400. 深層学習の説明性

秋葉洋哉

2024 年 7 月 15 日

1　解釈性の重要さ

1.1　概要

ディープラーニング活用の難しいことの一つは「ブラックボックス性」にあるといえる。つまり、モデルの出力の根拠を説明することができないということである。例えば、医療現場においては、患者の診断結果を説明できないモデルを用いることは困難である。そのため、モデルの解釈性を高めることが求められている。

ディープラーニングモデルの解釈に使われる四つの手法として、

- CAM
- Grad-CAM
- LIME
- SHAP

がある。以下では、それぞれの手法について説明する。

1.2　CAM

CAM(Class Activation Map) は、正則化の役割のあった Global Average Pooling を再検討し、CNN が潜在的に注目している部分を可視化できるようにする役割を持っていることが明らかになったことで、提案された手法である。直観的には、出力層の重みを畳み込み特徴マップに投影することで、画像領域の重要性を識別することができる (図 1)。ただし、CAM が利用できるネットワークは、CNN を大部分で使用しており、出力層前に GAP を実行していることが前提となる。数式で表すと、

$$M_c(x,y) = \sum_k w_k^c f_k(x,y) \tag{1}$$

で表される。CAM を用いてモデルがどこに注目しているかを可視化した図を、図 2 に示す。

1.3　Grad-CAM

Grad-CAM とは、CAM と同様、CNN モデルに判断根拠を持たせ、モデルの予測根拠を可視化する手法である。Grad-CAM は、CAM での GAP の代わりに、最後の CNN 層の勾配を用いて、特徴マップを抽出する。この性質から、最後に GAP を用いないネットワークにも適用することができる。勾配が大きいピクセル

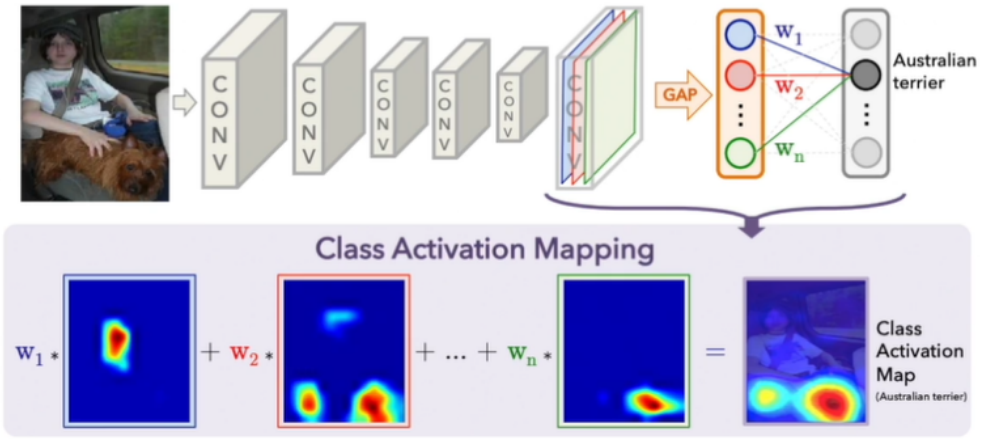


図 1: GAP 後の各セルの重みと、GAP 直前の CNN 層のフィルタの重みを掛け合わせることで、Class Activation Map が出力される。

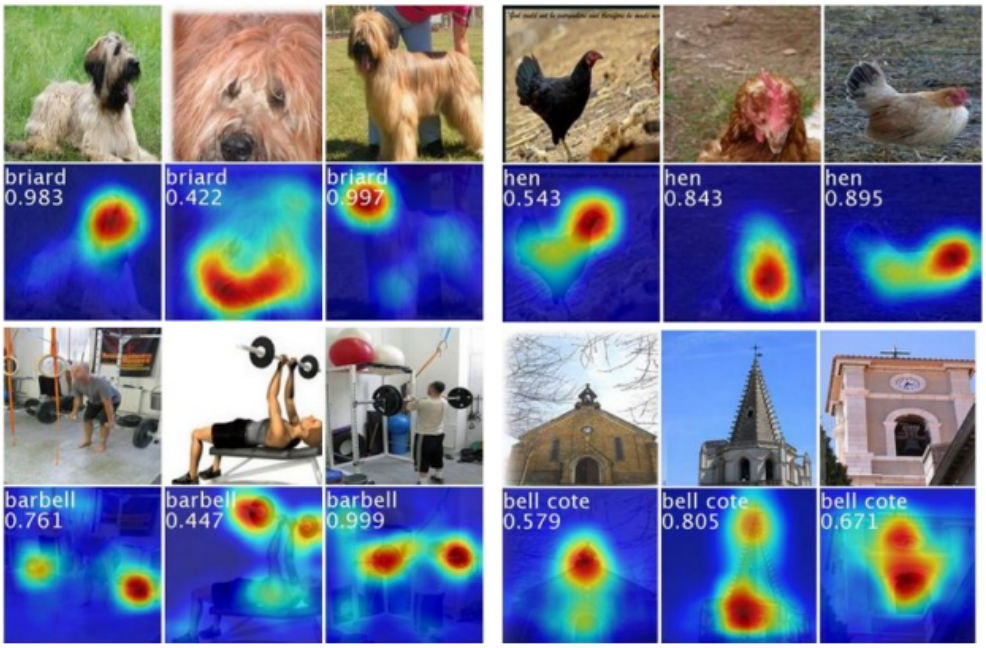


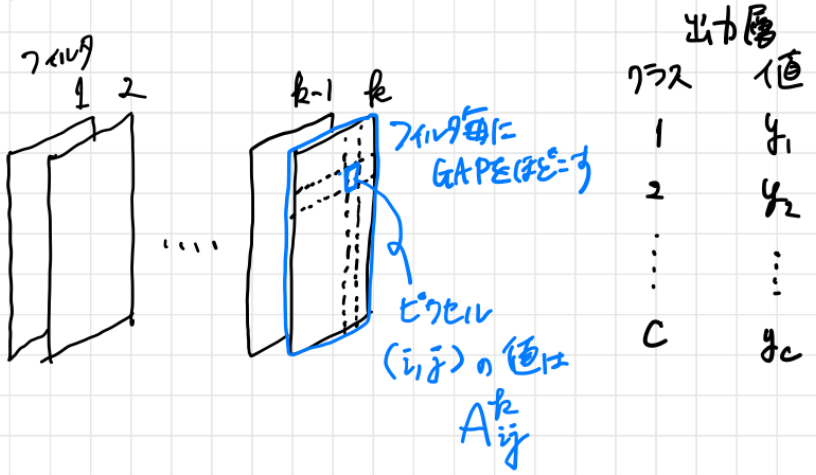
図 2: CAM を用いてモデルがどこに注目しているかを可視化した図

ほど、モデルが注目していると考えられる。

また、Grad-CAM は、CAM と比較して、ネットワークの構造に依存しないため、画像分類だけでなく、物体検知やセグメンテーションにも適用できる。

Grad-CAM のヒートマップがどのように生成されるかを示した図を、図 3 に示す。

Grad-CAM



・y_cのA^kにおける勾配をGAPで表すと以下のようになる。

$$\frac{\partial y_c}{\partial A^k} = \frac{1}{Z}\sum_i\sum_j \frac{\partial y_c}{\partial A^k_{ij}}$$

(フィルタk内ですべてのピクセルについて足し合わせ)

・この時、ヒートマップの値は以下で示せる。

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \frac{\partial y_c}{\partial A^k} \cdot A^k\right)$$

(フィルタ毎のGAPの重みと元々の値を線型結合して、和を取ったものをReLUに入力)

[batch_size, height, width, channel]

⊙ [batch_size, channel, 1]

= [batch_size, height, width]

図 3: Grad-CAM のヒートマップ生成過程

2 LIME

LIME(Local Interpretable Model-agnostic Explanations) は、特定の入力データに対する予測について、その判断根拠を解釈・可視化するツールのことを指す。例えば、表形式データを入力したとき、どの変数が予測に寄与しているか、また、画像データを入力したとき、どのピクセルが予測に寄与しているかを可視化することができる。

LIME は、複雑なモデルをシンプルなモデル (決定木等) で線形近似し、その近似モデルを解釈することで、モデルの予測根拠を解釈する。LIME の手法は、以下の手順で行われる。

1. 入力：一つの個別の予測結果
2. 近似モデル：入力データに類似したデータ (領域削除・反転・ランダム置換等) をサンプリングし、そのデータを教師データとして、データ空間の対象範囲内でのみ有効な線形近似モデルを作成
3. 出力：近所用モデルから予測に寄与した特徴量を選び解釈を行うことで、本来の難解なモデルの予測根拠を解釈 (したとみなす)

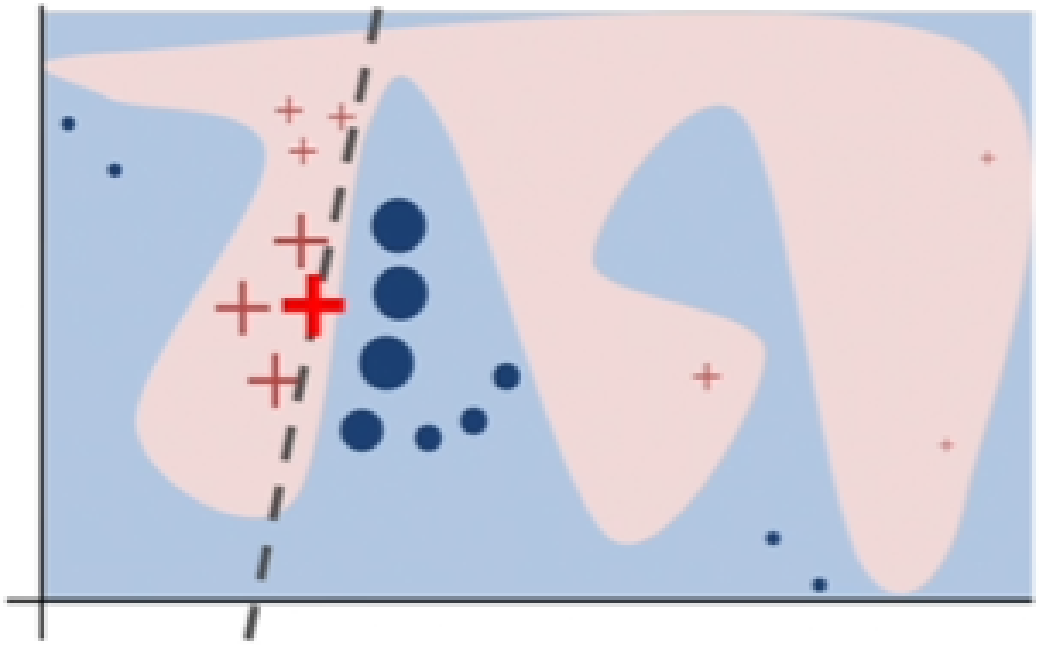


図 4: 赤い領域と青い領域の境界が関数 f の分類境界を表す。太赤十字が LIME への入力データであり、その周辺の十字が作成されたデータである。LIME においては、点線のようにそのデータ付近の説明を端的に表す。

LIME においては、以下の数式を解くことで、近似モデルを求める。

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{2}$$

ここで、$\mathcal{L}$ は損失関数、Ω はモデルの複雑さを表す正則化項、π_x は入力データ x の近傍データを生成する関

数である。また、f は解釈したいモデル、g は近似モデルである。π_x は以下で定義される。

$$\pi_x(z) = \exp\left(-D(x,z)^2/\sigma^2\right) \tag{3}$$

ここで、$D(x,z)$ は x と z の距離を表し、σ はカーネル幅を表す。これは、入力データ x と近傍データ z の距離が近いほど、そのデータの重みが大きくなることを示している。損失関数 $\mathcal{L}$ は、以下のように定義される。

$$\mathcal{L}(f,g,\pi_x) = \sum_{z,z' \in \mathcal{Z}} \left(f(z) - g(z')\right)^2 \pi_x(z) \tag{4}$$

これは、近似モデル g が元のモデル f をどれだけ近似できるかを表す。つまり、解釈したいモデルに z を入力した結果と、近似モデルに z' を入力した結果の差に、z と x の類似度で重みづけを行っていることを示している。LIME の実装は、https://github.com/marcotcr/lime を参考にすると良い。

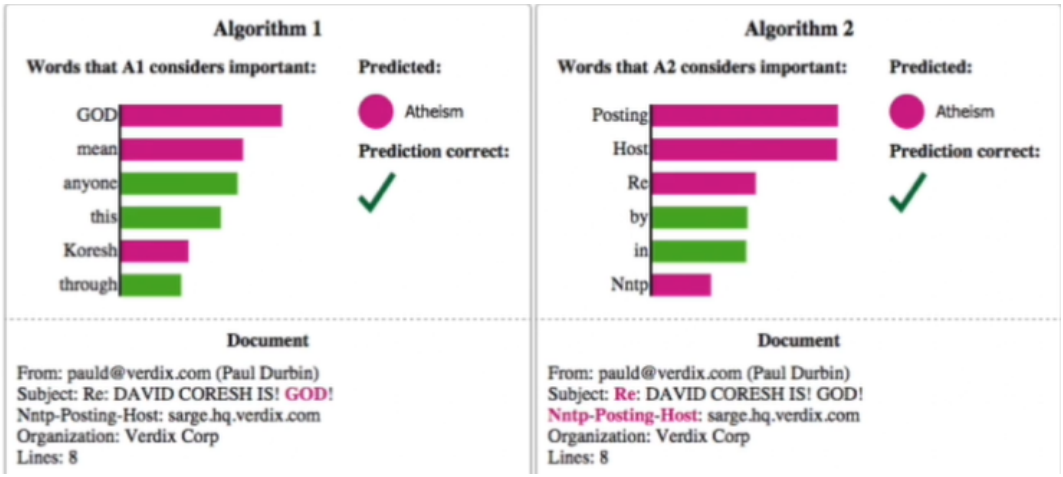


図 5: LIME を用いて、モデルの予測根拠を解釈した図：入力にテキストデータを与え、「無神論」と「キリスト教」を分類するタスクにおいて、Algorithm 1 と Algorithm 2 に LIME を適用した結果を表している。Algorithm 2 は Posting, Host, Re といった単語に着目しており、これらは無神論にもキリスト教にも関係のない単語であることから、学習データに問題があったということが分かった。

3 SHAP

SHAP(Shapley Additive Explanations) は、ゲーム理論に基づいたモデル解釈手法であり、モデルの予測根拠を解釈するための手法である。SHAP は、特徴量の重要度を評価するために、Shapley 値を用いる。Shapley 値は、特徴量の重要度を評価するための指標であり、特徴量の組み合わせに対する貢献度を評価することができる。

SHAP の概念を、A, B, C の三人で得られる報酬額を計算する例で説明する。A, B, C の三人が得られる報酬額を計算する際、A, B, C 一人ひとりが得られる報酬額を計算する (図 6a)。次に、A, B, C の二人組が得られる報酬額を計算する。この 6 通りの情報を元に、「限界貢献度」を計算する。例えば、A が得られる報酬額を計算する際、A, B, C の三人で得られる報酬額から、A 以外の二人で得られる報酬額を引いたものが A の限界貢献度となる。その後、B と C の二人で得られる報酬額から、B 一人の報酬額を引いたものが、C の限界貢献度となる。さらに、B の限界貢献度は、B 一人での報酬額となる (順序によって異なる結果となる)。これらの限界貢献度をすべての通りで計算したものが、図 6b である。こうして求まった限界貢献度の算術平均が、shapley 値となる。

Aさん	4万円
Bさん	7万円
Cさん	5万円
AさんとBさん	10万円
AさんとCさん	16万円
BさんとCさん	20万円
AさんとBさんとCさん	35万円

(a) A, B, C の三人で得られる報酬額

	A	B	C
A -> B -> C	4万円	6万円	25万円
B -> A -> C	3万円	7万円	25万円
B -> C -> A	15万円	7万円	13万円
C -> B -> A	15万円	15万円	5万円
A -> C -> B	4万円	19万円	12万円
C -> A -> B	11万円	19万円	5万円

(b) 限界貢献度のマトリクス：平均的な限界貢献度は、A=8.7, B=12.2, C=14.2 となる。(算術平均)

図 6: SHAP の概念

SHAP を機械学習に適用するときは、特徴量が上述の例の A, B, C に対応する。SHAP は、特徴量の組み合わせに対する貢献度を評価することができるため、特徴量の重要度を評価することができる。SHAP の特徴は、特徴量の重要度を評価するだけでなく、特徴量の組み合わせに対する貢献度を評価することができる点である。

SHAP を用いて、Boston データセットを解釈した結果を図 7 に示す。

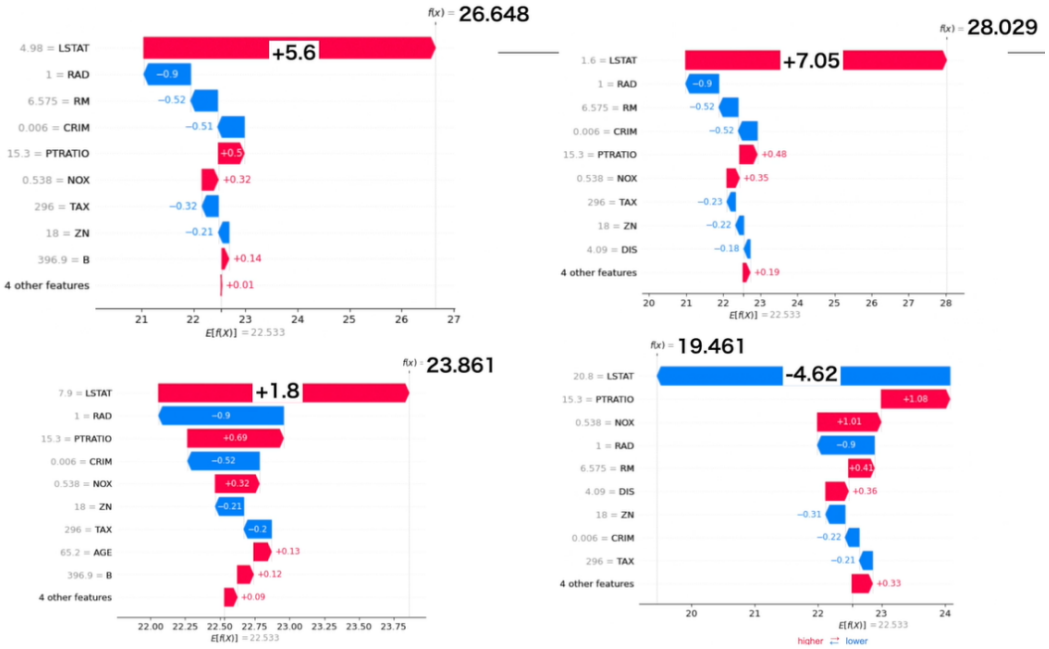


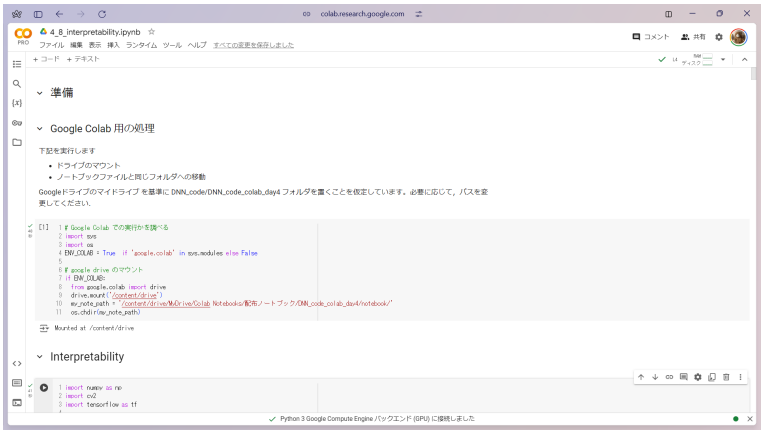
図 7: Boston データセットを用いて、SHAP を適用した結果：左上は LSTAT(低所得者率)=4.98 の場合、右上 LSTAT=1.6, 左下 LSTAT=7.9, 右下 LSTAT=20.8 の場合の shapley 値 (限界貢献度)$f(x)$ を示している。LSTAT が低いほど、住宅価格を高く予測するのに働き、LSTAT が高いほど、住宅価格を低く予測するのに働いていることが読み取れる。

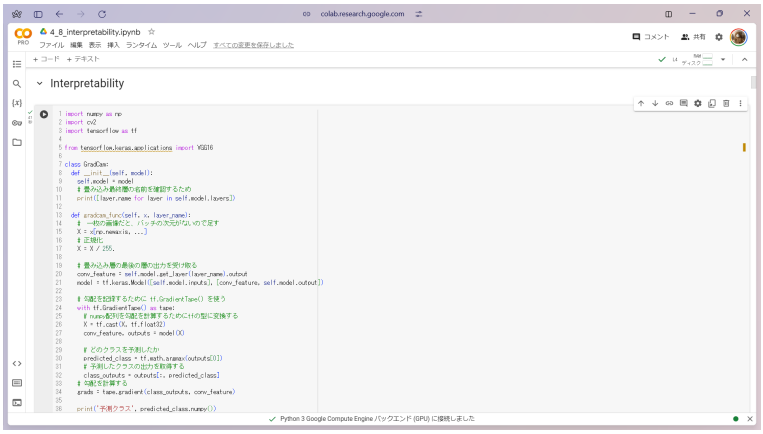
■参考文献

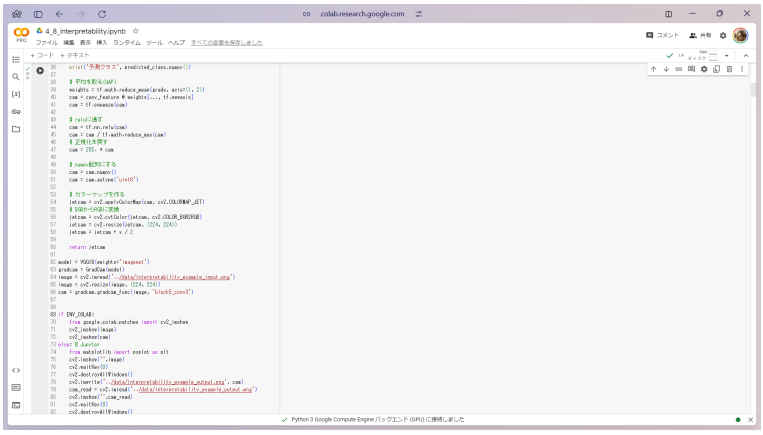
1. 岡谷貴之/深層学習 改訂第 2 版 [機械学習プロフェッショナルシリーズ]/ 講談社サイエンティフィク/ 2022-01-17
2. LIME で機械学習の予測結果を解釈してみる https://qiita.com/fufufukakaka/items/d0081cd38251d22ffebf
3. 機械学習モデルの局所的な解釈 (LIME と SHAP) https://horomary.hatenablog.com/entry/2019/09/16/000110

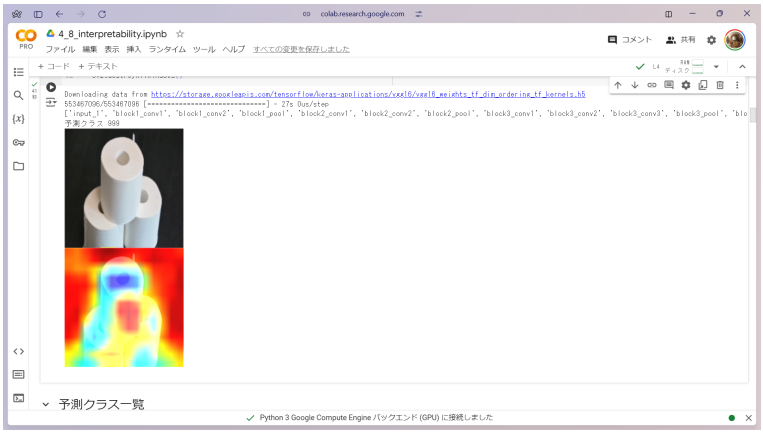
4　実装演習キャプチャ

4.1　CAM








colab.research.google.com
4_8_interpretability.ipynb
ファイル 編集 表示 挿入 ランタイム ツール ヘルプ すべての変更を保存しました
コメント
共有
+ コード + テキスト
Downloading data from https://storage.googleapis.com/tensorflow/keras-applications/vgg16/vgg16_weights_tf_dim_ordering_tf_kernels.h5
553467096/553467096 [==============================] - 27s 0us/step
['input_1', 'block1_conv1', 'block1_conv2', 'block1_pool', 'block2_conv1', 'block2_conv2', 'block2_pool', 'block3_conv1', 'block3_conv2', 'block3_conv3', 'block3_pool', 'blo
予測クラス 999
予測クラス一覧
Python 3 Google Compute Engine バックエンド (GPU) に接続しました