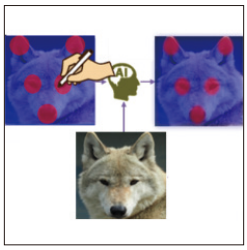


# 人の知見を反映した外観検査 AI

AI for Visual Inspection Reflecting Human Expertise

**松本　悠希**
Yuki Matsumoto

当社では製品の外観検査を行うAIの開発・運用を進めている。そのようなAIを製造現場に導入するうえで、課題を二点挙げることができる。まずAIの開発に必要な画像データとその注釈（教師データ）の準備に多くの人手を要すること、次にAIの検査結果の根拠が可視化されず製造現場での信頼が得にくいことである。前者については、ChatGPTに も採用されたTransformerというAIに教師データを極力必要としない自己教師あり学習を応用し、後者については、AIの注視領域を明示する技術であるアテンションを応用して人の知見をAIに反映させる機能を開発したので報告する。アテンションを活用する方法として弊社独自のシグモイドアテンションを採用したことで、より明確にAIの判断根拠を可視化しつつ性能向上にも寄与することができたため、あわせて報告する。

We are developing and operating AI systems for automated visual inspection of products using image data. Implementing such AI in manufacturing environments presents two main challenges: First, substantial manpower is required to prepare the necessary image data and its accompanying annotations (labeled data) for AI development. Second, the basis of AI inspection results is not visualized, hindering trust in the factory setting. To address the first challenge, we applied self-supervised learning to the Transformer AI architecture, a method also adopted by ChatGPT, minimizing the need for extensive annotations. For the second challenge, we developed a function to incorporate human knowledge into AI by utilizing our novel sigmoidal attention mechanisms to clarify the areas of focus. Our original sigmoidal attention adopted as a method of utilizing attention not only enhances the visualization of the grounds for AI judgments but also contributes to performance improvement. We report our solutions to the two challenges in detail.

キーワード：外観検査、自己教師あり学習、Transformer、アテンション

## 1. 緒　　言

当社では製品の外観検査を行う自動外観検査AIの開発が行われており、多くの工程にて運用が進んでいる。製造現場での外観検査をAIに代替させるにあたって、不良流出防止などの現場の基準を担保することが前提のもと、そのようなAIを製造現場に円滑に導入するにあたり、次の二つの課題を解決することが求められることが多い。

一つ目は、一般的なAIの開発には画像データとその注釈（教師データ）を必要とすることである。自動外観検査AIにはディープラーニングを用いるが、運用に耐える高い性能を得るには、数千枚以上の学習用画像と教師データが必要となるため、それらの準備に多くの人手を要することが製造現場では大きな課題となる。

二つ目はAIの検査結果の根拠が可視化されず製造現場での信頼が得にくいことである。一般的なAIの機能は良否判定の結果のみを通知することに限定されることが多く、なぜ不良と判定したかの根拠を提示することは考慮されていない。そのため不良と判定された画像が製造現場の検査基準と合致しているかを確認するのは難しく、さらにはその検査基準を満足するようにAIを調整することも一般的には困難となる。

そこで我々は、教師データを極力必要としない学習方法として近年注目を集めている自己教師あり学習を導入し、大量の学習データを用意する製造現場の負担軽減を試みた。また、検査結果の根拠が示されない課題に対しては、AIの注視領域を学習する技術であるアテンションを活用して検査結果の根拠を可視化するとともに、このアテンションを経由して製造現場の知見をAIに直接反映させることにより、より製造現場の理解を得やすい自動外観検査AIの実現を目指した。アテンションを活用する方法として弊社独自の技術であるシグモイドアテンションを採用したことで、より明確に判断根拠を可視化しつつ性能向上を実現した。

本稿では2章で自己教師あり学習について、3章では人の知見を反映させるAIの実現方法について、4章で本技術を適用した結果を報告する。

## 2. 自己教師あり学習

### 2−1　教師あり学習の課題

一般的なAI学習では、大量の画像を収集し各画像に注釈を付けて作成した学習用データ（教師データ）を用いた教師あり学習が採用されることが多いが、多くの時間と労力を必要とすることから製造現場の自動外観検査AIの導入の障壁となることが多い。また現場で収集された大量の画像に偏りがあった場合、AIに誤った根拠で判断させる悪影響が起こる可能性も指摘されている。例えば、**図1**の赤い部分は、AIが画像を「狼」と判断した際に最も注視した領域を表わしているが、AIは「狼」を背景の雪山を根拠として

図1　AIの判断根拠の可視化の例

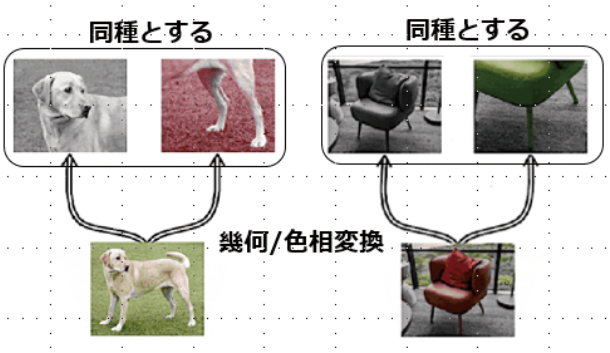



図2　自己教師の生成例[2]

判断していることが示されている[1]。この要因は雪山と狼がセットで現れる学習画像が多いことに起因している。

### 2−2　自己教師あり学習

自己教師あり学習は、注釈が付けられていない大量の画像のみの学習データからAIが有用な特徴量の表現を学習する機械学習の手法である。2−1で前述した従来の教師あり学習では注釈付きの学習データを必要とするが、自己教師あり学習ではデータ自体から教師信号を生成する。例えば**図2**に示すように、オリジナルの学習画像に幾何変換や色相変換を行った二次学習画像と、同じ犬の画像から派生した二次学習画像は互いに同じ画像とみなすという教師信号をAI自身が生成し、これらの学習を繰り返すことで自己教師あり学習が実現される。「犬」あるいは「椅子」という注釈はなくとも、それぞれの二次学習画像は同種とみなす自己教師により、「犬」と「椅子」を判別するために必要な表現方法が学習されていく。

この自己教師あり学習にはSimCLR[2]、BYOL[3]、など多数のAIモデルがあるが、本稿では他モデルより分類精度の面で優れるDINO[4]を採用した。DINOはChatGPTにも採用されているTransformerを画像向けに応用したものである。

## 3.　人の知見を反映させるAIの実現

### 3−1　アテンションの応用

AIの判断根拠を可視化し、さらに製造現場の知見をAIに直接反映させる機能を実現するために、DINOに元々備わっているアテンションという機能を応用することにした。アテンションは、画像の分類精能向上を目的として画像のどの部分を重要視すべきかをAIが学習する技術である。画像の注視すべき領域を学習し、その注視すべき領域と一致する画像特徴を採用してAIの分類性能を向上させる。ただし、人に提示することを目的として構成されていないため、我々はこのアテンションを可視化してAIの判断根拠を人に理解しやすく明示するだけでなく、アテンションに直接介入して人の知見をAIに教える機能の開発を試みた。この試みによりAIの判断根拠を製造現場に明示してAIに対する信頼感を得るだけでなく、人の知識を直接取り込むことで自動外観検査の性能向上も同時に期待できる。

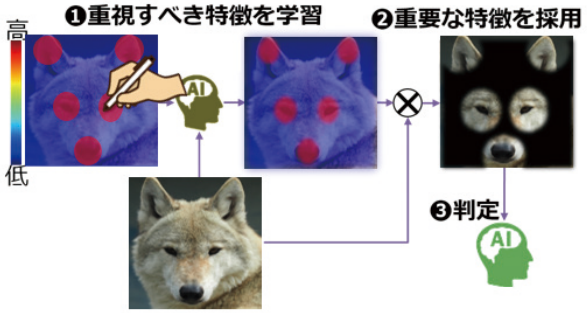


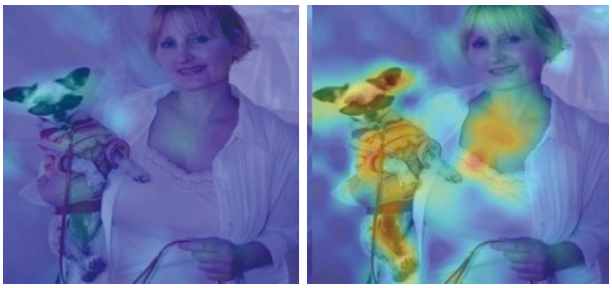

図3　人の知見を模倣するAI

**図3**は、狼を見分けるためにその目や鼻に注視するべきという人の知見を、AIのアテンションに直接書き込む手順を示している。例えば、狼の目など画像を見分けるために重要度を高くしたい部分には、例えばタッチパネルなどを使って人が直接指示内容を書き込むようにする。その指示内容に従ってアテンションを模倣して学習したAIは、狼の「目」を重要視して判断するようになり、人と同じ根拠で見分けることが可能となる。

### 3−2　シグモイドアテンションの導入

3−1で述べたようにアテンションは人との親和性を考慮した技術ではないため、それ単体では人にとって扱いやすい機能として成立しない。**図4 (a)** はアテンション単体の時の可視化結果を示している。この例ではチワワに注視していることを明示することが望まれるが、どこを重要視しているかの強調があいまいである。この課題を解決して人との親和性を高めるために我々はシグモイド関数を導入したアテンション、シグモイドアテンションを開発した。

シグモイド関数を**図5**（実線）に示す。導入の理由は、人にとって扱いやすい範囲に写像可能であり、かつ、AIの学習が微分を前提にした数学的な連鎖を基に理論が構成されているためである。すべての入力が（0、1）の出力に写像されるため人の直感的な重要度とも一致しやすく、また微分可能であるため前記連鎖に与える影響も少ない。また



（a）アテンション単体　　（b）シグモイド

**図4　アテンションの例**

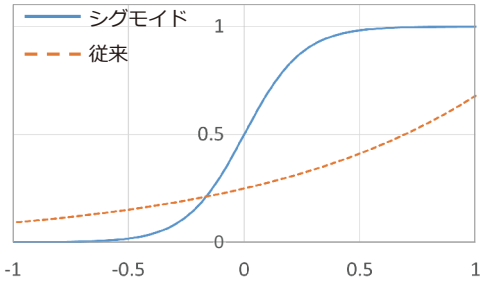



**図5　シグモイド関数**

従来のアテンション（**図5**点線）をシグモイドアテンションと比較すると、シグモイドよりも変動が少ないため強弱のつきにくい変換となり、可視化したとしても視認性の低いアテンションとなってしまう。

シグモイドアテンションの例を**図4 (b)** に示す。アテンション単体とは異なり、AIがチワワ全体に注視していることが明瞭にわかるようになり、また本来注視すべきでない人間の首付近も重要度が高いことが確認できる。

## 4. 実　　験

### 4－1　学習データ

**図6**に実験に用いるワイヤーハーネスの例を示す。これらの画像は黒色の蛇腹管に黒色のビニールテープが巻かれているか否か、および巻かれ方によって分類される。ビニールテープが巻かれていないものを「not_tape」、隙間なく巻かれているものを「tape」、巻かれているが一部蛇腹が露出しているものを「mix」として3クラスに分類する。

実験ではこれらの画像を240枚ずつの計720枚を学習データとした。2－2で述べたようにこれらのデータにはテープ巻き種の注釈は付随されておらず、自己教師あり学習を用いAIを学習している。AIの性能を確認するためのテストデータとして計7,800枚（各クラス2,600枚）を準備した。

### 4－2　実験方針

最初に**図6**に示すハーネス画像に対して自己教師あり学習によるAIの学習を行う。この時点ではシグモイドアテンションは導入していないため、自己教師あり学習の従来法として本稿におけるベンチマークに設定する。以降はこのベンチマークとなるAIを「通常学習」と呼称し、後述のシグモイドアテンションを用いた提案手法と区別する。

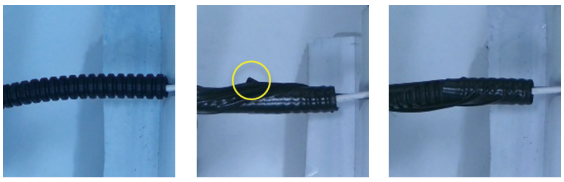



**図6　ハーネステープ巻きの種類**

次に通常学習の段階で不正解であった画像と、正解ではあるが背景に映り込んだゴミへの注視が強いことでAIの判断に悪影響を及ぼす可能性のある画像をそれぞれ1枚ずつ学習データの中から抽出した。この2つの学習データにおいてシグモイドアテンションを導入した上で、人の知見を基に作成した注視領域をAIに模倣させる学習を追加で実施した。

### 4－3　通常学習の結果

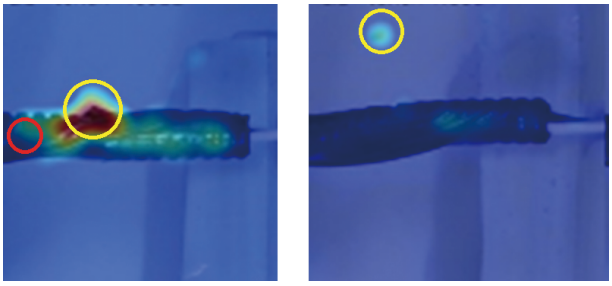
10回の通常学習を実施し、テストデータにおけるテープ巻きの分類精度を検証したところその正解率は平均83.57%であり、最大値は84.26%、最小値は83.06%であった。**表1**に最大精度時の正誤数を記録した混合行列を示す。

**表1**からnot_tapeは不正解がないことから、黒テープが全く巻かれていないnot_tapeと一部あるいは全部が黒テープで覆われているその他のクラスは、分離可能な特徴量を自己教師あり学習によりAIが獲得できたことが示されている。一方、テープを一部巻いたmixはその約45%がtapeと誤認識している。

**図7 (a)** には通常学習において不正解となった学習データ1のアテンションを示す。学習データ1においては、ビニールテープの山型状になる箇所への（黄丸）注視が強すぎる一方で、赤丸で示す蛇腹の露出部への注視が弱すぎることによって必要以上に黒テープへの注視が強調されていることが確認できる。実際の生産現場において、人がtapeとmixを見分ける際には、この蛇腹部分の露出の有無に着目するはずなので、このアテンションは不適切と判断でき

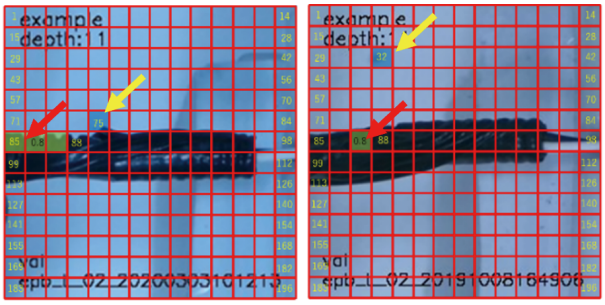
**表1　通常学習の最大精度時の混合行列**

| | 正解 | | | | 計 |
|---|---|---|---|---|---|
| 予測 | クラス | mix | not_tape | tape | |
| | mix | 1435 | 0 | **63** | 1498 |
| | not_tape | 0 | 2600 | 0 | 2600 |
| | tape | **1165** | 0 | 2537 | 3702 |
| 計 | | 2600 | 2600 | 2600 | 7800 |



（a）学習データ 1　　（b）学習データ 2

**図7　通常学習時点でのアテンション**




（a）学習データ 1　　（b）学習データ 2

**図8　シグモイドアテンションの指示方法**

る。また背景にゴミが写った学習データ2（**図7 (b)**）のアテンションでは、黄色で示したゴミ部分への注視が強いために、そもそものターゲット（ハーネス）への注視が弱くなるという悪影響が見て取れる。

### 4－4　提案手法による追加学習

4－3で指摘した人の知見から得られた改善点をAIに反映すべく、提案手法であるシグモイドアテンションを用いた追加学習を行う。**図8**に人の知見を反映させる方法を示す。提案手法のシグモイドアテンションにおいては、14×14のパッチに区切った状態で人の知見を反映させたい部分のパッチを選択することで、AIに追加学習を促すような実装を採用している。

学習データ1の黄色矢印のパッチは、テープの突起への注視を弱めるためにこのパッチの目標値を0.1とするようAIに指示し、一方で蛇腹に注視することが重要という人の知見に基づいて蛇腹が露出する赤色矢印は0.8を目指すように設定した。同様に学習画像2ではゴミがある黄色矢印部分の目標値は0.1とし、蛇腹の露出部を示す赤色矢印では0.8とした。それ以外の箇所は人からの指示はなくAIの出力をそのまま用いることを表す。

前述の学習条件にて10回の提案手法の追加学習を実施した結果、テスト画像に対する分類精度は平均84.49%、最大値87.03%、最小値82.23%であった（**表2**）。最小値は通常学習が高いものの、提案手法は通常学習と比較して、平均が約0.92%、最大値は約2.8%向上した。追加学習に使用した学習データは2枚と少ないにも関わらず平均して70枚以上が正解に転じるという大きな改善効果を得た。**表3**に提案手法置ける最大精度時の混合行列を示す。通常学習の**表1**と比較すると、tapeの誤認識が増えた一方で、追加学習を導入したmixの誤認識の改善効果が高く、結果として精度が向上したことがわかる。

また**図9**は通常学習（上段）から提案手法（下段）によるアテンションの変遷を示している。人の知見を基に注視を弱めるべきと設定したテープの突起部分とゴミの位置（黄丸）においては、提案手法の時点では指示通り注視度合が低減されていることが確認された。対して蛇腹の露出部分を注視してテープの巻き方を区別する人の知見を反映させた追加学習の部分（赤丸）では、該当箇所の注視が強くなっており人の知見に即した視覚的説明性を獲得していることが確認された。

**表2　各手法のテストデータにおける分類精度（%）**

| | 最小 | 最大 | 平均 |
|---|---|---|---|
| 通常学習 | 83.06 | 84.62 | 83.57 |
| 提案手法 | 82.23 | 87.03 | 84.49 |

**表3　提案手法の最大精度時の混合行列**

| | | 正解 | | | 計 |
|---|---|---|---|---|---|
| | クラス | mix | not_tape | tape | |
| 予測 | mix | 1749 | 0 | **161** | 1910 |
| | not_tape | 0 | 2600 | 0 | 2600 |
| | tape | **851** | 0 | 2439 | 3290 |
| 計 | | 2600 | 2600 | 2600 | 7800 |

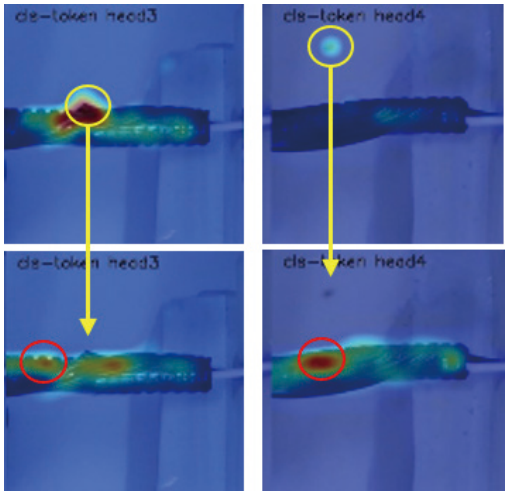



**図9　通常学習（上段）から提案手法（下段）へのアテンションの変遷**

## 5. 結　言

本稿では自己教師あり学習のアテンションに対して柔軟な追加学習が可能なシグモイドアテンションを導入することを提案するとともに、人の知見を反映させた追加学習を行った後の分類精度やアテンションの視覚的説明性について検証した。シグモイドアテンションによる追加学習後は分類精度が平均で84.49%、最大値が87.03%であり、通常学習と比較すると平均が約0.92%、最大値が約2.8%向上した。また、アテンションの視覚的説明性については、通常学習時点では映り込んだゴミやテープの極端な尖り部分への注視が強く、人が見て重要と考えられる箇所（蛇腹の露出部）への注視が弱かった。一方、追加学習後は人の知見がAIに反映されたことで上記への注視が薄れ、蛇腹の露出部へ注視を誘導することができ、より人の知見に即したアテンションが獲得できた。提案手法は導入コストが低いにも関わらず、分類精度やアテンションの視覚的説明性が向上する点で大きなメリットと言える。



**参 考 文 献**

(1) M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?" explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144 (2016)
(2) Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton, A Simple Framework for Contrastive Learning of Visual Representations, arXiv:2002.05709 (1 Jul. 2020)
(3) Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H.Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, MohammadGheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, Michal Valko: Bootstrap Your Own Latent A NewApproach to Self-Supervised Learning, arXiv:2006.07733v3 [cs.LG] (10 Sep. 2020)
(4) Mathilde Caron, Hugo Touvron, Ishan Misra, Herv´e Jegou, Julien Mairal, Piotr Bojanowski, Armand Joulin "Emerging Properties in Self-Supervised Vision Transformers," arXiv: 2104.14294v2 [cs.CV] (24 May. 2021)

**執 筆 者**

**松本　悠希**　：DX研究開発センター　主席

＊主執筆者