

# Grad-CAM : 勾配ベースのローカリゼーションによるディープネットワークからの視覚的説明

Ramprasaath R. Selvaraju Michael Cogswell Abhishek Das Ramakrishna  
Vedantam Devi Parikh Dhruv Batra

要約 畳み込みニューラルネットワーク (CNN) ベースのモデルの大規模なクラスからの意思決定のための「視覚的な説明」を作成する手法を提案し、それらをより透明で説明しやすくします。

私たちのアプローチ-勾配加重クラスアクティベーションマッピング (Grad-CAM) は、最終的な畳み込み層に流れ込む任意のターゲット概念 (分類ネットワークの「犬」またはキャプションネットワークの単語のシーケンスなど) の勾配を使用して、粗いものを生成します概念を予測するための画像内の重要な領域を強調するローカリゼーションマップ。

以前のアプローチとは異なり、Grad-CAM はさまざまな CNN モデルファミリに適用できます: (1) 完全に接続された層を持つ CNN (VGG など)、(2) 構造化出力に使用される CNN (キャプションなど)、(3) CNN マルチモーダル入力 (視覚的な質問応答など) または強化学習を伴うタスクで使用されます。これらはすべて、アーキテクチャの変更や再トレーニングなしで行われます。Grad-CAM を既存のきめ細かい視覚化と組み合わせ、高解像度のクラス識別視覚化、ガイド付き Grad-CAM を作成し、それを画像分類、画像キャプション、および視覚的な質問応答 (VQA) モデルに適用します。ResNet ベースのアーキテクチャを含みます。

画像分類モデルのコンテキストでは、私たちの視覚化は、(a) これらのモデルの故障モードへの洞察を提供し (一見不合理な予測には合理的な説明があることを示しています)、(b) ILSVRC-15 の弱く監視された以前の方法よりも優れていますローカリゼーションタスクは、(c) 敵対的な摂動に対してロバストであり、(d) 基礎となるモデルにより忠実であり、(e) データセットのバイアスを特定することによってモデルの一般化を実現するのに役立ちます。

画像のキャプションと VQA の場合、私たちの視覚化は、注意に基づかないモデルでさえ、入力画像の識別領域をローカライズすることを学習することを示しています。

Grad-CAM を介して重要なニューロンを識別し、それをニューロン名と組み合わせる方法を考案します[4]モデル決定

のための tex-tual 説明を提供します。最後に、人間の研究を設計および実施して、Grad-CAM の説明が、ユーザーが深いネットワークからの予測に適切な信頼を確立するのに役立つかどうかを測定し、Grad-CAM が、訓練を受けていないユーザーが「より強力な」深いネットワークをうまく識別できるようにすることを示します。両方が同じ予測を行う場合でも、「弱い」もの。私たちのコードはで利用可能です <https://github.com/ランブ/grad-cam/>、CloudCV のデモと一緒に[2]<sup>1</sup>、とビデオ <youtu.be/COJUB9Izk6E>。

1 はじめに

畳み込みニューラルネットワーク (CNN) に基づくディープニューラルモデルは、画像分類から、さまざまなコンピュータビジョンタスクで前例のないブレイクスルーを可能にしました。[33, 24]、オブジェクト検出[21]、セマンティックセグメンテーション[37]画像のキャプション[55, 7, 18, 29]、視覚的な質問応答[3, 20, 42, 46]そして最近では視覚的な対話[11, 13, 12]と具体化された質問応答[10, 23]。一方これらのモデルは優れたパフォーマンスを可能にし、個々に直感的なコンポーネントへの分解性がないため、解釈が困難です[36]。その結果、今日の聡明なシステムが失敗すると、警告や説明なしに見事に恥ずかしそうに失敗することが多く、ユーザーは一貫性のない出力を見つめ、なぜシステムが何をしたのか疑問に思います。

解釈可能性が重要です。インテリジェントシステムへの信頼を構築し、日常生活への有意義な統合に向けて進むためには、予測内容を予測する理由を説明できる「透明な」モデルを構築する必要があることは明らかです。大まかに言えば、この透明性と説明能力は、人工知能 (AI) の進化の 3 つの異なる段階で役立ちます。まず、AI が人間よりも大幅に弱く、まだ確実に展開できない場合 (たとえば、視覚的な質問応答[3])、透明性と説明の目的は、故障モードを特定することです[1, 25]、それによって研究者が最も実り多い研究の方向性に彼らの努力を集中するのを助けます。第二に、AI が人間と同等であり、確実に展開可能である場合 (たとえば、画像分類[30]十分なデータでトレーニングされている)、目標はユーザーに適切な信頼と信頼を確立することです。第三に、AI が人間よりも大幅に強い場合 (例: チェスや囲碁[50])、説明の目標は機械教育にあります[28]

つまり、より良い意思決定を行う方法について人間に教える機械。

通常、正確さと単純さの間にはトレードオフが存在します。簡潔さまたは解釈可能性。従来のルールベースまたはエキスパートシステム[tems [26]は非常に解釈可能ですが、あまり正確ではありません（または堅牢ではありません）。各ステージが手作業で設計された分解可能なパイプラインは、個々のコンポーネントが自然で直感的な説明を前提としているため、より解釈しやすいと考えられています。深いモデルを使用することにより、より優れた抽象化（より多くのレイヤー）とより緊密な統合（エンドツーエンドのトレーニング）によってパフォーマンスを向上させる、解釈不可能なモジュールの解釈可能なモジュールを犠牲にします。最近導入されたディープ残差ネットワーク（ResNets）[24]は200層以上の深さであり、いくつかの困難なタスクで最先端のパフォーマンスを示しています。このような複雑さにより、これらのモデルの解釈が困難になります。そのため、深いモデルは解釈可能性と正確さの間のスペクトルを探索し始めています。

周ら。[59]最近、完全に接続された層を含まない制限されたクラスの画像分類 CNN によって使用される識別領域を識別するためのクラスアクティベーションマッピング（CAM）と呼ばれる手法を提案しました。本質的に、この作業はモデルの複雑さとパフォーマンスをトレードオフし、モデルの作業の透明性を高めます。対照的に、既存の最先端のディープモデルは、アーキテクチャを変更せずに解釈可能にするため、解釈可能性と精度のトレードオフを回避できます。私たちのアプローチは CAM の一般化です[59]そしてかなり広範囲の CNN モデルファミリーに適用可能：（1）完全に接続された層を持つ CNN（VGG など）、（2）構造化出力に使用される CNN（キャプションなど）、（3）タスクで使用される CNN アーキテクチャの変更や再トレーニングを必要とせずに、マルチモーダル入力（VQA など）または強化学習を使用します。

**何が良い視覚的説明になりますか？** 画像分類を検討する[14] -ターゲットカテゴリを正当化するためのモデルからの「適切な」視覚的説明は、（a）クラス識別（つまり、画像内のカテゴリをローカライズする）および（b）高解像度（つまり、きめ細かい詳細をキャプチャする）である必要があります。

図. 1 は、「tigercat」クラス（上）と「boxer」（犬）クラス（下）のいくつかの視覚化からの出力を示しています。ガイド付きバックプロパゲーションなどのピクセル空間勾配の視覚化[53]およびデコンボリューション[57]は高解像度であり、画像内のきめ細かい詳細を強調しますが、クラスを区別するものではありません（図. 1b および図. 1時間 非常に似ています）。

対照的に、CAM や提案された方法である勾配加重クラスアクティベーションマッピング（Grad-CAM）のようなローカリゼーションアプローチは、クラスを高度に識別します（「猫」の説明は「猫」領域のみを強調し、「犬」は強調しません）図の領域. 1c、およびその逆. 1i）。

両方の長所を組み合わせるために、既存のピクセル空間勾配視覚化を Grad-CAM と融合して、高解像度でクラス識別性のあるガイド付き Grad-CAM 視覚化を作成できることを示します。その結果、図に示すように、画像に複数の可能な概念の証拠が含まれている場合でも、関心のある決定に対応する画像の重要な領域が高解像度の詳細で視覚化されます。1d そして 1j。Guided Grad-CAM は、「虎猫」で視覚化すると、猫の領域を強調表示するだけでなく、猫の縞模様も強調表示します。これは、特定の種類の猫を予測するために重要です。

要約すると、私たちの貢献は次のとおりです。

- （1）アーキテクチャの変更や再トレーニングを必要とせずに、CNN ベースのネットワークの視覚的な説明を生成するクラス識別ローカリゼーション手法である Grad-CAM を紹介します。Grad-CAM のローカリゼーションを評価します（Sec. 4.1）、およびモデルへの忠実性（秒. 5.3）、ベースラインを上回っています。
- （2）Grad-CAM を既存の最高の分類に適用します-フィクション、キャプション（秒. 8.1）、および VQA（Sec. 8.2）モデル。画像分類については、私たちの視覚化により、現在の CNN の障害に関する洞察が得られます（秒. 6.1）、一見不合理な予測には合理的な説明があることを示しています。キャプションと VQA の場合、私たちの視覚化により、一般的な CNN + LSTM モデルは、接地された画像とテキストのペアでトレーニングされていないにもかかわらず、識別可能な画像領域のローカル化に驚くほど優れていることがよくあります。
- （3）解釈可能な Grad-CAM 視覚化が、データセットのバイアスを明らかにすることにより、障害モードの診断にどのように役立つかについての概念実証を示します。これは、一般化だけでなく、社会のアルゴリズムによってますます多くの決定が行われるため、公平で偏見のない結果にとっても重要です。
- （4）ResNets の Grad-CAM ビジュアライゼーションを紹介します[24]画像分類と VQA に適用されます（秒. 8.2）。
- （5）Grad-CAM のニューロンの重要性和[4]そしてモデル決定のためのテキストによる説明を入手する（秒. 7）。
- （6）私たちは人間の研究を行います（秒. 5） Guided Grad-CAM の説明はクラス識別的であり、人間が信頼を確立するのに役立つだけでなく、両方が同じ予測を行った場合でも、訓練を受けていないユーザーが「強い」ネットワークと「弱い」ネットワークをうまく区別するのに役立ちます。

**紙の組織：**残りの論文は次のように構成されています。セクション 3 では、Grad-CAM と Guided Grad-CAM のアプローチを提案します。セクション 4 と 5 では、Grad-CAM のローカリゼーション能力、クラス識別性、信頼性、忠実性を評価します。セクション 6 では、画像分類 CNN の診断や

データセットのバイアスの特定など、Grad-CAM の特定の使用例を示します。セクション 7 では、Grad-CAM を使用してテキストによる説明を取得する方法を提供します。セクション 8 では、Grad-CAM を視覚モデルと言語モデル（画像のキャプションと視覚的な質問応答（VQA））に適用する方法を示します。

## 2 関連作業

私たちの仕事は、CNN の視覚化、モデルの信頼性評価、および弱く監視されたローカリゼーションにおける最近の仕事を利用しています。

**CNN の視覚化。**以前の作品の数[51、53、57、19]「重要な」ピクセルを強調表示することにより、CNN 予測を視覚化しました（つまり、これらのピクセルの強度の変化が予測スコアに最も影響を与えます）。具体的には、Simonyan et al. [51]予測されたクラススコアの偏導関数をピクセル強度で視覚化し、ガイド付きバックプロパゲーション[53]およびデコンボリューション[57]質的な改善をもたらす「生の」グラデーションに変更を加えます。

これらのアプローチは[40]。製品にもかかわらずきめ細かい視覚化を行うため、これらのメソッドはクラスを区別しません。異なるクラスに関する視覚化はほぼ同じです（図を参照）1b そして 1 時間）。

他の視覚化方法では、画像を合成してネットワークユニットを最大限にアクティブ化します[51、16]または潜在的な表現を反転します-表現[41、15]。これらは高解像度でクラスを区別することができますが、単一の入力画像に固有ではなく、モデル全体を視覚化します。

**モデルの信頼性の評価。** 解釈可能性の概念に動機付けられている[36]そしてモデルへの信頼を評価する[47]、[]と同様の方法で Grad-CAM の視覚化を評価します。[47]人間の研究を通じて、自動化されたシステムを評価し、信頼を置くためのユーザーにとって重要なツールになり得ることを示しています。

**グラデーションベースの重要性の調整。** Selvaraju et al. [48]は、私たちの研究で導入された勾配ベースのニューロンの重要性を使用し、それを人間からのクラス固有のドメイン知識にマッピングして、新しいクラスの分類子を学習するアプローチを提案しました。将来の仕事では、セルバ Raju 等。[49]は、視覚と言語モデルを接地するために、勾配ベースの重要性を人間の注意マップに合わせるアプローチを提案しました。

**弱く監視されたローカリゼーション。** もう 1 つの関連する作業は、CNN のコンテキストでの弱く監視されたローカリゼーションです。このタスクでは、全体的な画像クラスラベルのみを使用して画像内のオブジェクトをローカライズします[8、43、44、59]。

私たちのアプローチに最も関連するのは、ローカリゼーションへのクラスアクティベーションマッピング（CAM）アプローチです[59]。このアプローチは、画像分類 CNN アーキテクチャを変更して、完全に接続されたレイヤーを畳み込みレイヤーとグローバル平均プーリングに置き換えます[34]、したがって、クラス固有の機能マップを実現しま

す。他の人は、グローバル最大プーリングを使用して同様の方法を調査しました[44]および log-sum-exp プーリング[45]。

CAM の欠点は、ソフトマックス層の直前に特徴マップが必要なことです。そのため、予測の直前に畳み込みマップ上でグローバル平均を実行する特定の種類の CNN アーキテクチャにのみ適用できます（つまり、conv 特徴マップのグローバル平均プーリング→ソフトマックス層）。このようなアーキテクチャは、一部のタスク（画像分類など）で一般的なネットワークと比較して精度が低い場合や、他のタスク（画像キャプションや VQA など）に適用できない場合があります。ネットワークアーキテクチャの変更を必要としない勾配信号を使用して特徴マップを組み合わせる新しい方法を紹介します。これにより、私たちのアプローチを、画像のキャプションや視覚的な質問応答など、既成の CNN ベースのアーキテクチャに適用できます。完全畳み込みアーキテクチャの場合、CAM は Grad-CAM の特殊なケースです。

他の方法は、入力画像の摂動を分類することによってローカリゼーションにアプローチします。Zeiler と Fergus [57]パッチを遮蔽し、遮蔽された画像を分類することによって入力を混乱させます。通常、これらのオブジェクトが遮蔽されると、関連するオブジェクトの分類スコアが低くなります。この原則は、[でのローカリゼーションに適用されます。5]。Oquab et al. [43]ピクセルを含む多くのパッチを分類し、これらのパッチごとのスコアを平均して、ピクセルのクラスごとのスコアを提供します。これらとは異なり、私たちのアプローチはワンショットでローカリゼーションを実現します。画像ごとに 1 回の順方向パスと部分的な逆方向パスのみが必要であるため、通常は 1 桁効率的です。最近の仕事では、張等。[58]対照的な限界勝率（c-MWP）を導入します。これは、識別領域を強調できる神経分類モデルのトップダウン注意をモデル化するための確率的勝者-テイク-オール定式化です。これは Grad-CAM よりも計算コストが高く、画像分類 CNN でのみ機能します。さらに、Grad-CAM は、定量的および定性的評価において c-MWP よりも優れています（セクションを参照）。4.1 および秒。D）。

## 3 卒業生-CAM

以前の多くの作品は、CNN のより深い表現がより高いレベルの視覚的構成をキャプチャすると主張しています[6、41]。さらに、畳み込み層は自然に空間を保持します完全に接続されたレイヤーで失われる情報。したがって、最後の畳み込みレイヤーには、高レベルのセマンティクスと詳細な空間情報の間で最良の約束があります。これらの層のニューロンは、画像内のセマンティッククラス固有の情報（オブジェクトパーツなど）を探します。Grad-CAM は、CNN の最後の畳み込み層に流入する勾配情報を使用して、関心のある特定の決定のために各ニューロンに重要度の値を割り当てます。私たちの手法は、ディープネットワークの任意のレイヤーでのアクティベーションを説明するために使用できるという点でかなり一般的ですが、この作業で



は、出力レイヤーの決定のみを説明することに焦点を当てます。

図に示すように、2、クラス識別を取得するために

ローカリゼーションマップ Grad-任意のクラスの幅と高さの CAM、最初にクラスのスコアの勾配を計算します

$L_{Grad-CAM}^c \in \mathbb{R}^{u \times v}$  uvcc、 $y^c$  (ソフトマックスの前)、に関して 畳み込み層の特徴マップのアクティベーション、つまり。

$A^k \frac{\partial y^c}{\partial A^k}$  逆流するこれらの勾配は、グローバル平均プールの次元 (およびインデックス付け)  $ij$  それぞれ) ニューロンの重要度の重みを取得するには:  $\alpha_k^c$

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

活性化に関して勾配を逆伝播している間の計算中、正確な計算は、勾配が伝播される最終的な畳み込み層まで、重み行列と活性化関数に関する勾配の連続する行列積になります。したがって、この重みは、A の下流のディープネットワークの部分的な線形化を表し、ターゲットクラスの特徴マップの「重要性」をキャプチャします。  $\alpha_k^c \alpha_{kc}^c$

フォワードアクティベーションマップの重み付けされた組み合わせを実行し、それに続いて ReLU を取得して、

$$L_{Grad-CAM}^c = ReLU \left( \sum_k \alpha_k^c A^k \right)$$

これにより、畳み込み特徴マップと同じサイズの粗いヒートマップが生成されることに注意してください (VGG の最後の畳み込み層の場合 [14 × 14] および AlexNet [33] ネットワーク) 3。マップの線形結合に ReLU を適用するのは、関心のあるクラスにプラスの影響を与える特徴、つまり、を増やすために強度を増やす必要があるピクセルにのみ関心があるためです。負のピクセルは、画像内の他のカテゴリに属する可能性があります。予想どおり、この ReLU がないと、ローカリゼーションマップは、目的のクラス以上のものを強調表示し、ローカリゼーションでパフォーマンスが低下することがあります。数字  $y^c$  1c、1f そして 1i、1l それぞれ「タイガーキャット」と「ボクサー (犬)」の Grad-CAM ビジュアライゼーションを表示します。アプレーション研究は秒で利用可能です。B。

一般に、画像分類 CNN によって生成されたクラススコアである必要はありません。それは、キャプションからの単語や質問への回答を含む、差別化可能な活動である可能性があります。  $y^c$

### 3.1 Grad-CAM は CAM を一般化します

このセクションでは、Grad-CAM とクラスアクティベーションマッピング (CAM) の関係について説明します [59]、そして正式に、Grad-CAM が CAM をさまざまな CNN ベー

スのアーキテクチャに一般化することを証明します。CAM は、グローバル平均プールの畳み込み特徴マップが直接ソフトマックスに供給される特定の種類のアーキテクチャを備えた画像分類 CNN のローカリゼーションマップを作成することを思い出してください。具体的には、最後から 2 番目のレイヤーで K 個の特徴マップを生成します。各要素には、インデックスが付けられます。つまり、特徴マップの場所でのアクティベーションを指します  $A^k \in \mathbb{R}^{u \times v}$   $i, j$   $A_{ij}^k(i, j)$   $A^k$ 。次に、これらの特徴マップは、グローバル平均プーリング (GAP) を使用して空間的にプールされ、線形変換されて、各クラスのスコアが生成されます。  $Y^c$

$$Y^c = \sum_k w_k^c \frac{1}{Z} \sum_i \sum_j A_{ij}^k$$

$F^k$  をグローバル平均プール出力として定義しましょう。

$$F^k = \frac{1}{Z} \sum_i \sum_j A_{ij}^k$$

CAM は次の方法で最終スコアを計算します

$$Y^c = \sum_k w_k^c \cdot F^k$$

ここで、は特徴マップとクラスを接続する重みです。取得した特徴マップに関するクラスのスコアの勾配を取得すると、  $w_k^c k^{th} c^{th} c(Y^c) F^k$

$$\frac{\partial Y^c}{\partial F^k} = \frac{\frac{\partial Y^c}{\partial A_{ij}^k}}{\frac{\partial F^k}{\partial A_{ij}^k}}$$

(の偏導関数を取る 4) に関して  $A_{ij}^k$ 、私たちはそれを見ることができます  $\frac{\partial F^k}{\partial A_{ij}^k} = \frac{1}{Z}$ 。これを (6) に代入すると、次のようになります。

$$\frac{\partial Y^c}{\partial F^k} = \frac{\partial Y^c}{\partial A_{ij}^k} \cdot Z$$

(5) から、次のようになります。したがって、  $\frac{\partial Y^c}{\partial F^k} = w_k^c$

$$w_k^c = Z \cdot \frac{\partial Y^c}{\partial A_{ij}^k}$$

すべてのピクセルにわたって (8) の両側を合計します。  $(i, j)$

$$\sum_i \sum_j w_k^c = \sum_i \sum_j Z \cdot \frac{\partial Y^c}{\partial A_{ij}^k}$$

に依存しないので、これを次のように書き直します  $Z w_k^c(i, j)$

$$Zw_k^c = Z \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k}$$

ご了承ください  $Z$  は特徴マップ（または）のピクセル数です。したがって、用語を並べ替えて、それを確認できます。 $Z = \sum_i \sum_j 1$

$$w_k^c = \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k}$$

視覚化中に正規化される比例定数まで、の式は **Grad-CAM** で使用されるものと同じです  $((1/Z)w_k^c \alpha_k^c)$ 。したがって、**Grad-CAM** は厳密な一般化です-CAM の一般化。この一般化により、画像のキャプションや VQA (Sec. 8.2)。

### 3.2 ガイド付き卒業生-CAM

**Grad-CAM** はクラスを識別し、関連する画像領域をローカライズしますが、ピクセル空間のグラデーション視覚化手法（ガイド付きバックプロパゲーション[ガイド付きバックプロパゲーション[53]、デコンボリューション[57]）。ガイド付きバックプロパゲーションは、**ReLU** レイヤーをバックプロパゲーションするときに負の勾配が抑制されている画像に関する勾配を視覚化します。直感的には、これはニューロンを抑制するピクセルではなく、ニューロンによって検出されたピクセルをキャプチャすることを目的としています。図を参照してください 1c、**Grad-CAM** は猫を簡単に見つけることができます。ただし、粗いヒートマップから、ネットワークがこの特定のインスタンスを「タイガーキャット」として予測する理由は不明です。両方の最良の側面を組み合わせるために、要素ごとの乗算を介してガイド付きバックプロパゲーションと **Grad-CAM** の視覚化を融合します ( $L_{Grad-CAM}^c$  最初に、双一次内挿を使用して入力画像の解像度にアップサンプリングされます)。図. 2 左下はこの融合を示しています。この視覚化は、高解像度（対象のクラスが「虎猫」の場合、縞模様、先のとがった耳や目などの重要な「虎猫」の特徴を識別します）とクラス識別（「虎猫」を強調表示しますが、'ボクサー（犬）'）。ガイド付きバックプロパゲーションをデコンボリューションに置き換えると同様の結果が得られますが、デコンボリューションの視覚化にはアーティファクトがあり、ガイド付きバックプロパゲーションのノイズは一般的に少ないことがわかりました。

### 3.3 反事実的説明

**Grad-CAM** にわずかな変更を加えることで、ネットワークの予測を変更させる領域のサポートを強調する説明を取得できます。結果として、これらの地域で発生する概念を排除すると、モデルはその予測についてより自信を持つようになります。この説明モダリティを反事実的説明と呼びます。

具体的には、畳み込み層の特徴マップに関して（クラスのスコア）の勾配を否定します。したがって、重要度の重みは次のようになります  $y^c c A \alpha_k^c$

$$w_k^c = \frac{1}{Z} \sum_i \sum_j - \frac{\partial y^c}{\partial A_{ij}^k}$$

（のように 2）、フォワードアクティベーションマップの加重和を加重で取得し、それに続いて **ReLU** を実行して、図に示すように反事実的説明を取得します。  $A \alpha_k^c$  3.

## 4 Grad-CAM のローカリゼーション能力の評価

### 4.1 弱教師ありローカリゼーション

このセクションでは、画像分類のコンテキストで **Grad-CAM** のローカリゼーション機能を評価します。ImageNet ローカリゼーションの課題[14]は、分類ラベルに加えて境界ボックスを提供するアプローチを必要とします。分類と同様に、評価は上位 1 と上位 5 の両方の予測カテゴリに対して実行されます。

画像が与えられると、最初にネットワークからクラス予測を取得し、次に予測されたクラスごとに **Grad-CAM** マップを生成し、最大強度の 15% のしきい値でそれらを 2 値化します。これにより、ピクセルのセグメントが接続され、最大の単一セグメントの周囲に境界ボックスが描画されます。これは弱く監視されたローカリゼーションであることに注意してください。トレーニング中にモデルがバウンディングボックスの注釈にさらされることはありませんでした。

既製の事前トレーニング済み VGG-16 を使用して **Grad-CAM** ローカリゼーションを評価します[52]、AlexNet [33] および GoogleNet [54]（カフェから入手[27]動物園）。ILSVRC-15 の評価に続いて、表の値セットで上位 1 と上位 5 の両方のローカリゼーションエラーを報告します。1。**Grad-CAM** ローカリゼーションエラーは、c-MWP によって達成されるエラーよりも大幅に優れています[58] および Simonyan et al. [51]、グラブカットを使用して画像を後処理します-年齢空間の勾配をヒートマップに変換します。VGG-16 用の **Grad-CAM** は、CAM よりも優れたトップ 1 ローカリゼーションエラーも実現します[59]、モデルアーキテクチャの変更が必要であるため、再トレーニングが必要であり、それによって分類エラーが悪化します（トップ 1 が 2.98% 悪化）が、**Grad-CAM** は分類パフォーマンスを約束しません。

### 4.2 弱教師ありセグメンテーション

セマンティックセグメンテーションには、画像内の各ピクセルにオブジェクトクラス（または背景クラス）を割り当てるタスクが含まれます。やりがいのある作業であるため、これには高価なピクセルレベルの注釈が必要です。弱教師

ありセグメンテーションのタスクには、画像レベルの注釈のみを使用してオブジェクトをセグメント化することが含まれます。これは、画像分類データセットから比較的安価に取得できます。最近の仕事では、Kolesnikov 等。[32]弱教師あり画像セグメンテーションモデルをトレーニングするための新しい損失関数を導入しました。それらの損失関数は、3つの原則に基づいています。1) 弱いローカリゼーションキューでシードし、セグメンテーションネットワークがこれらのキューに一致するように促します。2) 画像内で発生する可能性のあるクラスに関する情報に基づいて、オブジェクトシードを妥当なサイズの領域に拡張します。3) セグメンテーションをオブジェクト境界に制約し、トレーニング時にすでに不正確な境界の問題を軽減します。彼らは、上記の3つの損失からなる、提案された損失関数がより良いセグメンテーションにつながることを示しました。

ただし、それらのアルゴリズムは弱いローカリゼーションシードの選択に敏感であり、それがないとネットワークはオブジェクトを正しくローカライズできません。彼らの研究では、フォアグラウンドクラスを弱くローカライズするためのオブジェクトシードとして使用される VGG-16 ベースのネットワークからの CAM マップを使用しました。CAM マップを標準の VGG-16 ネットワークから取得した Grad-CAM に置き換え、PASCAL VOC 2012 セグメンテーションタスクで 49.6 (CAM で取得した 44.6 と比較して) の Intersection over Union (IoU) スコアを取得しました。図. 4 いくつかの定性的な結果を示しています。

#### 4.3 ポインティングゲーム

張ら。[58]は、シーン内のターゲットオブジェクトをローカライズするためのさまざまな視覚化方法の識別性を評価するためのポインティングゲーム実験を導入しました。彼らの評価プロトコルは、最初にグラウンドトゥールズオブジェクトラベルを使用して各視覚化手法をキューに入れ、生成されたヒートマップ上で最大にアクティブ化されたポイントを抽出します。次に、ポイントがターゲットオブジェクトカテゴリの注釈付きインスタンスの1つ内にあるかどうかを評価し、それによってヒットまたはミスとしてカウントします。

次に、ローカリゼーションの精度は次のように計算されます。
$$Acc = \frac{\#Hits}{\#Hits + \#Misses}$$
ただし、この評価では、視覚化手法の精度のみを測定します。プロトコルを変更してリコールも測定します—CNN 分類器から上位5クラスの予測のローカリゼーションマップを計算します<sup>4</sup>マップ内で最大にアクティブ化されたポイントがしきい値を下回っている場合、つまり視覚化によって存在しない予測が正しく拒否された場合、モデルからの上位5つの予測のいずれかを拒否する追加オプションを備えたポインティングゲームセットアップを使用してそれら进行评估します。グラウンドトゥールズカテゴリ、それはヒットとしてそれを取得します。Grad-CAM が c-MWP よりも優れていることがわかりました[58]か

りの差で (70.58% 対 60.30%)。c-MWP を比較する定性的な例[58]および Grad-CAM on は、セクション 2 にあります。D<sup>5</sup>。

#### 5 視覚化の評価

このセクションでは、モデル予測へのアプローチの解釈可能性と忠実度のトレードオフを理解するために実施した人間の研究と実験について説明します。私たちの最初の人間の研究は、私たちのアプローチの大前提を評価します—Grad-CAM の視覚化は、以前の手法よりもクラスの識別力がありますか？それを確立したら、エンドユーザーが視覚化されたモデルを適切に信頼できるかどうかを理解することになります。これらの実験では、PASCAL VOC 2007 トレインで微調整された VGG-16 と AlexNet、および val で評価された視覚化を比較します。

##### 5.1 階級差別の評価

Grad-CAM がクラス間の区別に役立つかどうかを測定するために、PASCAL VOC 2007 値セットから画像を選択します。これには、正確に2つの注釈付きカテゴリが含まれ、それぞれに視覚化が作成されます。VGG-16 と AlexNetCNN の両方について、これらの各メソッドのデコンボリューション、ガイド付き逆伝播、および Grad-CAM パージョン (デコンボリューション Grad-CAM およびガイド付き Grad-CAM) の4つの手法を使用してカテゴリ固有の視覚化を取得します。これらの視覚化を Amazon Mechanical Turk (AMT) の43人のワーカーに示し、「2つのオブジェクトカテゴリのどちらが画像に描かれていますか？」と尋ねます。(図に示されています。5)。

直感的には、適切な予測の説明は、関心のあるクラスの識別可能な視覚化を生成するものです。実験は、90の画像カテゴリペア (つまり、360の視覚化) に対して4つの視覚化すべてを使用して実施されました。各画像について9匹のラットを収集し、グラウンドトゥールズに対して評価し、平均して表の精度を得ました。2. Guided Grad-CAM を表示すると、人間の被験者は、61:23%のケースで視覚化されているカテゴリを正しく識別できます (Guided Backpropagation の 44:44%と比較して、Grad-CAM は人間のパフォーマンスを 16:79%向上させます)。同様に、Grad-CAM は、Deconvolution をよりクラス識別的にするのに役立つこともわかりました (53:33% ! 60:37%から)。Guided Grad-CAM は、すべての方法の中で最高のパフォーマンスを発揮します。

興味深いことに、私たちの結果は、デコンボリューションがガイド付きバックプロパゲーションよりもクラス識別的であることを示しています (53:33% 対 44:44%) が、ガイド付きバックプロパゲーションはより審美的に満足のいくものです。私たちの知る限りでは、私たちの評価はこの微妙な違いを最初に定量化したものです。



## 5.2 信頼の評価

2つの予測の説明を前提として、どちらがより信頼できると思われるかを評価します。AlexNetとVGG-16を使用して、ガイド付きバックプロパゲーションとガイド付き Grad-CAM の視覚化を比較します。VGG-16 は AlexNet よりも信頼性が高く、精度は 79:09 mAP (vs. 69:20 mAP) であることがわかっています。PASCAL 分類。視覚化されているモデルの精度から視覚化の有効性を区別するために、両方のモデルがグラウンドトゥールースと同じ予測を行った場合のみを考慮します。AlexNetとVGG-16からの視覚化、および予測されたオブジェクトカテゴリを考慮して、54 人の AMT ワーカーに、モデルの信頼性を明らかに信頼性の高い/低い (+/- 2) のスケールで評価するように指示しました。)、わずかに信頼性が高い/低い (+/- 1)、および同等に信頼性がある (0)。このインターフェースを図 1 に示します。5。バイアスを排除するために、VGG-16 と AlexNet は、ほぼ等しい確率で「モデル 1」に割り当てられました。驚くべきことに、表に見られるように。2、両方のモデルが同じ予測を行っているにもかかわらず、人間の被験者は予測の説明から簡単に、より正確な分類器 (AlexNet 上の VGG-16) を識別できることがわかりました。Guided Backpropagationを使用すると、人間は VGG-16 に平均スコア 1:00 を割り当てます。これは、AlexNet よりもわずかに信頼性が高いことを意味します。Guided Grad-CAM は 1:27 の高いスコアを達成し、VGG-16 は明らかに信頼性が高くなります。したがって、私たちの視覚化は、個々の予測の説明に基づいて、ユーザーがより一般化するモデルに信頼を置くのに役立ちます。

## 5.3 忠実性と解釈可能性

モデルに対する視覚化の忠実性は、モデルによって学習された機能を正確に説明する能力です。当然のことながら、視覚化の解釈可能性と忠実度の間にはトレードオフが存在します。より忠実な視覚化は通常、解釈可能性が低く、その逆も同様です。実際、完全に忠実な説明はモデルの説明全体であり、深いモデルの場合は解釈できない視覚化するのが容易ではないと主張することができます。前のセクションで、視覚化が合理的に解釈できることを確認しました。ここで、基礎となるモデルにどれだけ忠実であるかを評価します。1つの期待は、説明が局所的に正確である必要があることです。つまり、入力データポイントの近くでは、説明はモデルに忠実である必要があります[47]。

比較のために、現地の忠実度が高い参考説明が必要です。このような視覚化の 1 つの明白な選択は、画像のオクルージョンです[57]、入力画像のパッチがマスクされたときの CNN スコアの差を測定します。興味深いことに、CNN スコアを変更するパッチは、Grad-CAM と Guided Grad-CAM が高強度を割り当て、ランク相関 0 : 254 と 0 : 261 を達成するパッチでもあります (対 0 : 168, 0 : 220, 0 : 208 を達成) ガイド付きバックプロパゲーションによると、それぞれ c-MWP と CAM) は、PASCAL 2007val セットで 2510 枚

以上の画像を平均しました。これは、Grad-CAM が以前の手法と比較して元のモデルにより忠実であることを示しています。ローカリゼーション実験と人間の研究を通して、Grad-CAM の視覚化がより解釈しやすく、オクルージョンマップとの相関を通じて、Grad-CAM がモデルにより忠実であることがわかります。

## 6 Grad-CAM を使用した画像分類 CNN の診断

このセクションでは、Imagenet で事前トレーニングされた VGG-16 のコンテキストで、画像分類 CNN の障害モードの分析、敵対的ノイズの影響の理解、データセットのバイアスの特定と除去における Grad-CAM の使用についてさらに説明します。

### 6.1 VGG-16 の故障モードの分析

ネットワークがどのような間違いを犯しているかを確認するために、最初にネットワーク (VGG-16) が正しく分類できない例のリストを取得します。これらの誤分類された例では、Guided Grad-CAM を使用して、正しいクラスと予測されたクラスの両方を視覚化します。図に見られるように。6、一部の失敗は、ImageNet 分類に固有のあいまいさによるものです。また、一見不合理に見える予測には合理的な説明があることがわかります。これも HOGgles で行われた観察です[56]。Guided Grad-CAM の視覚化が他の方法よりも優れている主な利点は、その高解像度とクラス識別機能により、これらの分析が容易に可能になることです。

### 6.2 VGG-16 に対する敵対的ノイズの影響

Goodfellow et al. [22]は、敵対的な例に対する現在の深いネットワークの脆弱性を示しました。これは、入力画像のわずかな知覚できない摂動であり、ネットワークをだまして、高い信頼性でそれらを誤分類させます。ImageNet で事前トレーニングされた VGG-16 モデルの敵対的な画像を生成し、高い確率 (> 0 : 9999) を割り当てます。画像に存在しないカテゴリと、存在するカテゴリの確率が低い。次に、存在するカテゴリの Grad-CAM 視覚化を計算します。図に示すように。7、ネットワークがこれらのカテゴリ (「タイガーキャット」と「ボクサー」) がないことを確信しているにもかかわらず、Grad-CAM visualizations はそれらを正しくローカライズできます。これは、Grad-CAM が敵対的なノイズに対してかなり堅牢であることを示しています。

### 6.3 データセットのバイアスの特定

このセクションでは、Grad-CAM の別の使用法を示します。トレーニングデータセットのバイアスを特定して削減します。バイアスのかかったデータセットでトレーニングされたモデルは、実際のシナリオに一般化されない場合があります。さらに悪いことに、バイアスとステレオタイプ (性別、人種、年齢など) が永続する場合があります。「医師」と「看護師」の二項分類タスク用に、ImageNet で事前トレーニングされた VGG-16 モデルを微調整します。人気のあ

る画像検索エンジンからの上位 250 の関連画像（クラスごと）を使用して、トレーニングと検証の分割を作成しました。また、テストセットは、2つのクラス間での性別の分布のバランスが取れるように制御されました。トレーニングされたモデルは優れた検証精度を達成しますが、一般化はうまくいきません（82%のテスト精度）。

モデル予測の Grad-CAM 視覚化（赤いボックスを参照）<sup>6</sup> 図の中央の列の領域。<sup>8</sup> モデルは、看護師と医師を区別するために人の顔/髪型を見る方法を学習し、したがって性別のステレオタイプを学習したことを明らかにしました。実際、このモデルでは、数人の女性医師を看護師に、男性看護師を医師に誤分類していました。明らかに、これには問題があります。画像検索結果は性別に偏っていたことが判明しました（医師の画像の 78%が男性で、看護師の画像の 93%が女性でした）。

Grad-CAM の視覚化から得られたこれらの直感を通じて、以前と同じクラスあたりの画像数を維持しながら、男性看護師と女性医師の画像を追加することで、トレーニングセットのバイアスを減らしました。再トレーニングされたモデルは、一般化が向上するだけでなく（90%のテスト精度）、適切な領域も確認します（図の最後の列）。<sup>8</sup>。この実験は、Grad-CAM がデータセットのバイアスを検出して除去するのに役立つという概念実証を示しています。これは、一般化を改善するだけでなく、社会でよりアルゴリズム的な決定が行われるため、公正で倫理的な結果にとっても重要です。

## Grad-CAM による 7 つのテキストによる説明

方程式。（<sup>1</sup>）は、特定のクラスの畳み込み層内の各ニューロンのニューロンの重要度を取得する方法を提供します。文献に提示された仮説があります[60、57] そのニューロンは概念「検出器」として機能します。より高いポジティブニューロンの重要度の値は、その概念の存在がクラススコアの増加につながることを示し、負の値が高いほど、その概念の不在がクラススコアの増加につながることを示します。

この直感を踏まえて、テクスチャルな説明を生成する方法を調べてみましょう。最近の仕事では、バウ等。[4]は、訓練されたネットワークの任意の畳み込み層のニューロンに自動的に名前を付けるアプローチを提案しました。これらの名前は、ニューロンが画像内で探す概念を示しています。彼らのアプローチを使用します。最初に、最後の畳み込み層のニューロン名を取得します。次に、クラス固有の重要度スコアに基づいて、上位 5 ニューロンと下位 5 ニューロンを並べ替えて取得します。これらのニューロンの名前は、テキストの説明として使用できます。 $\alpha_k$

図。9 Places365 データセットでトレーニングされた画像分類モデル（VGG-16）の視覚的およびテキストによる説明の例をいくつか示します[61]。（a）では、（<sup>1</sup>）クラス

「書店」を示す本や棚などの直感的な概念を探します。また、ネガティブに重要なニューロンは、「書店」の画像にはない、空、道路、水、車などの概念を探すことに注意してください。（b）では、「滝」を予測するために、視覚的説明とテキストによる説明の両方で、「滝」の画像を説明する「水」と「層化」が強調表示されています。（e）は、ロープがない場合にネットワークが「ロープブリッジ」を予測したための誤分類による障害ケースですが、それでも重要な概念（水と橋）は予測されたクラスを示しています。（f）では、Grad-CAM が紙のドアと階段を正しく見て「エレベーターのドア」を予測しているのに、ドアを検出するニューロンが IoU しきい値を超えていませんでした。<sup>7</sup> 0.05（ニューロン名のノイズを抑制するために選択）であるため、テキストによる説明の一部ではありません。より定性的な例は、セクションで見つけることができます。F。

## 画像キャプションおよび VQA 用の 8Grad-CAM

最後に、Grad-CAM を画像キャプションなどの視覚および言語タスクに適用します[7、29、55]および視覚的質問回答-ing（VQA）[3、20、42、46]。Grad-CAM は、予測の変化によって目立って変化しないベースラインの視覚化と比較して、これらのタスクの解釈可能な視覚的説明につながるということがわかりました。既存の視覚化手法は、クラスを区別しない（ガイド付きバックプロパゲーション、デコンボリューション）か、これらのタスク/アーキテクチャに使用できないか、またはその両方（CAM、c-MWP）であることに注意してください。

### 8.1 画像のキャプション

このセクションでは、Grad-CAM を使用して画像キャプションモデルの空間サポートを視覚化します。公開されている neuraltalk2 の上に Grad-CAM を構築します<sup>8</sup> 実装[31]画像に微調整された VGG-16CNN と LSTM ベースの言語モデルを使用します。このモデルには明示的な注意メカニズムがないことに注意してください。キャプションを指定して、CNN の最後の畳み込み層（VGG-16 の場合は conv5\_3）の対数確率 wrt 単位の勾配を計算します。16 セクションで説明されているように、Grad-CAM ビジュアライゼーションを生成します。<sup>3</sup> 図を参照してください。<sup>10a</sup>。最初の例では、生成されたキャプションの Grad-CAM マップは、比較的小さいサイズにもかかわらず、人々の両方のすべての出現をローカライズします。次の例では、Grad-CAM はピザと男性を正しく強調表示しますが、キャプションに「女性」が記載されていないため、近くの女性を無視します。より多くの例は秒にあります。C。

密なキャプションとの比較。ジョンソン等。[29]最近、特定の画像内の顕著な領域を共同でローカライズしてキャプションを付けるシステムを必要とする Dense Captioning（DenseCap）タスクが導入されました。彼らのモデルは、関心領域の境界ボックスを生成する完全畳み込みローカライゼーションネットワーク（FCLN）と、関連するキャプショ



ンを生成する LSTM ベースの言語モデルで構成されています。これらはすべて 1 回のフォワードパスで行われます。DenseCap を使用して、画像ごとに 5 つの地域固有のキャプションを生成し、グラウンドトゥールズバウンディングボックスを関連付けます。全画像キャプションモデル (neuraltalk2) の Grad-CAM は、領域キャプションが生成された境界ボックスをローカライズする必要があります。これを図に示します。10b。ボックスの内側と外側の平均活性化の比率を計算することにより、これを定量化します。キャプションが生成された領域への注意が強いことを示すため、比率が高いほど優れています。画像全体を均一に強調表示すると、ベースライン比が 1 : 0 になりますが、Grad-CAM では達成されます。高解像度の詳細を追加すると、(Guided Backpropagation) のベースラインが改善され、(Guided Grad-CAM) で最適なローカリゼーションが得られます。したがって、Grad-CAM は、全体的なキャプションモデルがバウンディングボックスアノテーションでトレーニングされていなくても、DenseCap モデルが記述する画像内の領域をローカライズできます。3.27 ± 0.182.32 ± 0.086.38 ± 0.99

### 8.1.1 キャプションの個々の単語の Grad-CAM

私たちの実験では、Show and Tell モデルを使用します[55] Inception から取得した視覚的表現を介して微調整することなく、MSCOCO で事前トレーニングされています[54] 建築。グラウンドトゥールズキャプション内の個々の単語の Grad-CAM マップを取得するために、対応するタイムステップで各視覚単語をワンホットエンコードし、式 (1) を使用してニューロンの重要度スコアを計算します。(1) そして、式 (1) を使用して畳み込み特徴マップと結合します。(2)。

人間の注意との比較<person>などのオブジェクトカテゴリを["child", "man", "woman", ...]などの潜在的なきめ細かいラベルのリストにマッピングするオブジェクトカテゴリから単語へのマッピングを手動で作成しました。COCO キャプションに存在する合計 830 の視覚的な単語を 80 の COCO カテゴリにマッピングします。次に、一致する単語のこのサブセットに対する人間の注意として、80 のカテゴリのセグメンテーション注釈を使用します。

次に、[からのポインティング評価を使用します58]。キャプションからの視覚的な単語ごとに、Grad-CAM マップを生成し、最大にアクティブ化されたポイントを抽出します。次に、ポイントが対応する COCO カテゴリの人間の注意マップセグメント内にあるかどうかを評価し、それによってヒットまたはミスとしてカウントします。次に、ポインティング精度はとして計算されます。この実験は、COCO データセットからランダムにサンプリングされた 1000 枚の画像に対して実行され、30.0%の精度が得られます。いくつかの定性的な例を図に示します。 $Acc = \frac{\#Hits}{\#Hits + \#Misses}$  11。

### 8.2 視覚的な質問応答

典型的な VQA パイプライン[3、20、42、46]画像を処理するための CNN と質問のための RNN 言語モデルで構成されます。画像と質問の表現は、通常 1000 通りの分類 (1000 は回答スペースのサイズ) で回答を予測するために融合されます。これは分類の問題なので、答えを選びます (y<sup>c</sup>3) そして、そのスコアを使用して、画像上で Grad-CAM の視覚化を計算し、答えを説明します。視覚的コンポーネントとテキストコンポーネントの両方を含むタスクの複雑さにもかかわらず、(Lu らの VQA モデルの説明[38]) 図に記載されています。12 驚くほど直感的で有益です。節のように、オクルージョンマップとの相関を介して Grad-CAM のパフォーマンスを定量化します。5.3. Grad-CAM は (オクルージョンマップを使用して) のランク相関を達成しますが、Guided Backpropagation はを達成し、Grad-CAM 視覚化の忠実度が高いことを示します。0.60 ± 0.0380.42 ± 0.038

人間の注意との比較。Das et al. [9] VQA データセットのサブセットの人間の注意マップを収集しました[3]。これらのマップは、視覚的な質問に答えるために人間が画像を見る場所で強度が高くなっています。人間の注意マップは、[からの VQA モデルの Grad-CAM 視覚化と比較されます。38] [からの 1374val 質問画像 (QI) ペア 3] [のように順位相関評価プロトコルを使用する 9]。Grad-CAM と人間の注意マップの相関は 0.136 であり、偶然またはランダムな注意マップ (ゼロ相関) よりも高くなっています。これは、接地された画像とテキストのペアについてトレーニングを受けていないにもかかわらず、注意を払っていないことを示しています。ベースの CNN + LSTM ベースの VQA モデルは、特定の回答を予測するための領域のローカライズに驚くほど優れています。

ResNet ベースの VQA モデルを共同注意を払って視覚化します。Lu et al. [39] 200 層の ResNet を使用する[24]画像をエンコードし、質問と画像の階層的注意メカニズムを共同で学習します。図。12b は、このネットワークの Grad-CAM 視覚化を示しています。ResNet のより深い層を視覚化すると、ほとんどの隣接する層で Grad-CAM に小さな変化が見られ、次元削減を伴う層間で大きな変化が見られます。ResNet のその他の視覚化については、セクション 2 を参照してください。G。私たちの知る限り、ResNet ベースのモデルからの決定を視覚化したのは私たちが初めてです。

### 9 結論

この作業では、視覚的な説明を作成することにより、CNN ベースのモデルをより透明にするための新しいクラス識別局所化手法である勾配加重クラスアクティベーションマッピング (Grad-CAM) を提案しました。さらに、Grad-CAM のローカリゼーションを既存の高解像度の視覚化手法と組み合わせて、高解像度でクラスを区別するガイド付き Grad-CAM の視覚化という両方の長所を活用しました。私たちの視覚化は、解釈可能性と元のモデルへの忠実さという両方の軸で既存のアプローチよりも優れています。広範

な人間の研究は、私たちの視覚化がクラスをより正確に区別し、分類器の信頼性をよりよく明らかにし、データセットのバイアスを特定するのに役立つことを明らかにしています。さらに、**Grad-CAM** を介して重要なニューロンを識別する方法を考案し、モデル決定のためのテキストによる説明を取得する方法を提供します。最後に、画像分類、画像キャプション、視覚的な質問への回答などのタスクのためのさまざまな既製のアーキテクチャへの **Grad-CAM** の幅広い適用性を示します。真の **AI** システムは、インテリジェントであるだけでなく、人間がそれを信頼して使用するための信念と行動について推論できる必要があると私たちは信じています。今後の作業には、強化学習、自然言語処理、ビデオアプリケーションなどのドメインの深いネットワークによって行われた決定の説明が含まれます。真の **AI** システムは、インテリジェントであるだけでなく、人間がそれを信頼して使用するための信念と行動について推論できる必要があると私たちは信じています。今後の作業には、強化学習、自然言語処理、ビデオアプリケーションなどのドメインの深いネットワークによって行われた決定の説明が含まれます。真の **AI** システムは、インテリジェントであるだけでなく、人間がそれを信頼して使用するための信念と行動について推論できる必要があると私たちは信じています。今後の作業には、強化学習、自然言語処理、ビデオアプリケーションなどのドメインの深いネットワークによって行われた決定の説明が含まれます。

## 10 謝辞

この作品は、DB と DP への NSF CAREER 賞、DB と DP への DARPA XAI 賞、DP と DB への ONR YIP 賞、DB への ONR Grant N00014-14-1-0679、DP へのスローンフェローシップ、ARO によって部分的に資金提供されました。 DB と DP への YIP 賞、Paul G. Allen Family Foundation からの DP への AllenDistinguished Investigator 賞、DB と DP への ICTAS Junior Faculty 賞、DP と DB への Google Faculty Research Awards、DP と DB への Amazon Academic Research Awards、AWS in Education Research は DB に助成し、NVIDIAGPU は DB に寄付します。ここに含まれる見解と結論は著者のものであり、米国政府またはスポンサーの公式の方針または承認を必ずしも表すものとして解釈されるべきではありません。

# 付録

## 付録の概要

付録では、以下を提供します。

- I-私たちのデザインの選択を評価するアブレーション研究
- II -画像分類、キャプション、VQA のより定性的な例

- III -ポインティングゲームの評価手法の詳細
- IV-既存の視覚化技術との定性的比較-niques
- V-テキストによる説明のより定性的な例

B アブレーション研究

Grad-CAM 視覚化を計算するための設計の選択を調査および検証するために、いくつかのアブレーション研究を実行します。これには、ネットワーク内のさまざまなレイヤーの視覚化、(2)、さまざまなタイプのグラジエント (ReLU バックワードパスの場合)、およびさまざまなグラジエントプーリング戦略を分析します。

1.さまざまなレイヤーの Grad-CAM

AlexNet と VGG-16 のさまざまな畳み込み層での「tiger-cat」クラスの Grad-CAM 視覚化を示します。予想通り、図からの結果。13 以前のコンボリューションレイヤーに移動すると、ローカリゼーションが徐々に悪化することを示しています。これは、後の畳み込み層が、受容野が小さく、局所的な特徴のみに焦点を当てている前の層よりも空間情報を保持しながら、高レベルの意味情報をより適切にキャプチャするためです。

2.デザインの選択

方法	Top-1Loc エラー
Grad-CAM	59.65
Eq.1 の ReLU なしの Grad-CAM	74.98
絶対グラジエントを使用した Grad-CAM	58.19
GMP グラジエントを使用した Grad-CAM	59.96
DeconvReLU を使用した Grad-CAM	83.95
ガイド付き ReLU を備えた Grad-CAM	59.14

表 3：アブレーションの ILSVRC-15val でのローカリゼーションの結果。この評価は 10 作物を超えていますが、視覚化は単一作物であることに注意してください。

ILSVRC-15 値セットのトップ 1 ローカリゼーションエラーを介して、さまざまな設計の選択を評価します[14]。表を参照してください。3。

2.1。（の ReLU の重要性 3）。

ReLU の削除（3）エラーが 15.3%増加します。Grad-CAM の負の値は、複数発生するクラス間の混乱を示します。

2.2。グローバル平均プーリングとグローバル最大プーリング

畳み込み層への入力勾配である GlobalAverage Pooling（GAP）の代わりに、Global MaxPooling を試しました。



(GMP)。GMPを使用すると、Grad-CAMのローカリゼーション能力が低下することがわかります。例を図に示します。15 未満。これは、max が平均化された勾配と比較してノイズに対して統計的にロバスト性が低いという事実が原因である可能性があります。

## 2.3. Grad-CAM に対する異なる ReLU の影響

Guided-ReLU を試してみます[53]および Deconv-ReLU [57] ReLU のバックワードパスへの変更として。

Guided-ReLU : Springenberg et al. [53]ガイド付きバックプロパゲーションが導入されました。ReLU のバックワードパスは、正の活性化の領域にのみ正の勾配を渡すように変更されています。この変更を Grad-CAM の計算に適用すると、図 1 に示すように、クラス識別能力が低下します。16、ただし、表に示すように、ローカリゼーションのパフォーマンスはわずかに向上します。3。

Deconv-ReLU : デコンボリューション[57]、Zeiler と Fergus は、正の勾配のみを通過させるように ReLU の逆方向パスに変更を導入しました。この修正を Grad-CAM の計算に適用すると、結果が悪化します（図。16）。これは、負の勾配もクラスの識別性に関する重要な情報を持っていることを示しています。

## C 視覚および言語タスクの定性的結果

このセクションでは、画像分類、画像キャプション、および VQA のタスクに適用された Grad-CAM および GuidedGrad-CAM のより定性的な結果を提供します。

### 1.画像分類

Grad-CAM と GuidedGrad-CAM を使用して、特定の予測をサポートする画像の領域を視覚化します。図に報告された結果。17 VGG-16 に対応[52] ImageNet でトレーニングされたネットワーク。

図。 17 COCO からランダムにサンプリングされた例を示しています[35]検証セット。COCO 画像は通常、画像ごとに複数のオブジェクトを持ち、Grad-CAM の視覚化は、モデルの予測をサポートするための正確なローカリゼーションを示します。

Guided Grad-CAM は、小さなオブジェクトをローカライズすることもできます。たとえば、私たちのアプローチは、予測されたクラス「トーチ」を正しくローカライズします（図。17.a）そのサイズと画像内の奇妙な位置にもかかわらず。私たちの方法もクラス識別的であり、人気のある ImageNet カテゴリ「犬」が画像に存在する場合でも、「便座」にのみ注意を向けます（図。17.e）。

また、ILSVRC13 検出値セットからの画像について、Grad-CAM、ガイド付きバックプロパゲーション（GB）、デコンボリューション（DC）、GB + Grad-CAM（ガイド付き Grad-CAM）、DC + Grad-CAM（デコンボリューション Grad-CAM）を視覚化しました。それぞれに少なくとも2つの一意のオブジェクトカテゴリがあります。上記のクラスの視覚化は、次のリンクにあります。

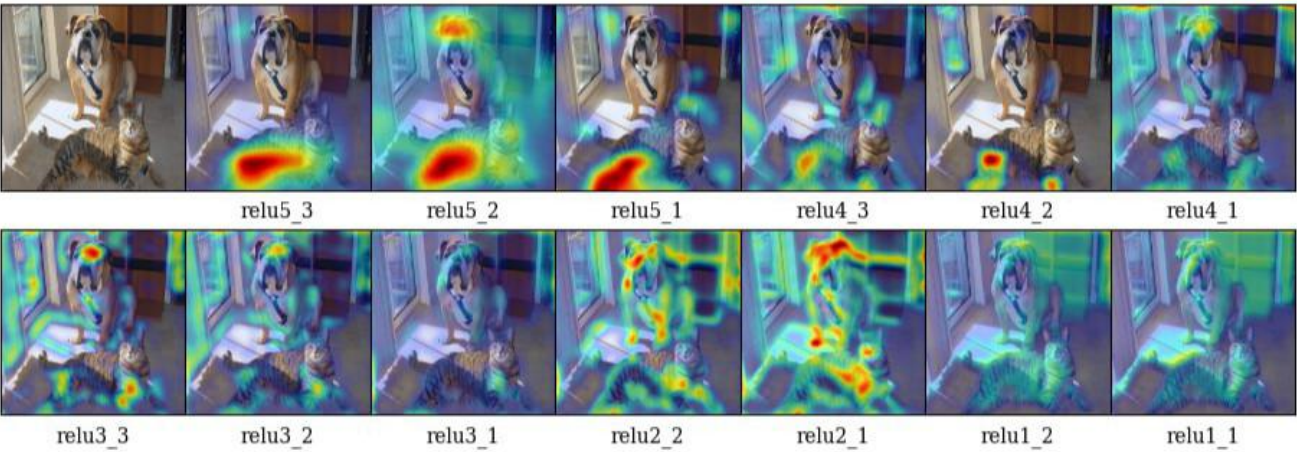


図 13 : 「タイガーキャット」クラスのさまざまな畳み込み層での Grad-CAM。この図は、CNN (VGG16 [VGG16 [VGG16 [VGG16 [52])。ネットワーク内で最も深い畳み込み層の後で最も見栄えの良い視覚化が得られることが多く、浅い層ではローカリゼーションが徐々に悪化することがわかります。これは、メインペーパーのセクション 3 で説明されている、より深い畳み込み層がより多くのセマンティック概念をキャプチャするという直感と一致しています。

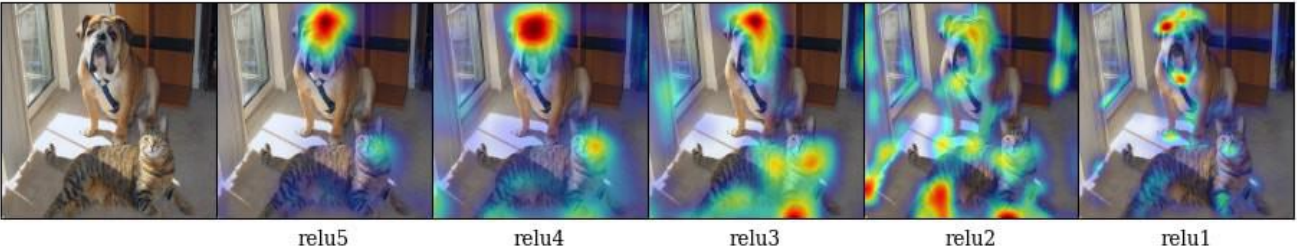


図 14 : AlexNet のさまざまな修正畳み込み層特徴マップの「タイガーキャット」カテゴリの Grad-CAM ローカリゼーション。  
「サングラス、ダークグラス、シェード」クラス : <http://i.imgur.com/a1C7DGh.jpg>

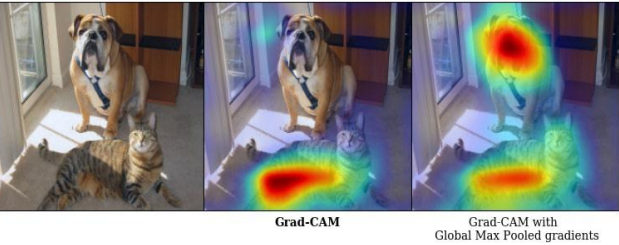


図 15 : グローバル平均プーリングとグローバル最大プーリングを使用した「タイガーキャット」カテゴリの Grad-CAM 視覚化。

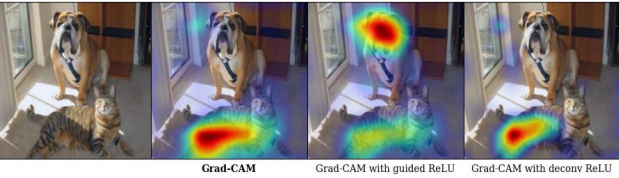


図 16 : ReLU バックワードパスへのさまざまな変更に対する「タイガーキャット」カテゴリの Grad-CAM 視覚化。Grad-CAM の計算中に実際の勾配を使用すると、最良の結果が得られます。

「コンピューターキーボード、キーパッド」クラス : <http://i.imgur.com/QMhsRzf.jpg>

## 2.画像のキャプション

公開されている **Neuraltalk2** コードとモデルを使用します<sup>9</sup> 画像キャプション実験用。モデルは **VGG-16** を使用して画像をエンコードします。画像表現は、最初のタイムステップで入力として **LSTM** に渡され、**LSTM** は画像のキャプションを生成します。モデルは、**COCO** を使用した **CNN** の微調整とともにエンドツーエンドでトレーニングされます[35]データセットのキャプション。画像を画像キャプションモデルにフィードフォワードして、キャプションを取得します。**Grad-CAM** を使用して粗いローカリゼーションを取得し、それをガイド付きバックプロパゲーションと組み合わせて、生成されたキャプションのサポートを提供する画像内の領域を強調表示する高解像度の視覚化を取得します。

## 3.視覚的な質問応答 (VQA)

**Grad-CAM** と **GuidedGrad-CAM** を使用して、公開されている **VQA** モデルの理由を説明します[38]それが答えたものに答えた-答えた。

Lu らによる **VQA** モデル。標準の **CNN** とそれに続く完全に接続されたレイヤーを使用して、質問の **LSTM** 埋め込みに一致するように画像を **1024-dim** に変換します。次に、変換された画像と **LSTM** 埋め込みはポイントごとに行われます

---

<sup>9</sup> <https://github.com/karpathy/neuraltalk2>



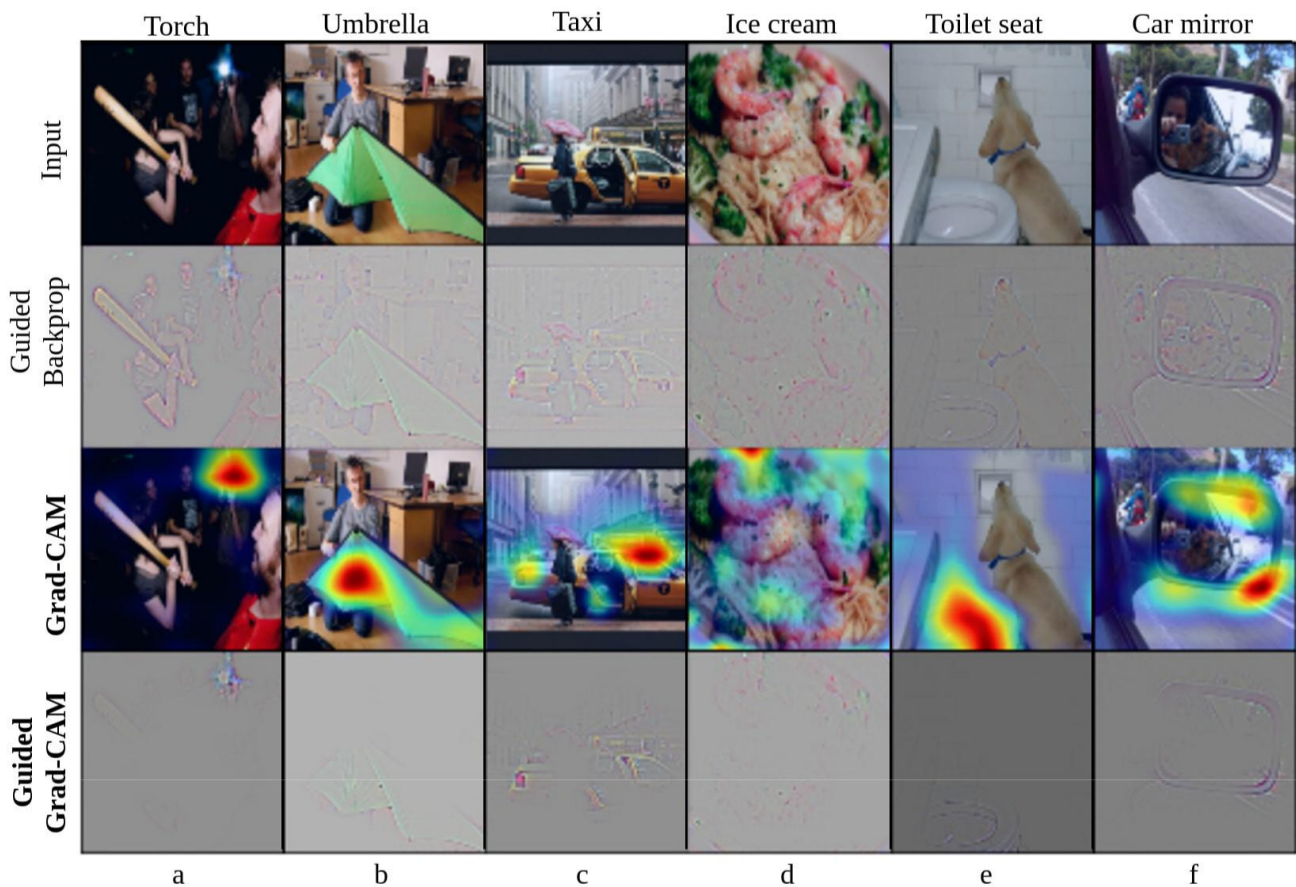


図 17 : COCO 検証データセットからランダムにサンプリングされた画像の視覚化。予測されるクラスは、各列の上部に記載されています。

たちの視覚化は、モデルの予測を説明し、助けるための正確な方法を提供します

画像と質問の組み合わせ表現を取得するために乗算され、多層パーセプトロンが 1000 の回答のうちの 1 つを予測するように上部でトレーニングされます。3つの異なる CNN でトレーニングされた VQA モデルの視覚化を示します-AlexNet [33]、VGG-16 および VGG-19 [52]。CNN は VQA のタスク用に微調整されていませんが、モデルが見ている領域のローカライズされた高解像度の視覚化を提供することにより、これらのネットワークをよりよく理解するためのツールとして私たちのアプローチがどのように役立つかを見るのは興味深いことです。これらのネットワークは、明示的な注意メカニズムが適用されていない状態でトレーニングされていることに注意してください。

図の最初の行に注目してください。19、「人は波に乗っているのか」という質問に対して、AlexNet と VGG-16 を使用した VQA モデルは、波ではなく主に人に集中しているため、「いいえ」と回答しました。一方、VGG-19 は「はい」と正しく答え、男性の周りの地域を調べて質問に答えました。2行目では、「打っている人は何ですか？」という質問に対して、AlexNet でトレーニングされた VQA モデルは、ボールを見ずにコンテキストだけに基づいて「テニスボール」と答えました。このようなモデルは、実際のシナリオで使用する場合にリスクを伴う可能性があります。予測された答えだけでモデルの信頼性を判断することは困難です。私

アーキテクチャを変更したり、精度を犠牲にすることなく、信頼するモデルを決定します。図の最後の行に注目してください。19、「これは完全にオレンジですか？」という質問に対して、モデルはオレンジの周りの領域を探して「いいえ」と答えます。

#### D ポインティングゲームの詳細

[で 58]、ポインティングゲームは、グラウンドトゥールースカテゴリをローカライズするためのさまざまな注意マップの識別性を評価するために設定されました。ある意味で、これは視覚化の精度、つまり注意マップがグラウンドトゥールースカテゴリのセグメンテーションマップと交差する頻度を評価します。これは、視覚化手法が対象のカテゴリに対応しないマップを生成する頻度を評価しません。

したがって、上位 5 つの予測カテゴリの視覚化を評価するためのポインティングゲームの変更を提案します。この場合、ビジュアライゼーションには、CNN 分類器からの上位 5 つの予測のいずれかを拒否する追加のオプションが与えられます。Grad-CAM と c-MWP の 2 つのビジュアライゼーションのそれぞれについて、ビジュアライゼーションの最大値のしきい値を選択します。

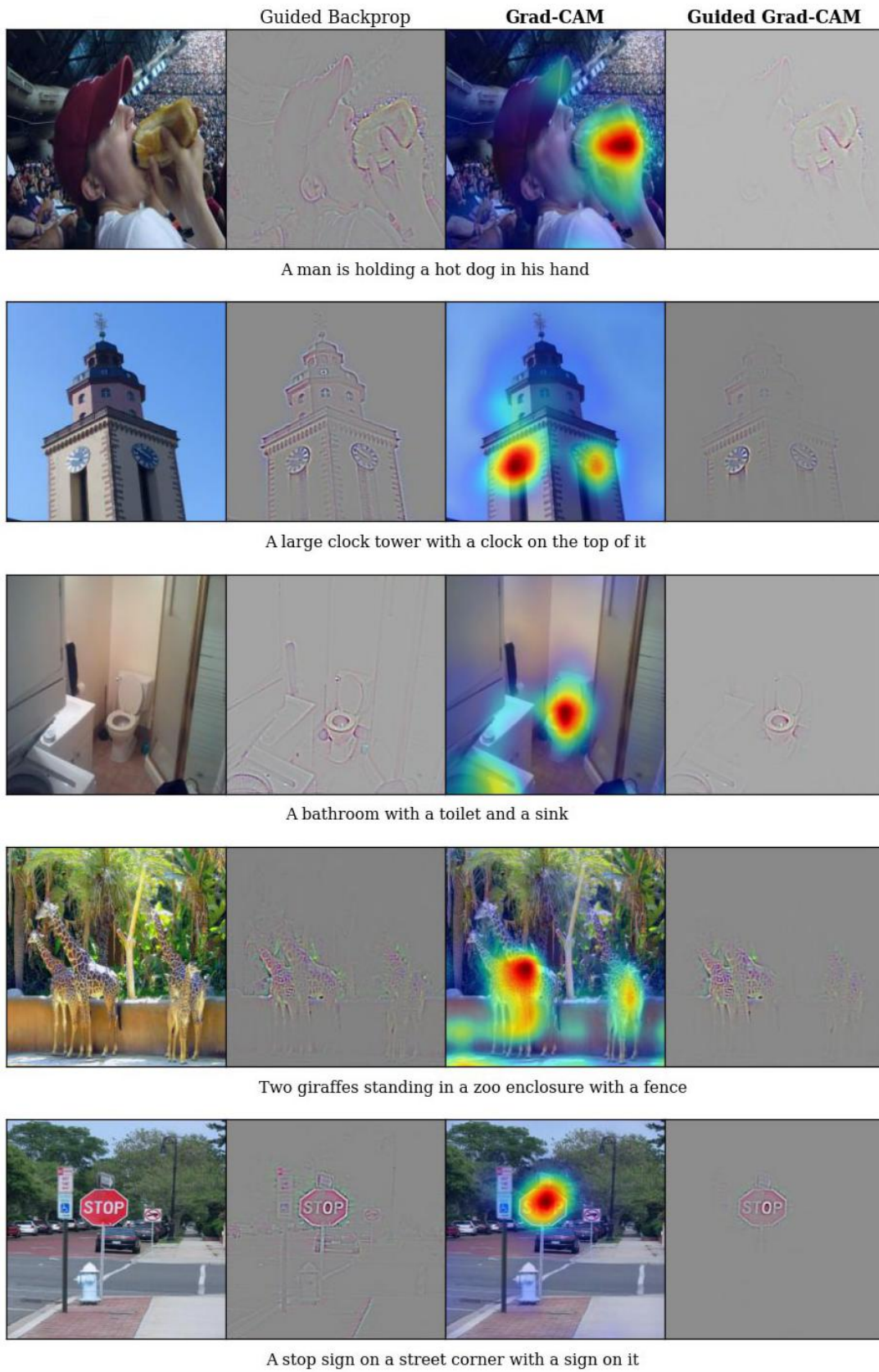


図 18 : Neuraltalk2 画像キャプションモデルによって生成されたキャプションのガイド付きバックプロパゲーション、Grad-CAM およびガイド付き Grad-CAM の視覚化。



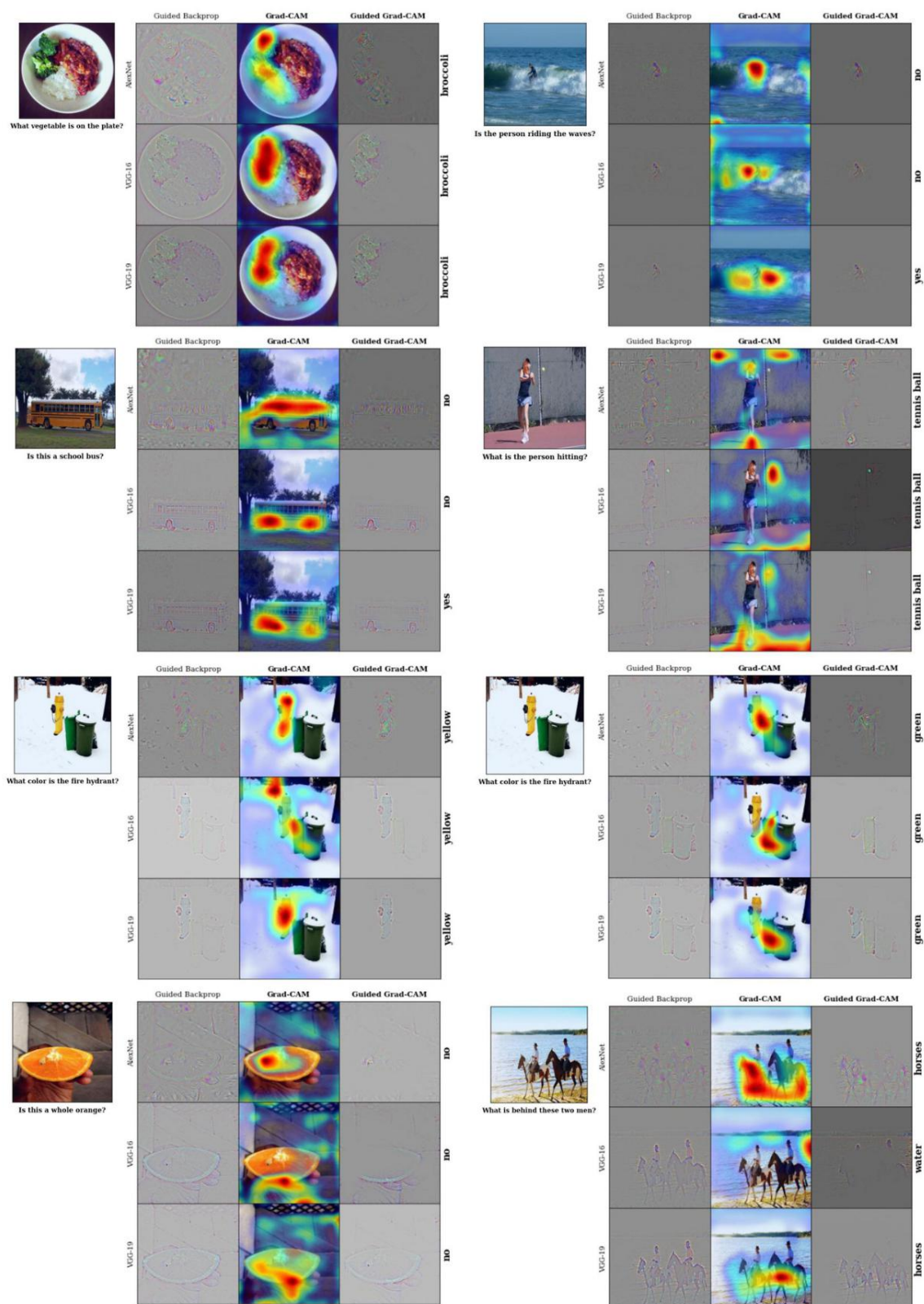


図 19 : VQA モデルからの回答に対するガイド付きバックプロパゲーション、Grad-CAM およびガイド付き Grad-CAM の視覚化。画像と質問のペアごとに、AlexNet、VGG-16、VGG-19 の視覚化を示します。答えを黄色から緑色に変更すると、3 行目で注意がどのように変化するか注目してください。

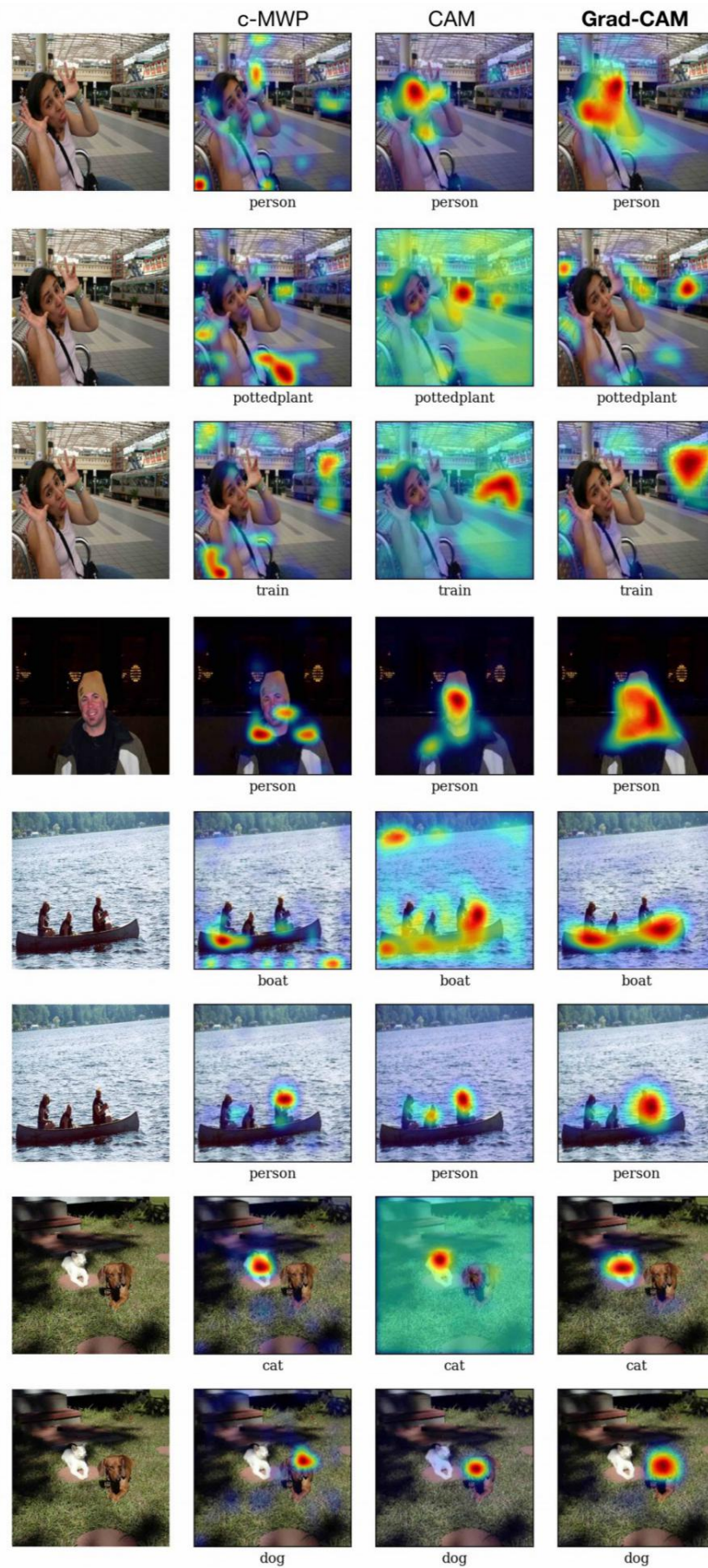


図 20 : PASCAL からサンプリングされた画像のグラウンドトゥールースカテゴリ（各画像の下に表示）の視覚化[17]検証セット。



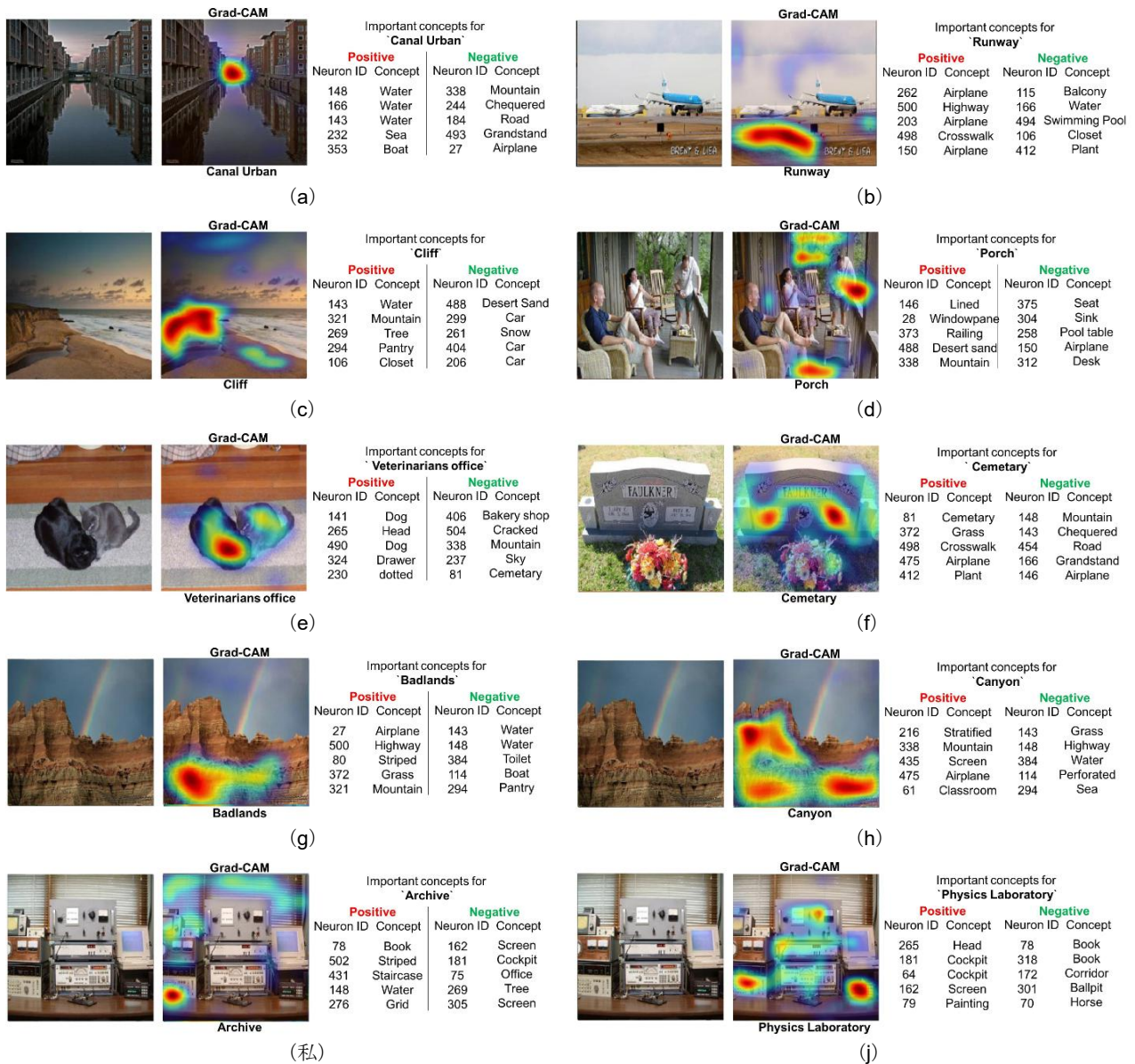


図 21 : Places365 データセットでトレーニングされた VGG-16 の視覚的な説明とテキストによる説明を示すより定性的な例 ([61])。テキストによる説明のために、予測されたクラスの最も重要なニューロンとその名前を示します。重要なニューロンは、説得力がある（正に重要）か、抑制的（負に重要）である可能性があります。最初の 3 行は肯定的な例を示し、最後の 2 行は失敗のケースを示しています。

これは、視覚化されているカテゴリが画像に存在するかどうかを判断するために使用できます。

上位 5 つのカテゴリのマップを計算し、マップの最大値に基づいて、マップが GT ラベルのものであるか、画像に存在しないカテゴリのものであるかを分類しようとしています。メインペーパーのセクション 4.2 で述べたように、私たちのアプローチ Grad-CAM は c-MWP を大幅に上回っています (VGG-16 の 60.30% に対して 70.58%)。

## E Excitation Backprop (c-MWP) および CAM との定性的比較

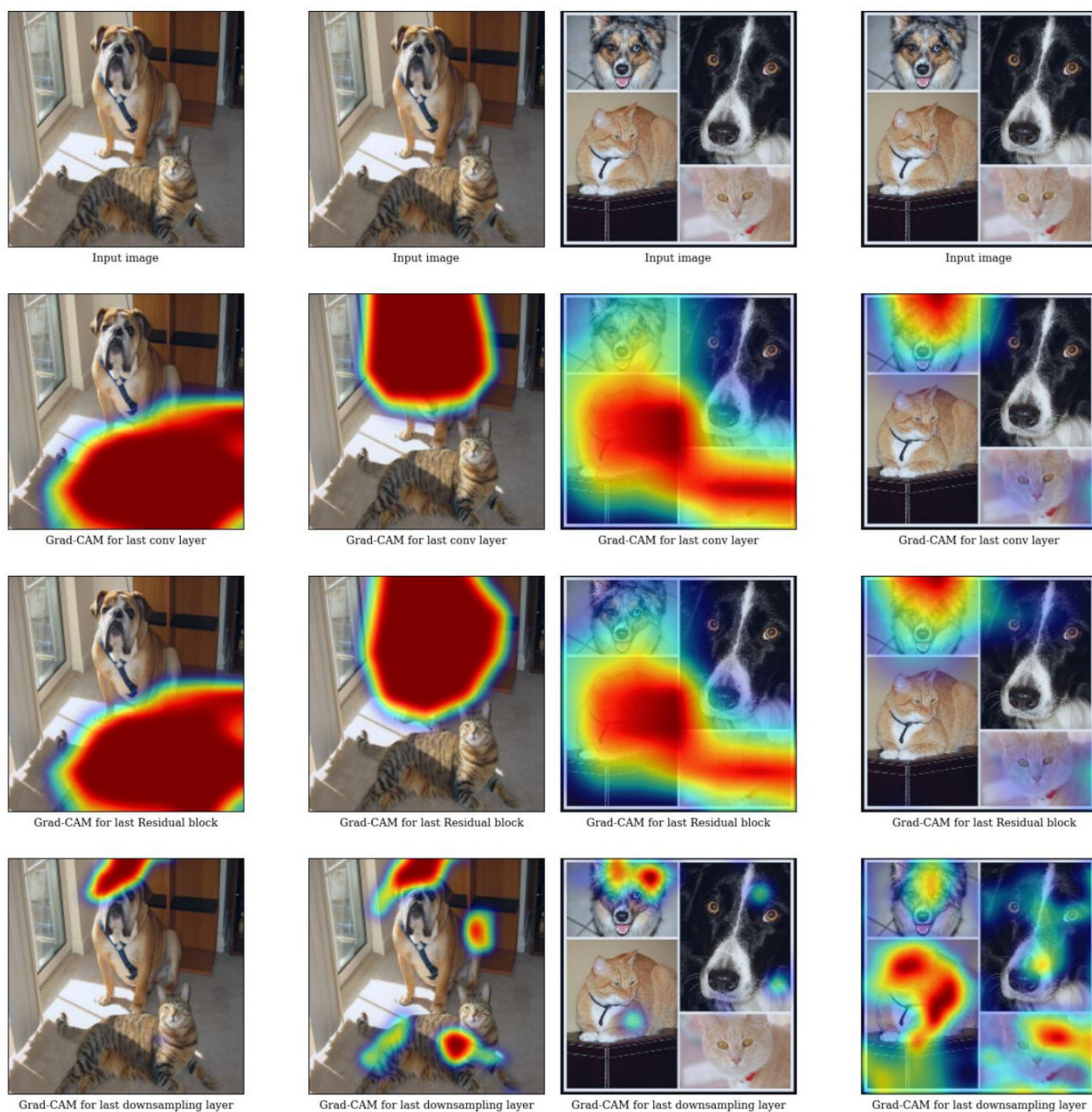
このセクションでは、Grad-CAM と CAM を比較したより定性的な結果を提供します[59]および c-MWP [58]パスカル[17]。

PAS-CAL VOC 2012 データセットで微調整された ImageNet でトレーニングされた VGG-16 モデルからの Grad-CAM、CAM、および c-MWP の視覚化を比較します。Grad-CAM および c-MWP の視覚化は既存のモデルから直接取得できますが、CAM はアーキテクチャの変更が必要であり、再トレーニングが必要であるため、精度が低下します。また、Grad-CAM とは異なり、c-MWP と CAM は画像分類ネットワークにのみ適用できます。グラウンドトゥールスカテゴリの視覚化は、図にありす。20。

F Places データセットの視覚的およびテキストによる説明図。21 視覚的およびテキストによる説明の例をさらに示します (セクション 7) Places365 データ



セットでトレーニングされた画像分類モデル (VGG-16) の場合 ([61]) 。



(a) ResNet-200 レイヤーアーキテクチャの Grad-CAM ビジュアライゼーション (b) ResNet-200 レイヤーアーキテクチャの Grad-CAM ビジュアライゼーション

「虎猫」(左)と「ボクサー」(右)のカテゴリ。

「ぶち猫」(左)と「ボクサー」(右)のカテゴリ。

図 22 : ダウンサンプリング層に遭遇すると、Grad-CAM の識別能力が大幅に低下することがわかります。

## G 残差ネットワークの分析

このセクションでは、残差ネットワーク (ResNets) で Grad-CAM を実行します。特に、ImageNet でトレーニングされた 200 層のアーキテクチャを分析します。

10。

現在の ResNets [24] 通常、残りのブロックで構成されます。ブロックの 1 つのセットは、ID スキップ接続 (同一の出力

di-を持つ 2 つのレイヤー間のショートカット接続) を使用します。

<sup>10</sup> からの 200 層の ResNet アーキテクチャを使用します

<https://github.com/facebook/fb.resnet.torch>。

メンション)。これらの残余ブロックのセットには、伝搬信号の次元を変更するダウンサンプリングモジュールが散在しています。図に見られるように、22 最後の畳み込み層に適用された視覚化により、猫と犬を正しくローカライズできます。

**Grad-CAM** は、最後のセットの残りのブロックで猫と犬を正しく視覚化することもできます。ただし、空間分解能が異なる以前の残余ブロックのセットに進むと、**Grad-CAM** が対象のカテゴリをローカライズできないことがわかります（図の最後の行を参照）。22）。他の **ResNet** アーキテクチャ（18層および50層）でも同様の傾向が見られます。

## 参考文献

- 1.1  
。 A. Agrawal, D. Batra, および D.Parikh. の動作の分析  
視覚的な質問応答モデル。EMNLP、2016 年。2
- 2.2 H. Agrawal, CS Mathialagan, Y. Goyal, N. Chavali, P. Banik,  
A.モハバトラ、A. オスマン、D. バトラ。CloudCV : 大規模クラウドサービスとしての分散コンピュータビジョン。モバイルクラウドの場合  
ビジュアルメディアコンピューティング、265~290 ページ。Springer、2015 年。1
- 3.3 S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, および D. Parikh. VQA : 視覚的な質問応答。ICCV では、  
2015 年。1、2、10、12
- 4.4 D. Bau, B. Zhou, A. Khosla, A. Oliva, および A. Torralba. 通信網  
解剖 : 深い視覚表現の解釈可能性の定量化  
。コンピュータビジョンとパターン認識、2017 年。1、3、10
- 5.5 L.バツァーニ、A. ベルガモ、D. アンゲロフ、L. トレサーニ。自己-  
ディープネットワークでオブジェクトのローカリゼーションを教えました。WACV、2016 年。  
4
- 6.6  
。 Y.ベンジオ、A. クールヴィル、P. ヴィンセント。表現学習 :  
パターン上の IEEE トランザクション  
レビューと新しい視点。  
分析と機械知能、35 (8) : 1798-1828、2013 年。4
- 7 X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar,  
と CLZitnick. Microsoft COCO のキャプション : データ収集と  
評価サーバー。arXiv preprint arXiv : 1504.0325、2015。1、10 31。
- 8.8 RG Cinbis, J. Verbeek, および  
。 C.Schmid。弱く監視されている  
マルチフォールドマルチインスタンス学習によるオブジェクトのローカリゼーション。IEEE  
パターン分析と機械知能に関するトランザクション、2016 年。3
- 9.9 A. Das, H. Agrawal, CL Zitnick, D. Parikh, および D.Batra.  
。 人間  
視覚的な質問応答における注意 : 人間と深いことを行う  
ネットワークは同じ地域を見ていますか? EMNLP、2016 年。12
10.  
10 A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, および D. Batra  
具現化された質問応答。IEEECon-の議事録  
コンピュータビジョンとパターン認識 (CVPR) に関する会議、  
2018 年。  
1
- 11 A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, JM Moura,  
D.バトリクと D.バトラ。ビジュアルダイアログ。IEEE の議事録  
コンピュータビジョンとパターン認識 (CVPR) に関する会議、  
2017 年。1
22. IJ グッドフェロー、J. シュレンズ、C. セゲディ。説明とハー-  
敵対的な例を理解する。統計、2015 年。9
23. D.ゴードン、A. ケンバビ、M. ラステガリ、J. レッドモン、  
D. フォックス、  
A.ファーハディ。Iqa : インタラクティブな環境での視覚的な質問  
間応答-  
メント。arXiv preprint arXiv : 1712.03316、2017。1
24. K. He, X. Zhang, S. Ren, および J.Sun. の深い残余学習  
画像認識。2016 年の CVPR で。1、2、13、21
25. D. Hoiem, Y. Chodpathumwan, および Q.Dai. のエラーの診断  
オブジェクト検出器。ECCV、2012 年。2
26. P.ジャクソン。エキスパートシステムの紹介。アディソン-ウェス  
リーロング-  
man Publishing Co., Inc.、米国マサチューセッツ州ボストン、第 3 版、1998 年。  
2
27. Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick,  
S.グアダラマ、T. ダレル。Caffe : 畳み込みアーキテクチャ  
高速機能埋め込み用。ACM MM、2014 年。6
28. E. Johns, O. MacAodha, および GJBrostow. エキスパートになる  
-インタラクティブなマルチクラスマシニング。  
CVPR、2015 年。2
29. J.ジョンソン、A. カルパシー、L. フェイフェイ。DenseCap :  
完全にコンボ-  
密なキャプションのための lutional ローカリゼーションネットワーク。CVPR では、  
2016 年。1、10、11
30. A.カルパシー。ConvNet との競合から学んだこと  
ImageNet で。http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet /、2014 年。2
31. A.カルパシーと L.フェイフェイ。の深い視覚的意味論的アライ  
ンメント  
画像の説明を生成します。CVPR、2015 年。10、11
32. A.コレスニコフと CH ランパート。シード、拡張、および制  
約 :  
弱教師あり画像セグメンテーションの 3 つの原則。に  
ECCV、2016 年。7
33. A. Krizhevsky, I. Sutskever, および GE ヒントン。Imagenet 分類-  
深い畳み込みニューラルネットワークを使用します。NIPS で  
は、2012 年。1、5、  
6、16
34. M.リン、Q. チェン、S. ヤン。ネットワーク内のネットワーク。  
ICLR、2014 年。  
3
35. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan,  
P.Dollar, および CLZitnick. Microsoft coco : の一般的なオブ  
ジェクト  
環境。ECCV で。2014 年。14、15
36. ZC リプトン。モデル解釈可能性の神話。ArXiv e-prints、  
2016 年 6 月。2、3



12. A. Das, S. Kottur, JM Moura, S. Lee, および D.Batra. 学習。深い強化学習を伴う協調的視覚対話エージェント。コンピュータに関する IEEE 国際会議の議事録ビジョン (ICCV) 、2017 年。1
13. H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, および AC クールビル。何だと思う？！視覚オブジェクトの発見マルチモーダルダイアログ。IEEE 会議の議事録コンピュータビジョンとパターン認識 (CVPR) 、2017 年。1
14. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, および L.Fei-Fei. Im-ageNet : 大規模な階層画像データベース。CVPR では、2009 年。2、6、14
15. A.Dosovitskiy と T.Brox。畳み込みネットワークの反転畳み込みネットワークを使用します。CVPR、2015 年。3
16. D. Erhan, Y. Bengio, A. Courville, および P.Vincent。視覚化。ディープネットワークの上位層の機能。モントリオール大学、1341、2009。3
17. M. Everingham, L. VanGool, CKI Williams, J. Winn, PASCAL ビジュアルオブジェクトクラスと A.ジサーマン。PASCAL 2007 (VOC2007) チャレンジの結果。http : //www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html、2009 年。19、20
18. H. Fang, S. Gupta, F. Iandola, RK Srivastava, L. Deng, P.Dollar, J. Gao, X. He, M. Mitchell, JC Platt, 他 キャプションからビジュアルコンセプトとバック。CVPR、2015 年。1
19. C.ガン, N. ワン, Y. ヤン, D.-Y. ヨン, AG ハウプトマン。Devnet : マルチメディアイベント検出と証拠の再集計。CVPR、2015 年。3
20. H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, および W. Xu あなたは機械と話している？多言語画像のデータセットと方法質問応答。NIPS では、2015 年。1、10、12
21. R. Girshick, J. Donahue, T. Darrell, および J.Malik。豊富な機能の階層-正確なオブジェクト検出とセマンティックセグメンテーションのためのアーチ。CVPR、2014 年。1
37. J.ロング, E. シェルハマー, T. ダレル。完全畳み込みネットワークセマンティックセグメンテーション用。CVPR、2015 年。1
38. J. Lu, X. Lin, D. Batra, および D.Parikh。より深い LSTM と正規化された CNN ビジュアル質問応答モデル。https : //github.com/VT-vision-lab/VQA\_LSTM\_CNN、2015 年。12、13、15
39. J. Lu, J. Yang, D. Batra, および D.Parikh。階層的な質問-視覚的な質問応答のための画像の共同注意。NIPS では、2016 年。13
40. A.マヘンドランと A.ヴェダルディ。顕著なデコンボリューションネットワーク。コンピュータビジョンに関するヨーロッパ会議、2016 年。3
41. A.マヘンドランと A.ヴェダルディ。深い畳み込みを視覚化する自然なプレイメージを使用したニューラルネットワーク。の国際ジャーナルコンピュータビジョン、2016 年 1~23 ページ。3、4
42. M. Malinowski, M. Rohrbach, および M.Fritz。ニューロンに聞いてください : A 画像に関する質問に答えるためのニューラルベースのアプローチ。ICCV、2015 年。1、10、12
43. M. Oquab, L. Bottou, I. Laptev, および J.Sivic。学習とトランスジェンダー畳み込みニューラルを使用した中間レベルの画像表現の取得ネットワーク。CVPR、2014 年。3、4
44. M. Oquab, L. Bottou, I. Laptev, および J.Sivic。オブジェクトのローカリゼーションですか無料？-畳み込みニューラルによる弱教師あり学習ネットワーク。CVPR、2015 年。3
45. POPinheiro と R.Collobert。画像レベルからピクセルレベルへ畳み込みネットワークによるラベリング。CVPR、2015 年。3
46. M.レン, R. キロス, R. ゼメル。のモデルとデータの調査画像の質問応答。NIPS では、2015 年。1、10、12
47. MT Ribeiro, S. Singh, および C.Guestrin。「なぜ私は信頼すべきなのかあなた？」：分類子の予測を説明します。SIGKDD では、2016 年。3、8
48. RR Selvaraju, P. Chattopadhyay, M. Elhoseiny, T. Sharma, D. Ba-tra, D. パリク, S. リー。ニューロンを選択してください : 組み込みニューロンを通じたドメイン知識-重要性。議事録

コンピュータビジョンに関する欧州会議 (ECCV) のページ  
526–541、2018。3

49. RR セルパラジュ、S。リー、Y。シェン、H。ジン、S。ゴーシュ、L。ヘック、D。バトラ、D。バリク。ヒントをとる：説明を活用して、ビジョンと言語モデルをより根拠のあるものにします。2019 年コンピュータビジョン国際会議 (ICCV) の議事録。3
50. D.シルバー、A。ファン、CJ マディソン、A。ゲス、L。シフレ、G。ヴァンデンドライシエ、J。シュリットウィーザー、I。アントノグロウ、V。パネルシェルバム、M。ランクト他 ディープニューラルネットワークとツリー検索で囲碁のゲームをマスターする。Nature, 529 (7587) : 484–489、2016 年。2
51. K. Simonyan, A. Vedaldi, および A. Zisserman。畳み込みネットワークの奥深く：画像分類モデルと顕著性マップの視覚化。CoRR, abs / 1312.6034、2013。3、6
52. K. シモニャンと A. ジサーマン。大規模画像認識のための非常に深い畳み込みネットワーク。ICLR では、2015 年。5、6、14、15、16
53. JT Springenberg, A。Dosovitskiy, T。Brox、および MARied-miller。シンプルさの追求：すべての畳み込みネットワーク。CoRR, abs / 1412.6806、2014。2、3、6、14
54. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, および Z. Wojna。コンピュータビジョンの開始アーキテクチャを再考する。コンピュータビジョンとパターン認識に関する IEEE 会議の議事録、2818~2826 ページ、2016 年。6、12
55. O. ヴィニヤルス、A。トシェフ、S。ベンジオ、D。エルハン。表示して伝える：ニューラル画像キャプションジェネレータ。CVPR、2015 年。1、10、12
56. C. Vondrick, A. Khosla, T. Malisiewicz, および A. Torralba。HOGgles：オブジェクト検出機能の視覚化。ICCV、2013 年。9
57. MD ツァイラーと R. フェーガス。会話型ネットワークの視覚化と理解。ECCV、2014 年。2、3、4、6、8、10、14
58. J. チャン、Z。リン、J。ブランド、X。シェン、S。スクラロフ。励起バックプロパゲーションによるトップダウンの神経的注意。2016 年の ECCV で。4、6、7、12、16、20
59. B. Zhou, A. Khosla, LA, A. Oliva, および A. Torralba。識別ローカリゼーションのためのディープ機能の学習。2016 年の CVPR で。2、3、5、6、20
60. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, および A. Torralba。オブジェクト検出器は、深いシーンの cnns に出現します。CoRR, abs / 1412.6856、2014 年。10
61. B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, および A. Torralba。場所：シーン認識用の 1,000 万の画像データベース。パターン分析とマシンインテリジェンスに関する IEEE トランザクション、2017 年。10、11、20