である.

連絡先: 菊田遥平, 有限責任監査法人トーマツ　デロイトアナリティクス, yohei.kikuta@tohmatsu.co.jp

** 本研究の内容は有限責任監査法人トーマツの公式見解を示すものではありません.

載画像で人が写り込むという違反を確認していたものに対して, recall で 0.963,precision で 0.868 という高い検出精度を達成した.

提案手法では, 図 1 で示した人が写っている矩形領域を選出する分析に加え, 中間層から抽出した feature map にクラスタリング手法を適用することで正常に写っている画像とそうでない画像を選別する分析も同時に実施する. この分析により, 新築戸建外観画像を完成済みのものと基礎工事の段階のものや建設中でシートが掛けられているものと大別したり, より細かく分類することで文字を含むチラシ画像やパース画像のクラスなどが得られ, 掲載時の付加情報として有用な結果が得られた.

以上のように, 異なる二つの分析を平行して走らせることで, 処理時間を犠牲にすることなく検出できる異常画像の種類を増やすことができる. 本稿における分析は基本的に CPU を用いて実施しているが, Amazon Web Service(AWS) の GPU インスタンスを用いたスケーリングの可能性とコストの検討も行った.

2. 問題設定

ウェブサービスの一般化や SNS サービスの発展により, いたるところに画像データがアップロードされ利用できるようになっている. このことは我々の生活に豊かさをもたらしてくれるものであるが, データ量が増加すると共に, 個人のプライバシー侵害や有用な情報が埋没してアクセスしづらくなるという問題が生じやすくなる. これらの問題はウェブサイト上でサービスを展開する企業にとって, コンプライアンスやユーザビリティの観点から看過できないものとなっている. しかしながら, サービスで扱う大量の画像を人手でチェックしたり選別したりすることはコストの観点から非現実的であり, ここに機械学習技術, 特に Deep Learning, が威力を発揮することが期待されている. ウェブ上には各種ライブラリや先端的モデルなどの利用可能な資源が数多く存在するため, 適切にビジネス課題を定義して資源を活用することで低いコストで高い成果を出すことが可能である.

本稿ではリクルート住まいカンパニーが運営する不動産ポータルサイト SUUMO の画像データを用いて, 掲載上不適切な画像を自動的に検出することを分析の目的とする.

2.1 SUUMO における問題点

大量の画像データを扱う SUUMO は正に上述の問題を抱えるウェブサイトの一つであり, 毎日アップされる膨大な量の画像に対して人手で十分な検閲を行うには多大なコストがかかる状態である. これらの画像には, 故意でなくとも誤って物件と関係のない人が写り込んでしまったり, 物件以外のものが物件としてアップロードされていてユーザにとって十分な情報を有さない画像が含まれている場合がある. このような画像の中には違反画像を人手で確認して報告がなされるものもあり, 報告があって初めて対応するというオペレーションにならざるを得なかった. また過去にはクライアントから提出された問題のある画像がそのままアップロードされたこともあり, 画像分析により効果的にこれらの問題を解決する方法が必要であった.

2.2 本稿における検出対象

上記の事情を考慮し, 本稿では画像分析で検出する異常画像として下記二つに対象を絞る.

・人が写り込んでいる画像
・無関係な画像や遮蔽物等で物件が十分に写っていない画像

前者は掲載上のリスクが高く対象も明確であるため独立して扱っている. 設定した閾値を上回る場合に検出した矩形部分に人が写っていると考えて該当領域を切り出し, スコアが十分高ければ自動でモザイク処理を施しそうでなければ人手でのチェックに回すことを想定している. 後者に関してはクラスタリングを実施した後に人手で正常なクラスタと物件が適切に写っていない異常クラスタのラベルを付与し, 一度ラベルを付与したあとはインプット画像が異常クラスタに振り分けられた場合に人手での確認を行うことを想定している.

3. 提案手法

本稿で提案する手法は単一の Deep Learning モデルで前述の二つの分析を平行して実施するものであり, ここでは全体構成とそれぞれの分析の特徴を概説する. Deep Learning の実装には (Py)Caffe[Jia 14] を用いている.

3.1 全体構成

Faster R-CNN をベースとして図 2 のようなアーキテクチャを構築した. オリジナルの R-CNN とそれを改良した Fast R-CNN[Girshick 15] は矩形領域の選出と分類が独立したアルゴリズムで実装されているが, Faster R-CNN では単一の Deep Learning モデルで完結しているため, 途中の情報が扱いやすく処理時間も高速化されている. 本稿の提案手法では CNN の途中の中間層からプロセスを分岐させ, 一方のプロセスでは人検出を実施してもう一方のプロセスではクラスタリングを実施する. 本稿では CNN のモデルとして VGG16[Simonyan 14] を採用する.

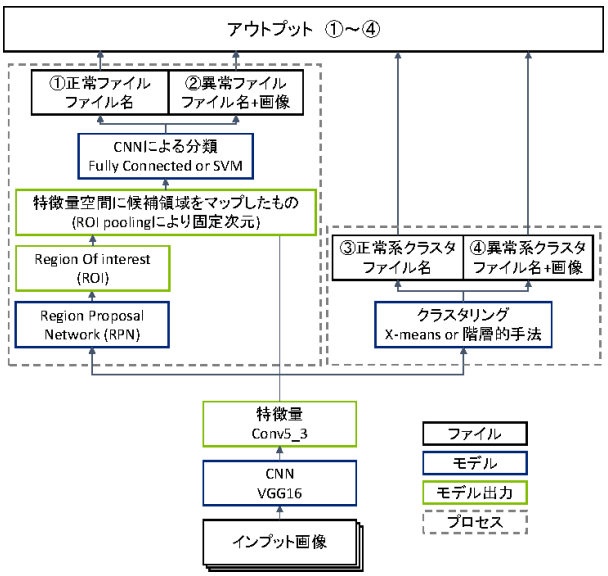

図 2: 全体的なアーキテクチャの概念図. 途中の中間層から並行したプロセスで矩形領域の判別・分類手法とクラスタリング手法を実施する.

3.2 人検出

Faster R-CNN においては, 注目矩形領域を検出した後に検出した矩形領域に対して human クラスを含む 21 クラスの判別問題を実施する. 本稿では human クラスのみに着目して, 設定した閾値よりも確率が高ければ矩形領域の描写を行う. その擬似コードを Algorithm1 に示す. ここでは描写した矩形領域全てにモザイク処理を施すコードになっているが, もう一つ閾値を導入することでその値を基準として前述のようにモザイク処理と人手での確認を切り替えることも可能である.

Algorithm 1 Human regions detecion

Require: threshold t ($e.g.$ $t = 0.3$)
Ensure: images having detected human regions
  for each image of the data set:
    detect high objectness regions by RPN
    for each region of the detected regions:
      compute class probabilities p_i by CNN
      if $p_{human} > \max(t, \max(p_{others}))$:
        draw the rectangular on the image
        pixelize the image within the rectangular
      else:
        continue
    return the image

3.3 クラスタリング

Deep Learning が抽出した中間層の feature map を活かしつつ様々な画像を選別するために, feature map をベクトル化したものを特徴量としてクラスタリング手法を適用する. 人検出での計算過程で得られる中間層情報を別プロセスに流すことで効率的な処理が可能であることに加え, 特徴量の設計なしに画像の高次特徴量を用いたクラスタリングが実行できる. 抽出する中間層として,RPN に投入する直前の conv5_3 の層を使用する. 抽出した feature map の次元は VGG16 では (512,h_{out},w_{out}) となっており, 512 はフィルタの数で h_{out}, w_{out} は height と width の次元で $h_{out} = ceil(h_{init}/16)$ となり w_{out} も同様である. ここで $ceil$ は天井関数であり,$h_{init}(w_{init})$ はインプット画像（もしくはそれを拡大・縮小してスケールしたもの）の次元である. この feature map を用いてクラスタリングのインプットとするためには feature map のテンソル f_{nhw} をベクトル化する必要があるため, ここでは単純に以下のようにマッピングを行う.

$$f_{nhw} \to f_N = f_{n*h*w}. \tag{1}$$

本稿においては $n = \{1, 2, \cdots, 512\}$ と 512 個全てのフィルタを抜きだしベクトル化する. convolutional layer から抽出したベクトルは疎ベクトルとなり, 本稿で扱うものの典型的な密度は4%程度である. これをクラスタリング手法のインプットして扱うために, 各行がインプット画像で各列が構成したベクトルからなる疎行列を構成し, Singular Value Decomposition(SVD) により密行列へ次元圧縮し d 次元の特徴ベクトルを構成する. 以降の分析においては d=100 と設定する.

適用するクラスタリング手法として, クラスタ数を自動決定する X-means[Pelleg 00] と階層的手法の二つを用いる. 階層的手法においては scipy.cluster の fcluster を用いた実装で method⊗metric⊗criterion = {ward, single, complete, average, weighted, centorid}⊗{euclidean, cosine, canberra}⊗{inconsistent, distance, maxclust} という組み合わせで可能なものを比較した. 結果から解釈可能性と有用性を検討し, 以降の分析においては, ward⊗euclidean⊗maxclust, の組み合わせを用いる.

4. 実験

実際に SUUMO で取り扱われる画像データに提案手法を適用してその結果を考察する. 扱えるデータソースと分析目的を考慮し, 下記二つの実験を実施する.

4.1 違反画像を対象とした人検出

賃貸物件領域において違反画像として確認された画像から一部抽出した 354 枚を対象として人検出を実施する. 違反画像には様々な要因があるが, その中で最も数が多いものが人が写り込んでいる画像である. そのため本稿では人検出をターゲットとし, 特に recall の値を高めるようにパラメタをチューニングし分析を実施する. recall の値に注目する理由は, 実際の運用を想定した場合に画像を見逃すことによるリスクが高いためである.

分析の結果は表 1 と表 2 にまとめられている. ここではモデルの予測として, 画像中に閾値を越える値で human クラスと予測された矩形領域が存在する場合を予測の 1 としてカウントし, 何も検出されなかった場合を予測の 0 としてカウントしている. 答えのラベルは目視でつけたものであり, 画像に一人でも人が写っていれば 1, それ以外は 0 とカウントしている. 一枚に複数の人が検出されることも少なくないためこの単純な {0,1} の指標では完全な評価とはならないが, 結果を確認したところ画像中に現れるほぼ全ての人が検出されていたため, その意味で有用な指標であると考えてよい. 人が写っているにも関わらず検出に失敗した (予測, 答え)=(0,1) の 3 件の画像に関しては, 人が非常に小さく写っており個人を特定できないレベルであった.

表 1: 人検出の混同行列

予測 \ 答え	1	0	計
1	79	12	91
0	3	260	263
計	82	272	354

表 2: 精度指標

指標	精度
accuracy	0.958
precision	0.868
recall	**0.963**
F_1 score	0.913

手法の比較として,OpenCV における HOG 特徴量をベースとした物体検出で同様の実験を実施した. デフォルトで実装されている SVM に基づく人検出をパラメタを調整して実行した結果, 正しく人領域を抽出できたのはわずか 18 枚であった. 検出器の学習における使用データなどが異なるため単純比較はできないが, それを差し引いても R-CNN による人検出が高い精度で実現されていると言ってよいだろう.

対象としたデータセットには同一の画像が複数枚含まれるものも多いためデータセットが変われば結果が変動しうる可能性はあるが, この性能は様々なビジネス利用に対して有益なインパクトを与えるものである. 一方で, モデルが人と誤認識した対象は工事現場のコーンのような細長いものが多く, これは分類クラス中に他に適したクラスが存在しないために人と誤認識されたものと考えられる. より高い precistion を達成するためには, 問題設定毎に誤認識した対象を含めて再学習を行うことが有効である.

処理時間は 24[sec/image] 程度だが, 後述するように GPU を用いることで 0.56[sec/image] 程度に短縮することが可能である. GPU による処理速度の向上により一台当たり一日数万枚程度の処理が可能となるので, 日々の定常運用にも耐えうるものとなっている.

4.2 新築戸建外観画像を対象としたクラスタリング

新築戸建領域において画像のサイズを (width,height)=(640,480) に固定した外観画像から一部抽出した 11,368 枚を対象に分析を実施した. また, 対象画像には人が写っているものも散見されるため, クラスタリングと並行して前述の人検出も行っている.

4.2.1 クラスタ数自動決定手法による結果

X-means によるクラスタリングの結果, 表 3 のように四つのクラスタに分けられた. これは大別して完成済みの物件全体像が写っているクラスタ 1,2 と物件の外観情報としては十分な情報を有さないクラスタ 3,4 に分けられている. クラスタ 3,4 に属する画像の典型的な例を図 3 に示す. 新築戸建なので掲載時にまだ物件の外観画像が存在しない場合は生じうるが, このようなクラスタリングに基づいて, 既に外観が見れる物件とそうでない物件というようなユーザに有用な情報を付与することが可能となる.

表 3: X-means によるクラスタリング結果

クラスタ	枚数	クラスタに属する画像の特徴
1	4408	適切に写っている画像
2	2316	適切に写っているが少し引きの画像
3	2608	家の前の玄関や土地のみの画像
4	2036	工事中の足場やブルーシートの画像

図 3: クラスタ 3,4 に属する画像の例. 左側がクラスタ 3 に属するもので右側がクラスタ 4 に属するものである. 画像はウェブ (http://www.photo-ac.com) から収集した著作権フリーのものだが, 実際のサービスでも現れるものに近い画像である.

4.2.2 階層的手法による結果

クラスタ数を変えながら得られるクラスタの特徴を確認した結果, クラスタ数を 10 程度に設定するとパース画像のクラスタや会社のキャラクターが写っているクラスタなどが構築された. クラスタ数を 100 程度に設定すると文字を含むチラシ画像のクラスタやこれまでの結果をより細かく分割するようなクラスタが得られた. これらの結果はビジネス目的と擦り合わせて有用なものとそうでないもののラベルを与えることで, ユーザにとって有用な付加情報として扱うことができる.

4.2.3 人検出

人検出の結果は, 目視により人が写っていると確認された 11,368 枚中 585 枚の画像に対して, 0.888 という recall を達成した. これは前述のユーザ違反報告画像の結果よりも低い値となっているが, 対象画像に含まれる人領域の多くは既にモザイク処理が施されているため, それを考慮すればやはり高い精度を実現している.

4.3 GPU を用いたスケーリング

これまでは手法の実験的要素が強かったため計算は全て CPU で実行していたが, その処理時間は約 24[sec/image] である. この処理時間は 10^4 を越えるような画像データ量に対して適用するには遅すぎるため, AWS の GPU インスタンス (g2.2xlarge) を用いた計算でどの程度速度が改善するかを検証する. サイズが (640,480) である画像に対して同様の分析を実施した結果, cuDNN[Chetlur 14] による実装で処理時間が約 0.56[sec/image] まで短縮された. これにより毎日一万枚の画像を分析する際に要求されるコストは 1.5 時間程度で月額 7,500 円程度に抑えられるため, 低コストで実運用が可能である. ここでの金額計算は Asia Pacific Region でオンデマンドインスタンスを使用する場合 (0.86[$/hour]) を想定している.

5. 結論と今後の展望

本稿では,Faster R-CNN をベースとした画像データに対する人検出と中間層の feature map を用いたクラスタリング手法を並行して実行する手法を提案した. 不動産ポータルサイト SUUMO において実際にサービスで使用している画像データに提案手法を適用した結果, 人検出では 0.963 という高い recall を達成し, クラスタリングでは物件の建設前・中・後の分類や物件外観と関係の薄いクラスタの選出に成功した. これらの結果はビジネスニーズに応えるポテンシャルを有するものであり, GPU 演算と合わせることでサービスの実運用において高い効果を発揮することが期待される.

この他にもカラーヒストグラムをインプットとした異常検知も実施し, 白飛び画像や漫画・アニメ画像が検出される結果が得られており, このような分析も組み合わせることでより広範な種類の異常画像を発見することが可能である.

拡張の方向性としてはまず転移学習による検出対象の任意クラスへの適応が考えられる. これは既に着手しており一定の性能が得られているが、データ準備や学習には一定の人的コストや計算資源が必要になる. その他には抽出する feature map のベクトル化に際した次元の取扱いや適用するクラスタリング手法の選択に試行錯誤の余地が残っている. また, 提案手法は domain specific な知識を用いるものではなく汎化性能が高いため, 異なるインダストリーのデータを対象に分析を実施していく方向性も意義深い.

参考文献

[Ren 15] Ren, Shaoqing, et al. "Faster R-CNN: Towards real-time object detection with region proposal networks." Advances in Neural Information Processing Systems. 2015.

[Jia 14] Jia, Yangqing, et al. "Caffe: Convolutional architecture for fast feature embedding." Proceedings of the ACM International Conference on Multimedia. ACM, 2014.

[Girshick 15] Girshick, Ross. "Fast r-cnn." Proceedings of the IEEE International Conference on Computer Vision. 2015.

[Simonyan 14] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[Pelleg 00] Pelleg, Dan, and Andrew W. Moore. "X-means: Extending K-means with Efficient Estimation of the Number of Clusters." ICML. Vol. 1. 2000.

[Chetlur 14] Chetlur, Sharan, et al. "cudnn: Efficient primitives for deep learning." arXiv preprint arXiv:1410.0759 (2014).