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て、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](#page22)]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/](https://github.com/ramprs/grad-cam/) [ランプ/ grad-cam /](https://github.com/ramprs/grad-cam/)、 CloudCVのデモと一緒に[[2](#page22)][1](#page1)、とビデオ [youtu.be/COjUB9Izk6E](file:///C:\Users\rurus\AppData\Local\Temp\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ークスルーを可能にしました。[33](#page22)、[24](#page22)]、オブジェクト検出[[21](#page22)]、セマンティックセグメンテーション[[37](#page22)]画像のキャプション[[55](#page23)、[7](#page22)、[18](#page22)、[29](#page22)]、視覚的な質問応答[[3](#page22)、[20](#page22)、[42](#page22)、[46](#page22)]そして最近では視覚的な対話[[11](#page22)、 [13](#page22)、[12](#page22)] と具体化された質問応答[[10](#page22)、[23](#page22)]。 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](#page22)]。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](#page22)]）、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](#page22)、[25](#page22)]、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](#page22)]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](#page23)]）、説明の目標は機械教育にあります[[28](#page22)] –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](#page22)]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](#page22)]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求し始めています。

周ら。[[59](#page23)]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](#page23)]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](#page22)] –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](#page2) 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](#page23)]およびデコンボリューション[[57](#page23)]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](#page3) および図。 [1時間](#page3) 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](#page3)、およびその逆。 [1i](#page3)）。

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](#page2) そして [1j](#page2)。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](#page6)）、およびモデルへの忠実性（秒。 [5.3](#page8)）、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-フィケーション、キャプション（秒。 [8.1](#page10)）、およびVQA（Sec。 [8.2](#page12)）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](#page9)）、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](#page22)]画像分類とVQAに適用されます（秒。 [8.2](#page12)）。
5. Grad-CAMのニューロンの重要性と[[4](#page22)]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](#page10)）。
6. 私たちは人間の研究を行います（秒。 [5](#page7)）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GuidedGrad-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](#page23)、[53](#page23)、[57](#page23)、 [19](#page22)] 「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、Si-monyan etal。[[51](#page23)]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](#page23)]およびデコンボリューション[[57](#page23)]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](#page22)]。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](#page2) そして [1時間](#page2)）。

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](#page23)、[16](#page22)]または潜在的な表現を反転します-表現[[41](#page22)、[15](#page22)]。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](#page22)]そしてモデルへの信頼を評価する[[47](#page22)]、[]と同様の方法でGrad-CAMの視覚化を評価します。[47](#page22)]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 etal。[[48](#page22)]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、セルバラジュ等。[[49](#page23)]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](#page22)、[43](#page22)、[44](#page22)、[59](#page23)]。

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](#page23)]。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](#page22)]、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](#page22)]およびlog-sum-expプーリング[[45](#page22)]。

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](#page23)]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](#page22)]。Oquab etal。[[43](#page22)]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](#page23)]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](#page6) および秒。 [D](#page16)）。

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](#page22)、 [41](#page22)]。 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](#page4)、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 、 （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ブ化、つまり。 逆流するこれらの勾配は、グローバル平均プールされます [2](#page4) 幅と高さの寸法（およびでインデックス付け） それぞれ）ニューロンの重要度の重みを取得するには：

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。

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、

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[[52](#page23)]およびAlexNet [[33](#page22)]ネットワーク） [3](#page5)。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[1c、1f](#page2) そして [1i、1l](#page2) 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。アブレーション研究は秒で利用可能です。[B](#page14)。

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。

3.1Grad-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](#page23)]、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。で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 。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。

Fkをグローバル平均プール出力として定義しましょう。

CAMは次の方法で最終スコアを計算します

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、

（の偏導関数を取る[4](#page5)） に関して 、 私たちはそれを見ることができます 。これを（6）に代入すると、次のようになります。

（5）から、次のようになります。したがって、

すべてのピクセルにわたって（8）の両側を合計します。

に依存しないので、これを次のように書き直します

ご了承ください 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。

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](#page4)）。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](#page12)）。

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](#page23)]、デコンボリューション[[57](#page23)]）。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](#page2)、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（ 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](#page4) 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削除すると、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

（のように[2](#page5)）、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。 [3](#page6)。

4Grad-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](#page22)]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](#page23)]、AlexNet [[33](#page22)]およびGoogleNet [[54](#page23)]（カフェから入手[[27](#page22)]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](#page6)。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](#page23)]およびSimonyanetal。[[51](#page23)]、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](#page23)]、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](#page22)]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](#page7) 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](#page23)]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ルース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。 。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[4](#page7)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を評価します。グラウンドトゥルースカテゴリ、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](#page23)]かなりの差で（70.58％対60.30％）。c-MWPを比較する定性的な例[[58](#page23)]およびGrad-CAMonは、セクション2にあります。 [D](#page16)[5](#page7)。

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します。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バージョン（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AmazonMechan-ical 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](#page8)）。

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、グラウンドトゥルースに対して評価し、平均して表の精度を得ました。[2](#page8)。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、Deconvo-lution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](#page8)。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](#page8)、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。GuidedGrad-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](#page22)]。

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](#page23)]、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が以前の方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。

6Grad-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。

6.1VGG-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](#page9)、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](#page23)]。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。

6.2VGG-16に対する敵対的ノイズの影響

Goodfellow etal。[[22](#page22)]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](#page9)、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-CAMvi-sualizations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）[6](#page9) 図の中央の列の領域。 [8](#page10)）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）。[8](#page10)）。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。（（[1](#page4)）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](#page23)、 [57](#page23)] 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の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、テクスチュア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](#page22)]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。

図。 [9](#page11) 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](#page23)]。（a）では、（[1](#page4)）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。[7](#page10) 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](#page20)。

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](#page22)、[29](#page22)、[55](#page23)]および視覚的質問回答-ing（VQA）[[3](#page22)、[20](#page22)、[42](#page22)、[46](#page22)]。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。

* 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[8](#page10) 実装[[31](#page22)]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-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。 [3](#page4)。図を参照してください。[10a](#page11)。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、凧と人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](#page14)。

密なキャプションとの比較。ジョンソン等。[[29](#page22)]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リ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、グラウンドトゥルース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](#page11)。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ではが達成されます。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。

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 andTellモデルを使用します[[55](#page23)] 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](#page23)] 建築。グラウンドトゥルース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](#page4)）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](#page5)）。

**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](#page23)]。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。[11](#page12)。

8.2視覚的な質問応答

典型的なVQAパイプライン[[3](#page22)、[20](#page22)、[42](#page22)、[46](#page22)]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（（[3](#page5)））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](#page22)]）図に記載されています。 [12](#page13) 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](#page8)。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。

**人間の注意との比較。** Das etal。[[9](#page22)] 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](#page22)]。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](#page22)] [からの1374val質問画像（QI）ペア[3](#page22)] [のように順位相関評価プロトコルを使用する[9](#page22)]。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 etal。[[39](#page22)] 200層のResNetを使用する[[24](#page22)]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](#page13) 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](#page21)。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 FamilyFoundation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

1. -画像分類、キャプション、VQAのより定性的な例

14

1. -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](#page5)）、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](#page15) 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](#page22)]。表を参照してください。[3](#page14)。

2.1。（のReLUの重要性[3](#page5)）。

ReLUの削除（（[3](#page5)））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](#page15)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](#page23)]およびDeconv-ReLU [[57](#page23)] ReLUのバックワードパスへの変更として。

Guided-ReLU：Springenberg etal。[[53](#page23)]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](#page15)、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](#page14)。

Deconv-ReLU：デコンボリューション[[57](#page23)]、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](#page15)）。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](#page16) VGG-16に対応[[52](#page23)] ImageNetでトレーニングされたネットワーク。

図。 [17](#page16) 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](#page22)]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](#page16).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Ima-geNet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](#page16).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。

15

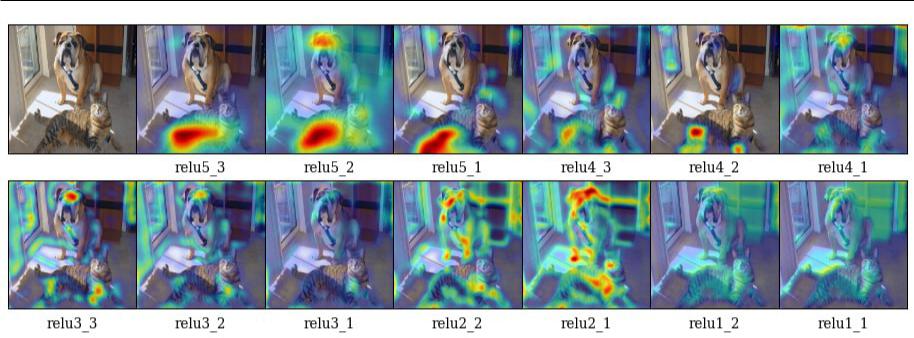


図13：「タイガーキャット」クラスのさまざまな畳み込み層でのGrad-CAM。この図は、CNN（VGG16 [VGG16 [VGG16 [VGG16 [[52](#page23)]）。ネットワーク内で最も深い畳み込み層の後で最も見栄えの良い視覚化が得られることが多く、浅い層ではローカリゼーションが徐々に悪化することがわかります。これは、メインペーパーのセクション3で説明されている、より深い畳み込み層がより多くのセマンティック概念をキャプチャするという直感と一致しています。

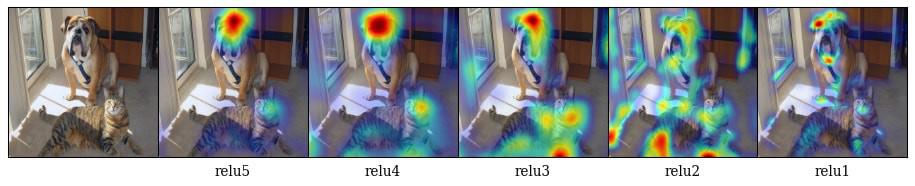


図14：AlexNetのさまざまな修正畳み込み層特徴マップの「タイガーキャット」カテゴリのGrad-CAMローカリゼーション。

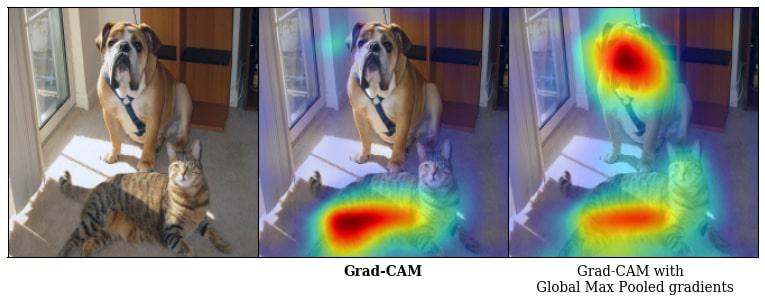


図15：グローバル平均プーリングとグローバル最大プーリングを使用した「タイガーキャット」カテゴリのGrad-CAM視覚化。

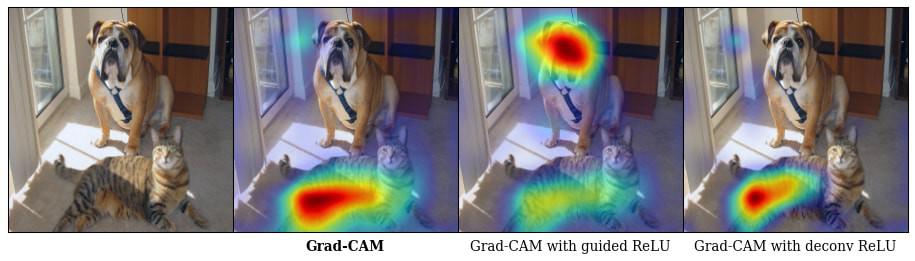


図16：ReLUバックワードパスへのさまざまな変更に対する「タイガーキャット」カテゴリのGrad-CAM視覚化。Grad-CAMの計算中に実際の勾配を使用すると、最良の結果が得られます。

「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

「サングラス、ダークグラス、シェード」クラス：

http://i.imgur.com/a1C7DGh.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[9](#page15)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](#page22)]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](#page22)]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

* <https://github.com/karpathy/neuraltalk2>

16

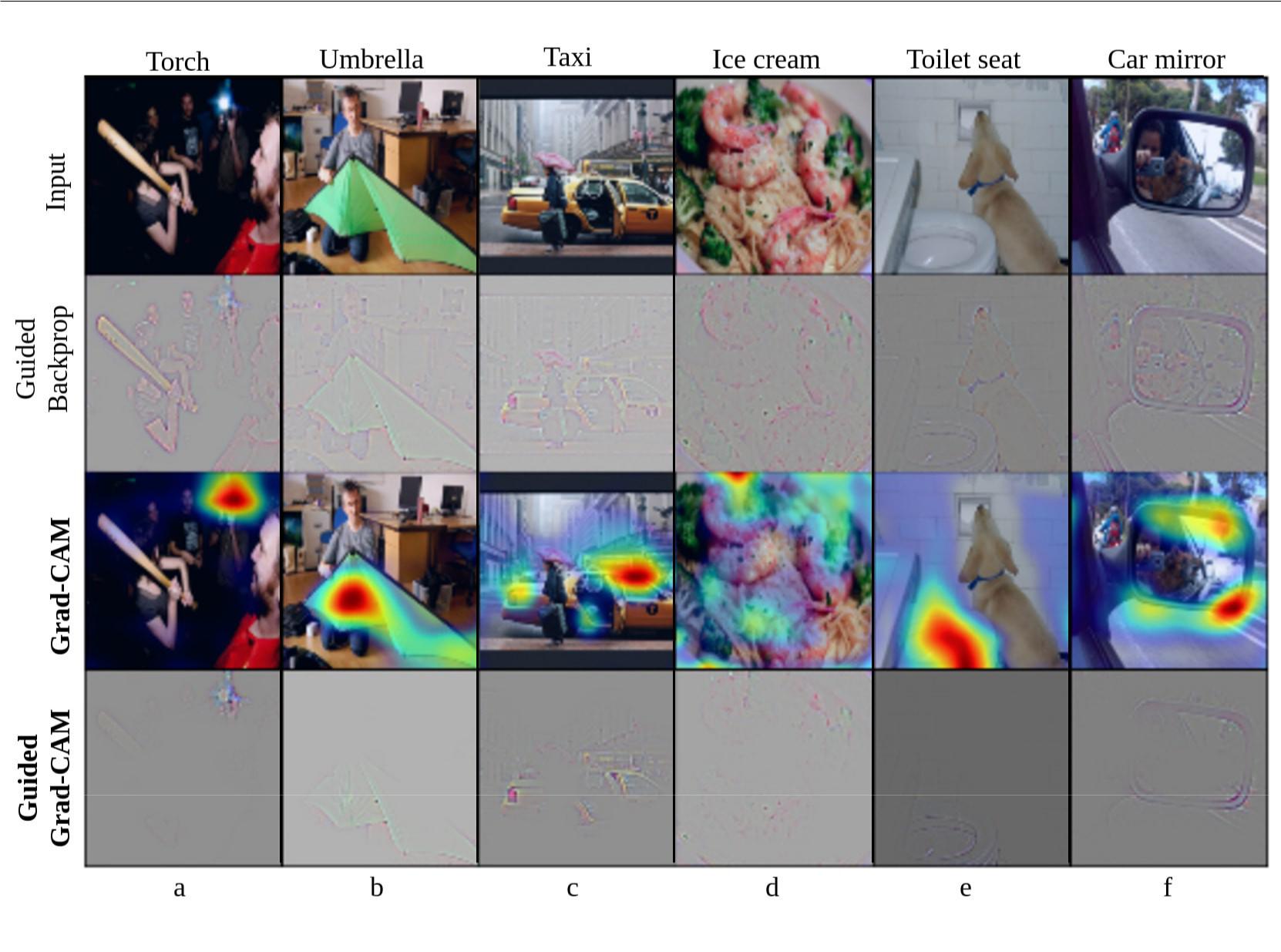


図17：COCO検証データセットからランダムにサンプリングされた画像の視覚化。予測されるクラスは、各列の上部に記載されています。

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](#page22)]、VGG-16およびVGG-19 [[52](#page23)]。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](#page18)、「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

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](#page18)、「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](#page23)]、ポインティングゲームは、グラウンドトゥ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。

17

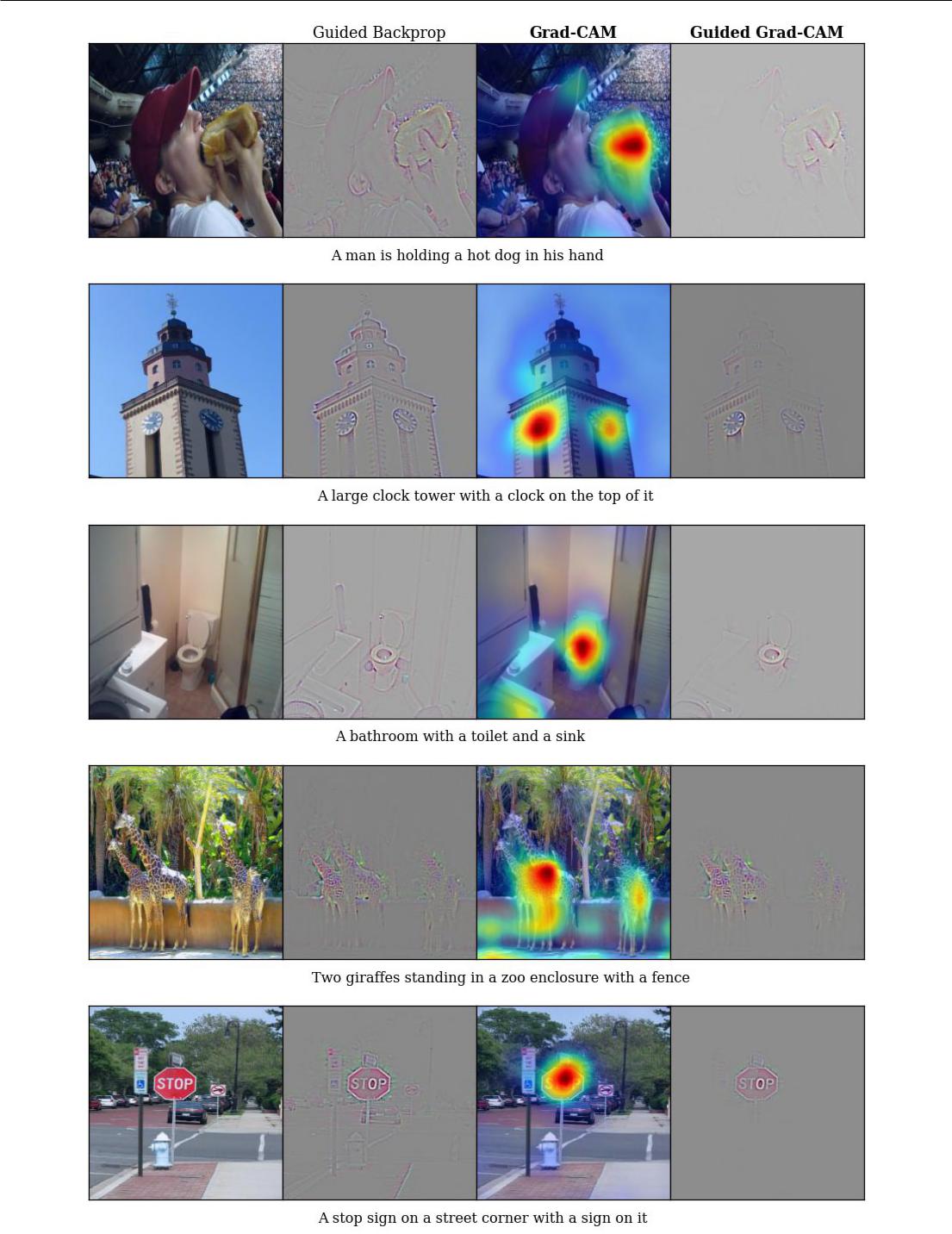


図18：Neuraltalk2画像キャプションモデルによって生成されたキャプションのガイド付きバックプロパゲーション、Grad-CAMおよびガイド付きGrad-CAMの視覚化。

18

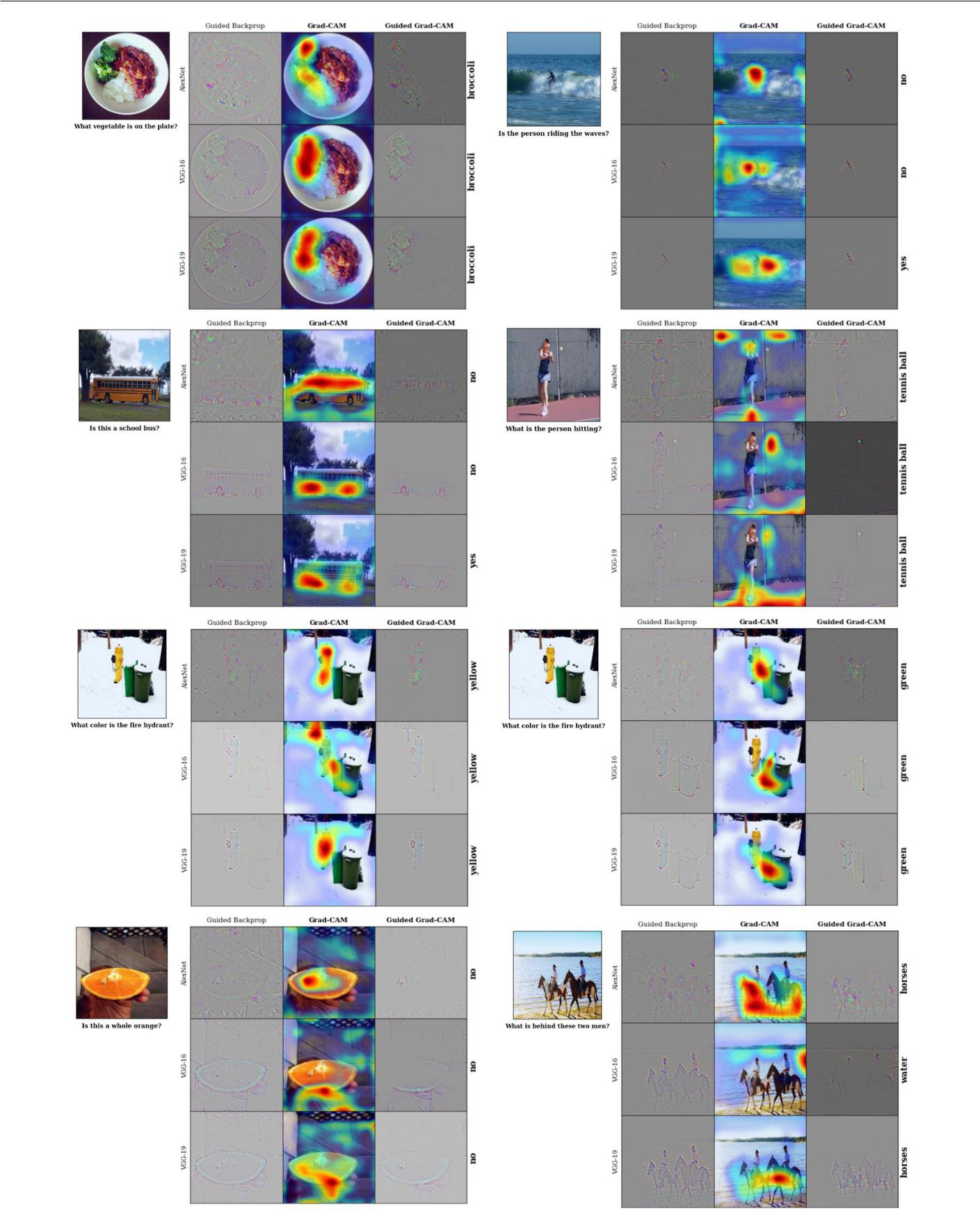


図19：VQAモデルからの回答に対するガイド付きバックプロパゲーション、Grad-CAMおよびガイド付きGrad-CAMの視覚化。画像と質問のペアごとに、AlexNet、VGG-16、VGG-19の視覚化を示します。答えを黄色から緑色に変更すると、3行目で注意がどのように変化するかに注目してください。

19

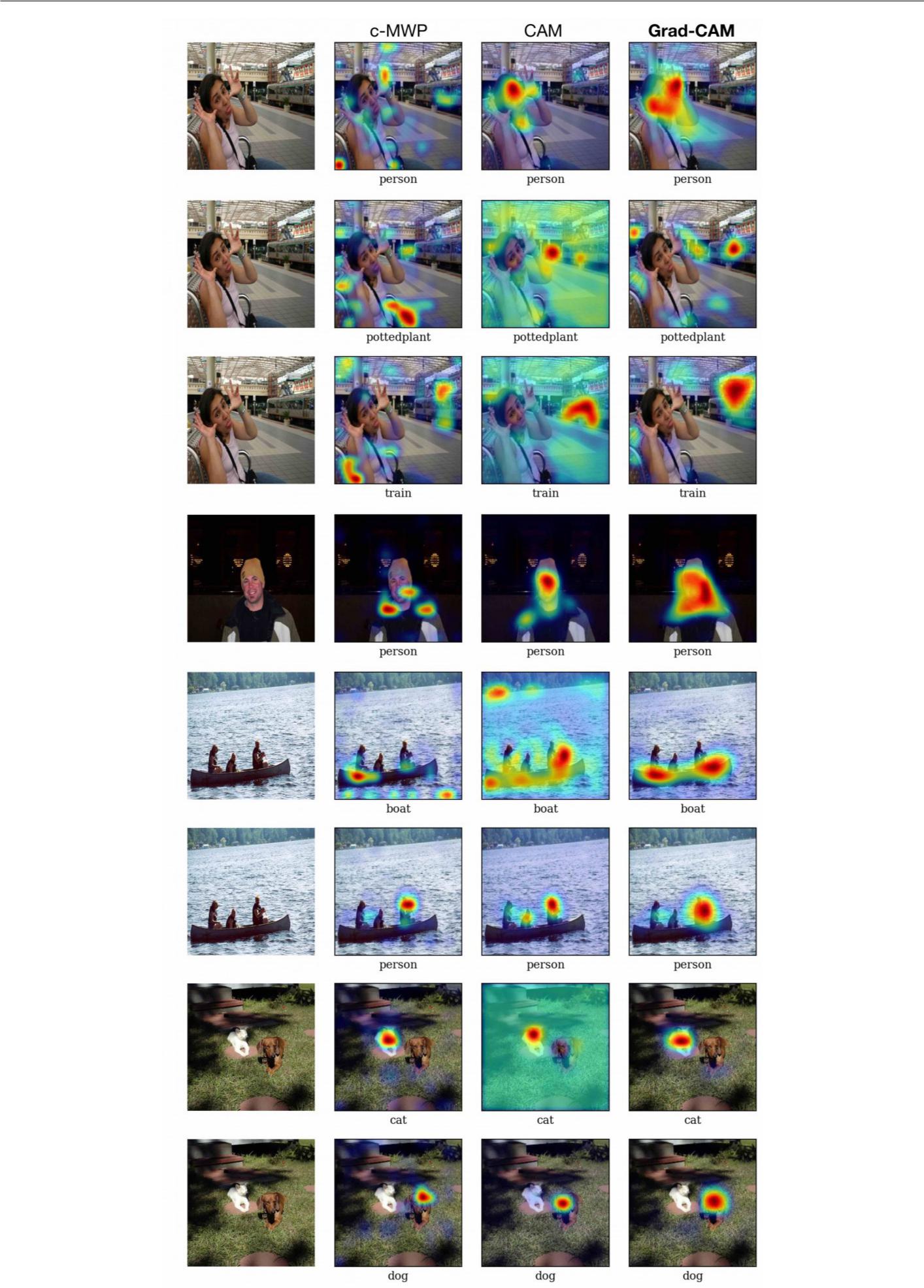


図20：PASCALからサンプリングされた画像のグラウンドトゥルースカテゴリ（各画像の下に表示）の視覚化[[17](#page22)]検証セット。

20



（a） （b）



（c） （d）



（e） （f）



（g） （h）



（私） （j）

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](#page23)]）。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％）。

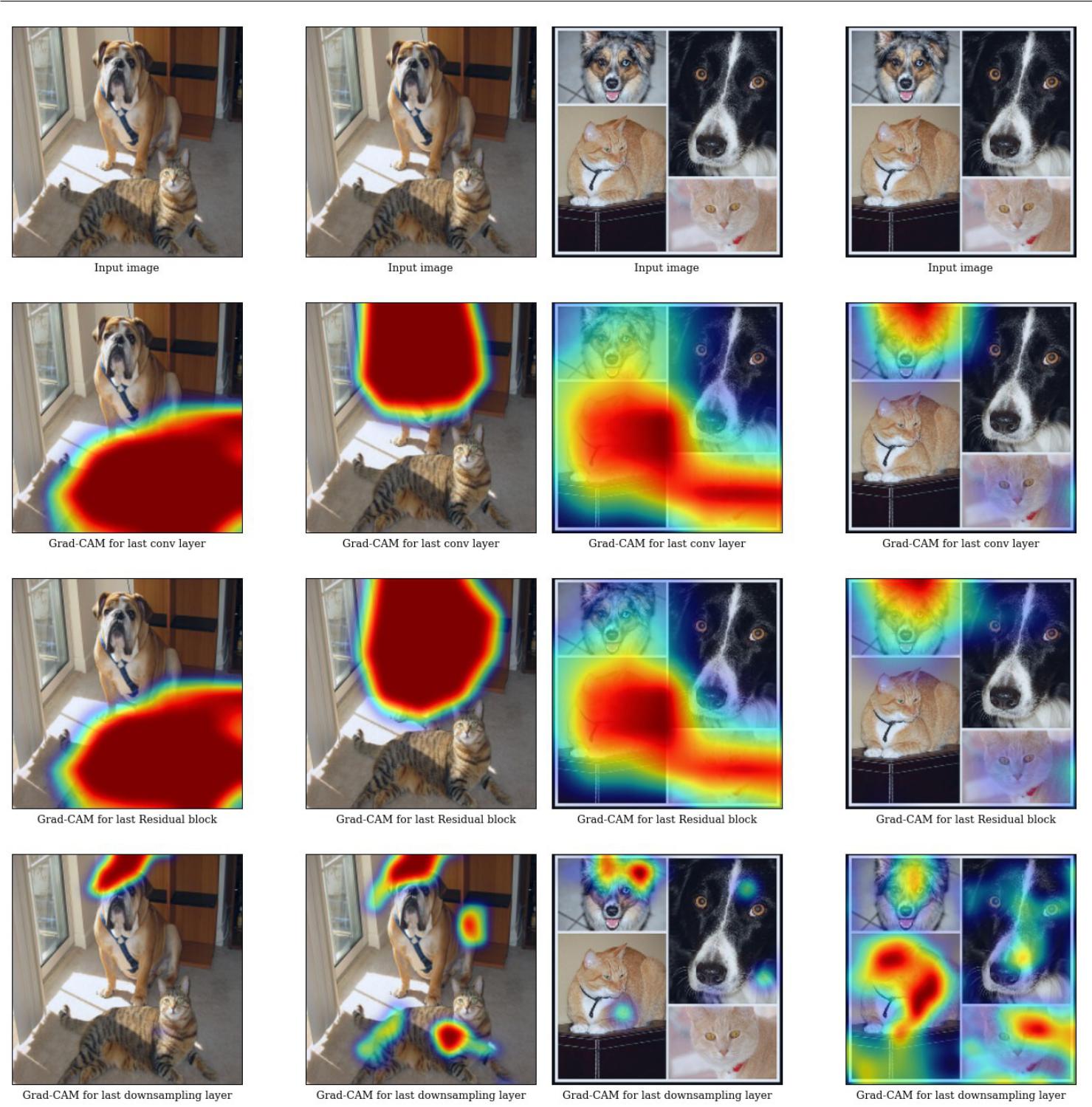
* Excitation Backprop（c-MWP）およびCAMとの定性的比較

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](#page23)]およびc-MWP [[58](#page23)]パスカル[[17](#page22)]。

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。グラウンドトゥルースカテゴリの視覚化は、図にあります。[20](#page19)。

* 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](#page20) 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](#page10)）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](#page23)]）。

21



（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](#page21)。

現在のResNets [[24](#page22)]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。

10 からの200層のResNetアーキテクチャを使用します [https：//](https://github.com/facebook/fb.resnet.torch.)

[github.com/facebook/fb.resnet.torch。](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](#page21) 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](#page21)）。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。

22

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 参考文献 | |  |  |  |  | 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](#page9) | |  |
| 1.1。 | A. Agrawal、D。Batra、およびD.Parikh。の動作の分析 | | | | | 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、 | |  |
|  | 視覚的な質問応答モデル。EMNLP、2016年。[2](#page2) | | | | |  | 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応答- | |  |
| 2.2。 | H. Agrawal、CS Mathialagan、Y。Goyal、N。Chavali、P。Banik、 | | | | |  | メント。arXiv preprint arXiv：1712.03316、2017。[1](#page1) | |  |
|  | A.モハパトラ、A。オスマン、D。バトラ。CloudCV：大規模 | | | | | 24。 | K. He、X。Zhang、S。Ren、およびJ.Sun。の深い残余学習 | |  |
|  | 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 | | | | | 25。 | 画像認識。2016年のCVPRで。[1](#page1)、 [2](#page2)、 [13](#page13)、 [21](#page21) | |  |
|  | 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](#page1) | | | | | D. Hoiem、Y。Chodpathumwan、およびQ.Dai。のエラーの診断 | |  |
| 3.3。 | S. Antol、A。Agrawal、J。Lu、M。Mitchell、D。Batra、C。LawrenceZit- | | | | |  | オブジェクト検出器。ECCV、2012年。[2](#page2) |  |  |
|  | ニック、およびD.パリク。VQA：視覚的な質問応答。ICCVでは、 | | | | | 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- | |  |
|  | 2015年。 [1](#page1)、 [2](#page2)、 [10](#page10)、 [12](#page12) |  |  |  |  |  | man Publishing Co.、Inc。、米国マサチューセッツ州ボストン、第3版、1998年。 | |  |
| 4.4。 | D. Bau、B。Zhou、A。Khosla、A。Oliva、およびA.Torralba。通信網 | | | | |  | [2](#page2) |  |  |
|  | 解剖：深い視覚表現の解釈可能性の定量化- | | | | | 27。 | Y. Jia、E。Shelhamer、J。Donahue、S。Karayev、J。Long、R。Girshick、 | |  |
|  | 。コンピュータビジョンとパターン認識、2017年。[1](#page1)、 [3](#page3)、 | | | | |  | S.グアダラマ、T。ダレル。Caffe：畳み込みアーキテクチャ | |  |
|  | [10](#page10) |  |  |  |  |  | 高速機能埋め込み用。ACM MM、2014年。[6](#page6) | |  |
| 5.5。 | L.バッツァーニ、A。ベルガモ、D。アンゲロフ、L。トレサーニ。自己- | | | | | 28。 | E. Johns、O。MacAodha、およびGJBrostow。エキスパートになる | |  |
|  | 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。 | | | | | 29。 | -インタラクティブなマルチクラスマシンティーチング。CVPR、2015年。[2](#page2) | |  |
|  | [4](#page4) |  |  |  |  | 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- | |  |
| 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： | | | | |  | 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、 | |  |
|  | レビューと新しい視点。 | | | パターン上のIEEEトランザクション | |  | 2016年。 [1](#page1)、 [10](#page10)、 [11](#page11) |  |  |
|  | 分析と機械知能、35（8）：1798–1828、2013年。 [4](#page4) | | | | | 30。 | A.カルパシー。ConvNetとの競合から学んだこと | |  |
| 7。 | X. Chen、H。Fang、T.-Y。Lin、R。Vedantam、S。Gupta、P.Dollar、 | | | | |  | ImageNetで。http://karpathy.github.io/2014/09/02/what-i-learned- | |  |
|  | とCLZitnick。Microsoft COCOのキャプション：データ収集と | | | | |  | from-competing-against-a-convnet-on-imagenet /、2014年。 [2](#page2) | |  |
|  | 評価サーバー。arXiv preprint arXiv：1504.0325、2015。[1](#page1)、 [10](#page10) | | | | | 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 | |  |
| 8.8。 | RG Cinbis、J。Verbeek、およびC.Schmid。 | | | | 弱く監視されている |  | 画像の説明を生成します。CVPR、2015年。[10](#page10)、 [11](#page11) | |  |
|  | 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 | | | | | 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： | |  |
|  | 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](#page3) | | | | |  | 弱教師あり画像セグメンテーションの3つの原則。に | |  |
| 9.9。 | A. Das、H。Agrawal、CL Zitnick、D。Parikh、およびD.Batra。人間 | | | | |  | ECCV、2016年。 [7](#page7) |  |  |
|  | 視覚的な質問応答における注意：人間と深いことを行う | | | | | 33。 | A. Krizhevsky、I。Sutskever、およびGEヒントン。Imagenet分類- | |  |
|  | ネットワークは同じ地域を見ていますか？EMNLP、2016年。[12](#page12) | | | | |  | 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](#page1)、 [5](#page5)、 | |  |
| 10.10。 | A. Das、S。Datta、G。Gkioxari、S。Lee、D。Parikh、およびD. Batra | | | | |  | [6](#page6)、 [16](#page16) |  |  |
|  | 具現化された質問応答。IEEECon-の議事録 | | | | | 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。 | |  |
|  | コンピュータービジョンとパターン認識（CVPR）に関する会議、2018年。 | | | | |  | [3](#page3) |  |  |
|  | [1](#page1) |  |  |  |  | 35。 | T.-Y. Lin、M。Maire、S。Belongie、J。Hays、P。Perona、D。Ramanan、 | |  |
| 11.11。 | A. Das、S。Kottur、K。Gupta、A。Singh、D。Yadav、JM Moura、 | | | | |  | P.Dollar、およびCLZitnick。Microsoft coco：の一般的なオブジェクト | |  |
|  | D.パリクとD.バトラ。ビジュアルダイアログ。IEEEの議事録 | | | | |  | 環境。ECCVで。2014年。[14](#page14)、 [15](#page15) |  |  |
|  | コンピュータビジョンとパターン認識（CVPR）に関する会議、 | | | | | 36。 | ZCリプトン。モデル解釈可能性の神話。ArXiv e-prints、 | |  |
|  | 2017年。 [1](#page1) |  |  |  |  |  | 2016年6月。 [2](#page2)、 [3](#page3) |  |  |
| 12.12。 | A. Das、S。Kottur、JM Moura、S。Lee、およびD.Batra。学習 | | | | | 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 | |  |
|  | 深い強化学習を伴う協調的視覚対話エージェント。 | | | | |  | セマンティックセグメンテーション用。CVPR、2015年。[1](#page1) | |  |
|  | コンピュータに関するIEEE国際会議の議事録 | | | | | 38。 | J. Lu、X。Lin、D。Batra、およびD.Parikh。 | より深いLSTMと |  |
|  | ビジョン（ICCV）、2017年。 [1](#page1) | |  |  |  |  | 正規化されたCNNビジュアル質問応答モデル。 [https：](https://github.com/VT-vision-lab/VQA_LSTM_CNN) | |  |
| 13.13。 | H. de Vries、F。Strub、S。Chandar、O。Pietquin、H。Larochelle、および | | | | |  | [//github.com/VT-vision-lab/VQA\_LSTM\_CNN](https://github.com/VT-vision-lab/VQA_LSTM_CNN)、 2015年。 | |  |
|  | ACクールビル。 | 何だと思う？！視覚オブジェクトの発見 | | | |  | [12](#page12)、 [13](#page13)、 [15](#page15) |  |  |
|  | マルチモーダルダイアログ。IEEE会議の議事録 | | | | | 39。 | J. Lu、J。Yang、D。Batra、およびD.Parikh。 | 階層的な質問- |  |
|  | コンピュータービジョンとパターン認識（CVPR）、2017年。 [1](#page1) | | | | |  | 視覚的な質問応答のための画像の共同注意。NIPSでは、2016年。 | |  |
| 14.14。 | J. Deng、W。Dong、R。Socher、L.-J。Li、K。Li、およびL.Fei-Fei。Im- | | | | |  | [13](#page13) |  |  |
|  | ageNet：大規模な階層画像データベース。CVPRでは、 | | | | | 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。 | |  |
|  | 2009年。 [2](#page2)、 [6](#page6)、 [14](#page14) |  |  |  |  |  | コンピュータビジョンに関するヨーロッパ会議、2016年。 [3](#page3) | |  |
| 15。 | A.DosovitskiyとT.Brox。畳み込みネットワークの反転 | | | | | 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 | |  |
|  | 畳み込みネットワークを使用します。CVPR、2015年。[3](#page3) | | | | |  | 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 | |  |
| 16.16。 | D. Erhan、Y。Bengio、A。Courville、およびP.Vincent。視覚化 | | | | |  | コンピュータビジョン、2016年1〜23ページ。 [3](#page3)、 [4](#page4) |  |  |
|  | 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、 | | | | | 42。 | M. Malinowski、M。Rohrbach、およびM.Fritz。ニューロンに聞いてください：A | |  |
|  | 1341、2009。 [3](#page3) |  |  |  |  |  | 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。に | |  |
| 17.17。 | M. Everingham、L。VanGool、CKI Williams、J。Winn、 | | | | |  | ICCV、2015年。 [1](#page1)、 [10](#page10)、 [12](#page12) |  |  |
|  | とA.ジサーマン。 | | PASCALビジュアルオブジェクトクラス | | | 43。 | M. Oquab、L。Bottou、I。Laptev、およびJ.Sivic。学習とトランスジェンダー | |  |
|  | チャレンジ2007 | （VOC2007） | | 結果。 | http：//www.pascal- |  | 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 | |  |
|  | network.org/challenges/VOC/voc2007/workshop/index.html、2009年。 | | | | |  | ネットワーク。CVPR、2014年。[3](#page3)、 [4](#page4) |  |  |
|  | [19](#page19)、 [20](#page20) |  |  |  |  | 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 | |  |
| 18.18。 | H. Fang、S。Gupta、F。Iandola、RK Srivastava、L。Deng、P.Dollar、 | | | | |  | 無料で？–畳み込みニューラルによる弱教師あり学習 | |  |
|  | J. Gao、X。He、M。Mitchell、JC Platt、他 キャプションから | | | | |  | ネットワーク。CVPR、2015年。[3](#page3) |  |  |
|  | ビジュアルコンセプトとバック。CVPR、2015年。[1](#page1) | | | |  | 45。 | POPinheiroとR.Collobert。画像レベルからピクセルレベルへ | |  |
| 19。 | C.ガン、N。ワン、Y。ヤン、D.-Y。ヨン、AGハウプトマン。 | | | | |  | 畳み込みネットワークによるラベリング。CVPR、2015年。[3](#page3) | |  |
|  | Devnet：マルチメディアイベント検出と | | | | | 46。 | M.レン、R。キロス、R。ゼメル。のモデルとデータの調査 | |  |
|  | 証拠の再集計。CVPR、2015年。[3](#page3) | | | |  |  | 画像の質問応答。NIPSでは、2015年。[1](#page1)、 [10](#page10)、 [12](#page12) | |  |
| 20。 | H. Gao、J。Mao、J。Zhou、Z。Huang、L。Wang、およびW. Xu あなたは | | | | | 47。 | MT Ribeiro、S。Singh、およびC.Guestrin。「なぜ私は信頼すべきなのか | |  |
|  | 機械と話している？多言語画像のデータセットと方法 | | | | |  | あなた？」：分類子の予測を説明します。SIGKDDでは、 | |  |
|  | 質問応答。NIPSでは、2015年。[1](#page1)、 [10](#page10)、 [12](#page12) | | | | |  | 2016年。 [3](#page3)、 [8](#page8) |  |  |
| 21。 | R. Girshick、J。Donahue、T。Darrell、およびJ.Malik。豊富な機能の階層- | | | | | 48。 | RR Selvaraju、P。Chattopadhyay、M。Elhoseiny、T。Sharma、D。Ba- | |  |
|  | 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。 | | | | |  | トラ、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 | |  |
|  | CVPR、2014年。 [1](#page1) |  |  |  |  |  | ニューロンを通じたドメイン知識-重要性。議事録 | |  |

23

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](#page3)

1. RRセルバラジュ、S。リー、Y。シェン、H。ジン、S。ゴーシュ、L。ヘック、D。バトラ、D。パリク。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](#page3)
2. 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ェ、J。シュリットウィーザー、I。アントノグロウ、V。パネールシェルバム、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](#page2)
3. 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](#page3)、 [6](#page6)
4. 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](#page5)、 [6](#page6)、 [14](#page14)、 [15](#page15)、 [16](#page16)
5. 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。CoRR、abs / 1412.6806、2014。[2](#page2)、 [3](#page3)、 [6](#page6)、 [14](#page14)
6. 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](#page6)、 [12](#page12)
7. O.ヴィニャルス、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](#page1)、 [10](#page10)、 [12](#page12)
8. 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](#page9)
9. MDツァイラーとR.ファーガス。会話型ネットワークの視覚化と理解。ECCV、2014年。[2](#page2)、 [3](#page3)、 [4](#page4)、 [6](#page6)、 [8](#page8)、 [10](#page10)、 [14](#page14)
10. J.チャン、Z。リン、J。ブラント、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](#page4)、 [6](#page6)、 [7](#page7)、 [12](#page12)、 [16](#page16)、 [20](#page20)
11. 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](#page2)、 [3](#page3)、 [5](#page5)、 [6](#page6)、 [20](#page20)
12. B. Zhou、A。Khosla、À。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](#page10)
13. 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](#page10)、 [11](#page11)、 [20](#page20)