

視覚変換器を用いた簡単なオープンボキャブラリーオブジェクトの検出

Matthias Minderer^{*}, Alexey Gritsenko^{*},
オースティン・ストーン、マキシム・ノイマン、ダーク・ヴァイセンボーン、アレクセイ・ドソヴィツキー。
Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen,
Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby

Google リサーチ

{[mjlm](mailto:mjlm@google.com),[agritsenko](mailto:agritsenko@google.com)}@google.com

概要 単純なアーキテクチャと大規模な事前学習を組み合わせることで、画像分類の大幅な改善がもたらされている。しかし、物体検出においては、事前学習とスケールアップアプローチはあまり確立されておらず、特に、学習データが比較的少ないロングテールやオープンボキャブラリーセッティングにおいて顕著である。本論文では、画像-テキストモデルをオープンボキャブラリーオブジェクト検出に変換するための強力なレシビを提案する。我々は標準的なVision Transformerアーキテクチャを用い、最小限の変更、対照的な画像テキストの事前学習、およびエンドツーエンドの検出の微調整を行う。このセットアップのスケールアップ特性の分析により、画像レベルの事前学習とモデルサイズを大きくすることで、下流の検出タスクに一貫した改善をもたらすことが示された。我々は、ゼロショットテキスト条件付きおよびワンショット画像条件付き物体検出で非常に高い性能を達成するために必要な適応戦略と正則化を提供します。コードとモデルはGitHubで公開されている¹。

キーワード：オープンボキャブラリー検出、変換器、視覚変換器、ゼロショット検出、画像条件付き検出、ワンショット物体検出、対比学習、画像-テキストモデル、基礎モデル、CLIP

1 はじめに

物体検出はコンピュータビジョンにおける基本的なタスクである。最近まで、物体検出モデルは一般的に少数の固定された意味カテゴリに制限されていた。これは、大きなラベル空間やオープンラベル空間を持つ局所的な学習データを得るにはコストと時間がかかるためである。しかし、強力な言語エンコーダと対照的な画像・テキスト学習の開発により、この問題は解決された。これらのモデルは、ウェブ上に豊富に存在する緩く整列した画像とテキストのペアから、画像とテキストの共有表現を学習する。大量の画像-テキストデータを活用することで、対照学習はゼロショット分類や他の言語ベースのタスクに大きな改善をもたらしました[33,19,44]。

^{*} コンセプトと技術的な貢献が同等であること。

¹ github.com/google-research/scenic/tree/main/scenic/projects/owl_vit

最近の多くの研究は、これらのモデルの言語能力を物体検出に転換することを目的としている[12,26,45,46,20]。これらの手法は、例えば、画像クロップの埋め込みに対する歪曲[12]、画像レベルのラベルによる弱い監視[46]、または自己学習[26,45]を用いる。本論文では、これらの手法を用いずに、学習中に見出されなかったカテゴリに対しても強力なオープンボキャブラリ検出を実現する、シンプルなアーキテクチャとエンドツーエンドの学習方法を提供する。

我々は、拡張性が高いことが示されているVision Transformerアーキテクチャ[22]から開始し、大規模な画像-テキストデータセット[44,19]で対比的に事前学習する。このモデルを検出に移行するために、我々は最小限の変更を行う。最終的なトークンプーリング層を削除し、その代わりに軽量の分類とボックスヘッドを各変換器出力トークンに付加する。分類層の重みをテキストモデル[2]から得られるクラス名の埋め込みに置き換えることで、オープンボキャブラリー分類を可能にする(図1)。我々は、二分割マッチング損失[6]を用いて、標準的な検出データセット上で事前学習したモデルを微調整している。画像モデルとテキストモデルの両方がエンド・ツー・エンドで微調整される。

この手法のスケーリング特性を分析し、モデルサイズと事前学習時間を増やすことで、200億画像・テキストペアを超えても検出性能の向上が続くことを見出した。これは、検出データとは対照的に画像-テキストペアが豊富であり、さらなるスケーリングが可能であるため、重要なことである。

我々のモデルの主な特徴はそのシンプルさとモジュール性である。我々のモデルは画像とテキストを融合していないため、クエリの表現源に依存しない。そのため、本モデルをそのまま一発検出学習器として利用することができ、画像由来の埋め込みを問い合わせるだけで利用できる。ワンショット物体検出は、対象物を示すクエリ画像パッチのみに基づいて新規物体を検出する困難な問題である[16,4,31]。画像条件付き一発検出は、テキスト条件付き検出の強力な拡張であり、特殊な技術部品のような、テキストで説明することが困難な(しかし画像で捉えることは容易な)オブジェクトを検出することができる。この問題に特化していない汎用的なアーキテクチャを用いたにもかかわらず、未見のCOCOカテゴリ(学習時に除外)に対する一発検出の技術水準を、AP50 26.0→41.8 と72%向上させることに成功した。

オープンボキャブラリーテキスト条件付き検出では、我々のモデルはLVISデータセットにおいて、全体で34.6%のAP、未見クラスで31.2%のAPを達成した。^{rare}

まとめると、以下のような貢献をすることになります。

1. 画像レベルの事前学習をオープンボキャブラリー物体検出に移行するためのシンプルで強力なレシピ。
2. 一発(画像条件)検出で大差をつける最先端。
3. 設計を正当化するための詳細なスケーリングとアブレーションスタディ。

このモデルは、様々なフレームワークに容易に実装できる強力なベースラインとして、また、オープンボキャブラリーローカリゼーションを必要とするタスクの将来の研究のための柔軟な出発点として機能すると考えている。我々はこの手法を*Vision Transformer for Open-World Localization*、略して**OWL-ViT**と呼んでいる。

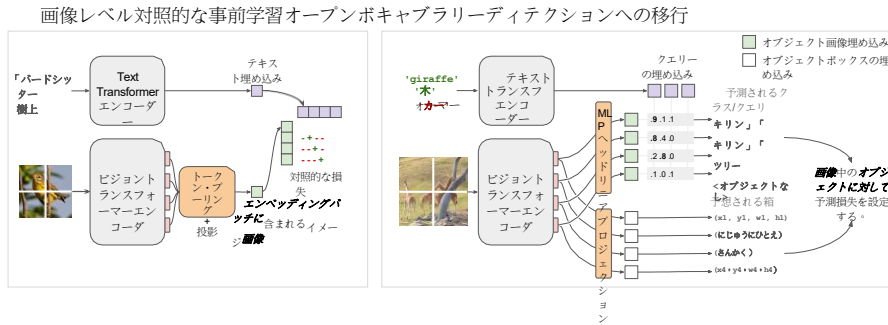


図1. 本手法の概要。左：CLIP [33], ALIGN [19], LiT [44]と同様に、画像とテキストのペアを用いて、画像とテキストのエンコーダを対比的に事前学習させる。右図次に、トークンのプーリングを除去し、軽量オブジェクト分類と局所化ヘッドを画像エンコーダ出力トークンに直接取り付けることにより、事前学習したエンコーダをオープンボキャブラリーオブジェクト検出に移行させる。オープンボキャブラリー検出を実現するために、クエリ文字列はテキストエンコーダで埋め込まれ、分類に使用されます。このモデルは標準的な検出データセットで微調整される。推論時には、オープンボキャブラリー検出にはテキスト由来の埋め込みを、少数ショットの画像条件付き検出には画像由来の埋め込みを用いることができる。

2 関連作品

対照的な視覚-言語事前学習。 画像とテキストを共有空間に埋め込むというアイデアは、長い間「ゼロショット」汎化を達成するために使用されてきた[10,36,40]。対照的な損失とより良いアーキテクチャの革新のおかげで、最近のモデルは、人間の明示的な注釈を必要とせずに、ウェブ由来の画像とテキストのペアから一貫した視覚と言語の表現を学習することができます。これにより、利用可能な学習データが大幅に増加し、ゼロショット分類ベンチマーク[33,19,44,32]の大幅な改善につながった。最近の画像-テキストモデルのどれもが我々のアプローチと互換性がありますが、我々のモデルとデータセットはLiT [44]とALIGN [19]に最も類似しています。

閉語的な物体検出。 物体検出モデルは、従来から閉語彙の設定に対して定式化されてきた。当初はSSD [28]やFaster-RCNN [34]などの「1段階」「2段階」検出器が開発された。さらに最近では、DETR [6]が、物体検出を集合予測問題として構成し、二分割マッチングで学習させ、競合する結果を得ることができることを示した。注目すべきは、このようなアーキテクチャでは、領域提案の生成や非最大限の抑制を必要としないことである。後続の研究は、「デコーダステージ」を持たないアーキテクチャ[9]を含む、より効率的なDETRの変形を提案している[48,41,37]。我々の研究は、デコーダを使用しないという点で DETR を単純化する。追加の「検出」トークンを用いる[9]と比較して、我々は各画像トークンから直接1つのオブジェクトインスタンスを予測することでモデルをさらに簡略化する。

ロングテールおよびオープンボキャブラリーオブジェクトの検出。 クローズドボキャブラリーを超えるために、固定された分類レイヤーを言語エミュレータで置き換えることができる。

を構築している[2]。最近、対照的に学習された画像-テキストモデルと古典的なオブジェクト検出器を組み合わせることにより、オープンボキャブラリーオブジェクト検出が大きく発展している[12,20,26,45,46,42]。このタスクの主な課題は、希少なクラスに対する局所的な注釈が乏しいにもかかわらず、画像テキストのバックボーンの画像レベルの表現を検出にどのように移行させるかということである。画像テキストの事前学習を効率的に利用することは、高価な人間のアノテーションを必要とせずにスケーリングが可能になるため、非常に重要である。これまで、様々なアプローチが提案されてきた。**ViLD** [12]は、切り出した画像領域に対してCLIPやALIGNを適用して得られた埋め込み情報を、クラスにとらわれない領域提案ネットワーク(RPN)から抽出するものである。しかし、RPNは新規物体に対する汎化性能に限界があり、これはViLDの2段階の蒸留-学習過程によって悪化する。また、**RegionCLIP**では、キャプションデータから擬似ラベルを生成し、領域-テキスト対比の事前学習を行い、検出へと移行する多段学習が行われている。これに対し、本手法は、一般に公開されている検出データセット上で画像モデルとテキストモデルをエンドツーエンドで微調整することで、学習を簡略化し、未知のクラスへの汎化を向上させる。**MDETR** [20]と**GLIP** [26]は画像全体に対して単一のテキストクエリを使用し、検出をフレーズ接地問題として定式化している。このため、1回のフォワードパスで処理できるオブジェクトカテゴリの数が制限される。我々のアーキテクチャは画像とテキストの融合を行わず、複数の独立したテキストまたは画像由来のクエリを扱えるという点でよりシンプルかつ柔軟である。**OVR-CNN** [42]は限られた語彙での検出のために画像テキストモデルを微調整し、オープンな語彙への汎化のために画像テキストの事前学習に依存するという点で我々のアプローチと最も類似しています。しかし、我々はすべてのモデリングと損失関数の選択において異なっている。我々はResNet [15]の代わりにViT [22]を、Faster-RCNN [34]の代わりにDETRに似たモデルを、PixelBERT [18]と視覚接地損失の代わりにLiT [44]のように画像テキストの事前学習を使用する。我々のアプローチと直交する**Detic** [46]は、画像レベルのアノテーションのみが利用可能な例に対して分類ヘッドのみを学習することにより、弱い監視下でロングテール検出性能を向上させます。

オープンボキャブラリー検出の定義では、オブジェクトカテゴリは検出トレーニングとテストの間で重複する可能性があることに注意する。また、学習時に局所的なインスタンスが見られなかったカテゴリを検出することを特に指す場合、**ゼロショット**という用語を用いる。

画像条件付き検出。オープンボキャブラリー検出と関連して、画像条件付き検出のタスクがあります。これは、問題のカテゴリのオブジェクトを示す単一のクエリ画像に一致するオブジェクトを検出する能力を指します[4,16,7,31]。このタスクは、クエリ画像が本質的に単一の学習例であるため、**ワンショット・オブジェクト検出**とも呼ばれる。画像ベースの問い合わせは、例えば、ユニークな物体や特殊な技術部品など、物体の**名前**さえ不明な場合に、オープンワールドの検出を可能にする。我々のモデルは、テキスト由来の埋込みの代わりに画像由来の埋込みをクエリとして用いるだけで、このタスクを修正することなく実行することができる。この問題に対する最近の先行研究は、例えば、クエリとターゲット画像の間の洗練された形のクロスアテンションを用いるなど、主にアーキテクチャ上の革新に焦点を当てている[16,7]。我々のアプローチは、単純だが大規模なモデルと、広範な画像-テキスト事前学習に依存している。

3 方法

我々の目標は、シンプルでスケーラブルなオープンボキャブラリーオブジェクト検出器を作成することである。我々は、スケーラビリティ[22]と閉語検出の成功[6]のため、標準的なTransformerベースのモデルに焦点を当てる。我々は2段階のレシピを提示する。

1. 大規模な画像・テキストデータで画像・テキストエンコーダを対比的に事前学習する。
2. 検出ヘッドを追加し、中程度の検出データで微調整を行う。
このモデルを様々な方法で照会し、オープンボキャブラリーや少数点検出を実行することができる。

3.1 モデル

アーキテクチャ我々のモデルは画像エンコーダとして標準的なVision Transformerを用い、テキストエンコーダとして同様のTransformerアーキテクチャを用いる (Fig.1)。検出のために画像エンコーダを適応させるために、トークンプールと最終投影層を削除し、代わりに分類のためにオブジェクトごとの画像埋め込みを得るために各出力トークン表現を線形に投影する (図1、右)。したがって、予測されるオブジェクトの最大数は、画像エンコーダのトークン数 (シーケンス長) に等しくなる。我々のモデルのシーケンス長は少なくとも576 (入力サイズ768×768のViT-B/32) であり、これは今日のデータセットの最大インスタンス数 (例えばLVIS [13]では294インスタンス) より大きいので、これは実際のところボトルネックにはなっていない。ボックス座標はトークン表現を小さなMLPに通過させることで得られる。我々の設定はDETR[6]に似ているが、デコーダを削除することで簡略化されている。

オープンボキャブラリーオブジェクトの検出。検出されたオブジェクトのオープンボキャブラリー分類のために、我々は先行研究に従い、分類ヘッドの出力層に学習されたクラス埋め込みではなく、テキスト埋め込みを用いる[2]。テキスト埋め込み (我々はクエリーと呼ぶ) は、カテゴリ名や他のテキストオブジェクトの説明をテキストエンコーダーに渡すことによって得られる。そして、このモデルのタスクは、各オブジェクトについて、各クエリがそのオブジェクトに適用されるバウンディングボックスと確率を予測することになる。クエリーは画像ごとに異なる可能性があります。つまり、各画像はテキスト文字列の集合によって定義される独自の判別可能なラベル空間を持つ。このアプローチは、古典的な閉語彙物体検出を、物体カテゴリ名の完全なセットを各画像のクエリーセットとして使用する特別な場合として包含する。

他のいくつかの方法[26,20]とは対照的に、我々は画像に対するすべての問合せを一つのトークンシーケンスにまとめることはしない。その代わり、各クエリは個々のオブジェクトの説明を表す別々のトークン列からなり、テキスト・エンコーダによって個別に処理される。さらに、我々のアーキテクチャでは、画像エンコーダとテキストエンコーダの間の融合は行われません。早期の融合は直感的に有益に思えるが、クエリのエンコードには画像モデル全体を通過する必要があり、画像とクエリの組み合わせごとに繰り返す必要があるため、推論効率を劇的に低下させる。私たちは、画像とは別にクエリの埋め込みを計算することができるので、画像ごとに何千ものクエリを使用することができ、これはearly fusion [26]で可能な数よりはるかに多い。

1ショットまたは数ショットの転送。本システムでは、問い合わせの埋め込みがテキスト由来である必要はない。画像エンコーダとテキストエンコーダの融合がないため、分類ヘッドへのクエリとして、テキストではなく画像由来の埋め込みを、モデルを修正することなく供給することが可能である。このように、典型的な物体画像の埋め込みを問合せとして用いることで、本モデルは画像条件付きの一発物体検出を行うことができる。このように、画像埋め込みを問い合わせとして用いることで、テキストでは表現しにくい物体を検出することができる。

3.2 トレーニング

画像レベルの対比的事前学習。44]と同じ画像-テキストデータセットと損失を用いて、画像とテキストのエンコーダを対照的に事前学習する(図1、左)。画像表現とテキスト表現に対して、対照的な損失を用いてランダムな初期化で両方のエンコーダをゼロから学習する。画像表現には、マルチヘッド・アテンション・プーリング(MAP) [25,43]を用いてトークン表現を集約している。テキスト表現はテキストエンコーダの最後のEOS (End-of-Sequence) トークンから得られる。また、一般に公開されている事前学習済みのCLIPモデル[33]を利用する(詳細は付録A1.3)。

エンコーダのみのアーキテクチャの利点は、モデルのほぼすべてのパラメータ(画像およびテキストエンコーダ)が画像レベルの事前学習から恩恵を受けることができることです。検出専用ヘッドは、モデルのパラメータの最大1.1%(モデルサイズに依存)を含みます。

検出器のトレーニング 分類のための事前学習済みモデルの微調整はよく研究されている問題である。分類器、特に大規模な変換器は、うまく機能させるために正則化とデータ補強を十分に調整する必要があります。分類器の学習のためのレシピは、現在、文献[39,38,3]でよく確立されています。ここでは、オープンボキャブラリ検出のための同様の微調整のレシピを提供することを目的とする。

本モデルの一般的な検出学習方法は、閉語彙検出器とほぼ同じであるが、各画像に対して物体カテゴリ名の集合をクエリとして与えている点が異なる。そのため、分類器は固定されたグローバルなラベル空間ではなく、クエリによって定義された画像ごとのラベル空間上でロジットを出力する。

DETR [6]によって導入された二分割マッチング損失を用いるが、以下のようにロングテール/オープンボキャブラリーディテクションに適応させる。検出用データセットを網羅的にアノテーションするのに必要な労力のため、多数のクラスを持つデータセットはフェデレート方式でアノテーションされる[13,24]。このようなデータセットではラベル空間が非分離型であり、各オブジェクトが複数のラベルを持ちうることになる。そのため、分類損失として、ソフトマックスクロスエントロピーの代わりにフォーカルシグモイドクロスエントロピー[48]を使用する。さらに、全ての画像に全てのオブジェクトカテゴリが注釈されているわけではないので、連携データセットでは、各画像に対してポジティブ(存在する)注釈とネガティブ(存在しないことが分かっている)注釈の両方が提供される。学習時には、与えられた画像に対して、その正負のアノテーションを全てクエリとして使用する。さらに、データ中の頻度に比例してカテゴリをランダムに抽出し、「擬似ネガティブ」として追加し、1画像あたり少なくとも50個のネガティブを持つようにする[47]。

最大規模の連合検出データセットでさえ、 $\approx 10^6$ の画像しか含まれておらず、数十億の画像レベルの弱いラベルが存在するのは対照的です。

のような大規模なTransformerを学習させることができます [29,43,33,19]。このサイズのデータセット (ImageNet-1kなど) で学習した大規模Transformerが良い性能を発揮するためには、慎重に調整された正則化とデータ増強が必要であることが知られています [39,38,3]。我々は、検出学習においても同様のことが言えると考え、セクション4.6において、大規模Transformerで非常に高い性能を達成するために必要な補強と正則化の詳細な内訳を提供しています。

4 実験風景

4.1 モデル詳細

画像モデルには、標準的なVision Transformers [22]を使用する。モデルサイズ、パッチサイズ、トランスフォーマー対ハイブリッドアーキテクチャについては[22]の命名法に従っている。例えば、B/32はパッチサイズ32のViT-Baseを指し、R50+H/32はストライド32のResNet50 + ViT-Hugeのハイブリッドを指している。

テキストモデルには、画像モデルと同様のTransformerアーキテクチャを使用しています。特に断りのない限り、12層、512の隠れサイズ (D)、2048のMLPサイズ、8ヘッド (これはBより小さい) のテキストモデルを使用します。

画像とテキストのモデルは、まず画像レベルで事前学習され、次にオブジェクトレベルのアノテーションで微調整される。事前学習はLiT[44] (彼らの表記ではuu) のように、36億の画像-テキストペアのデータセットに対してゼロから行われます。

事前学習後、トークンプーリングを除去し、検出ヘッドを追加する (セクション3.1および図1参照)。このモデルは各出力トークンに対して1つのボックスを予測する。予測されたボックスの座標にバイアスを加え、トークン列を2次元グリッドとして配置したとき、各ボックスはデフォルトでこのボックスが予測されたトークンに対応する画像パッチを中心とするようにする。したがって、このモデルは、Region Proposal Networks [34]が予め定義されたアンカーに対するオフセットを予測するのと同様に、そのデフォルト位置からの差異を予測する。画像パッチとトークン表現の間には厳密な対応はないが、このようにボックス予測にバイアスをかけることで、学習のスピードアップと最終的な性能の向上が図れる (セクション4.6)。

ほとんどのモデルで、事前学習には 224×224 の画像サイズを用い (Appendix A1.3 参照)、検出の微調整や評価にはより大きなサイズを用いる (表1に規定)。事前学習後にモデルの入力サイズを変更するために、画像位置の埋め込みを線形補間でサイズ変更する。モデルの微調整は、バッチサイズ256で最大14万ステップ (大規模なモデルではより少ないステップ) 行う。このモデルはJAX[5]とScenicライブラリ[8]を用いて実装されている。

4.2 検出データ

本モデルのオープンボキャブラリーデザインにより、整数ラベルをクラス名文字列に置き換えることで、異なるラベル空間を持つデータセットを容易に組み合わせることができる。オブジェクトレベルの学習には、合計約200万画像の一般公開されている検出データセット (OpenImages V4 (OI) [24], Objects 365 (O365) [35], 及び/又はVisual Genome (VG) [23], as indicated) を使用します。評価はCOCO[27]、LVIS[13]、O365で行っている。データセットの詳細については、付録 A1.2 を参照のこと。

表1. LVIS v1.0 valのオープンボキャブラリーとゼロショット性能。我々のモデルでは、 AP^{LVIS} 、ゼロショット性能を測定するように、すべての検出トレーニングデータセットからLVISレアカテゴリ一名に一致するアノテーションを削除している。灰色の数字は、LVISの頻度の高いアノテーションと共通のアノテーション（「基本」）で学習したモデルを示しています。参考までに、ViT-B/32はResNet50と同等の推論計算量（139.6対141.5GFLOPs）である。また、3回の微調整を行った場合の平均性能を報告する。COCOとO365の結果は付録A1.8に記載した。

方法	バックボーン	画像レベル	オブジェクトレベル	AP^{COCO}	AP^{LVIS}	希少
LVISのベストトレーニング。						
1 ViLD-ens [12]	ResNet50	クリップ	LVISベース	1024	25.5	16.6
2 ViLD-ens [12]	EffNet-b7	ALIGN	LVISベース	1024	29.3	26.3
3 Reg.CLIP [45]	R50-C4	CC3M	LVISベース	?	28.2	17.1
4 レジCLIP [45]	R50x4-C4	CC3M	LVISベース	?	32.3	22.0
5 OWL-ViT(当社比)	ViT-H/14	LiT	LVISベース	840	35.3	23.3
6 OWL-ViT(当社製品)	ViT-L/14	クリップ	LVISベース	840	34.7	25.6
無制限のオープンボキャブラリートレーニング。						
7 GLIP [26] (グリップ)	スウィンT	キャップフ オーエム	O365、GoldG、 ...	?	17.2	10.1
8 GLIP [26] (グリップ)	スウィンL	CC12M、 SBU	OI, O365, VG, ...	?	26.9	17.1
9 OWL-ViT (当社)	ViT-B/32	LiT	O365、VG	768	23.3	19.7
11 OWL-ViT (当社)	R26+B/32	LiT	O365、VG	768	25.7	21.6
10 OWL-ViT (当社製品)	ViT-B/16	LiT	O365、VG	768	26.7	23.6
12 OWL-ViT (当社製品)	ViT-L/16	LiT	O365、VG	768	30.9	28.8
13 OWL-ViT (当社)	ViT-H/14	LiT	O365、VG	840	33.6	30.6
14 OWL-ViT (当社)	ViT-B/32	クリップ	O365、VG	768	22.1	18.9
15 OWL-ViT (当社)	ViT-B/16	クリップ	O365、VG	768	27.2	20.6
16 OWL-ViT (当社)	ViT-L/14	クリップ	O365、VG	840	34.6	31.2

OI、VG、O365、画像レベルの事前学習データには、COCO / LVISにも含まれる画像があるため、学習に使用するすべてのデータセットからCOCOまたはLVISのテスト画像と検証画像を削除する厳格な重複排除手順を使用している（詳細は付録A1.2参照）。特に断りのない限り、我々の実験では検出学習にOIとVGを70%対30%の割合でランダムに混合している。表 1 では、先行研究との比較のため、LVIS をベースとした学習、または O365 と VG を 80%から 20%の割合で使用した。また、様々な画像やラベルの補強を行ったが、これについては4.6節で説明する。

4.3 オープンボキャブラリー検出性能

LVIS v1.0 val [13]の主なベンチマークは、このデータセットが希少なカテゴリのロングテールを持ち、オープンボキャブラリーの性能を測定するのに適しているためである。評価には、各画像に対して全てのカテゴリ名をクエリとして用いる。

すなわち、LVISでは1画像あたり1203のクエリである。クラス予測はセクション4.6で説明したように7つのプロンプトテンプレートに対してアンサンブルされます。LVISのいくつかのカテゴリは我々が学習に使用するデータセットに含まれている。このため、LVISの「まれな」カテゴリのいずれかに一致するラベルを持つボックスアノテーションをすべて学習データから削除し、未知のカテゴリに対する性能を測定する。したがって、 AP^{LVIS} メトリックは以下のように測定する。

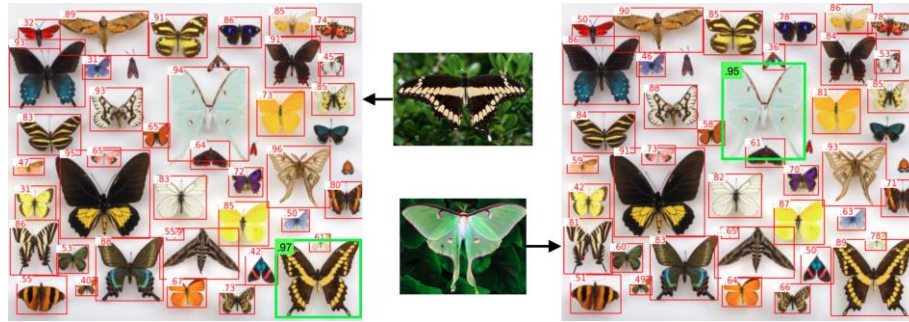


図2.ワンショット画像条件付き検出の例。中央の画像はクエリとして使用され、左と右はターゲット画像上でのそれぞれの検出結果を示している。いずれの場合も、クエリに一致する種のインスタンスに最高スコアが与えられる。一方、テキストベースのクエリ（図示せず）では、上の例（「アゲハチョウ」）のみ正しい種が検出され、下の例（「ルナガン」）は検出されない。

は、これらのカテゴリのローカライズされたアノテーションを見たことがないという意味で、我々のモデルの「ゼロショット」性能である。

表1は我々のモデルと先行研究によるLVISの結果である。LVISの全データセットで学習しないオープンボキャブラリーモデルと比較している。LVISの一部（例えば「ベース」カテゴリ[12]）で学習した結果はグレーで示されている。我々の手法は、オープンボキャブラリーシナリオ（ AP^{LVIS} ）とゼロショットシナリオ（ AP^{LVIS} ）の両方において、アーキテクチャサイズにより高い競争力を有している。我々の最良のモデルは31.2%の AP^{LVIS} を達成し、一般に利用可能なCLIPバックボーンを使用しています。

先行研究との比較のため、MS-COCO 2017とObjects 365での結果も示しています。これらの評価では、汎化性を測定するために、O365+VGの代わりにOI+VGでモデルを訓練しています。しかし、ほとんどのCOCOとO365のカテゴリは学習データに存在し、利用可能なアノテーションの大きな割合を構成しているため、それらを削除しないようにした。したがって、COCOとO365の結果は「ゼロショット」ではなく、我々のモデルのオープンボキャブラリートランスファー能力をテストしています。我々の最良のモデル（CLIP L/14；表1参照）は43.5%の AP^{COCO} を達成した。O365を使わずに学習したモデルのバージョンは15.8%の AP^{O365} を達成した（さらなる再結果は付録A1.8を参照）。

4.4 少数撮影画像による条件付き検出性能

3.1節で述べたように、本モデルでは、テキスト由来のクエリ埋め込みを画像由来のクエリ埋め込みに置き換えるだけで、1ショットまたは数ショットの物体検出を行うことができる。数ショット検出の場合、オブジェクトを囲む枠の付いたクエリ画像が与えられる。目標は、新しいターゲット画像から例と同じカテゴリのオブジェクトを検出することである。クエリの埋め込みを得るために、まずクエリ画像に対して推論を行い、クエリボックスとボックスの重なりが大きい予測検出を選択する（いくつかのフィルタリングの後；詳細は付録A1.7を参照のこと）。そして、その予測の画像埋め込みをテスト画像に対するクエリとして用いる。

表2・COCO AP50における1ショットおよび数ショットの画像条件付き検出性能。我々の手法（R50+H/32アーキテクチャ）は先行研究を強く上回り、また、条件付けクエリの数を $k=10$ に増やすと顕著な改善が見られる。COCOのカテゴリ分割は[16]と同様です。評価は確率的であるため、我々の結果については、3回の実行の平均を報告する。

方法	スプリット1	スプリット2	スプリット3	スプリット4	平均値
1 ショット	サイアムマスク【30】	38.9	37.1	37.8	36.6
	CoAE [16] の場合	42.2	40.2	39.9	41.3
	AIT [7] (エイト)	50.1	47.2	45.8	46.9
	OWL-ViT (当社)	49.9	49.1	49.2	48.2
	OWL-ViT ($k=10$; 我々)	54.1	55.3	56.2	54.9
10 ショット	サイアムマスク【30】	15.3	17.6	17.4	17.0
	CoAE [16] の場合	23.4	23.6	20.5	20.4
	AIT [7] (エイト)	26.0	26.4	22.3	22.6
	OWL-ViT (当社)	43.6	41.3	40.2	41.9
	OWL-ViT ($k=10$; 我々)	49.3	51.1	42.4	44.5

このタスクの評価は、[16]に記載されている手順に従う。検出学習時に、評価対象となるCOCOカテゴリを抽出し、さらに検出学習データに現れる同義・意味的に下位のカテゴリをすべて抽出する。

画像-テキスト事前学習段階は変更しない。このタスクのために特別に設計されたわけではないにもかかわらず、表2に示すように、我々のモデルは4つのCOCO分割において、72%の-marginでタスク固有の最高の先行研究を強く上回る性能を発揮する。先行研究とは異なり、我々のモデルは推論中にクエリ画像とターゲット画像の特徴を取り込まないため、何千もの異なる画像埋め込みに対して同時にモデルを実行することが可能である。

を効率的に行い、実用性を高めています。

単一のクエリ例（ワンショット）から数ショットの予測に移行するには、各カテゴリの複数のクエリ例に対する画像埋め込みを単純に平均化すればよい。これにより、さらに大きな改善効果が得られる（表2最下段）。

4.5 画像レベルの事前トレーニングのスケールアップ

本手法が強力なオープンボキャブラリ、ゼロショット、画像条件付き検出性能を達成することを確認した後、次にそのスケールアップ特性と設計選択を分析する。本節では、画像レベルの事前学習に着目する。セクション4.6では、事前学習されたモデルをうまく検出に移行させるために必要な微調整の方法について説明する。

画像レベルの事前学習が最終的な検出性能とどのように関連するかを理解するために、事前学習時間、モデルサイズ、モデルアーキテクチャの次元を系統的に検討した。これらのパラメータは構成によって最適な設定が異なるため、各構成について、学習率と重みの減少の範囲にわたって複数のモデルを事前学習し、微調整を行った（対象設定の一覧は付録 A1.3 を参照）。

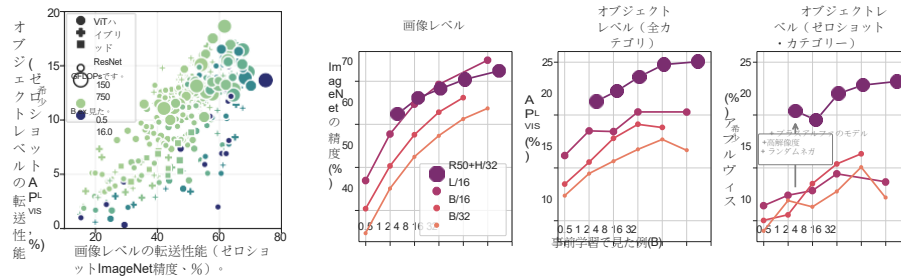


図3.画像レベルの事前学習が検出へ移行する。左：事前学習した画像テキストモデルの画像レベル性能（事前学習後のゼロショットImageNet精度）とオブジェクトレベル性能（検出微調整後の AP^{Lvis} ）の関係の概要。各ドットは、1つの事前学習設定と、学習率とウェイトの減衰の範囲におけるその最高の検出性能を表しています。構成は、エンコーダアーキテクチャ（ViT/Hybrid/ResNet）、モデルサイズ（検出推論計算の順番で異なる。モデルサイズ（検出推論計算順：R50、B/32、R26+B/32、R101、L/32、B/16、H/32、R50+H/32、L/16）、事前学習時間（繰り返しを含む数十億例；36億ユニーク例）が異なります。画像レベルの性能は物体レベルの性能に必要なが、十分ではない（ピアソンの $r = 0.73$ 、一方、画像レベルの伝達性能は事前学習課題の性能と良い相関がある： $r = 0.98$ ）。右図モデルサイズによらず、画像レベルの事前学習が長いほど、物体レベルの性能が向上する。また、微調整の規模を大きくすることで、検出性能のさらなる向上が期待できる。

まず、画像レベルの事前学習が一般的な検出にどの程度移行するのかを検討します。図3は、我々の研究でカバーした全てのアーキテクチャ、サイズ、事前学習期間の構成について、画像レベルの性能（ゼロショットImageNet精度）とオブジェクトレベルの性能（ゼロショット AP^{Lvis} ）の関係を示しています（学習率と重み減衰に渡る最良の結果を示しています）。我々は、最良の物体レベルのモデルは一般的に画像レベルの性能も高いが、逆は真ならず、画像レベルのタスクに優れたモデルの多くは検出にはうまく移行しないことを見出した。言い換えれば、画像レベルの性能が高いことは、検出への強力な移行のために必要ですが、十分ではありません。

どのような要因が強力な転移に寄与しているのでしょうか？分類に関する先行研究では、最適な転送を実現するためには、事前学習とモデルサイズを一緒にスケールする必要があることがわかりました。我々は、この効果が検出への移植においてさらに強いことを発見した。事前学習量を増やすと、検出性能は最初上昇しますが、その後ピークに達し、画像レベルの性能は上昇し続けます（図3、右）。しかし、モデルサイズを大きくし、検出の微調整を改善することで、事前学習による検出性能の正のトレンドは拡大することができる（図3右、R50+H/32）。

モデルサイズを大きくすることで性能が向上することを考えると、どのアーキテクチャが最も好ましいスケーリング特性を持つかということが重要な課題である。トランスフォーマーに基づくアーキテクチャは、ResNetsよりも事前学習の計算効率が高く、ResNetとトランスフォーマーのハイブリッドアーキテクチャは、少なくとも計算量の少ない芽では最も効率が高いことが分かっている。

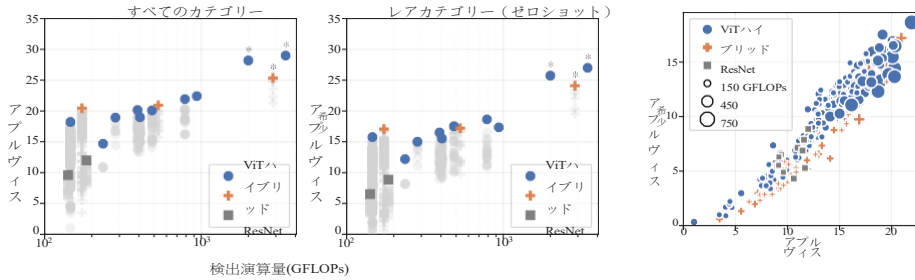


図4.図 4：検出性能に対するモデルアーキテクチャの影響。左：ハイブリッド・アーキテクチャ（ \times ）ヒット・アーキテクチャは、小さなモデルでは純粋なトランスフォーマーよりも効率的です。モデルサイズが大きくなると（検出推論FLOPsの観点から）、純粋なViTは、全体およびゼロショット性能の両方でハイブリッドよりも良好にスケールします。純粋なResNetsは、我々のセットアップでは性能が低い。色のついたマーカーは、探索された全てのハイパーパラメータにおいて、与えられたサイズの最適なモデルを示し、薄い灰色のマーカーは最適でないハイパーパラメータを示す。アスタリスク（*）はランダムなネガティブラベルで学習させたモデルを示す。右図アーキテクチャもタスクのどの側面をモデルが学習するか。純粋なViTは、ゼロショット検出（ AP^{Lvis} ）において、与えられたオブジェクトレベルの総合性能（ AP^{Lvis} ）において、ハイブリッドアーキテクチャよりも系統的に優れた性能を発揮する。我々は、ViTは意味的汎化の学習に偏り、ResNets/Hybridは既知のクラスのローカライズの学習に偏っていると推測している。この差は、モデルサイズと性能が大きくなるにつれて小さくなる。

を取得する[22]。また、ResNets は事前学習データが少ない場合に優れているが、利用可能なデータが増加すると Transformers に追い越されることが分かっている[22,38]。検出に関しても同様の分析を行った。検出の推論量をモデルサイズの指標とし、各サイズに最適なハイパーパラメータと事前学習時間を選択したところ、モデルサイズが小さい場合はハイブリッドモデルが純粋なViTよりも効率的であり、ResNetsは我々のセットアップでは性能が低いことがわかった（図4）。しかし、大規模なモデルでは、純粋なViTがハイブリッドを追い越す。この違いを説明するために、オーバーオールとゼロショット検出の性能を比較したところ、ハイブリッドと純粋なトランスフォーマーの間に明確な違いがあることがわかりました（少なくとも小さなモデルサイズにおいて；図4、右）。これはおそらく、Transformerがハイブリッド構造よりも（高いゼロショット性能に必要な）意味的汎化の学習に偏っていることを示しており、大規模な事前学習が可能な場合には有益である可能性があります。全体として、我々の発見は分類のためのものを超え、さらなるスケールアップの努力は純粋なTransformerアーキテクチャに焦点を当てるべきであることを示唆する。

4.6 事前トレーニングで検出の可能性を引き出す方法

セクション4.5では、強力な検出性能には強力な画像レベルの性能が必要であるが、十分ではないことを明らかにした。ここでは、画像レベルの事前学習後に強力なオープン語彙検出性能を得るための我々のレシピを説明する。本レシピのすべての構成要素は、比較的少数の検出用アノテーションとアノテーションによってカバーされる小さな意味的ラベル空間に対するオーバーフィットを減少させることを目的としている。我々のアプローチは

(i)最適化を安定させるための方策、(ii)利用可能な検出方法の慎重な利用

表 3.画像-テキストモデルの検出への移行を成功させるために必要な主な方法論の改善に関するアブレーションスタディ。簡略化のため、ベースラインに対するAPの差分を示している。LVISレアラベルを再トレーニングする実験（最後の行）を除いて、すべての差は負になることが予想される。分散を減らすために、すべての結果は2つの複製で平均化されている。すべてのアブレーションはViT-R26+B/32モデルで実施され、特に指定がない限り70Kステップのトレーニングスケジュールを使用した。

アブレーション	AP ^{LVIS}	AP ^{LVIS}	AP ^{COCO}	AP ^{OI}
ベースライン	21.0	18.9	30.9	54.1
(1) VGはトレーニングにのみ使用する	-14.5	-14.0	-23.6	-38.3
(2) OIはトレーニングにのみ使用する	-6.9	-5.7	-4.2	0.3
(3) 画像符号化装置と文字符号化装置は同じLR	-3.0	-8.5	-0.5	0.4
(4) 推論時の迅速なアンサンブルなし	-2.8	-5.5	-5.9	-0.1
(5) プロンプトなし（電車や推論など）	-1.2	-1.3	-0.6	-6.3
(6) ランダムネガがないこと	-1.0	-2.8	-0.4	1.0
(7) モザイクがない	-2.3	-1.5	-1.7	-0.7
(8) モザイクなし、2倍速の列車。	-2.9	-2.8	-1.8	-0.7
(9) モザイクなし、3倍速の列車。	-3.4	-3.6	-1.8	-0.8
(10) 重複するインスタンスをマージしないこと	-0.8	-1.3	-0.6	-0.7
(11) ボックス予測因子における位置の偏りがないこと	-1.2	-1.1	-1.3	-1.0
(12) 切り取ったボックスをフィルタリングしない	-0.1	0.0	0.1	-0.1
(13) 切り取ったボックスをすべてフィルタリングする	-0.1	-0.6	0.1	0.2
(14) OIクラウドインスタンスを削除しないでください。	0.0	0.7	-0.4	3.0
(15) LVISレアラベルは剥がさないでください。	0.1	0.2	-0.1	1.1

トレーニングデータ、および(iii)様々なデータ補強を行った。イタリック体の数字（例えば(15)）は表3の個々のアブレーション実験に言及している。重要なことは、ゼロショット性能の最適レシビ(AP^{LVIS})は、必ずしも配給内性能(AP^{OI})を最大化しないことである。この発見とさらなるアブレーションについては付録A1.9で説明する。

安定化最適化。微調整の目的は、事前学習で学習した表現を破壊することなく、利用可能な検出データから学習することである。そのために、次のような工夫をする。まず、微調整の際に**テキストエンコーダの学習率を 2×10^{-6}** （すなわち、画像エンコーダの学習率より100倍小さくする）(3)。これは、検出ラベルの小さな空間で微調整を行う際に、テキストエンコーダが事前学習で学習したセマンティクスを「忘れる」のを防ぐためと思われる。興味深いことに、テキストエンコーダを完全にフリーズさせると、結果が悪くなる。第二に、セクション3.1で述べたように、**予測ボックス座標(11)を2Dグリッド上の対応するトークンの位置を中心とするように偏らせる**。これは学習を高速化し、最終的な性能を向上させる。おそらく、損失で用いられる二分割マッチングの際に対称性を破ることによってであろう。第三に、大規模なモデルに対しては、画像とテキストエンコーダの両方で確率0.1の**確率的深度正則化** [17,1] を用い、**学習スケジュールを短くする**（セクションA1.3）。

利用可能な検出データの慎重な使用。我々のアブレーションが示すように（表3）、検出学習データの量は、性能の制限要因である。

のデータセットを組み合わせた。そこで、本研究のほとんどのモデル（1-2）ではOI+VG、最大規模のモデルではO365+VGという**複数のデータセットを組み合わせています**（表1参照）。さらに、利用可能なアノテーションをノイズのない状態に保つよう配慮している。**グループ」アノテーションや「網羅的にアノテーションされていない」カテゴリ（14）を、**そのようなアノテーションを示すデータセット（例：OI）から**削除**している。これらのアノテーションは、どのアノテーションが網羅的で、どれがそうでないかを（記憶する以外）学習できないため、モデルに対して相反する監視を与える。これらを除くことで、より大きなモデルの性能を向上させることができる。さらに、オブジェクトの大部分が実際に切り取られている場合、**ランダムな切り抜き補強によって残された部分的なボックスも矛盾した監視を提供する可能性があるため、再移動させることにした**。元の面積の60%以上のインスタンスを保持することで、全てのインスタンス（12）や切り取られていないインスタンス（13）だけを保持するよりも良い結果が得られる。

補強。最後に、画像とクエリの両方を補強することにより、利用可能な検出ラベルを充実させる。画像については、**ランダムな切り出し**（前述のように部分的に切り出されたボックスを削除する）を行う。さらに、「**ラージスケールジッター**」[11]に類似した**画像スケール拡張**を使用する。しかし、単に画像をリサイズしてパディングするのではなく、複数のダウンスケール画像を1つの大きな「モザイク」画像にタイル化します。単一画像、2×2グリッド、3×3グリッドをそれぞれ0.5、0.33、0.17の確率でランダムにサンプリングします（7-9）。クエリ（カテゴリ名）を補強するために、学習時には**ランダムプロンプト**を用い、評価時には**複数のプロンプトに対する予測値のアンサンブル**を行う（4-5）。また、80のCLIPプロンプトを訓練に使用し、7つの「最良」CLIPプロンプト（[33]で定義）に対するアンサンブルを評価に使用する。最後に、少なくとも50のネガティブラベルが存在するまで、各画像の**擬似ネガティブラベル**をランダムにサンプリングする[47]。さらなる実装の詳細は付録A1.5とA1.6に記載されている。

5 結論

我々は、対照的に学習された画像-テキストモデルを検出に移行するための簡単なレシピを提示した。本手法は、LVISベンチマークにおいて、より複雑なアプローチと競合するゼロショット検出を達成し、画像条件付き検出において既存の手法を大きく上回る性能を示した。この結果は、数十億の画像-テキスト例に対する事前学習が、比較的限られたオブジェクトレベルのデータ（数百万例）しか利用できない場合でも、検出に移行できる強力な汎化能力を与えることを示唆している。また、より多くのデータに対して単純でスケラブルなアーキテクチャを事前学習することで、ゼロショット検出が可能になることを示す。我々は、このモデルがオープンワールド検出のさらなる研究のための強力な出発点となることを期待している。

謝辞 DETRの実装を手伝ってくれたSunayana RaneとRianne van den Berg、データ重複排除のコードを提供してくれたLucas Beyer、有益なアドバイスをくれたYi Tayに感謝します。

参考文献

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucey, M., Schmid, C. (アルナブ、A)。このように、「映像の視覚化」を行うことで、「映像の視覚化」を実現している。において。ICCV. pp.6836-6846 (2021年10月)
2. Bansal, A., Sikka, K., Sharma, G., Chellappa, R., Divakaran, A.: ゼロショット物体検出.In:ECCV (2018年9月)
3. このような場合、「ResNets: Improved training and scaling strategies (ResNetsの再考察：学習とスケーリング戦略の改善)」を参照。NeurIPS **34** (2021)
4. Biswas, S.K., Milanfar, P.です。ラプシアンオブジェクトと高速マトリックスコサイン類似度を用いたワンショット検出。IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(3), 546-562 (2016)
5. Bradbury, J., Frostig, R., Hawkins, P., Johnson, M.J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., Zhang, Q.: JAX: composable transformations of Python+NumPy programs (2018), <http://github.com/google/jax>
6. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end Object Detection with transformers.において。ECCV. pp.213-229.シュプリンガー・インターナショナル・パブリッシング、チャム (2020)
7. Chen, D.J., Hsieh, H.Y., Liu, T.L.: Adaptive image transformer for one-shot object detection (ワンショット物体検出のための適応的画像変換器) .In:CVPR. pp.12242-12251 (2021)
8. Dehghani, M., Gritsenko, A.A., Arnab, A., Minderer, M., Tay, Y.です。SCENIC: コンピュータビジョンの研究とその先のためのJAXライブラリ.arXiv preprint arXiv:2110.11403 (2021)
9. Fang, Y., Liao, B., Wang, X., Fang, J., Qi, J., Wu, R., Niu, J., Liu, W.: You only look at one sequence: また、このような「俯瞰的な視点」を持つことが重要である。In:NeurIPS. vol.34 (2021)
10. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: 視覚的意味深層埋め込みモデル.In:NeurIPS.vol.26 (2013)
11. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation.In:CVPR. pp.2918-2928 (2021)
12. 郭 洙 (Gu, X., Lin, T.Y., Kuo, W., Cui, Y.)。視覚と言語知識蒸留によるオープンボキャブラリーオブジェクト検出。
13. Gupta, A., Dollar, P., Girshick, R.: LVIS: A dataset for large vocabulary instance segmentation.In:CVPR (2019年6月)
14. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask R-CNN.にて。ICCV (2017)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (画像認識のための深い残差学習) .In:CVPR (2016年6月)
16. Hsieh, T.I., Lo, Y.C., Chen, H.T., Liu, T.L.: Co attention and co-excitation with One-shot object detection.In:NeurIPS. vol.32.カランアソシエイツ, Inc. (2019年)
17. Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth (確率的深度を持つディープネットワーク) .において。ECCV. pp.646-661.Springer International Publishing, Cham (2016)
18. Huang, Z., Zeng, Z., Liu, B., Fu, D., Fu, J.: Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers. arXiv preprint arXiv:2004.00849 (2020)
19. において。ICML. vol.139, pp.4904-4916.PMLR (2021)

20. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: MDETR。
- エンドツーエンドのマルチモーダル理解のための変調された検出。
In:ICCV. pp.1780-1790 (2021)
21. コレスニコフ、A、ベイヤー、L、ザイ、X、プイグセルバー、J、ユン、J、ゲリー、S、ホールズビー、N。
ビッグトランスファー(BiT)。一般的な視覚表現学習。In:ECCV.491-507.シュ
ブリンガー・インターナショナル・パブリッシング、チャム (2020)
22. コレスニコフ、A、ドソヴィツキー、A、ヴァイセンボーン、D、ハイゴールド、G
、ウスコレイト、J、バイヤー。
L., Minderer, M., Dehghani, M., Houlisby, N., Gelly, S., Unterthiner, T., Zhai, X. で
す。画像は16x16語の価値がある。このように、画像認識のためのトラ
ンスフォーマーには、様々な種類がある。In:ICLR (2021)
23. クリシュナ、R、Zhu、Y、Groth、O、Johnson、J、Hata、K、Kravitz、J、Chen、S、
Kalantidis, Y., Li, L.J., Shamma, D.A., et al. Visual genome: クラウ
ドソースの高密度画像アノテーションを用いた言語とビジョンの接続。イン
ターナショナル・ジャーナル・オブ・ . . コンピュータ・ビジョン
123(1), 32-73 (2017)
24. クズネツォフ、A、ロム、H、アルルドリン、N、ウイリング、J、クラシン、I、
ボン＝テュセ、J、カ。
mali, S., Popov, S., Mallocci, M., Kolesnikov, A., Duerig, T., Ferrari, V.: The Open
Images Dataset V4. (オープン画像データセットV4)。国際コンピュータ
ビジョン学会誌 **128**(7), 1956-1981 (Mar 2020)
25. Lee, J., Lee, Y., Kim, J., Kosiorek, A.R., Choi, S., Teh, Y.W.: Set transformer (セット・ト
ランスファー): A
注意に基づく順列不変のニューラルネットワークのためのフレームワーク
。でICML.Proceedings of Machine Learning Research, vol.97, pp.3744-3753.PMLR
(2019)
26. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan. (以下、L.H.と略
す)。
L., Zhang, L., Hwang, J.N., et al. です。Grounded language-image pre-training. arXiv
preprint arXiv:2112.03857 (2021)
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár,
P., Zitnick, C.L.: Microsoft COCO: Common objects in context.In:ECCV. pp. 740-755.
シュブリンガー・インターナショナル・パブリッシング、チャム (2014)
28. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.:
SSD シングルショット・マルチボックス・ディテクタ。において。ECCV.
pp.21-37.シュブリンガー・インターナショナル・パブリッシング、チャム
(2016)
29. マハジャン、D、ギルシク、R、ラマナサン、V、ヘー、K、パルリ、M、リー、Y
、バランベ。
A., van der Maaten, L.: 弱教師付きプリトレーニングの限界を探索。In:ECCV.
pp.185-201.シュブリンガー・インターナショナル・パブリッシング、チャ
ム (2018)
30. ミカエリス、C、ウスチジャニノフ、ベスゲ、M、エッカー、A.S:ワンショットイ
ンスタンスセグメンテーション
mentation. arXiv preprint arXiv:1811.11507 (2018)である。
31. Osokin, A., Sumin, D., Lomakin, V.: OS2D: One-stage one-shot object detection
by matching anchor features.アンカー特徴のマッチングによるワンステー
ジワンショットオブジェクト検出.In:ECCV. pp.635-652.Springer
International Pub- lishing, Cham (2020).
32. Pham, H., Dai, Z., Ghiasi, G., Liu, H., Yu, A.W., Luong, M.T., Tan, M., Le, Q.V.:
ゼロショット転移学習における結合スケリング arXiv preprint
arXiv:2111.10050 (2021)
33. ラドフォード、A、キム、J.W、ハラシー、C、ラメッシュ、A、ゴー、G、アガルワ
ル、S、サストリー、G。
アスケル、A、ミシキン、P、クラーク、J、クルーガー、G、スツキーバー
、I.: 自然言語監視から転送可能な視覚モデルを学習する。において。
ICML. vol.139, pp.8748- 8763.PMLR (2021年7月18日～24日)
34. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN:リアルタイムの物体
領域提案ネットワークによる検出。In:NeurIPS. vol.28.カランアソシエイ
ツ、Inc. (2015年)
35. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365:
物体検出のための大規模かつ高品質なデータセット.での話。ICCV.
pp.8429- 8438 (2019)

36. Socher, R., Ganjoo, M., Manning, C.D., Ng, A.: Cross modal transfer によるゼロショット学習.NeurIPS **26** (2013)
37. Song, H., Sun, D., Chun, S., Jampani, V., Han, D., Heo, B., Kim, W., Yang, M.H.: ViDT: An efficient and effective fully transformer-based object detector.In:ICLR (2022)
38. Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., Beyer, L.: How to train your ViT? data, augmentation, and regularization in vision transformers. arXiv preprint arXiv:2106.10270 (2021)
39. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jegou, H.: Training data-efficient image transformers and distillation through attention.データ効率的な画像変換と注意による蒸留のトレーニング.において。ICML. vol.139, pp.10347-10357 (2021年7月)
40. Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.。ゼロショット学習-良いもの、悪いもの、醜いものの総合的な評価。IEEE transactions on pattern analysis and machine intelligence **41**(9), 2251-2265 (2018).
41. Yao, Z., Ai, J., Li, B., Zhang, C.です。効率的な detr: 密な事前分布を持つエンドツーエンドのオブジェクトデテクタの改善。 arXiv preprint arXiv:2104.01318 (2021)
42. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions.In:CVPR. pp.14393-14402 (2021年6月)
43. Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling Vision Transformers.arXiv preprint arXiv:2106.04560 (2021)
44. Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L.: LiT: Zero-shot transfer with locked-image text tuning.arXiv preprint arXiv:2111.07991 (2021)
45. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al. RegionCLIP: arXiv preprint arXiv:2112.09106 (2021)
46. また、このような場合、「曖昧さ」を解消するために、「曖昧さ」を解消するために、「曖昧さ」を解消するために、「曖昧さ」を解消するために、「曖昧さ」を解消するために、「曖昧さ」を解消するために、「曖昧さ」を解消する必要があります。In: arXiv preprint arXiv:2201.02605 (2021)
47. Zhou, X., Koltun, V., Kr"ahenbu"hl, P.です。確率的2段検出法。
48. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable transformers for end-to-end object detection (変形可能なDETR: エンドツーエンド物体検出のための変形可能な変換器)。In:ICLR (2021)

付録

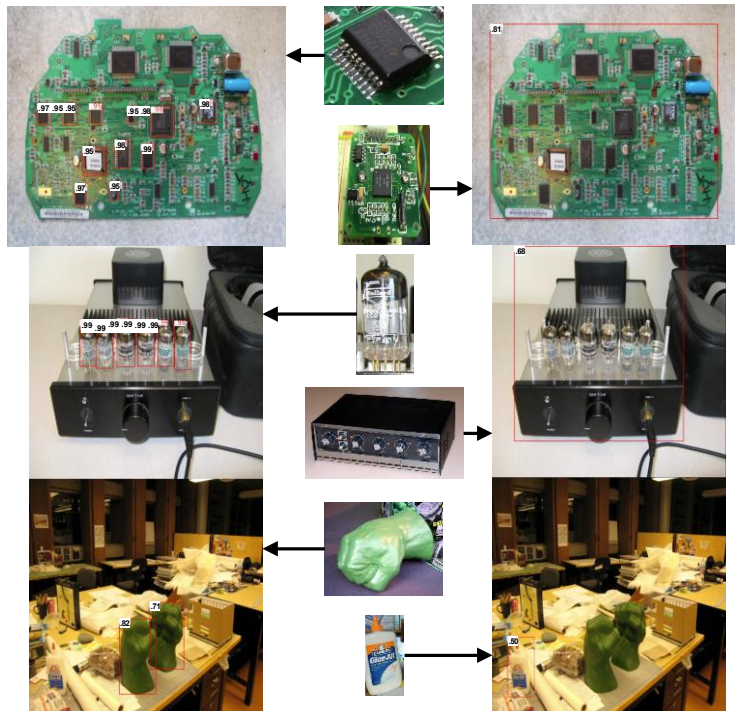
付録には、追加の例、結果、方法論の詳細が記載されています。残りの質問は github.com/google-research/scenic/tree/main/scenic/projects/owl_vit にあるコードを参照してください。

A1.1 定性的な例



図A1. テキストコンディショニングの例。プロンプト「の画像」, ここで $\{ \}$ は、本棚、デスクランプ、コンピュータのキーボード、バインダーのいずれかに置き換えられます。

パソコン、マウス、モニター、椅子、引き出し、グラス、iPod、ピンクの本、黄色の本、カーテン、赤いリンゴ、バナナ、青リンゴ、オレンジ、グレープフルーツ、ジャガイモ、看板、車のホイール、車のドア、車のミラー、ガソリタンク、蛙、ヘッドライト、ナンバープレート、ドアハンドル、テールライト。



図A2. 画像の条件付けの例。中央の列はクエリパッチ、外側の列は検出結果を類似度スコアとともに示している。

A1.2 検出データセット

この研究では、オブジェクト検出のアノテーションを持つ5つのデータセットが、微調整と評価のために使用されました。表 A1 に各データセットの関連統計量を示す。

ms-coco (coco) [27]。Microsoft Common Objects in Contextデータセットは、中規模のオブジェクト検出データセットである。80のオブジェクトカテゴリに対して約900kのバウンディングボックスのAn-notationがあり、1画像あたり約7.3のアノテーションがあります。最も利用されているオブジェクト検出データセットの一つであり、その画像は他のデータセット（VGやLVISを含む）内でよく利用される。本作品では2017年のtrain, validation, testの分割を使用している。

Visual Genome (VG) [23]は、各画像内のオブジェクト、リージョン、オブジェクトの属性、およびそれらの関係に対する高密度なアノテーションを含んでいる。VGはCOCO画像に基づいており、1画像あたり平均35個のオブジェクトに対してフリーテキストのアノテーションを再付与している。すべての実体はWordNetのシンジセットに正規化されている。このデータセットからオブジェクトのアノテーションのみを使用し、属性、関係、地域のアノテーションを使用してモデルを学習しない。

Objects 365 (O365) [35]は、365のオブジェクトカテゴリを持つ大規模なオブジェクト検出データセットである。我々が利用するバージョンは、10M以上のバウンディングボックスがあり、約1画像あたり15.8個のオブジェクトアノテーション。

LVIS [13]。Large Vocabulary Instance Segmentationデータセットは、1000以上のオブジェクトカテゴリを持ち、いくつかのカテゴリは数例しかないロングテール分布をしている。VGと同様に、LVISはCOCOと同じ画像を使用し、より多くのオブジェクトカテゴリで再アノテーションしている。COCOやO365とは異なり、LVISは連合データセットであるため、各画像にはカテゴリのサブセットのみがアノテーションされている。そのため、アノテーションには、存在するオブジェクトと存在しないカテゴリに対して、それぞれ正と負のオブジェクトラベルが含まれる。また、LVISのカテゴリはペアワイズディスジョイントではないため、同じオブジェクトが複数のカテゴリに属することがある。

OpenImages V4 (OI)[24]は、現在最大の公共オブジェクト検出データセットで、約14.6個のバウンディングボックス注釈（画像あたり約8個の注釈）があります。LVISと同様に、このデータセットもフェデレートされたデータセットです。

表A1.本研究で使用した物体検出データセットの統計値。

名称	電車	ヴァル	テスト	カテゴリー
ms-coco 2017 [27]	118k	5k	40.1k	80
ビジュアルゲノム【23	84.5k	21.6k	-	-
オブジェクト365【35	608.5k	30k	-	365
LVIS [13]（エルビス	100k	19.8k	19.8k	1203
OpenImages V4 [24]（オープンイメージ	1.7M	41.6k	125k	601

重複排除 我々の検出モデルは通常、OpenImages V4 (OI) と Visual Genome (VG) データセットの組み合わせで微調整され、MS-COCO 2017 (COCO) と LVIS で評価されています。いくつかの実験では、我々のモデルはさらに Objects 365 (O365) で訓練されています。我々は COCO と LVIS のデータセットで訓練することはないが、我々の訓練データセットの公開版には COCO と LVIS の検証セットと同じ画像の一部が含まれている。我々の model が訓練中に検証画像を見ないようにするため、OI、VG、O365 の訓練分割から、LVIS と COCO の検証およびテスト分割にも現れる画像を、[21] と同じ手順に従ってフィルタリングしている。重複排除の統計量は表 A2 に示すとおりである。

表A2. トレーニングデータセットの重複排除の統計。例」は画像、「インスタンス」はバウンディングボックスを指す。

	オリジナル		複製		残っている名	
前	例	例	例	例	例	例
OpenImages V4	1.7M	14.6M	948	6.4k	1.7M	14.6M
ビジュアルゲノム	86.5k	2M	6.7k	156k	79.8K	1.9M
オブジェクト 365	608.6k	9.2M	147	2.4k	608.5k	9.2M

A1.3 ハイパーパラメーター

表 A3 は、我々の主要な実験に使用したハイパーパラメータの設定の概要を示している。これより先、我々は

- は、コサイン学習率の減衰を利用した。
- $\alpha=0.3$ 、 $\gamma=2.0$ の焦点距離のものを使用。
- は、バウンディングボックス、gIoU、分類ロスに等しい重みを設定する[6]。
- は Adam オプティマイザーを使用し、 $\beta_1=0.9$ 、 $\beta_2=0.999$ としました。
- は、サンプルごとのグローバルノルム勾配クリッピングを用いた（セクション A1.9 参照）。
- LIT と CLIP の両モデルとも、テキストエンコーダの入力長を 16 トークンに限定しました。

CLIPベースのモデル 一般に公開されている CLIP モデルのビジュアルエンコーダは、画像埋め込み特徴に加えて、クラストークンを提供しています。クラストークンの情報が検出の微調整に有用かどうかを評価するために、このトークンを削除するか、他の特徴量マップトークンと乗算して統合するかを検討しました。その結果、クラストークンと特徴量マップトークンを乗算し、その後にレイヤノルムを乗算する方法が、大半のアーキテクチャで最も有効であることがわかったので、この方法を全体に使用することにした。CLIP モデルの微調整に使用した他のハイパーパラメータを表 A3 に示す。

表 A3. 論文で示した全モデルに使用したハイパーパラメータのリスト。アスタリスク（*）は掃引で変化させたパラメータを示す。MAPとGAPは、画像レベルの表現集約のためのマルチヘッド注目プーリングとグローバル平均プーリングの使用を示す。ドロップピング率に2つの数値が与えられている場合、1つ目は画像エンコーダとテキストエンコーダー用の2つがあります。

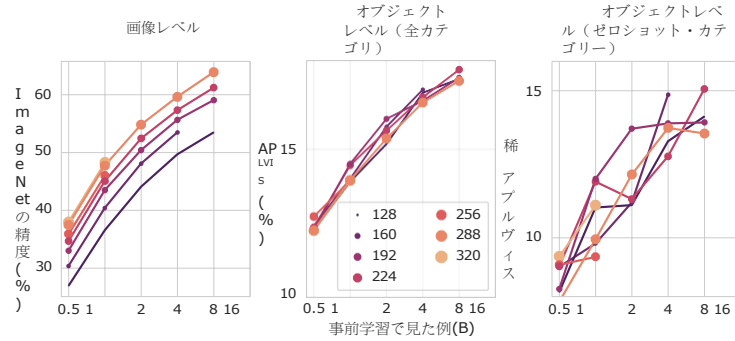
ランダムネガ	モザイクの割合	データ数比率	トレーニングデータセット	画像サイズ	ドロップパー率	重量減衰	学習率	バッチサイズ	トレーニングステップ	プーリングタイプ	画像サイズ	重量減衰	学習率	バッチサイズ	事前学習検出の微調整	モデル
			O365、VG	768	.2/.1	0	5×10^{-5}	256	140k							表1よりCLIPベースのOWL-ViTモデル。
は	.4/.3/.3	.8/.2	O365、VG	768	.2/.1	0	5×10^{-5}	256	140k							B/32
は	.4/.3/.3	.8/.2	O365、VG	768	.2/.1	0	5×10^{-5}	256	140k							B/16
は	.4/.3/.3	.8/.2	O365、VG	840	.2/.1	0	2×10^{-5}	256	70k							L/14
			O365、VG	768	0.0	0	2×10^{-4}	256								表1より、LiTベースのOWL-ViTモデル。
は	.4/.3/.3	.8/.2	O365、VG	768	0.0	0	2×10^{-4}	256								B/32
は	.4/.3/.3	.8/.2	O365、VG	768	0.0	0	2×10^{-4}	256								B/16
は	.4/.3/.3	.8/.2	O365、VG	768	0.0	0	2×10^{-4}	256								R26+B/32
は	.4/.3/.3	.8/.2	O365、VG	768	0.0	0	5×10^{-5}	256								L/16
は	.4/.3/.3	.8/.2	O365、VG	840	.1/.0	0	5×10^{-5}	256								H/14
			OI, O365, VG	960	0.1	0	2×10^{-4}	256								一発検出に使用したモデル（表2）。
は	.5/.33/.17	.4/.4/.2	OI, O365, VG	960	0.1	0	2×10^{-4}	256								R50+H/32
			OI、VG	768	0.0	0	2×10^{-4}	256								アブレーション試験のベースライン・モデル（表3・A5）。
は	.5/.33/.17	.7/.3	OI、VG	768	0.0	0	2×10^{-4}	256								B/32
は	.5/.33/.17	.7/.3	OI、VG	768	0.0	0	2×10^{-4}	256								R26+B/32
			OI, VG	768	0.0	0										スケールングスタディに使用したモデル（図3・図4）。
no	.5/.33/.17	.7/.3	OI, VG	768	0.0	0										*
あり	.5/.33/.17	.7/.3	OI, VG	960	0.0	0	2×10^{-4}	256								R50+H/32

A1.4 事前学習用画像の解像度

画像テキストの事前学習で使用する画像サイズが、ゼロショット分類と検出性能に与える影響を調査しました（図A3）。乱雑さを避けるため、結果はViT-B/32アーキテクチャのみにについて示されていますが、得られた傾向はハイブリッドトランスフォーマーを含む他のアーキテクチャにも当てはまります。事前学習で大きな画像を使用すると、ゼロショット分類には一貫して効果がありますが、検出性能には大きな違いがありません。そのため、事前学習では一般的に使用されている224×224の解像度をデフォルトとしています。ハイブリッドトランスフォーマーモデルの実験では、288×288の解像度を使用しました。

A1.5 ランダム・ネガ

我々のモデルは連合データセットで学習される。このようなデータセットでは、すべての画像にすべてのカテゴリが網羅的に注釈されているわけではありません。その代わりに、各画像には



図A3.画像レベルの事前学習で使用する画像サイズがゼロショット分類と検出性能に与える影響（ViT-B/32アーキテクチャの場合）を示す。

の数、ラベル付きバウンディングボックス（正のカテゴリの集合を構成する）、および画像に存在しないことが分かっているカテゴリのリスト（すなわち、負のカテゴリ）である。他のすべてのカテゴリについては、画像におけるその存在は未知である。負のラベルの数は少ないので、先行研究は、各画像の「擬似的な負の」ラベルをランダムにサンプリングし、それらをアノテーションに追加することが有益であることを発見した[47]。我々は同じアプローチに従い、少なくとも50のネガティブカテゴリが存在するまで、各画像の本当のネガティブにランダムにサンプリングされた擬似ネガティブを追加する。[47]とは対照的に、我々は全データセットにおける頻度按比例してカテゴリをサンプリングする（すなわち、OI、VG、および潜在的にO365の重み付けされた組み合わせ）。我々は、与えられた画像に対して陽性の中にあるカテゴリを標本から除外する。

A1.6 画像の拡大縮小

物体サイズに対する検出モデルの不変性を向上させるために、先行研究では、学習中に画像スケールの強いランダムジッタリングを使用することが有益であることがわかった[11]。我々は、同様のアプローチを用いるが、画像のパディングを最小化する2段階の戦略に従っている。

まず、各トレーニング画像をランダムに切り出します。サンプリング手順は、アスペクト比が0.75から1.33の間、面積が元画像の33%から100%の間で切り抜くように制約される。バウンディングボックスのアノテーションは、少なくとも60%の領域が切り抜き後の画像の領域内にある場合に保持される。クロップ後、画像の下端または右端にグレーの画素を追加して、アスペクト比が正方形になるようにパディングされます。

第二に、複数の画像を様々な大きさのグリッド（「モザイク」）に組み立て、モデルが見る画像スケールの範囲をさらに拡大する。特に断りのない限り、単一画像、2×2モザイク、3×3モザイクを、それぞれ確率0.5、0.33、0.17でランダムにサンプリングします（図A4）。この手順により、学習時の過剰なパディングやモデル入力サイズの可変性を回避しながら、大きく変化する画像スケールを使用することができ



図A4.学習用画像の例。グラントゥールズボックスを赤で示す。左から、1枚の画像、2×2のモザイク、3×3のモザイクを示す。正方形でない画像は下と右にパディングされている（灰色）。

A1.7 ワンショット（画像条件付き）検出の詳細

クエリとして用いる画像埋め込みの抽出。我々は、新しいターゲット画像 I から類似パッチを検出したいクエリー画像パッチ Q を与えられる。

Q 一般に、我々のモデルは多くの重複する境界ボックスを予測し、そのうちのいくつかは Q と高い重複を持つことになる。予測された各境界ボックス b_i は、対応するクラスヘッド特徴 z_i を持つ。我々の DETR スタイルの二分割マッチング損失により、我々のモデルは一般にオブジェクトに対して単一の前景埋め込みを予測することになる。

と、それに隣接する多くの無視すべき背景埋め込みがあります。背景埋め込みはすべて互いに類似しており、単一の前景埋め込みとは異なるので、前景埋め込みを見つけるには、対応するボックスが Q で $\text{IoU} > 0.65$ となるクラス埋め込み群の中から最も異質なクラス埋め込みを探索することになります。

したがって、我々は他のクラス埋め込みに対する類似度は、 $f_j(z_i) = \sum_{j=0}^{N-1} z_i \cdot z_j^T$ となります。した

は、 I の推論を行う際に、最も異質なクラスの埋め込み $\text{argmin}_i f(z_i)$ をクエリ特徴として用いる。

このような場合は、「ある物体の画像」というテキストクエリに対する埋め込みに戻ることになります。

画像条件付き評価プロトコル。我々は[16]の評価プロトコルに従う。評価の際、保留された MS-COCO のカテゴリを少なくとも1つ含む対象画像と、同じ保留されたカテゴリを含むクエリ画像パッチをモデルに提示する。ターゲット画像とクエリ画像は検証セットから抽出される。対象画像における検出結果の AP50 を報告する。一般的なオブジェクト検出とは異なり、ターゲット画像内に少なくとも1つのクエリ画像カテゴリのインスタンスが存在することが仮定されていることに注意。先行研究と同様に、我々は Mask-RCNN[14] を用いて、小さすぎる、あるいはクエリオブジェクトを明確に示していないクエリパッチをフィルタリングしている。また、検出学習時には、保持された分割の中の任意のカテゴリに関連する全てのカテゴリを保持するように注意した。に一致するラベルのアノテーションを削除した。

表A4. COCO および O365 データセットにおけるオープン語彙の検出性能。この結果は、学習に使われなかったデータセットに対する我々のモデルの開放語彙汎化能力を示している。対象データセットで学習したモデルの結果はグレーで表示されています。ここに示した我々のモデルのほとんどは、COCOやO365で直接学習したものではありません（表1のモデルとは異なります）。しかし、我々はCOCOやO365のオブジェクトカテゴリを学習データから削除していないため、これらの数値は「ゼロショット」ではない。我々のモデルについては、3回の微調整を行った際の平均性能を報告する。

方法	バックボーン	画像レベル	オブジェクトレベル	Res.	apcoco	ap50coco	ap0365	ap50o365
ViLD 【12】	ResNet50	クリップ	LVISベース	1024	36.6	55.6	11.8	18.2
Reg.CLIP [45]	R50-C4	CC3M	COCOベース	?	-	50.4	-	-
Reg.CLIP [45]	R50x4-C4	CC3M	COCOベース	?	-	55.7	-	-
GLIP [26] (グリップ)	スウィンT	キャップフォーエム	O365、GoldG、...	?	46.7	-	-	-
GLIP [26] (グリップ)	スウィンL	CC12M、SBU	OI, O365, VG, ...	?	49.8	-	-	-
デティック 【46】	R50-C4	クリップ、COCO-キャップ	COCOベース	1333	-	45.0	-	-
デティック 【46】	スウィンピー	クリップ、I21K	LVISベース	869	-	-	21.5	-
OWL-ViT (当社製品)	ViT-B/32	クリップ	OI、VG	768	28.1	44.7	-	-
OWL-ViT (当社製品)	ViT-B/16	クリップ	OI、VG	768	31.7	49.2	-	-
OWL-ViT (当社)	ViT-L/14	クリップ	O365、VG	840	43.5	64.7	-	-
OWL-ViT (当社)	ViT-B/32	LiT	OI、VG	768	28.0	44.4	9.4	15.2
OWL-ViT (当社)	ViT-B/16	LiT	OI、VG	768	30.3	47.4	10.7	17.0
OWL-ViT (当社)	R26+B/32	LiT	OI、VG	768	30.7	47.2	11.1	17.4
OWL-ViT (当社)	ViT-L/16	LiT	OI、VG	672	34.7	53.9	13.7	21.6
OWL-ViT (当社)	ViT-H/14	LiT	OI、VG	840	36.0	55.3	15.5	24.0
OWL-ViT (当社)	ViT-H/14	LiT	O365、VG	840	42.2	64.5	-	-

また、「少女」というラベルは「人」の子孫にあたる。また、それ以外にも、ホールドアウトされたカテゴリに類似するラベルを手動で削除した。なお、ホールドアウトされたラベルは、コードの公開と同時にすべて公表する予定である。

A1.8 COCOとO365での検出結果

表A4では、COCOおよびO365データセットに対する評価結果を追加で示す。これらの結果は、我々のアプローチのオープンボキャブラリーゼロ化の能力を示している。我々はCOCOやO365で直接モデルを訓練していないが（特に断りのない限り）、我々の訓練データセットにはCOCOやO365と重なるオブジェクトカテゴリが含まれているため、これらの結果は我々の定義によれば「ゼロショット」ではない。このように、様々な評価手法が存在するため、既存手法との比較は困難である。表 A4 では、公正な比較のために必要な差異を記すことに努めた。

A1.9 Extended Ablation Study (拡張アブレーション試験)

表A5は、本文の表3に提供されたアブレーション結果を拡張する。それは、表3に概説されたのと同じトレーニングおよび評価プロトコルを使用するが、研究で考慮された設定およびアーキテクチャ（ViT-B/32およびViT-R26+B/32）の範囲においてさらに踏み込んでいる。以下、追加のアブレーションについて説明する。

データセットの比率我々の実験の大部分では、OIとVGのデータセットをトレーニングに使用しています。本文で紹介したアブレーション研究（表3）では、より多くのトレーニングデータ（つまり、VGとOIの両方でトレーニング）を持つことで、ゼロショット性能が向上することを示しました。ここで、これらのデータセットを混合する最適な比率をさらに検討し、7:3=OI:VGの比率が最も効果的であることを見出した。これは、これらのデータセットの相対的なサイズに比べ、VGを大幅にオーバーしていることに注意する。VGはOIよりも大きなラベル空間を持つため、VGの例はOIの例よりも価値ある意味での監視を提供するため、VGを重視することは有益である可能性がある。

また、VGの "object "と "region "アノテーションの相対的な価値も検証した。VGでは、標準的な単一オブジェクトのアノテーションとは対照的に、「領域」アノテーションは画像領域全体の自由なテキストによる説明を提供する。興味深いことに、regionアノテーションを用いた学習はモデルの汎化能力を低下させることが分かったので、我々は学習にregionアノテーションを用いないことにした。

損失正規化およびグラディエントクリッピング。DETR [6]はその公式実装において、ローカル（つまりデバイス毎）の損失正規化を使用しており、したがって（ローカル）バッチサイズに敏感です。我々は、このことが実際には重要な詳細であり、性能に大きく影響することを発見した。我々は、ボックス、giou、分類損失を画像内のインスタンス数で正規化するか、バッチ全体のインスタンス数で正規化する方がパフォーマンスが良いかを調査した。我々の実験によると、サンプル毎の正規化が最も良い性能を示すが、それは**サンプル毎の勾配クリッピングと組み合わせた場合**、つまり、バッチ全体の勾配を蓄積する前に、各エクスプレスの勾配ノルムを個別に1.0にクリッピングした場合のみであることがわかった。我々は、サンプル単位のクリッピングが学習の安定性を向上させ、全体的な損失の低減につながり、より大きなバッチサイズでモデルを学習できることを発見しました。

インスタンスのマージ。OIのような連合データセットには非分離ラベル空間があり、（ほぼ）同義のラベル（例：「Jug」と「Mug」）、または非分離概念（例：「Toy」と「Elephant」ラベルは両方ともおもちゃの象に適用する）により、複数のラベルが同じオブジェクトに適用できることを意味している。一度に1つのラベルを考慮するアノテーション手順のため、1つのオブジェクトに複数の類似した（しかし同一ではない）バウンド・ボックスがアノテーションされることがある。我々は、このようなインスタンスを1つのマルチラベルインスタンスに統合することが有用であることを見出した。マルチラベルのアノテーションは連合アノテーションの非分離の性質と一致し、我々はこれがモデルにより効率的なスーパービジョンを提供すると推測している。このインスタンスマージがなければ、モデルはオブジェクトに適用される各ラベルに対して個別のボックスを予測する必要があり、これは明らかに無数の可能なオブジェクトラベルに一般化することができない。

重複するインスタンスをマージするために、各画像に対して以下のステップで無作為化された反復手順を使用する。

1. バウンディングボックスのオーバーラップが最も大きい2つのインスタンスを選びます。
2. IoU(intersection over union)が所定の閾値以上である場合。

- 2.1. 両者のレーベルを統合する。
- 2.2. 元のバウンディングボックスの1つを、マージされたインスタンスのバウンディングボックスとしてランダムに選択します。

次に、選択されたインスタンスは削除され、十分に高いIoUを持つインスタンスがなくなるまでこの手順が繰り返される。複数のIoU閾値を調査した結果、非常に類似したバウンディングボックスを持つインスタンスをマージしないことは、それらをマージするよりも明らかに悪いことであり、0.7-0.9という中程度の高閾値が実際に最も効果的であることが分かった。

学習率。表3は、画像とテキストのエンコードに同じ学習率を用いることは明らかに最適ではなく、テキストエンコードをより低い学習率で学習させる必要があることを示している。これは、対照的な事前学習段階でモデルが獲得した幅広い知識の壊滅的な忘却を防ぐのに役立つと思われる。ここでは、テキストエンコードの学習率の範囲を調べ、テキストエンコードの学習率を画像エンコードの学習率よりずっと低く（例えば100倍）しないと、良好なゼロショット転送ができないことを実証します（ $AP^{L_{VIS}}$ ）。しかし、テキストエンコードを完全にフリーズさせる（学習率0）のもうまいきません。配信内性能を測定する AP^{OI} は逆の挙動をします。イメージエンコードとテキストエンコードの学習率を同じにすると、 $AP^{L_{VIS}}$ は大きく低下するが、 AP^{OI} は上昇する。このことは、ゼロショット転送の最適なレシビ（ $AP^{L_{VIS}}$ ）が、そのようなものでないことを示している。

は必ずしも配信内パフォーマンスを最大化するものではありません（ AP ）。^{OI}

クロッピングバウンディングボックスフィルタリング我々はモデルを学習する際に、ランダムな画像のクロッピングを使用する。結果として得られた画像とバウンディングボックスを手動で検査したところ、元のインスタンスラベルと一致しなくなった縮退したバウンディングボックスを持つインスタンスが頻繁に発生することに気づいた（例えば、画像から人物をほとんど切り取った結果、「人物」のラベルを持つ手の周りのバウンディングボックスが発生するなど）。このようなインスタンスを記憶することによるモデルのオーバーフィットの可能性を減らすために、ボックス領域の大部分がランダムな切り抜き領域の外にある場合、オブジェクトのアノテーションを削除しています。最適な領域閾値は40%から60%の間にあり、すべてのボックスを保持することも、切り取られていないボックスのみを保持することも、同様にうまく機能する（表3、A1.9）。

モザイク付録 A1.6 で述べたように、複数の小画像を一つの大きな「モザイク」に並べることで、画像の拡大縮小を行う。我々は4×4までのモザイクサイズを調査し、単一画像に加えて2×2モザイクのみを使用することは、より大きなモザイクを含むよりも明らかに悪いことを発見した、考慮された解像度とパッチサイズでは、大きなモザイク（すなわち、小さなモザイクタイル）の使用の利点は、3×3または4×4モザイクを含むと飽和する。我々はモザイク比の広範なスイープを行っておらず、グリッドサイズが1×1（すなわち単一画像）から $M \times M$ までのモザイクについては、より小さいモザイクがより多くサンプリングされるように、 $k \times k$ のタイルを確率 $\frac{2 \cdot (M-k+1)}{M^2}$ でサンプリングするという発見的な方法を使用している。

は、モザイクサイズに比例して、大きなモザイクよりも頻度が高い。

プロンプトの作成。テキストクエリを生成するために、先行研究と同様に、対象物のカテゴリ名を "a photo of a {}" (ここで画像レベルの事前学習と検出の微調整の間の分布のずれを少なくするために、{}はカテゴリ名で置き換えられる)。プロンプトテンプレートはCLIP[33]で提案されたものを使用する。このため、CLIPのプロンプトテンプレートは、画像内の各カテゴリが同じプロンプトを持つが、カテゴリ間や画像間でプロンプトテンプレートが異なるように、学習時に80個のCLIPプロンプトテンプレートからランダムにサンプリングする。このようなプロンプトのテンプレートは、カテゴリごと、画像ごとに異なる。テストでは、「ベスト7」のCLIPプロンプトごとにモデルを評価し、その結果を平均して予測確率をアンサンブルする。表 A5 の結果は、プロンプトを使用しない場合、特に分布内 AP^{OI} メトリックにおいて、あまりうまくいかないことを示している。また、学習時にランダムプロンプトを使用した場合、テスト時プロンプトアンサンブルはより効果的に機能する。場合によっては、プロンプトはモデル・アーキテクチャによって異なる効果を持つことがある。例えば、VGデータセットにランダムプロンプトを適用すると、B/32モデルの性能は向上するが、R26+B/32モデルの性能は低下する傾向がある。これはプロンプトテンプレート数が少ないためであり、プロンプトテンプレートの数を増やせば、より一貫した効果が得られると考えられる。このため、OIデータセットでは訓練時ランダムプロンプトのみを使用し、一貫した効果が得られるようにした。

位置のバイアス本文で述べたように、ボックス予測に対応する画像パッチの位置にバイアスをかけると、学習速度と最終的な性能が向上します。この効果は純粋なTransformerアーキテクチャ（表A1.9のViT-B/32）において特に大きく、バイアスを除去すると AP^{LVIS} と AP^{LVIS} においてほぼ3ポイント性能が低下するが、ハイブリッドR26+B/32ではわずか1ポイント強の低下で済む。したがって、ハイブリッドの畳み込みコンポーネントの空間誘導バイアスは、位置バイアスと同様の機能を果たすと推測される。

表A5・追加アブレーション。VG(obj)とVG(reg)はそれぞれVi-sual Genomeのオブジェクトとリージョンのアノテーションを意味する。

レーシジョン レビス アブココ アボイ	ViT-B/32				ViT-R26+B/32 アブ			
	アブレビス アブココ		アボイ		アブレビス アブココ		アボイ アブ	
	レア				希少			
ベースライン	15.7	14.1	24.1	48.5	21.0	18.9	30.9	54.1
データセット比率。ベースラインはOI:VG(obj)=7:3								
OI:VG(オブジェ)=2:8	-1.9	-2.7	-2.4	-4.8	-4.2	-4.1	-4.7	-4.8
OI:VG(オブジェ)=3:7	-1.0	-1.9	-1.2	-3.1	-3.0	-3.0	-3.3	-2.9
OI:VG(オブジェ)=4:6	-0.6	-1.8	-0.4	-1.7	-2.2	-3.6	-2.2	-1.5
OI:VG(オブジェ)=5:5	0.0	-0.5	0.1	-0.6	-1.0	-1.1	-1.0	-1.1
OI:VG(オブジェ)=6:4	0.1	-0.6	0.1	-0.3	-0.3	-1.4	-0.4	-0.2
OI:VG(オブジェ)=8:2	-0.7	-0.9	-0.6	-0.1	-0.4	-0.3	0.2	0.4
OI:VG(obj) = 9:1	-1.8	-1.1	-1.6	0.1	-1.8	-1.8	-1.1	0.3
OI:VG(obj, reg) = 7:3	-0.6	0.0	-0.9	-3.3	-1.2	-0.5	-0.8	-3.6
OI:VG(reg) = 7:3	-2.1	-1.4	-2.3	-2.5	-2.9	-2.3	-2.2	-2.2
OIのみ	-4.9	-3.2	-3.5	-0.5	-6.9	-5.7	-4.2	0.3
VG(obj)のみ	-8.0	-8.4	-14.2	-28.5	-14.5	-14.0	-23.6	-38.3
グラデーションのクリッピング。ベースラインは、サンプルごとのクリッピングとサンプルごとの正規化を使用します。								
グローバルクリップ、グローバルノーム	-1.0	-2.0	-1.4	-4.9	-2.3	-2.9	-2.8	-5.4
グローバルクリップ、パーエクスノーム	-4.0	-2.6	-5.3	-4.7	-5.0	-5.0	-5.7	-5.7
インスタンスのマージ。ベースラインはIoU≥0.9で重複するインスタンスをマージする								
マージなし	-0.8	-1.2	-0.3	-1.2	-0.8	-1.3	-0.6	-0.7
IoU ≥ 0.7	0.2	0.3	-0.2	0.1	0.2	0.2	0.0	0.6
IoU ≥ 0.8	0.0	0.4	0.0	0.4	0.0	-1.3	0.1	0.4
IoU ≥ 0.95	-0.1	-0.1	0.0	-0.7	-0.5	-1.3	-0.2	-0.5
テキストエンコーダの学習率。ベースラインは画像LR 2 × 10 ⁻⁴ とテキストLR 2 × 10 ⁻⁶ を使用。								
LR 2 × 10 ⁻³	-5.1	-10.3	-0.8	-0.6	-7.1	-14.1	-1.4	-0.5
LR 2 × 10 ⁻⁴	-2.3	-6.7	-0.7	0.2	-3.0	-8.5	-0.5	0.4
LR 2 × 10 ⁻⁵	-1.1	-3.8	-0.5	0.6	-1.2	-3.2	-0.4	0.9
テキストエンコードの微調整は行わないでください。	-1.8	-1.2	-1.9	-0.7	-1.5	-2.3	-0.6	1.2
クロップドボックスフィルタリング。ベースラインは、元の面積の60%以上のボックスを保持します。								
ボックス領域のフィルタリングなし	-0.1	-0.3	-0.2	-0.2	-0.1	0.0	0.1	-0.1
≥ 20%エリア	-0.3	-1.7	0.0	-0.3	-0.2	-0.8	-0.2	-0.1
≥ 40%以上の面積	0.1	0.0	0.0	0.2	0.1	0.9	0.1	-0.2
フルボックスのみ	-0.2	-0.9	-0.3	-0.2	-0.1	-0.6	0.1	0.2
モザイク。ベースラインは、1対3サイズのモザイクを0.5 : 0.33 : 0.17の比率で使用します。								
1-2 @ 2:1	-0.4	-1.1	-0.1	0.4	-0.5	0.3	-0.5	0.0
1-4 @ 4:3:2:1	0.1	0.3	0.0	-0.3	0.0	-0.8	0.1	-0.3
モザイクなし	-1.4	-1.6	-1.5	-0.4	-2.3	-1.5	-1.7	-0.7
モザイクなし、電車2倍速	-1.0	-1.8	-0.3	1.2	-2.9	-2.8	-1.8	-0.7
モザイクなし、3倍速の列車スケジュール。	-1.2	-3.4	0.3	1.1	-3.4	-3.6	-1.8	-0.8
プロンプトを表示します。ベースラインはOIのための訓練プロンプトとテストアンサンブル(ens.)プロンプトを使用します。								
トレーニング：なし、テスト：なし	0.0	-0.1	0.8	-10.2	-1.2	-1.3	-0.6	-6.3
Train: なし; test: ens.	-2.6	-2.2	-7.3	-11.1	-4.5	-5.0	-10.0	-6.6
トレーニングOI+VG、test : ens。	0.8	1.3	0.9	-0.1	-0.7	-0.7	-0.4	-0.2
電車VG、test : ens。	-0.8	-1.1	-2.9	-7.8	-3.1	-4.0	-7.8	-5.6
その他ベースラインはロケーションバイアスを使用し、50のランダムなネガをサンプリングし、LVISのレアラベルを削除しています。								
位置の偏りがなくランダムネガキャンはしない	-2.8	-2.9	-3.7	-2.6	-1.2	-1.1	-1.3	-1.0
	-1.2	-3.7	-0.8	-0.4	-1.0	-2.8	-0.4	1.0

LVISの希少性を保つ		0.1	シンプルなおープンボキャブラリー オブジェクトの検出	0.1	0.2	-0.1	29 1.1
-------------	--	-----	-------------------------------	-----	-----	------	-----------