AUPIMO: ビジュアル異常検出ベンチマークの再定義

この記事は、Medium に公開されている「AUPIMO: Redefining Visual Anomaly Detection Benchmarks」の日本語参考訳です。原文は更新される可能性があります。原文と翻訳文の内容が異なる場合は原文を優先してください。

TL;DR

この記事では、**GSoC** 2023 中に提案された新しい**パフォーマンス・メトリック**である **AUPIMO** [1] の概要と、AUPIMO が**異常検出**における最先端の手法にどのような新たな視点をもたらすかについて説明します。

「AUPIMO」、「GSoC」、「異常検出」、「パフォーマンス・メトリック」の意味は、次のセクションで説明します 😉。

概要を説明した後、次のように、詳細な説明を行います。

- **まず、AUROC を定義します。**これは、AUPIMO の親である従来のメトリックです。
- **次に、AUPIMO を定義して**、その理論的根拠を説明します。
- **最後に、その成果を明らかにして**、ビジュアル異常検出の現状にどのような新たな視点をもたらすかについて説明します。

TL;ADR (A は「Almost (ほぼ)」の略)

GSoC とは…

…「Google Summer of Code」の略です。GSoC の定義は、Google で確認することにしましょう。

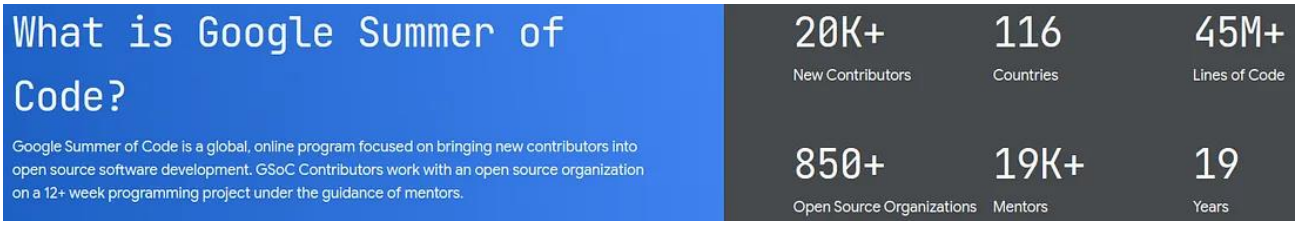

図 1. Google Summer of Code (GSoC)。
2023 年 12 月 14 日時点の summerofcode.withgoogle.com (英語)。

コードを記述するのが好きな学生の皆さんは、詳細を確認してみてください。

GSoC 2023 中に、インテルの @samet-akcay (英語) および @djdameln (英語) と協力して、OpenVINO™ (英語) 向けの異常検出モデルのライブラリー、Anomalib (英語) に貢献できたことを光栄に思っています。

異常検出は…

…教師なしマシンラーニング・タスクです。

何が特別なのでしょうか?

モデルは単一のクラス (つまり、アノテーションなし) を使用してトレーニングされ、課題は通常と異なるイベント (異常) を検出することです。

例えば、MVTec AD [2] のデータセット Transistor (図 2) には、同じ位置の機能するトランジスターのイメージが含まれています。これが要件となる工業生産ラインを考えてみてください。

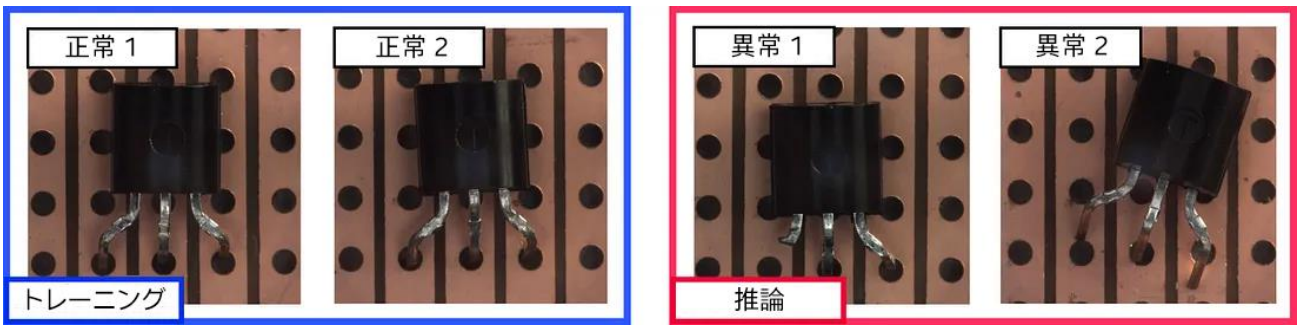

図 2. MVTec AD (英語) [2] のデータセット Transistor のイメージ。

推論時にモデルは、配置ミスやリード線の欠落などの不具合を、これまでにない状態で検出します。

ちなみに、イメージの話の続きとして、「**ビジュアル**異常検出」についても後で説明します。

Anomalib (英語) には、異常の存在だけでなく位置も検出できるモデルのコレクションがあり (増え続けてい) ます (図 3)。

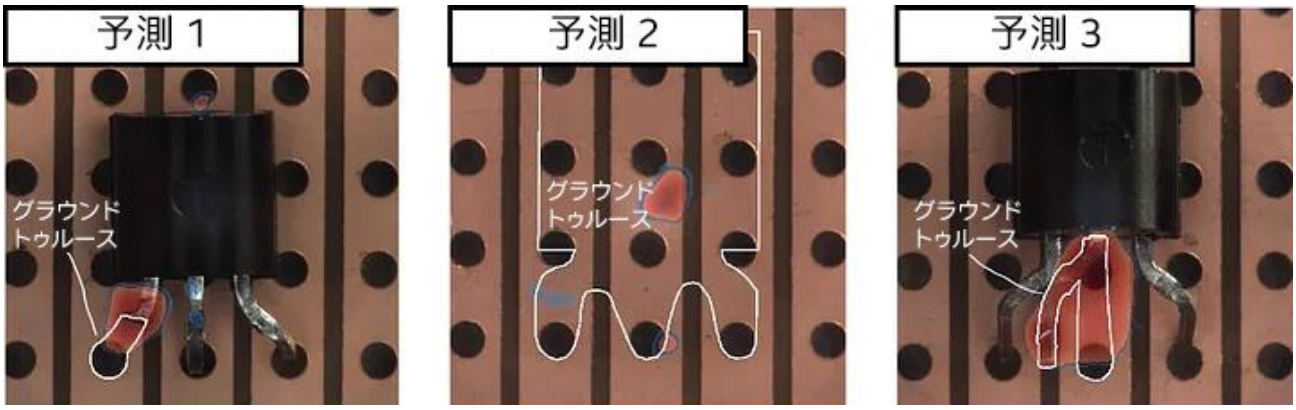

図 3. Anomalib (英語) のモデル PatchCore [3] からの予測。

どのモデルを選択すれば良いでしょうか?
Anomalib (英語) は、モデルのベンチマークを測定して比較することができます。

パフォーマンスは…

…通常、AUROC と AUPRO の 2 つのメトリックで測定されます (図 4)。

受信者動作特性曲線下の面積 (**AUROC**) は従来のメトリックです。AUROC について知りたい場合は、scikit-learn* の簡単な説明 (英語) [4] が役に立ちます。Tom Fawcett の詳細な説明 (2006) (英語) [5] も参考になるでしょう。

> **免責事項 AUPRO**
> この記事では PRO 曲線の下の面積 (**AUPRO**) について詳しく説明しませんが、必要に応じてヒントを提供しますのでご心配なく 😅。
> ここでは、「AUPRO の値は 0 から 1 であり大きいほど優れている」ということのみ知っておいてください。

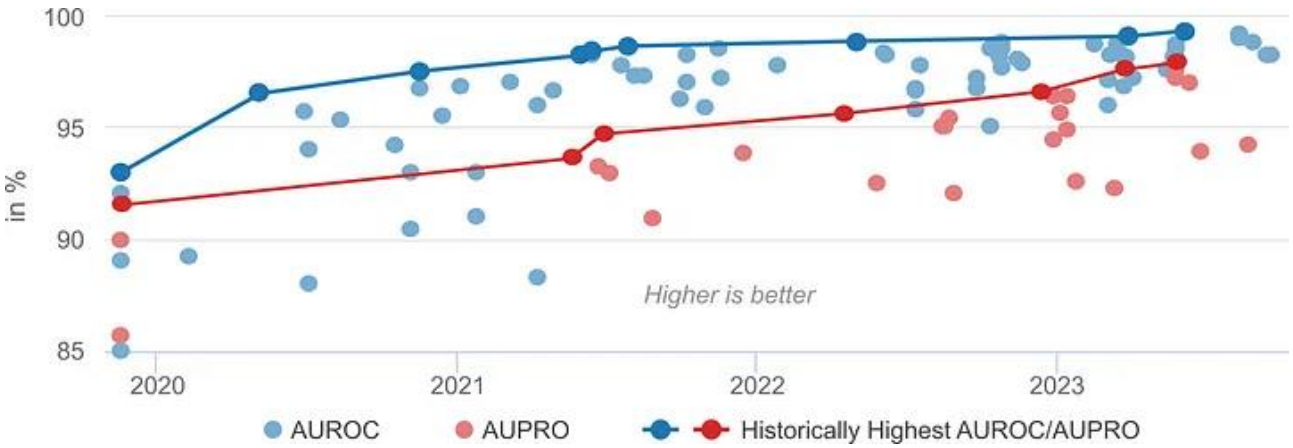

図 4. MVTec AD の 15 のデータセットの平均パフォーマンス。
すべてのモデルは Papers With Code のベンチマーク (英語) [6] でインデックスされています。

良いニュースは、研究が進んでいることです 😃。
悪いニュースは、ベンチマークが誤解を招きかねないことです 😟。

どのようにプラトーに達しているか注目してください。MVTec AD は「解決済み」のようです。

図 3 のモデルの AUROC は 98.3% です。しかし、予測 2 には非常に明白なミスがあります (トランジスターがありません)。

AUPIMO とは…

…**イメージごとのオーバーラップ領域下の面積 (Area Under the Per-IMage Overlap の略)** で、この問題に対処するために [1] で提案された新しいメトリックです。簡単に言うと、誤検出 (偽陽性) がほとんどないようにモデルが条件付けされている場合の、各イメージの再現率を測定します。

AUPIMO を使用して **27 のデータセット**で **8 つのモデル**のベンチマークを測定したところ、最先端の手法について全く新しいことが判明しました。

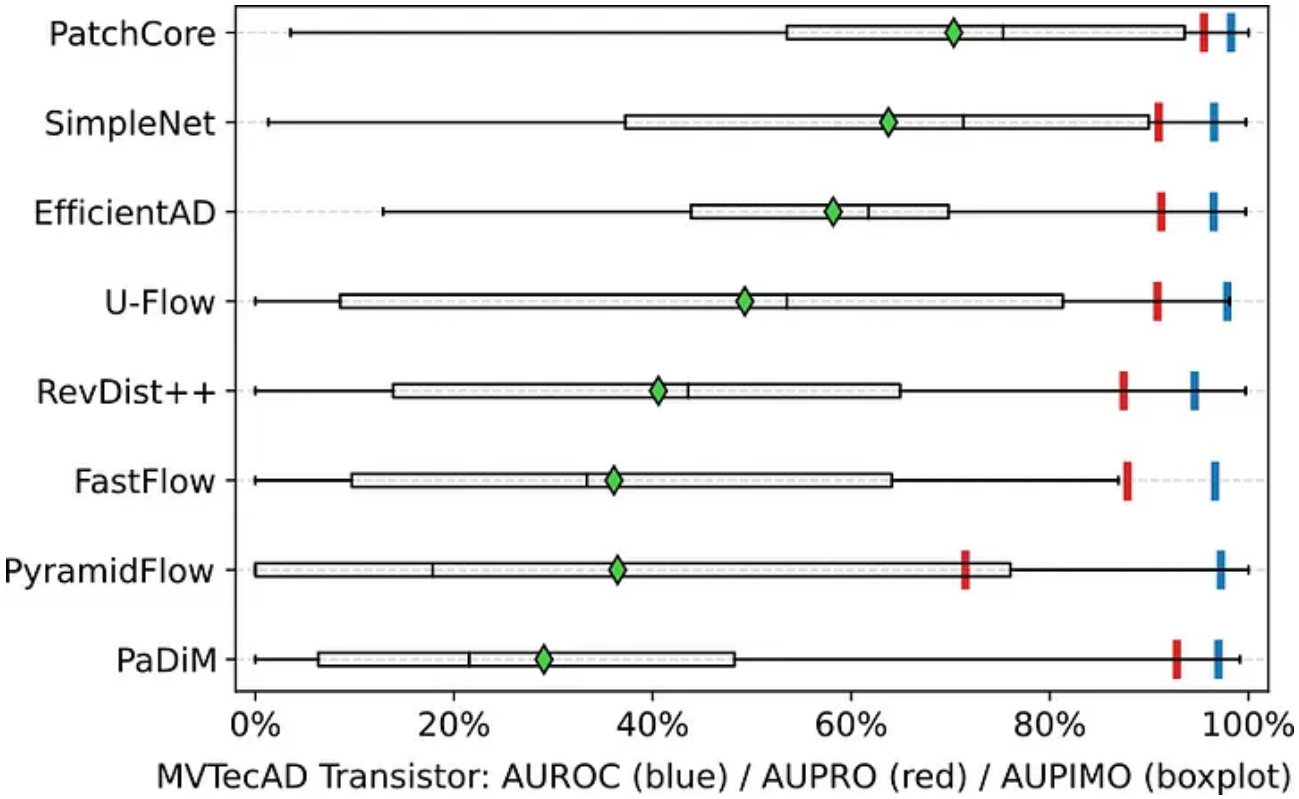

図 5. Transistor データセットのベンチマーク結果。
AUPIMO (イメージごとに 1 つのスコア) の箱ひげ図。ダイヤモンドは平均。

Transistor データセット (図 5) では、最高のモデル (PatchCore [3]) と最低のモデル (PaDiM [7]) の差は AUROC で 1.3% (AUPRO で 2.7%) です。

AUPIMO では、モデル間の差は非常に大きくなります。PatchCore は異常の 1/4 (<50% イメージ内再現率) を見逃していて、PaDiM は実にその 3 倍を見逃しています。

次の説明

ここまで読んでいただければ、AUPIMO に興味を持たれたことでしょう。
次のセクションでは、AUPIMO の定義について詳しく説明します。まず最初に、比較のために、親のメトリックである AUROC を定義します。
次に、AUROC からの変更点を説明します。
最後に、ビジュアル異常検出の最先端の手法についての新しいストーリーを紹介します。

準備

まず、いくつかの用語を定義します。

ビジュアル異常検出モデルは通常、異常スコアマップ (1 チャネルイメージ) を出力します。このスコアが高いほど「より異常」であることを意味し、しきい値を適用してマスクを生成します。

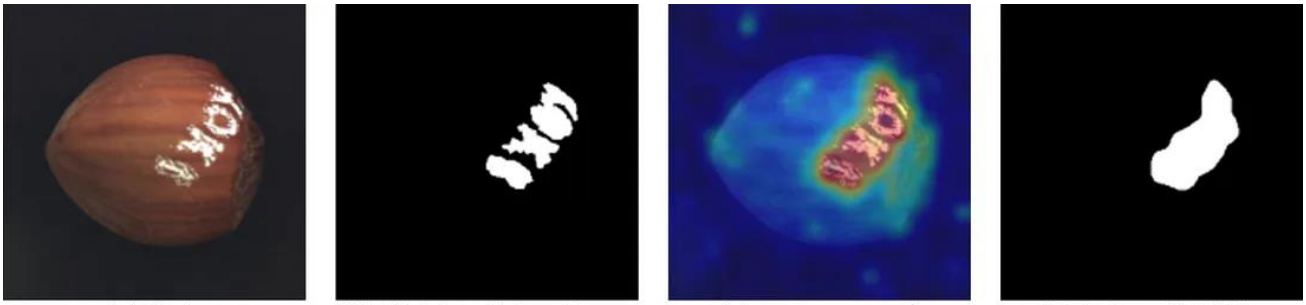

図 6. MVTec AD の Hazelnut データセットのイメージ。(a) 入力イメージ、(b) グラウンド・トゥルース・アノテーション、(c) 異常スコアマップ、(d) しきい値予測。出典: Anomalib (英語)。

図 6 は、MVTec AD コレクションの別のデータセットの例を示しています。スタンプは存在しないはずであり、異常です。

我々の目標は、モデルがイメージ内のすべての異常なピクセルを検出できるかどうかを測定することです。つまり、メトリックの仕事は、グラウンド・トゥルース・アノテーション (白の 1 は「異常」を意味します) と異常スコアマップを比較することです。

[AU]ROC から [AU]PIMO へ

AU = 曲線下の面積 (Area Under the curve の略)

簡単に言えば、PIMO 曲線は ROC 曲線の適応バージョンです。ROC (受信者動作特性) 曲線がどのように機能するか説明します。

> より正確には、タスクをピクセルの 2 値分類 (正常または異常) と見なす、**ピクセル単位の ROC** 曲線について話しています。

ROC

図 7 は、Toy データセットの 2 つの正常なイメージと 2 つのイメージの異常スコアマップを示しています。可能な各スコアしきい値で、偽陽性率 (FPR) と真陽性率 (TPR または再現率) (英語) がセット全体で測定されています (すべてのイメージが混同しています)。ROC 曲線は、すべての FPR/TPR ペアのセットです。

つまり、ROC とは正しい検出 (真陽性) と誤検出 (偽陽性) の歩み寄りを追跡したものと言えます。

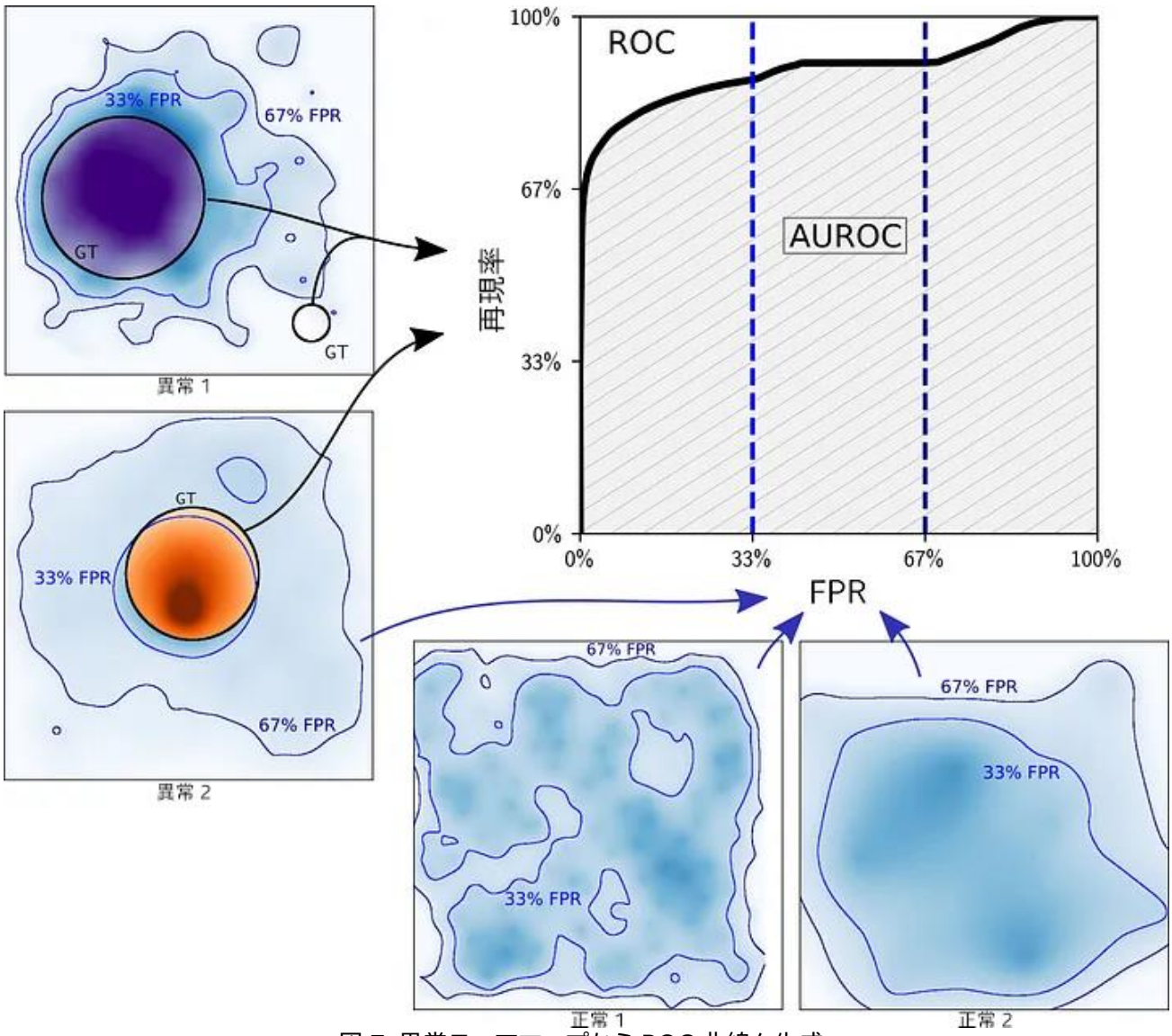

図 7. 異常スコアマップから ROC 曲線を生成。

曲線下の面積 (AUC) は、スコア内の曲線を要約したものです (図 8)。

2 値アノテーションのセット $\mathbf{Y}$ と異常スコアマップのセット $\mathbf{A}$ が与えられた場合、

$$F: \ t \ \mapsto \ \mathrm{FPR}(\mathbf{Y}, \ \mathbf{A} \geqslant t) \quad ,$$

$$T: \ t \ \mapsto \ \mathrm{TPR}(\mathbf{Y}, \ \mathbf{A} \geqslant t) \quad ,$$

ここで t はしきい値であり $\mathbf{A} \geqslant$ は $\mathbf{A}$ を t で 2 値化することを示す。この場合、

$$\mathrm{AUROC} = \int_0^1 T\left(F^{-1}(x)\right) dx \quad .$$

図 8. AUROC の定義。

完全なモデルの AUROC は 100%、ランダムモデル (比率に従って 0 と 1 をランダムに予測) の AUROC は 50% です。

積分が FPR のレベル (0 ~ 100%) で定義されていることに注目してください。次に、F を反転して、FPR レベルを生成するしきい値が出力されます。

結果的に、積分は通常のピクセルの最低スコアと最高スコアの間のすべてのしきい値を考慮します。これは、すべてのセグメンテーション・マスク (図 7 の青い等高線) をスキャンすることに相当します。

PIMO

AUROC は、FPR (x 軸) によりインデックスされたしきい値の領域における再現率関数 (y 軸) の平均であるという別の解釈もできます。

この見解に基づいて、ビジュアル異常検出にとってより有意義なメトリックになるように、PIMO (Per-IMage Overlap、イメージごとのオーバーラップ) 曲線を設計しました。

n アノテーションのセット $\mathbf{Y}$ と対応するスコアマップのセット $\mathbf{A}$ が与えられた場合、

$$\mathbf{YA}_0 = \{(y_i, a_i) \mid \forall i = 1 \ldots n \ \text{ if } \ y_i \ \text{ is normal}\}$$

は正常イメージのアノテーションとスコアマップの結合セットを示す。

x 軸は正常イメージの平均 FPR として定義される

$$F: \ t \ \mapsto \ \frac{1}{|\mathbf{YA}_0|} \sum_{(y,a) \ \in \ \mathbf{YA}_0} \mathrm{FPR}(y, \ a \geqslant t) \quad ,$$

ここで $|\cdot|$ は濃度オペレーターであり y 軸は y_i が異常であるすべての i について

$$T_i: \ t \ \mapsto \ \mathrm{TPR}(y_i, \ a_i \geqslant t)$$

イメージ内 TPR として再定義される。

図 9. PIMO 曲線の軸の定義。

各イメージには独自の曲線 (図 9 の下付き文字 *i*) があり、x 軸は対数目盛りで表現されていることに注意してください (図 10 を参照)。

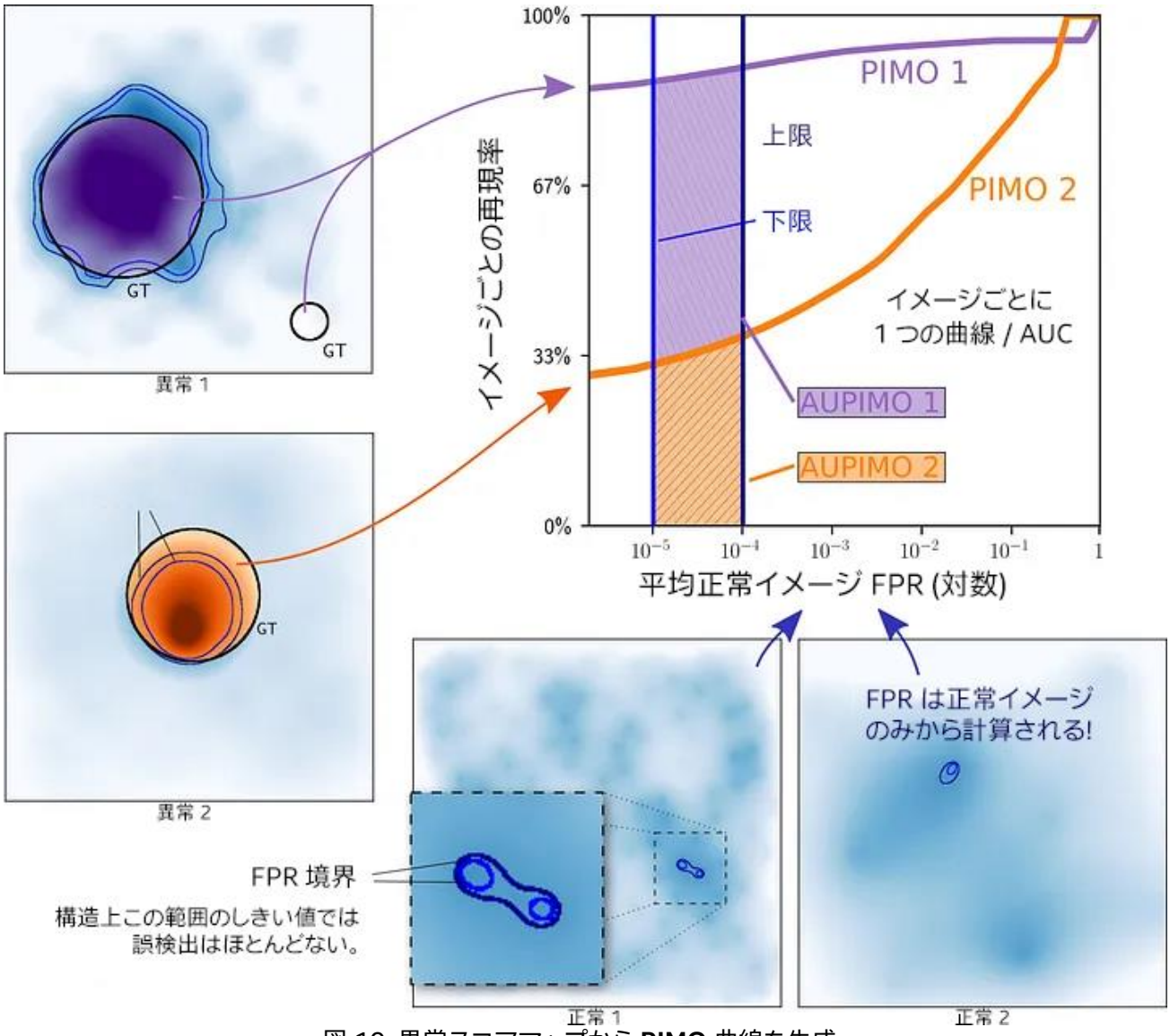

図 10. 異常スコアマップから **PIMO** 曲線を生成。

各イメージには独自の曲線下の面積 (AUC) があり、x 軸は対数目盛りで境界があります (デフォルトは 1e-5 および 1e-4)。

$$\text{AUPIMO}_i = \frac{1}{\log(U/L)} \int_{\log(L)}^{\log(U)} T_i\left(F^{-1}(x)\right) d\log(x)$$

ここで L と U は FPR の下限と上限*。

* より正確には正常テストイメージのイメージ内 FPR の平均。

図 11. AUPIMO の定義。

用語「1/log(U/L)」は、$0 \leq \text{AUPIMO} \leq 1$ のような正規化係数です。

異なる点

AUPIMO は、AUROC にいくつかの変更を加えています。
1. ピクセル比はイメージごとに計算されます。
2. x 軸は正常イメージのみから生成されます。
3. x 軸は対数目盛りで、AUC には境界があります。
4. 各イメージには独自のスコアがあります。

優れている理由

上記のリストを 1 つずつ見てみましょう。

1. ピクセル比はイメージごとに計算されます。

ROC 曲線の FPR/TPR メトリックは、テストセットのすべてのイメージのすべての正常/異常ピクセルを混合するため、イメージの構造は完全に無視されます。
PIMO では、類似したメトリックが各イメージのスコープ内で計算されるため、イメージの独立性が考慮されます。

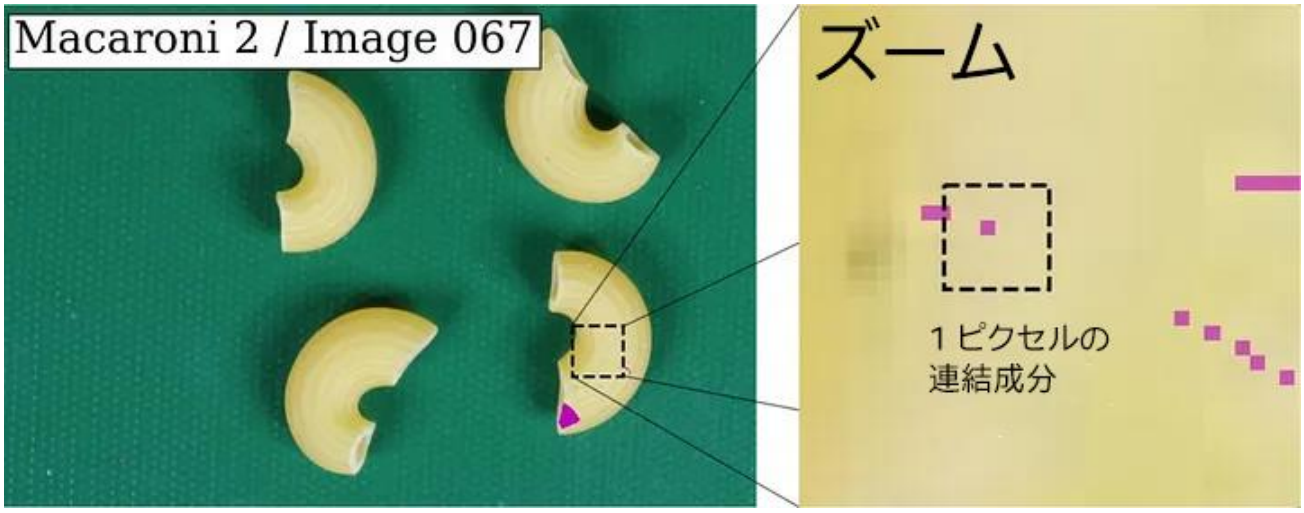

図 12. VisA [8] のデータセット Macaroni 2 のグラウンド・トゥルース・アノテーション内の小さな異常領域 (ピンク)。これらのノイズの多い領域は、AUPRO では非常に重視されますが、AUPIMO では無視されます。

> 各連結成分 (つまり、BLOB、領域) のスコープで TPR (再現率) を計算します。これが AUPRO の仕組みです。ただし、領域を見つけると計算は (大幅に) 遅くなります。
>
> また、AUPRO は小さな領域を重視しているため、アノテーションのミスに敏感です。もう 1 つの重要なデータセットのコレクションである **VisA** (Visual Anomaly Dataset) [8] には、ノイズの多いアノテーションが含まれています (図 12)。

2. x 軸は正常イメージのみから生成されます。

テストの正常イメージはトレーニングと同じ分布から仮定できるため、正常イメージのみを使用して x 軸を作成することは概念的に大きな違いを生みます。

x 軸のメトリックは、しきい値インデックス関数のように機能するため、非常に重要です。トレーニング・セットと同じ分布からサンプリングされたメトリックを使用すると、PIMO が代表的になり、利用可能な異常による偏りがなくなります。目標は、あらゆる種類の異常を検出することです。

異常イメージには正常ピクセルと異常ピクセルの両方があり、AUROC はそれらを混同しますが、モデルで異常に近い正常ピクセルのスコアが高くなることはよくあります。動作間の異常の境界は不明瞭な可能性があるため、その動作をメトリックで制限するべきではありません。

3. x 軸は対数目盛りで、AUC には境界があります。

関数 F のイメージ全体 (つまり、0 から 1、図 8 を参照) をスキャンすることにより、AUROC は無駄な動作点も考慮します。例えば、図 7 の 33% および 67% FPR レベルの等高線では非常に多くのピクセルが検出されます。

AUPRO [9] (この記事では取り上げていないため、論文 [1] を参照してください) と同様に、FPR に上限 (1e4) を設定して、メトリックを「十分に有用な」 (偽陽性が非常に少ない) しきい値に制限します。下限 (1e-5) は、x 軸の対数目盛りのために必要で、低レベルの FPR を「ズーム」するのに役立ちます。

この手法を使用すると、統合境界が何を表すかを、ビジュアルに、直感的に、簡単に作成できます。図 13 は、AUPIMO の境界 (1e-5 および 1e-4) でオブジェクトと比較して小さな偽陽性領域が生成される様子を示しています。

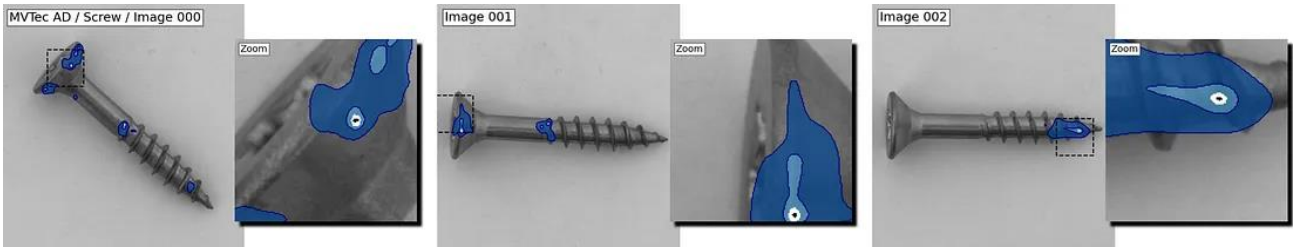

図 13. イメージ内の偽陽性率をビジュアルで直感的に把握。
MVTec AD のデータセット Screw のテストセット内の正常イメージの例。各色は、異なる FPR レベルのしきい値に対応しています (濃い青は 1e-2、明るい青は 1e-3、白は 1e-4、黒は 1e-5)。

3.5 検証評価フレームワーク

x 軸はトレーニング・セットと同じ分布のものであるため、FPR ベースの境界はモデル検証スキームのように機能します。

一方、y 軸は、評価に使用する、各イメージ内の異常なピクセルのピクセル単位の再現率を測定します。

我々は、この 1 つで 2 つの機能を持つスキームを「検証評価フレームワーク」と呼んでいます。

4. 各イメージには独自のスコアがあります。

イメージに個別のスコアを割り当てることには、重要な利点があります。
1. 複数のイメージ (例: 最低のケースと最高のケース) を比較することが可能です。
2. スコアの分布を含む詳細なベンチマークが表示されます。

次のセクションで実際に見てみましょう。

ベンチマーク

MVTecAD と VisA の 13 のモデルで 27 のデータセットのベンチマークを測定したところ、これまで詳細が不足していたために見つからなかったいくつかの情報が明らかになりました。

ここではモデルの詳細には注目しません。

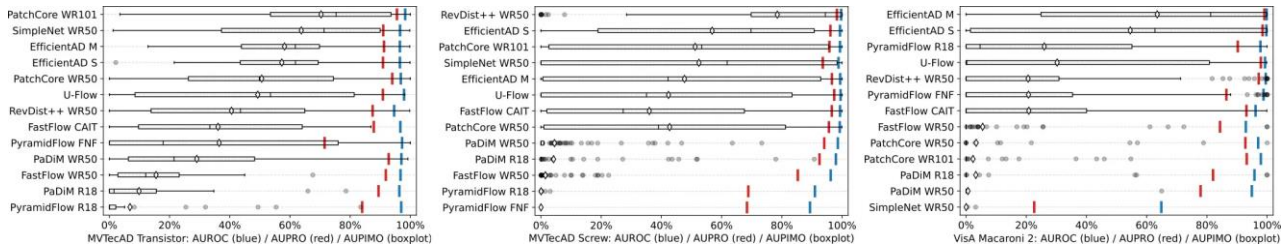

図 14. 3 つのデータセット (MVTec AD の Transistor および Screw、VisA の Macaroni 2) のベンチマーク。ダイヤモンドは平均の AUPIMO スコアです。

AUROC (青) と AUPRO (赤) ではモデル (特に最高のケース) 間の違いはほとんど示されていませんが、AUPIMO では平均値が近い場合でもパフォーマンスの変化が示されていることに注目してください。

AUPIMO により、最先端の手法に関する 2 つの事実が明らかになりました。
1. ほとんどの場合、クロスイメージの再現率は大きく異なります。
2. 最高のモデルであっても、一部の異常を再現できていません (図 14 の左のキュー)。

これは当初の認識 (図 4) と大きく異なっていることに注目してください。

ベンチマークのすべてのモデルとデータセットの概要を示した図 15 も確認してください。

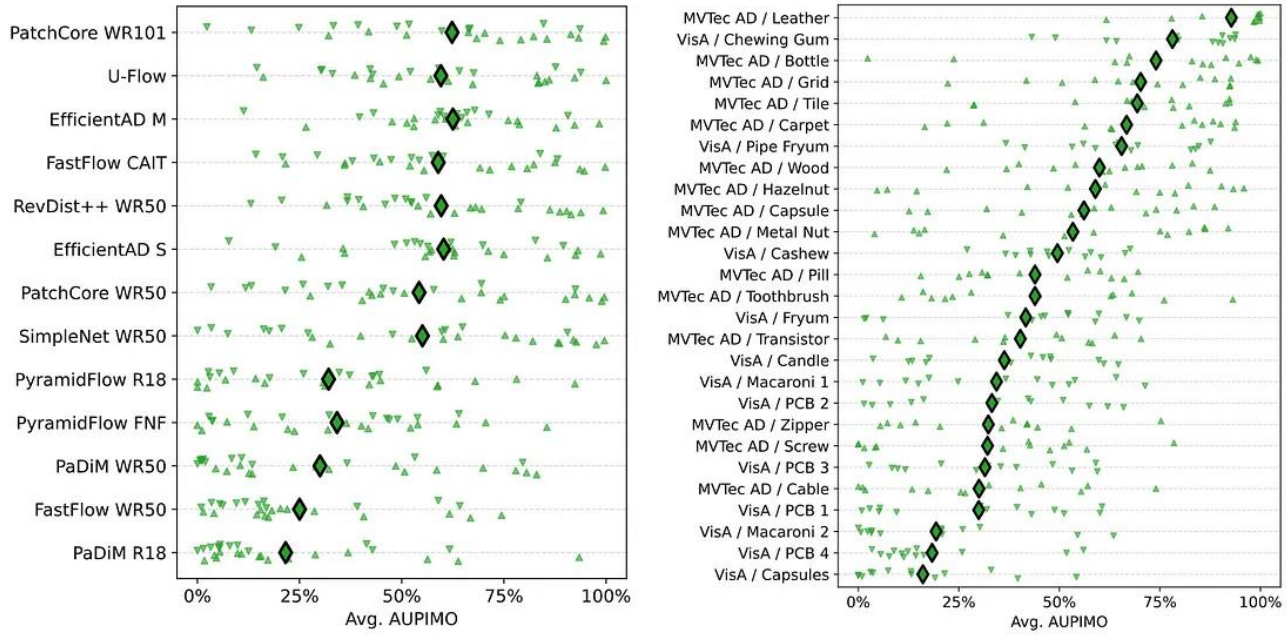

図 15. ベンチマークのサマリープロット (8 つのモデル X 27 のデータセット)。

ここで得られるもう 1 つの教訓は、すべてに優れたモデルはないということです。全体的に最高のモデル (PatchCore WR101) も、一部のデータセットではパフォーマンスが低くなっています。逆に、最低のモデル (PaDiM R18) でも、1 つのデータセットでは非常に優れたパフォーマンスを達成しています。

まとめ

この記事では、GSoC 2023 中に提案された新しいパフォーマンス・メトリックである **AUPIMO** の概要を説明しました。

従来の異常検出メトリックとは異なり、AUPIMO は個々のイメージにスコアを付けて、モデルのパフォーマンスの微妙な差異の評価を提供します。また、その検証評価フレームワークは、タスクに意味のある解釈を提供します。

> AUPIMO は、誤検出 (偽陽性) がほとんどないようにモデルが条件付けされている場合の、イメージの平均セグメンテーションの再現率です。

27 のデータセットと 15 のモデルでテストを行い、「解決済み」のデータセットのアイデア全体が現実のものであることを確認できました。お伝えするのを忘れていましたが、AUPIMO の計算は AUPRO よりもはるかに高速です。

> 詳細は、論文とリポジトリーで確認してください。

リンク (英語)

1. GSoC 2023 プロジェクトのページ: summerofcode.withgoogle.com/programs/2023/projects/SPMopugd
2. arXiv の論文: arxiv.org/abs/2401.01984
3. スタンドアロン・コード: github.com/jpcbertoldo/aupimo
4. GitHub* の OpenVINO™ ページ: github.com/openvinotoolkit/openvino
5. Anomalib: github.com/openvinotoolkit/anomalib

   anomalib との統合 (英語) は、anomalib v1 リリース (英語) で予定されています。

参考文献 (英語)

1. J. P. C. Bertoldo, D. Ameln, A. Vaidya, and S. Akçay, "AUPIMO: Redefining Visual Anomaly Detection Benchmarks with High Speed and Low Tolerance." arXiv, Jan. 03, 2024. doi: 10.48550/arXiv.2401.01984.
2. P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection," CVPR, 2019.
3. K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards Total Recall in Industrial Anomaly Detection," CVPR, 2022.
4. scikit-learn.org/stable/modules/model_evaluation.html#receiver-operating-characteristic-roc
5. T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, 2006, doi: 10/bpsghb.
6. paperswithcode.com/sota/anomaly-detection-on-mvtec-ad
7. T. Defard, A. Setkov, A. Loesch, and R. Audigier, "PaDiM: A Patch Distribution Modeling Framework for Anomaly Detection and Localization," ICPR, 2021.
8. Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, "SPot-the-Difference Self-Supervised Pre-training for Anomaly Detection and Segmentation," arXiv, 2022, doi: 10.48550/arXiv.2207.14315.
9. P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "The MVTec Anomaly Detection Dataset: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection," IJCV, 2021, doi: 10/gjp8bb.

OpenVINO™ ツールキットとは

AI を加速する無償のツールである OpenVINO™ ツールキットは、インテルが無償で提供しているインテル製の CPU や GPU、VPU、FPGA などのパフォーマンスを最大限に活用して、コンピューター・ビジョン、画像関係をはじめ、自然言語処理や音声処理など、幅広いディープラーニング・モデルで推論を最適化し高速化する推論エンジン/ツールスイートです。

OpenVINO™ ツールキット・ページでは、ツールの概要、利用方法、導入事例、トレーニング、ツール・ダウンロードまでさまざまな情報を提供しています。ぜひ特設サイトにアクセスしてみてください。

https://www.intel.co.jp/content/www/jp/ja/internet-of-things/openvino-toolkit.html