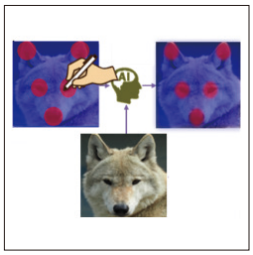


AI for Visual Inspection Reflecting Human Expertise

Yuki MATSUMOTO

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 sigmoid attention mechanisms to clarify the areas of focus. Our original sigmoid 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.

Keywords: visual inspection, self-supervised learning, transformer, attention

1. Introduction

Sumitomo Electric Industries, Ltd. is working on developing AI for automated visual inspection of products (automated visual inspection AI) and is using it in many of its manufacturing processes. For AI to replace the manufacturing staff in the visual inspection task, it is a prerequisite that it adheres to the worksite criteria, including those intended for preventing the escape of defects. To introduce such AI smoothly to the manufacturing site while meeting this prerequisite, the following two challenges need to be solved in many cases.

One issue is that, in general, the development of AI requires image data and corresponding annotations (labeled data). Automated visual inspection AI uses deep learning, which requires at least a few thousand of training images and labeled data, in order for it to acquire a high capability and be practically operational. Preparation of them requires a lot of labor, which is a major challenge to the manufacturing site.

The other issue is that AI produces inspection results without visualizing the grounds of the inspection results, making the results not readily trusted by the manufacturing staff. In general, AI functions are largely limited to mere notification pass/fail results and are not designed to present the reason or grounds used to reach the decision for rejection. This makes it difficult to verify whether or not the image used for determining the product to be defective agrees with the inspection criteria of the manufacturing site. Moreover, it is generally not easy to tune the AI system to be in accord with the inspection criteria.

As a solution, we introduced self-supervised learning, which has been recently gaining interest as a learning method requiring minimum labeled data, and made an attempt to reduce the workload of the manufacturing staff required for preparing a large volume of training data. The issue that no grounds of inspection results are presented was addressed by utilizing an attention mechanism, which is a technique of training AI's areas of focus, thereby enabling the inspection results to be visualized. In addition, by directly incorporating the expertise of the manufacturing staff via this attention mechanism, we aimed to achieve automated visual inspection AI that is more readily accepted by the manufacturing staff. We adopted sigmoid attention, our proprietary technology, as a method of utilizing attention and achieved successful performance improvements while making the grounds of decisions more clearly visible.

Chapter 2 of this paper reports on self-supervised learning, Chapter 3, on how to implement AI that incorporates human expertise, and Chapter 4, on the results of application of this technology.

2. Self-Supervised Learning

2-1 Challenges associated with supervised learning

General AI training in many cases adopts supervised learning, which uses training data (labeled data) created by collecting a huge number of images and adding an annotation to each image. However, this supervised learning requires a lot of time and labor, which often hinders the introduction of automated visual inspection AI to the manufacturing site. Moreover, it has been pointed out that if a large volume of images collected in the field is biased, it may adversely cause AI to make decisions on erroneous grounds. For example, the red zones in Fig. 1 correspond to areas which the AI system focused mostly on when determining that the image portrays a wolf. Figure 1 reveals that the AI system used the background snowy hill as the grounds for the decision.[1] This resulted from the abundance of training images containing a snowy hill and a wolf or wolves as a set.

Fig. 1. Examples of the visualization of the grounds for AI judgments

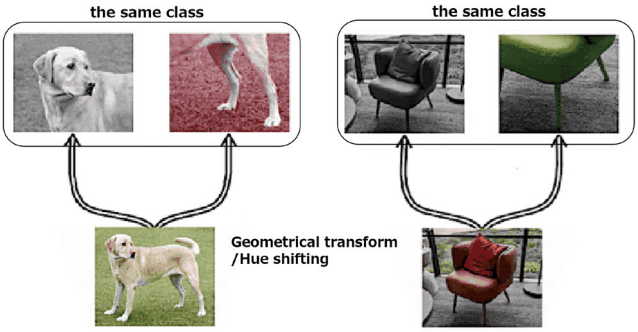


Fig. 2. Examples of self-supervised learning[2]

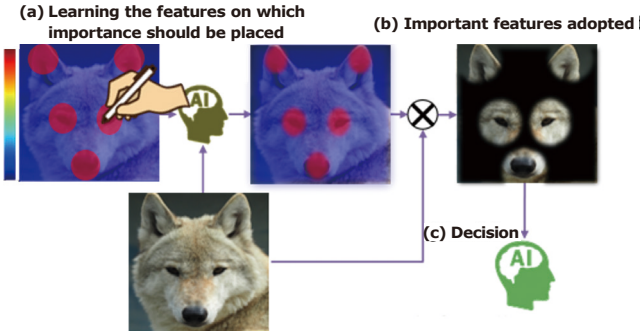


Fig. 3. AI imitating human expertise

2-2 Self-supervised learning

Self-supervised learning is a machine learning technique for AI to learn representations of useful features from training data only consisting of a large number of unannotated images. Traditional supervised learning described in 2-1 requires annotated training data; in contrast, self-supervised learning generates supervisory signals from the data itself. For example, as illustrated in Fig. 2, self-supervised learning is implemented through a course in which AI itself generates supervisory signals to treat secondary training images as the same images whether they have been geometrically transformed or hue-shifted from original training images or are derived from images of the same dog and through repetition of this learning course. Without annotations such as a "dog" or "chair," the self-supervised approach, which considers that these secondary training images are of the same kind, trains the AI system to acquire necessary representation methods required for distinguishing dogs from chairs.

While many AI models, such as a simple framework for contrastive learning of visual representations (SimCLR)[2] and Bootstrap Your Own Latent (BYOL),[3] are available for this self-supervised learning, for this report, we adopted Distillation with No Labels (DINO),[4] which is superior to other models in terms of classification accuracy. DINO was developed by adopting the Transformer, which is also used in ChatGPT, for use with images.

3. Implementing AI Incorporating Human Expertise

3-1 Application of attention

We decided to apply the attention mechanism originally provided in DINO in order to visualize AI's decision-making grounds and realize the functionality of incorporating the expertise of manufacturing staff directly in AI. Attention is a technique used with AI to learn on which parts of an image importance should be placed with the aim of improving AI's classification capability. AI learns the areas of focus in images and improves in classification performance by adopting such features of images that coincide with the areas of focus. However, it is not configured for presentation to humans. Therefore, we attempted to develop functions for not only clearly indicating AI's decision-making grounds by visualizing this attention but also teaching the AI system human expertise by directly intervening in the attention. This attempt is expected not only to gain the trust of the manufacturing staff by clearly indicating AI's decision-making grounds to them but at the same time to improve the automated visual inspection performance by directly incorporating human knowledge.

Figure 3 shows the steps followed to directly write, into AI attention, human knowledge that one needs to focus on a wolf's eyes and nose when identifying it as a wolf. For example, an operator directly writes instructions using a touch panel or the like for a wolf's eyes and other parts to be assigned greater importance so as to distinguish images. AI having imitated and learned attention according to the instructions begins attaching importance to the eyes of the wolf and acquires the ability to distinguish images on the same grounds as those used by humans.

3-2 Introduction of sigmoid attention

As described in 3-1, attention is not a technique developed with affinity for humans in mind. Accordingly, attention itself does not hold as a human-friendly function. Figure 4 (a) presents the results of visualization where the attention mechanism alone was used. This example reveals ambiguous highlighting made to indicate where importance was placed, although it is desirable that the AI system clearly indicates that its focus was placed on the Chihuahua. To solve this problem and to enhance the affinity of the attention mechanism for humans, we developed sigmoid attention, an attention mechanism incorporating a sigmoid function.

The sigmoid function is given in Fig. 5 (solid line). We introduced it because it enables mapping to a range that is easy for humans to handle and AI learning is theoretically configured based on mathematical relationships that are predicated on differentiation. Because all inputs are mapped to [0, 1] outputs, the sigmoid function aligns well

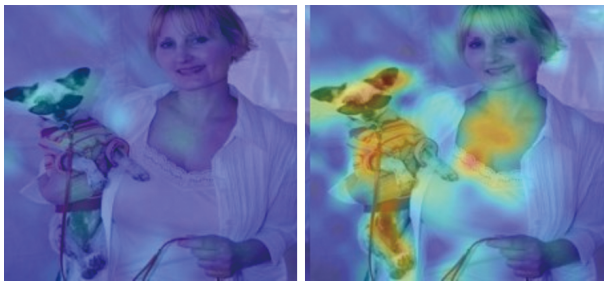


Fig. 4. Example of attention

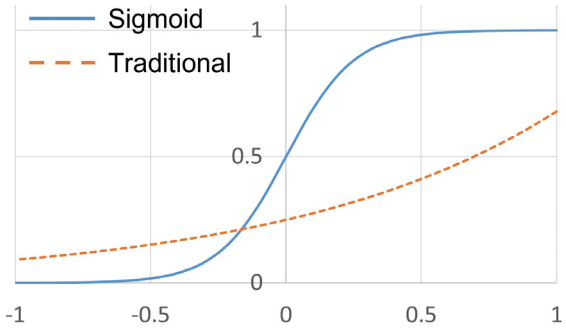


Fig. 5. Sigmoid function

with the intuitive importance assigned by humans. Additionally, because it is differentiable, it has a minimal impact on the aforementioned relationships. Meanwhile, compared to sigmoid attention, traditional attention (dashed line in Fig. 5) results in a transformation that is less accentuated due to its less radical nature, leading to low visibility even when visualized.

Figure 4 (b) presents an example of sigmoid attention. Unlike the attention mechanism alone, sigmoid attention makes it clear that the AI system focused the entire Chihuahua and indicates that the neck area received unexpected importance, which ideally should not attract attention.

4. Experiments

4-1 Training data

Figure 6 presents examples of wire harnesses used for the experiments. These images are categorized depending on whether or not the black corrugated tube is wrapped in black vinyl tape and how the tape is wrapped. The images are divided into three classes: the tube is not wrapped in vinyl tape (“not_tape”); the tube is wrapped in vinyl tape without gaps (“tape”); and the tube is wrapped in vinyl tape, but some sections of the corrugated tube are exposed (“mix”).

The experiment used 240 images of each class, totaling 720, as training data. As described in 2-2, these data had no annotations regarding the tape wrapping classes, and the AI system was trained using self-supervised learning. To verify the performance of the AI system, test data consisting of 7,800 images (2,600 for each class) were prepared.

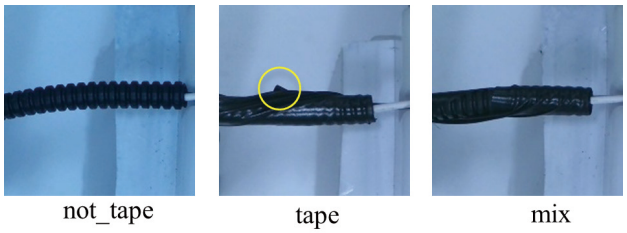


Fig. 6. Classes of wire harness tape wrapping

4-2 Experiment guidelines

First, the AI system was trained by means of self-supervised learning using the wire harness images shown in Fig. 6. Sigmoid attention was not introduced at this point, so the traditional self-supervised learning is set as the benchmark for this paper. In the following sections, this benchmark AI will be termed “regular learning” and distinguished from the proposed method, which uses sigmoid attention, which is described later.

Next, from the images incorrectly judged as a result of regular learning and from those that were correctly judged but could potentially negatively impact AI’s decisions due to strong focus on foreign matter visible in the background, one image from each category was selected from the training data. Using the training data consisting of these two images, we first introduced sigmoid attention and then had the AI system imitate and learn the areas of focus created based on human expertise.

4-3 Results of regular learning

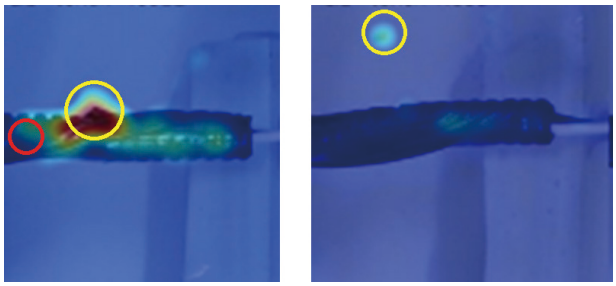
After 10 sessions of regular learning, the tape wrapping classification accuracy was verified using the test data. The correct answer rate was 83.57% on average, with the maximum and minimum being 84.26% and 83.06%, respectively. Table 1 presents a confusion matrix, which records the numbers of correct and incorrect answers at the maximum accuracy.

Table 1 reveals that no incorrect answers were produced for “not_tape.” This indicates that the AI system learned distinguishable features through self-supervised learning to separate “not_tape,” which had no black tape wrapped around, from the other classes, which had a black tape wrapped around either partially or entirely. Meanwhile, “mix,” which had a black tape partially wrapped around, was erroneously recognized as “tape” at a rate of 45%.

Figure 7 (a) illustrates attention during regular learning for training data 1, which was incorrectly classified. The figure reveals that, with training data 1, there was

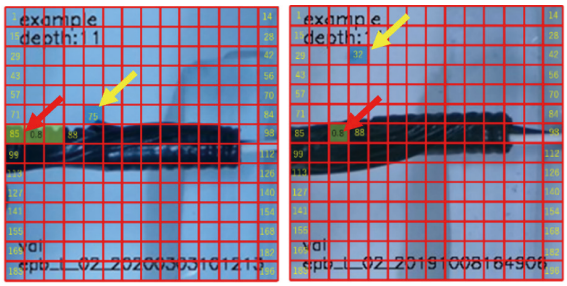
Table 1. Confusion matrix of regular learning at the maximum accuracy

	Ground Truth				Total
Prediction	Class	mix	not_tape	tape	
	mix	1435	0	63	1498
	not_tape	0	2600	0	2600
	tape	1165	0	2537	3702
Total		2600	2600	2600	7800



(a) Training data 1 (b) Training data 2

Fig. 7. Attention in regular learning



(a) Training data 1 (b) Training data 2

Fig. 8. Instruction method for sigmoid attention

an exaggerated focus on the black tape; that is, the focus was too strongly placed on the rising part of the vinyl tape (yellow circle), while the focus on the exposed part of the corrugated tube indicated by the red circle was too weak. This attention is deemed inappropriate because humans at an actual production site would pay attention to the exposure of the corrugated part when differentiating between "tape" and "mix." Attention regarding training data 2 [Fig. 7 (b)], which contained foreign matter in the background, produced adverse effects: the focus on the originally intended target (wire harness) became weak due to a strong focus on the yellow circle indicating the foreign matter.

4-4 Additional learning by the proposed method

Additional learning was conducted using sigmoid attention—the proposed method—to incorporate improvements derived from human expertise, as pointed out in 4-3, into the AI system. Figure 8 shows a method for incorporating human expertise. The proposed sigmoid attention technique utilizes an implementation designed to encourage additional training of the AI system by selecting patches intended to incorporate human expertise from images divided into patches of 14 × 14.

On the one hand, the AI system was instructed to set a target value of 0.1 for the patch indicated by a yellow arrow in training data 1 in order to weaken the focus on the protrusions of the tape; on the other hand, it was instructed to aim for 0.8 for exposure of the corrugation indicated by a red arrow, based on human expertise that emphasizes the importance of focusing on the corrugation. Similarly, with training data 2, a target value of 0.1 was assigned to the area indicated by a yellow arrow where foreign matter was found, while 0.8 was assigned to the red arrow indicating an exposed part of the corrugation. Other areas used the AI system's outputs as they were, without any human instructions.

Using the above-described training conditions, the AI system conducted 10 sessions of additional learning leveraging the proposed method and produced the following results: classification accuracy regarding the test images averaged 84.49%, with the maximum and minimum values being 87.03% and 82.23%, respectively (Table 2). Using the proposed method, the average and maximum figures improved by approximately 0.92% and 2.8%, respectively, compared to regular learning, although the minimum accuracy of regular learning was higher than that of the proposed method. Despite the small amount of data used for the additional learning—just two images—substantial improvements were achieved, resulting in an average increase of 70 images or more. Table 3 presents a confusion matrix corresponding to the maximum accuracy achieved by the proposed method. Compared to Table 1, which illustrates regular learning, erroneous recognition of "tape" increased, while significant improvements were made regarding "mix" after introducing additional learning, resulting in overall improved accuracy.

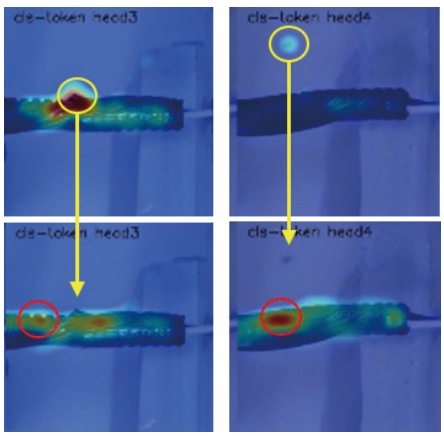
Figure 9 shows the changes in attention as a result of shifting from regular learning (top) to the proposed method

Table 2. Classification accuracy with test data used for different techniques (%)

	Min.	Max.	Ave.
Regular	83.06	84.62	83.57
Proposed	82.23	87.03	84.49

Table 3. Confusion matrix of proposed method at maximum accuracy

		Ground Truth			Total
	Class	mix	not_tape	tape	
Prediction	mix	1749	0	161	1910
	not_tape	0	2600	0	2600
	tape	851	0	2439	3290
Total		2600	2600	2600	7800



(a) Training data 1 (b) Training data 2

Fig. 9. Shifting of attention from regular learning (top) to proposed method (bottom)

(bottom). Regarding the tape projection area and the location of foreign matter (indicated by the yellow circle), for which settings were made to weaken the focus, the figure shows that the degree of focus decreased as instructed after the use of the proposed method. In contrast, the focus on the relevant locations became stronger for the areas used in additional learning that incorporated human expertise, which emphasizes the way of taping while focusing on the exposed parts of the corrugation (indicated by the red circles). Consequently, it was demonstrated that the AI system acquired visual explainability aligned with human expertise.

5. Conclusion

This paper proposed the introduction of sigmoid attention, which enables flexible additional learning for attention mechanisms used for self-supervised learning. At the same time, this paper verified classification accuracy after implementing additional learning incorporating human expertise, as well as the visual explainability of the attention mechanism. After additional learning using sigmoid attention, the classification accuracy averaged 84.49%, with the maximum being 87.03%. The proposed method showed improvements compared to regular learning, with the average and maximum increasing by approximately 0.92% and 2.8%, respectively. Meanwhile, regarding the visual explainability of attention, the focus was strong on visible foreign matter and exceptionally protruding parts of the tape at the point of regular learning, while it was weak on areas deemed important by human perception (exposed parts of the corrugation). However, after additional learning, which incorporated human expertise into the AI system, the focus became weaker on the aforementioned parts and was guided to the exposed parts of the corrugation, enabling the AI system to acquire attention closer to human expertise. The proposed method offers significant advantages by improving both the classification accuracy and visual explainability of the attention mechanism, while keeping introduction costs low.

References

(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 New Approach 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)

Contributor

Y. MATSUMOTO
• Assistant Manager, Digital Transformation Laboratory