Class Activation Mapping の安定性及び反応値の検証
－深層学習を用いた平均訪問意欲推定 AI を対象に－

Verification of Stability and Reaction Value of Class Activation Mapping
- For average visit motivation estimation AI using deep learning-

○大野耕太郎 [*1]，中西達也 [*2]，中村翔太 [*2]，山田悟史 [*3]
Kotaro ONO[*1],Tatsuya NAKANISHI[*2],Shota NAKAMURA[*2],and Satoshi YAMADA[*3]

*1 立命館大学 理工学研究科　環境都市専攻　博士前期課程
Graduate, Dept. of Architecture and Urban Design, Ritsumeikan Univ.
*2 立命館大学 理工学部建築都市デザイン学科
Dept. of Architecture and Urban Design, Ritsumeikan Univ.
*3 立命館大学 理工学部建築都市デザイン学科　任期制講師・博士（工学）
Lecturer, Dept. of Architecture and Urban Design, Ritsumeikan Univ., Dr.Eng.

キーワード：人工知能；AI; CAM; 注視点
Keywords: Artificial Intelligence(AI); Class Activation Mapping(CAM); Gaze Point

1. はじめに

深層学習を基盤とする AI 技術に対する社会的な関心は高まり、建築分野においても実際の業務や研究においてこうした技術を取り入れた実例は増えつつある。筆者らも建築・都市分野への適用可能性の検討として、街並み画像から街路名を推定する AI モデル、被験者 1 名を対象とした訪問意欲の有無とその度合を推定する AI モデルの作成を行い、一定の精度での分類・推定に成功した [1]。

深層学習により様々なデータに対して分類・推定が可能となった一方で、AI が学習を行う際に、対象データのどの部分に注目しているのか人間が判断できないという「ブラックボックス問題」がある。こうした問題を解決するために、AI が画像のどの部分に対して注視しているかを可視化する技術として CAM(Class Activation Mapping) [2] がある。

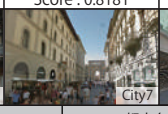
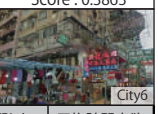
そこで本研究では、街並みに対する平均訪問意欲を推定する AI の作成と、CAM を用いた学習時の AI 注視領域の挙動と人間との認識の違いについての考察を行う。

2. 研究概要

2.1. 研究の流れ

本研究では街並み画像を対象に被験者実験を行い、各街路に対する被験者の平均訪問意欲を算出する。被験者実験では、AI と人間の注視領域の違いを検証するために注視点の計測とスケッチの描写を行う。被験者による注視点の計測結果の一例を図 6 に、スケッチ画像の一例を図 4 に示す。次に作成したデータを元に平均訪問意欲推定 AI の学習と検証を行う。その後、学習済み AI モデルに対して CAM による検証を行い、学習の各段階において AI の注視領域の変化がどのように遷移していくのかを確認する。また被験者実験で取得した注視点計測結果とス

Table 1 List of street names and average willingness to visit

平均訪問意欲が高い街並み画像の一例		平均訪問意欲が低い街並み画像の一例	
Score : 0.8863	Score : 0.8181	Score : 0.1136	Score : 0.3863
 City8	 City7	 City10	 City6

番号	国名	都市名	実験枚数	訪問したい	平均訪問意欲
1	アメリカ	ニューヨーク	45	26	0.5777
2	オーストラリア	メルボルン	44	16	0.3636
3	カナダ	セントジョーンズ	45	18	0.4
4	チェコ	プラハ	44	37	0.8409
5	イギリス	ロンドン	44	36	0.8181
6	香港	門框	44	17	0.3863
7	イタリア	フィレンツェ	44	36	0.8181
8	日本	京都	44	39	0.8863
9	日本	大阪	44	9	0.2045
10	日本	東京	44	5	0.1136
11	メキシコ	グァナファト	44	30	0.6818
12	ペルー	クスコ	44	32	0.7272
13	ポルトガル	ポルト	43	32	0.7441
14	ロシア	モスクワ	43	31	0.7209
15	スコットランド	エディンバラ	43	35	0.8139
16	南アフリカ	ケープタウン	43	27	0.6279
17	韓国	ソウル	43	17	0.3953
18	スペイン	バルセロナ	42	25	0.5952
19	台湾	九份	42	19	0.4523
20	タイ	バンコク	41	13	0.3170
21	アラブ首長国連邦	ドバイ	44	29	0.6590

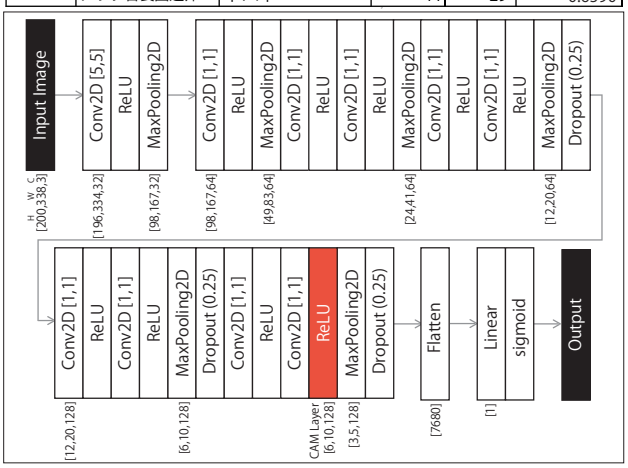


Fig1 Network structure of the model used in this study

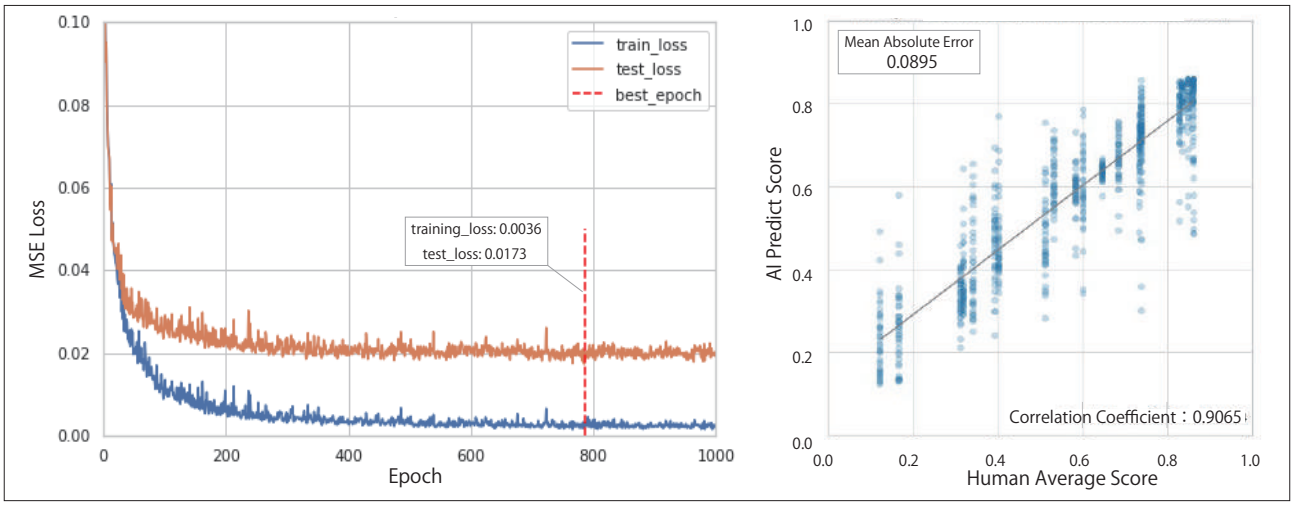


Fig2 Scatter plot of changes in learning loss, average visit motivation and AI estimates

ケッチ描写結果を各段階の CAM の結果と比較し、相関係数を測定することにより人間と AI との注視領域の比較を行う。最後に CAM を行う際に得られる出力値の強さと AI の平均訪問意欲推定結果を比較することにより AI の画像に対する反応と推定結果との関連性を考察する。

2.2. 研究対象・学習データセット

本研究の研究対象として用いるのは街並み画像である。データセットは大都市や観光情報を元に表 1 に示す 21 都市の街路から作成した。画像は画像が単一の建築物・地面・空に占められることのないように配慮しながら Google Earth のストリートビューから作成した。枚数は 1 街路 100 枚ずつの 2100 枚である。被験者実験の際には、街路の偏りと画像の被りがないように 1000 枚を選定した。また AI の学習のデータセットには、左右反転の水増し処理を行った 4200 枚の画像を用いた。

3. 平均訪問意欲を推定する AI の作成と検証

3.1. 被験者実験の概要

街並み画像に対する平均訪問意欲の算出と注視点・スケッチ画像の取得を行うために、建築系学生 100 名を対象とする被験者実験を行った。実験では先述したデータセットから被験者 1 名あたり 10 枚を選定した。

次に実験の流れを説明する。まず初めに被験者は椅子に座った状態でモニターに映し出された街並み画像を 1 枚当たり 30 秒注視する。この際画面の領域内における注視点の変化を PupilLabs 社の注視点計測デバイスである PupilCore を用いて測定する。次に被験者は追加情報が入らないように街並み画像を隠した状態で、2 分間で街並みのスケッチを行う。最後に被験者はその街並みに対して「訪問したい」「訪問したくない」かを回答する。各街並みに対して実験枚数中「訪問したい」と回答した割合をその街並みに対する平均訪問意欲とした。実験枚数 1000 枚のうち注視点計測が失敗したものや不適切な回答を省いた 914 枚を数値の算出に用いた。平均訪問意欲一覧と平均訪問意欲が高かった・低かった画像の一例を表 1 に示す。京都やフィレンツェなどの歴史的で整った街並みほど平均訪問意欲が高く、電線などの雑多な街並みほど低い傾向が見られた。

3.2. AI による平均訪問意欲推定

AI の学習モデルのネットワーク構造を図 1 に示す。高い精度で分類を行うモデルとして知られる「VGG」を参考に過学習が発生しないように Dropout 層を追加するなどの変更を行った。モデルの作成には深層学習のフレームワークである keras[3] を用いた。学習データセットは先述した 4200 枚の 3 チャンネル・縦 200 ピクセル＊横 338 ピクセルの街並み画像である。これらの画像を学習用データセット 3360 枚、検証用データセット 840 枚に分割した。また、学習データセットの正解ラベルとには各街並みの平均訪問意欲を正規化したものを付与した。本モデルでは入力画像を元に畳み込みを行い、最終的に 0 から 1 の範囲で値が出力される。出力値を再正規化することで AI の推定結果とした。AI の損失関数には平均二乗誤差 (MeanSquaredError) を用い、学習の最適化関数には Adam を、バッチサイズは 64 で 1000 エポック学習を行った。

学習時における学習用データセット・検証用データセットそれぞれに対する平均二乗誤差の推移の様子を図 2(左) に示す。1000 エポック学習を行った結果、400 エポック付近で学習が収束し、768 エポック目で検証用画像に対する平均二乗誤差が最小となり 0.0173 となった。(以降最良エポックと呼称)

最良エポックでの検証用画像に対する被験者の平均訪問意欲 (正解値) と AI による推定結果の散布図を図 2(右) に示す。両者の値に対して相関係数を算出した結果、0.9065 となった。その一方で検証用画像に対する正解値と推定結果との間の平均絶対誤差 (MeanAbsoluteError) は 0.0895 となった。

この結果から検証用画像全体としては高い精度での平均訪問意欲の推定を行う AI の作成に成功したが、個々の

画像に対しての推定の誤差にはまだ課題があると言える。

4　Grad-CAM による AI 注視領域の可視化と安定性の確認

4.1 CAM 詳細

前章では平均訪問意欲について高い精度での予測を行う AI モデルの作成を行った。そこで本章では AI が学習を行う際に画像のどの点に注目しているのかを CAM(ClassActivationMapping) を用いて分析を行う。CAM は主に画像認識のモデルで、特定のクラスに寄与したとされる入力領域をハイライトする手法である。本研究で使用するのは CAM 技術の一種である Grad-CAM[4] である。

AI は学習の際、誤差逆伝播法と呼ばれるネットワークの出口側から入口側に向かって各層の重みを更新することで精度を向上させていく。Grad-CAM はこの仕組みを利用し、対象となるクラスのみ 1、それ以外を 0 として逆伝播を行うことで勾配の GrobalAveragePooling(GAP) を求め重みとする。可視化したいレイヤーの特徴量マップに重みを掛け、足し合わせて活性化層を通すことでそのクラスに分類する際の特徴量マップのみ可視化することができる。本研究では最終活性化層（図 1 赤色部分）を対象とし、Grad-CAM を用いた可視化を行った。

4.2. CAM 出力結果の安定性確認

Grad-CAM の出力は図 5 のように特徴量マップのサイズ (6*10) 分の 60 個の数値であるが、本研究では視認性の観点からヒートマップを用いて元画像と重ねて図示する（後述する注視点画像・スケッチ画像も同様)。

最良エポック時の Grad-CAM と 1000 エポックまでの各 100 エポックごとの Grad-CAM 出力値を正規化しヒートマップ化した結果を図 5 に示す。ここで注目すべき点として各エポックにおいて Grad-CAM の出力結果が異なるという点が挙げられる。図 5 の例では 100 から 500 エポック周辺では画像の左下部分に大きく反応しているのに対して、それ以降のエポックでは安定して反応する回数が少なく、最良エポックでは入力画像の右上部分に対して強く反応していることが分かる。

図 6 に被験者実験に用いた 914 枚の画像を入力した際の、最良エポック Grad-CAM 出力値結果と 100 エポックごとの Grad-CAM 出力値結果との相関係数の箱ひげ図と平均値一覧を示す。結果としては 700 エポックとの相関の平均値が最も高く 0.7868 となった。それ以降のエポックでも 0.75 を上回り出力結果が最良エポックを過ぎた後でも安定していた。その一方で、前章で AI 学習時の損失の変化では 400 エポック時で学習が収束していたのに対し、Grad-CAM の出力値では相関が 0.6647 となり 700 エポック以降と比べて 0.10 程度下がった。このことより、AI の学習が安定する段階と Grad-CAM の出力結果が安定する段階にはずれがあり学習が安定した後でも CAM の出力値が安定しているかを確認する必要性があるといえる。

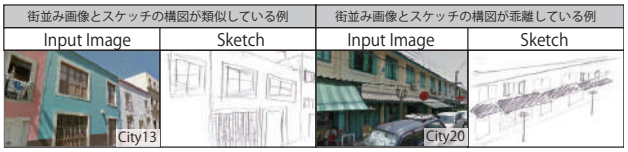


Fig3 An example of a sketch image

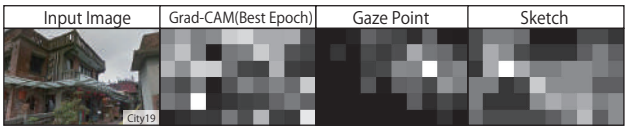


Fig4 Image of normalized image

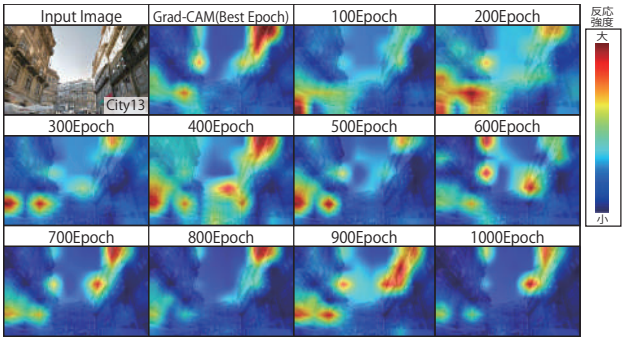


Fig5 An example of how the CAM changes at each epoch

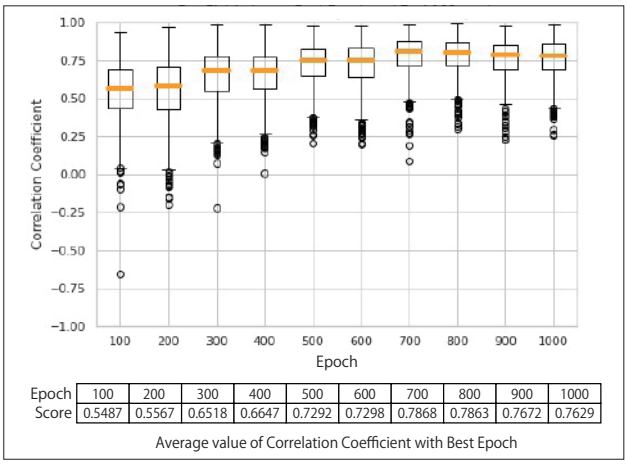


Fig6 Comparison with Best Epoch
Grad-CAM output value result correlation coefficient box

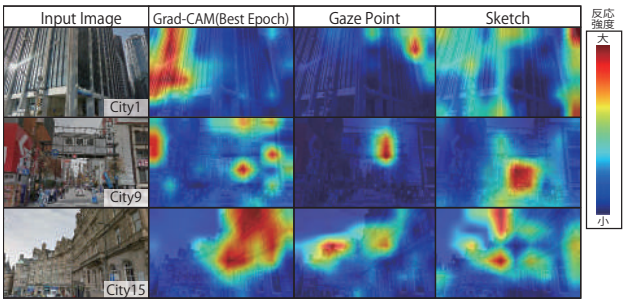


Fig7 An example of Grad-CAM, gazing point, and sketch results for the input image

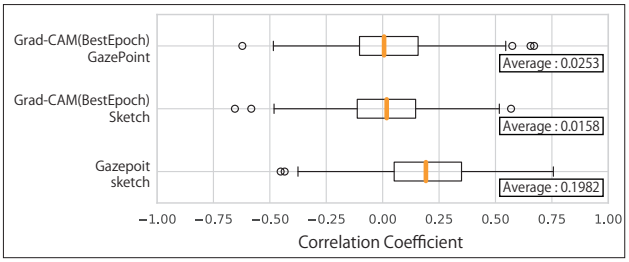


Fig8 Correlation coefficient boxplot between Grad-CAM, gazing point, and sketch

5. Grad-CAM の出力結果と人間の注視領域との比較

5.1. Grad-CAM の出力結果と人間の注視領域との比較

CAM により AI が注視している領域の可視化が可能となった一方で、AI の判断根拠が人間と一致しているかを確認する必要がある。そこで本章では、Grad-CAM による AI が街並み画像から平均訪問意欲を推定する際の注視領域と、被験者実験により得られた注視点領域・スケッチ描写との間の関係性の考察を行う。

前述した平均訪問意欲推定 AI では最良エポック付近での Grad-CAM の出力結果に高い相関があることが確認された。これを踏まえ注視点領域とスケッチ描写との Grad-CAM の注視領域との比較には最良エポックを用いた。

図 7 に入力画像に対する Grad-CAM・注視点・スケッチのヒートマップ画像を示す。入力画像には被験者実験に用いた 914 枚の街並み画像を用いる。先述したように注視点・スケッチの結果は Grad-CAM 特徴量マップのサイズ (6*10) に合わせ正規化を行った（図 4 参照)。ヒートマップ結果を見ると Grad-CAM では壁の模様や空に、被験者の注視点やスケッチでは画像の奥行きやシンボルとなような看板や建物の数値が高くなる傾向があった。

図 8 に被験者実験に用いた 914 枚の画像に対する「Grad-CAM と注視点」・「Grad-CAM とスケッチ」・「注視点とスケッチ」の入力画像に対する相関係数の箱ひげ図と平均値の一覧を示す。結果を見ると「Grad-CAM と注視点」、「Grad-CAM と注視点」の相関係数の平均値はそれぞれ 0.0253、0.0158 となり平均訪問意欲推定 AI と人間の注視領域には相関性は見られなかった。「注視点とスケッチ」でも相関係数の平均値は 0.1982 となり人間が街並み画像を認識する際の指標として用いた両者において相関係数が低い結果となった。原因の一つとしてスケッチの描写と入力画像との間でパースが大幅に異なっている場合があったことが挙げられる（図 3)。こうした問題点を解消することでより相関性が高まる可能性がある。

5.2. AI の推定と Grad-CAM 出力値との間の関係性

先項の内容では被験者の平均訪問意欲を学習する際に AI は人間とは異なる領域を注視していることが分かった。本研究では可視化する際に画像ごとに Grad-CAM の出力値を正規化してヒートマップを作製したが、実際には画像ごとに出力値は大きく異なる。

図 9 に被験者実験に用いた 914 枚の画像に対する Grad-CAM の正規化前の出力値を合計した値と、AI の平均訪問意欲推定結果を再正規化する前との散布図を示す。散布された数値は対応する街並みの番号に該当する（表 1)。これを見ると同じ番号の街並み同士が比較的近距離に分布しており全体としては弧を描くような分布となった。平均訪問意欲推定番号 11 番や 21 番といった街並みに対して強く反応しており、番号 7 番や 10 番といった街並みに対しては弱く反応している。また分布は推定結果 0 から約 0.7 付近までは平均訪問意欲推定結果と Grad-CAM 結果との間には正の相関関係が、約 0.7 から 1 までの平均訪問意欲推定結果と Grad-CAM 結果との間には負の相関関係が見られた。

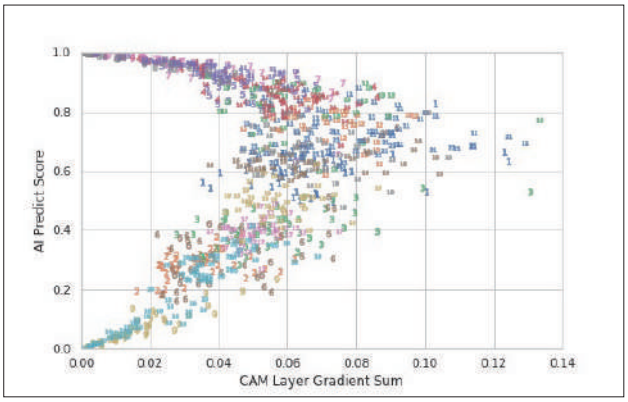


Fig9 Scatter plot of total Grad-CAM response and AI estimates

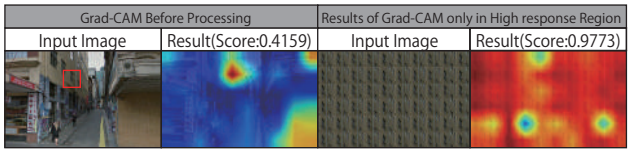


Figure 10 Changes in Grad-CAM and AI estimates during Image Conversion

6. まとめと考察

本研究では深層学習を用い街並み画像に対する平均訪問意欲推定 AI の作成を行い、高い精度での推定に成功した。また被験者実験により人間の注視点・スケッチと Grad-CAM 出力結果との比較を行い、人間と AI の注目領域の違いを考察した。さらに AI の推定と Grad-CAM の出力値を比較し、両者の関係性の可視化を行った。

CAM では画像の情報から AI がどういった点に着目して推定を行っているかを可視化できるが、その一方で人間感性との比較には課題が残る。図 10 の例では AI 最注目領域のみで構成された画像の推定値（再正規化前）が元画像よりも高くなり 0.97 となった。今後こうした問題点に対して追加実験やクラス分類時の CAM との比較などにより CAM の挙動と推定結果との関連性をする予定である。

[謝辞]
本研究の被験者実験は立命館大学建築計画研究室と共同で行った。

[参考文献]

1) 山田悟史，大野耕太郎：Deep Learning を用いた印象評価推定 AI の作成と検証，日本建築学会計画系論文集，第 84 巻，第 759 号，2019,5,pp.1323-1331. 山田悟史，大野耕太郎：Deep Learning を用いた デザイン AI の作成と検証 －街並みと建築物外観の画像生成を対象に－，日本建築学会計画系論文集，日本建築学会，第 85 巻，第 770 号，pp 987-995,2020.5
2) Bolei Zhou,Aditya Khosla,Agata Lapedriza,Aude Oliva,Antonio Torralba:Learning Deep Features for Discriminative Localization,2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2016,11,pp.2921-2929.
3) keras 公式ドキュメント :https://keras.io/ja/
4) R.R.Selvaraju,M.Cogswell,A.Das,R.Vedantam,D.Parikh,D.Batra,et al.:Grad-cam Visual explanations from deep networks via gradient-based localization,InICCV, pages 618-626, 2017.