物体検出モデルの判断根拠可視化における解釈性向上に関する研究

TP24010　仲井悠真　　指導教授：山下　隆義

1. はじめに

自動運転や医療画像解析において，深層学習による物体検出モデルは，高い性能だけでなく，高い信頼性が要求される．しかし，深層学習モデルは，その判断根拠がブラックボックスである．このような背景から，物体検出モデルの判断根拠を人間に理解可能な形で示す説明可能な AI（XAI）が注目されている．物体検出に特化した手法として ODAM[1] が提案されている．ODAM は勾配情報に基づいて可視化を行うため，入力画像に対する勾配消失や局所的な勾配ノイズの影響を受けやすいという課題がある．

そこで本研究では，ODAM が抱える勾配依存による課題を軽減するため，入力画像に関する情報を持たないベースライン画像から入力画像に至るまでの過程を考慮できる勾配計算法である Integrated Gradients[2] を導入する．さらに Integrated Gradients における補間画像の生成方法に起因する積分近似誤差および勾配ノイズの問題に着目する．これらを低減するため，勾配変動と空間的変化に基づいてサンプリング位置を適応的に制御する機構を導入した Adaptive IG-ODAM を提案する．

2. Integrated Gradients

XAI において，可視化結果が入力と予測の関係を適切に反映していることが求められる．既存の勾配ベース手法は，勾配消失により重要な特徴を捉えられないという感度の欠如が課題である．この課題に対し，ベースラインから入力までの直線経路に沿って勾配を積分し，各特徴量の寄与度を算出する Integrated Gradients が提案されている．画像タスクでは，ベースラインとして全画素がゼロの画像が用いられることが多く，経路全体の勾配情報を用いることで，単一の入力画像の勾配に依存しない忠実な寄与度推定が可能となる．

一方で，実装上は経路積分を有限個の補間点により近似するため，図 1 に示すように，補間経路を一様にサンプリングした場合には，勾配が急激に変化する区間に十分な補間点が割り当てられず，積分近似誤差や勾配ノイズが生じやすいという課題がある．特に，深層モデルにおいて勾配が非線形に変化する場合，この近似誤差は可視化結果の忠実性に影響を及ぼす可能性がある．

また，Integrated Gradients は主に単一の予測出力を対象とした設定を想定しており，物体検出のように複数のインスタンスや出力を同時に扱うマルチインスタンス環境への直接的な適用には，依然として課題が残されている．

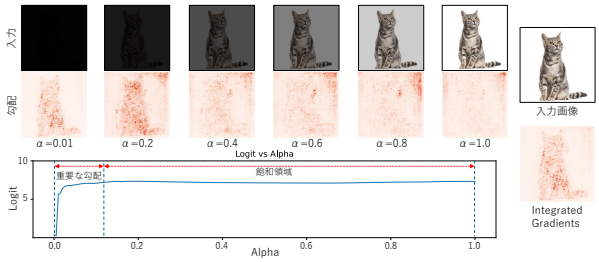

図 1: Integrated Gradients による寄与推定と補間経路上での勾配飽和の例

3. 提案手法: Adaptive IG-ODAM

本研究では，Integrated Gradients を物体検出の判断根拠可視化手法である ODAM に導入した IG-ODAM を提案する．さらに，一様サンプリングに起因する積分近似誤差や勾配ノイズを低減するため，補間経路上の重要区間にサンプリング点を適応的に配置する Adaptive IG-ODAM を提案する．

3.1. ODAM への Integrated Gradients の導入

IG-ODAM は，入力画像とベースライン画像を結ぶ補間経路全体の勾配情報を用いることで，単一の画像における勾配に依存する従来手法に見られる局所的な偏りの影響を低減する．さらに，物体検出特有のマルチインスタンス環境へ適用するために，IoU に基づく位置的類似度とクラススコアの類似度を統合したインスタンスマッチングを導入する．これにより，補間経路上において同一インスタンスを一貫して追跡しながら寄与度推定する．

物体 p に対する予測クラススコア $s^{(p)}(I)$ を寄与度推定の対象とし，ベースライン画像 I' から入力画像 I への補間経路 $I_\alpha = I' + \alpha(I - I')$（$\alpha \in [0,1]$）に沿って経路積分を行う．物体検出では，補間画像ごとに検出結果の数や順序が変化するため，単純な対応付けでは同一インスタンスを追跡できないという問題がある．そこで IG-ODAM では，入力画像 I における物体 p の BBox 座標と予測クラススコアから構成される検出結果 D_t を基準とし，m 番目の補間画像 $X_m = I_{\alpha_m}$ から得られる検出結果集合 $\phi(X_m)$ 内の各検出 $D_{j,m}$ との間で，位置類似度 s_{loc} およびクラススコア類似度 s_{cls} を用いた類似度を定義する．

$$\text{Sim}(D_t, D_{j,m}) = s_{\text{loc}}(D_t, D_{j,m}) \cdot s_{\text{cls}}(D_t, D_{j,m}) \quad (1)$$

各補間画像においては，$\text{Sim}(D_t, D_{j,m})$ が最大となる検出結果を対応インスタンス $\hat{d}_m$ として選択する．この対応付けにより，補間経路全体にわたって同一インスタンスを一貫して追跡しながら，寄与度推定を行うことが可能となる．

特徴マップ A_k に対するチャネル重み $w_k^{(p)}$ は，補間経路上の勾配を積分することで式 (2) のように定義される．

$$w_k^{(p)} = \int_0^1 \frac{\partial s^{(p)}(I_\alpha)}{\partial A_k}\, d\alpha \quad (2)$$

実装上は，補間経路を一様に分割し，有限個の補間点に基づく数値積分によって近似することで，インスタンス固有のヒートマップを生成する．

3.2. Spatial-Guided Adaptive Sampling

Integrated Gradients における一様サンプリングに起因する課題に対して，補間経路上のサンプリング点を動的に再配置する Spatial-Guided Adaptive Sampling を導入した Adaptive IG-ODAM を提案する．本手法は，物体検出モデルの出力が急激に変化する補間区間に重点的にサンプルを配置することで，積分近似誤差および勾配ノイズの低減を目的とする．図 3 に提案手法のモデル図を示す．

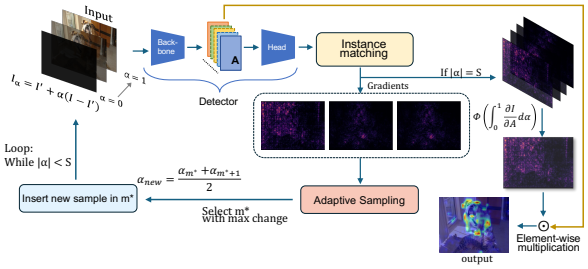

図 3: Adaptive IG-ODAM のモデル構造

Adaptive IG-ODAM では，補間経路上の連続するサンプリング点 α_m と α_{m+1} の間における重要度を評価し，重要度の高い区間を逐次的に細分化する．重要度評価には，勾配の変動量と予測 BBox の空間変動の両方を用いる．

まず，勾配変動 g_m を式 (3) により定義する．

$$g_m = \|G(\alpha_{m+1}) - G(\alpha_m)\|_1 \quad (3)$$

ここで，$G(\alpha_m)$ は補間画像 α_m における対象物体の検出スコアに対する特徴マップの勾配を表す．次に，連続する補間画像間における予測 BBox の空間変動 s_m を，IoU に基づいて式 (4) のように定義する．ここで，$B(\alpha_m)$ は補間画像 α_m に対する予測 BBox を表す．

$$s_m = 1 - \text{IoU}\big(B(\alpha_m), B(\alpha_{m+1})\big) \quad (4)$$

これらを重み係数 λ を用いて統合し，各補間区間の優先度スコア R_m を式 (5) により算出する．

$$R_m = \lambda g_m + (1 - \lambda) s_m \quad (5)$$

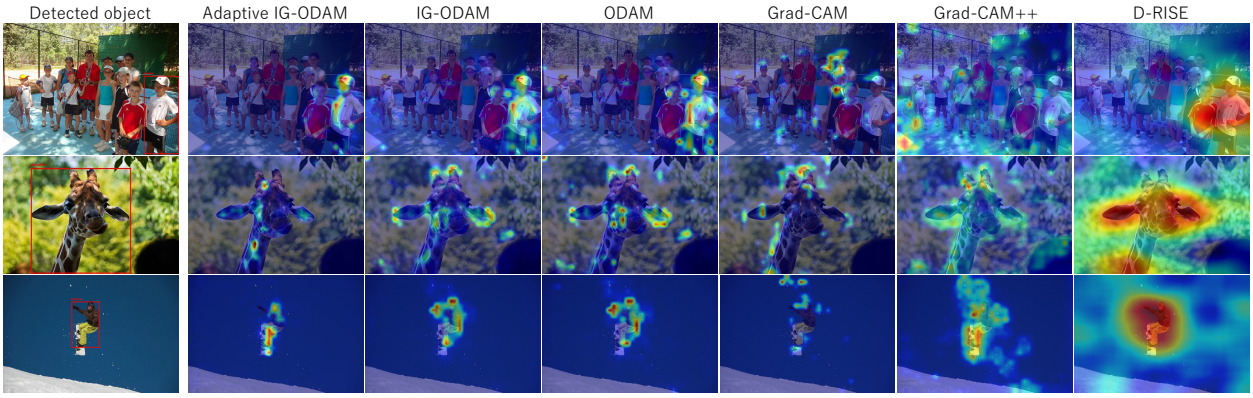

図 2: DETR による物体検出結果に対する判断根拠の可視化結果

優先度スコアの高い区間を逐次的に細分化することで，重要な補間区間にサンプルを集中的に割り当てる．Spatial-Guided Adaptive Sampling により得られた非一様な M 個の補間点に基づき，物体 p に対するチャネル k の重み $w_k^{(p)}$ を台形則に基づき式 (6) のように近似する．ここで，$G_k(\alpha_m)$ は補間画像 α_m におけるチャネル k に対応する勾配を表す．

$$w_k^{(p)} \approx \sum_{m=1}^{M-1} \frac{1}{2} \left[G_k(\alpha_m) + G_k(\alpha_{m+1}) \right] (\alpha_{m+1} - \alpha_m) \quad (6)$$

最後に，得られたチャネル重みを用いて，インスタンス固有のヒートマップ $H^{(p)}$ を生成する．ここで，A_k はチャネル k の特徴マップを表す．

$$H^{(p)} = \mathrm{ReLU}\left(\sum_k w_k^{(p)} \circ A_k \right) \quad (7)$$

4. 評価実験

提案手法の忠実性と空間識別能力を評価するため，比較実験を行う．判断根拠の忠実性は Deletion / Insertion テストの AUC により評価し，空間識別能力は Visual Explanation Accuracy（VEA）と Energy-based Pointing Game（EBPG）を用いて測定する．

物体検出モデルには Backbone に ResNet-50 を用いた DETR を使用し，MS COCO データセット上で Grad-CAM，Grad-CAM++，D-RISE，ODAM と比較する．なお，本実験では，勾配変動と空間変動の寄与を等しく考慮するため，重み係数 λ を 0.5 に設定する．

4.1. Deletion, Insertion

Deletion / Insertion テストは，可視化手法がモデル予測に重要な領域をどの程度正確に特定できるか，忠実度を評価する指標である．Deletion では，ヒートマップに基づき画素を重要度順にランダム値で置換し，予測スコアの低下を測定する．Insertion では，ベースライン画像に重要画素を順次追加し，予測スコアの上昇を測定する．本実験では，両テストを 100 ステップで実施し，信頼度推移から AUC を算出する．

実験結果を表 1 に示す．IG-ODAM は，従来の物体検出向け可視化手法である ODAM と比較して，Deletion スコアを 55.25 から 51.48 に低減し，Insertion スコアを 15.37 から 18.14 に向上させることで，忠実度の向上を示した．さらに，Adaptive IG-ODAM は，Deletion スコア 46.48，Insertion スコア 25.88 と最良の性能を示した．これは，Spatial-Guided Adaptive Sampling により経路積分に使用する補間画像が最適化され，積分近似誤差が低減されたためと考えられる．

4.2. VEA, EBPG

VEA は物体形状との一貫性を，EBPG は物体領域への局在精度をそれぞれ評価する指標である．実験結果を表 2 に示す．Adaptive IG-ODAM は，IG-ODAM と比較して，VEA を +0.0528 ポイント，EBPG を +0.1133 ポイント向上させ，空間的一貫性および局在精度の双方で性能向上を示した．これは，Spatial-Guided Adaptive Sampling により，補間経路上でモデル出力が大きく変化する区間に重点的なサンプリングが行われたためである．

表 1: 各手法の Deletion/Insertion 評価結果

Method	Deletion↓	Insertion↑
Grad-CAM	72.82	11.23
Grad-CAM++	72.60	11.04
D-RISE	57.57	13.23
ODAM	55.25	15.37
IG-ODAM	51.48	18.14
Adaptive IG-ODAM	**46.48**	**25.88**

表 2: VEA と EBPG の評価結果

Method	VEA ↑	EBPG ↑
Adaptive IG-ODAM	**0.1492**	**0.3934**
IG-ODAM	0.0964	0.2801

4.3. 定性的評価

図 2 に，各手法による可視化結果を示す．IG-ODAM は，従来手法と比較してノイズが低減され，物体境界をより正確に捉えている．一方，Grad-CAM および Grad-CAM++ では背景や他物体への注目が生じやすく，ODAM ではマルチインスタンス環境において注目領域の分散が見られる．さらに，Adaptive IG-ODAM はインスタンス固有の注目領域を明確に分離することで，従来手法で見られた注目の分散を最も効果的に抑制していることがわかる．

5. おわりに

本研究では，物体検出における説明可能性の向上を目的として，IG-ODAM および Adaptive IG-ODAM を提案した．IG-ODAM は，Integrated Gradients を ODAM に統合し，補間経路全体の勾配情報とインスタンスマッチングにより，マルチインスタンス環境におけるインスタンス単位の判断根拠可視化を実現した．さらに Adaptive IG-ODAM では，勾配変動と予測 BBox の空間変動に基づく Spatial-Guided Adaptive Sampling を導入することで，積分近似誤差およびノイズの削減を達成した．評価実験の結果，忠実度および空間識別能力の両面で既存手法を上回る性能を確認した．

今後は，得られたヒートマップを知識蒸留における教師信号として活用する手法を検討する．

参考文献

[1] Chenyang ZHAO, Antoni B. Chan, “ODAM: Gradient-based Instance-Specific Visual Explanations for Object Detection”, ICLR. 2023.

[2] Sundararajan, *et al.*, “Axiomatic Attribution for Deep Networks.” arXiv. 2017.

研究業績

[1] Nakai, *et al.*, “IG-ODAM: Instance-Aware Visual Explanations for Object Detection with Integrated Gradients,” MVA. 2025.

（他 3 件）