# Explaining object detectors: the case of transformer architectures

Baptiste Abeloos, Stéphane Herbin

# Explaining object detectors: the case of transformer architectures

Baptiste Abeloos and Stéphane Herbin

*ONERA/DTIS, Université Paris-Saclay*, F-91123 Palaiseau, France.
firstname.lastname@onera.fr

**Abstract.** Object detection is a complex visual function that has applications in many safety-critical domains such as autonomous driving or medical diagnosis. In this paper we examine how its behavior can be explained. More precisely, we discuss and analyze the specificity of object detection with respect to explainability, and describe an approach for explaining object location prediction from a popular and efficient attentional-based deep neural network architecture: DETR.

**Keywords:** Explainability · Object detection · Attentional models · Object localization · DetR.

## 1 Introduction

Developing the explainability of artificial intelligence (XAI) is now commonly seen as a way to improve the reliability of its usage. This goal is clearly motivated by ethical dimensions when people may be affected by the results of an AI-based system, but it can also be useful to other stakeholders such as engineers or authorities. In a certain way, an explanation can be considered as an argument provided by the predictive process that can be used to justify its output, the level of persuasiveness of the argument depending on the nature of its recipient.

Not all explanations have the same expressiveness, either because they do not show the same discriminating capacity, and also because they do not assume the same knowledge level of its recipient. Indeed, when addressing explainability issues, a series of questions have to be clarified: what is explained? – i.e. what feature or phenomenon the explanation is referring to – for whom? – end-user affected by the result, data scientist, auditor, authority, management, etc. – and for what purpose? – i.e. how will the explanation be used.

In this article we address the problem of explaining the predictions made by visual functions with a focus on object detection. Due to their high dimensional input, the decision structure of object detectors, and their output, are inherently complex and would benefit from explainability tools able to reveal how they operate.

A detection function has multiple objectives: it must reveal the existence of objects of interest – implying that not all potential objects are – estimate their location or pose and predict their category. The overall satisfaction of these three

objectives for a wide variety of input data is difficult and must take into account various trade-offs, making object detection a still unsolved problem with a large literature [27]. Explaining such complex processing chains appears to be very challenging, and requires new tools. Targeting the explainability of visual object detection is also an opportunity to question the way in which the explainability of image categorization, the main visual function studied in the literature, is currently carried out.

What we expect from an explanation is to give insight or justification of good behavior, but also to anticipate bad behavior. Its typical target audience is computer or data scientists. This means that we expect from the explanation recipient a rather good level of technical knowledge: he/she knows what are the basic components of a computer vision chain and how they function.

Attentional models such as the Transformer architecture [43] are now well established Deep Neural Network (DNN) architectures. They bring state of the art results in Natural Language Processing, but also more recently in many computer vision applications [24]. Regarding explainability, they are especially interesting as they are expected to provide *by design* attention weights that modulate inputs and other variables in the architecture. They can therefore potentially be used as feature importance measures. We will focus on one now well known attentional based architecture for object detection: the DEtection TRansformer (DETR) [6] and study how its attentional features can explain detection outputs.

In this article, we therefore address the design of explanations for object detectors that relies on an elementary knowledge of computer vision with the objective of understanding its behavior. The focus will be on explaining detection *locations*, a feature that has not yet been studied in the current literature, and on using attentional models – the actual trend of efficient computer vision models.

Our contributions are the following:

- we analyze the specificity of object detection with respect to explainability,
- show how modern attentional architectures are able to provide explainability by design for object detection,
- develop a simple model explaining object location prediction, and evaluate the fidelity of this model on the COCO dataset

## 2 Related work

### Object detection

Numerous DNN architectures have been proposed for object detection in the literature. They are often categorized as one stage or two stage architectures. YOLO [38] and SSD [28] are among the most popular for the former, Faster-RCNN [39] and mask-RCNN [17] for the latter. One stage architectures often produce outputs for a dense sub-sampling of possible object locations and assign

for each location a refined object extension description – a bounding box (BB) or a mask – objectness and category scores. Two-stage architectures first compute a population of possible object bounding boxes, then compute objectness and category scores to characterize each bounding box.

A more recent architecture, the DEtection TRansformer (DETR) [6], which can be considered as a one-stage approach, exploits a transformer-based [43] encoder-decoder principle to generate the full distribution of categorized bounding boxes. The main difference with previous one/two stage approaches is to view visual object detection as a direct set prediction problem – each element of the set being a categorized bounding box – where global interactions between objects can be exploited. This architecture gives state of the art performance on the COCO dataset. Note that, to accelerate the convergence, a modified architecture called Deformable DETR [50] applies the attention mechanism on a smaller number of target keys when computing attentional weights. One major impact of this modification is to decrease the training time from 2000 to 325 GPU hours on the COCO dataset.

### Attention-based XAI attention for detection

XAI applied to visual functions has mainly addressed image categorization, i.e. the answer of a *What/who?* questions, and mostly expressed in the form of a heat-map – improperly called attentional or saliency map – that can be superimposed on the input image to reveal its influential pixels. Object detection involves two other types of questions: *Is there?* and *Where?*.

The research field of explainability is rather recent and has not yet fixed its vocabulary and fundamental concepts in a unanimous way. Many concepts, methods and objectives under the expression explainability or interpretability are still debated [26, 11, 33], not to mention their philosophical foundations [32, section 2].

Explanations of a given prediction are often divided in two families: *post-hoc*, meaning that they require extra computation given a predictor state or *by design*, meaning that they can be provided at no cost from a *transparent* predictor [5, 15].

Explanation provided by attentional models are somehow between post-hoc and by-design: they rely on features that are natively designed to weigh several latent variables in the architecture – in that sense they transparently assign importance to those variables – but the way these variables are combined to build the final decision is complex and may not be readily interpretable.

Attentional latent representations, usually in the form of a 2D map that condition prediction, have been used for several visual functions: captioning [46], visual question answering [3, 47] or motion prediction [25]. The internal attentional 2D map can then be used as a feature attribution explanation.

Another motivation for exploiting attentional architecture, besides its pseudo-transparency objective, is the low computation capacity needed to provide an explanation: indeed, popular post-hoc methods that require data generation such

as SHAP [30], LIME [40] or RISE [36], for instance, are costly, especially for images.

The explanation of image-based detectors has attracted fewer studies. D-RISE [37] belongs to the family of post-hoc black-box perturbation-based feature attribution and uses input masking to generate saliency maps accounting for localization and objectness in addition to categorization. A score based on computing a similarity between two output detections encoded as a triple (objectness score, location, category probability distribution) is optimized to build each explanation. The computation of the explanation is very intensive (70s per image for the "fast" YOLOv3) and cannot target real-time monitoring applications such as autonomous driving [48]. Other works exploit LRP [4] or SHAP [30] approaches to explain the category of each detected object [42, 23, 14]

To the best of our knowledge, the only work that targets object detector exploiting attentional architectures is [8]. It extends prior work on computing attribution maps from attentional weights [9]. Basically, the idea of the proposed approach is to generalize the cumulative principle of LRP [4] to the attentional weights only.

## 3 Explaining detection

### 3.1 Structure of a detector

The output of a an object detector can be informally represented by statements like "there is a car on the left" or "there is a cat in the bounding box [15,15,40,80]". These statements imply that object detection expresses three different decisions:

- **Existence**, which is a binary decision answering to *Is there?*
- **Location**, which is a multivariate regression answering to *Where?*
- **Category**, which is a classification function answering to *What?*

Compared to data classification, the output of a detector is much complex and can be seen as a multi-task process: each task can be assessed individually, but also using a fused metric like what is done in academic benchmarks [1].

This multi-task objective is also revealed by the way the final decision process is generally performed. In a one stage detector like Yolo [38], for instance, the detection score characterizing each potential candidate is computed from scores characterizing objectness, category and location, usually followed by a non-maximal suppression step. The main differences between the various DNN architectures implementing a detector lie in what can be shared between these three tasks, what is specific to each, how they depend on each other and how they represent the three types of required output: existence, location and category.

[1] The Mean Average Precision computed for the COCO dataset `https://cocodataset.org` averages the area under the precision/recall curve for various localization precisions and classes.

### 3.2 What can be explained?

Explaining detection is difficult because the mechanisms used to compute the predictions are complex and do not lend themselves easily to intuitive interpretation, not only because they involve deep networks, but also because they exploit complex interactions between substructures whose role is often unclear.

The complexity of detection processes questions the use of feature attribution, the most popular type of explanation principle. When it comes to images, feature attribution is often expressed as saliency maps that can be super-imposed on the input image in order to reveal its influential pixels. What is the meaning of such saliency maps regarding detection is unclear: what do they give explanation of?

Indeed, using attention as an explanation is still controversial: it has been found that saliency heat-maps do not correlate well with feature importance either for NLP [21, 45] or vision [1, 2, 49] applications.

One proposed alternative to the causal attribution represented by feature importance representations is to consider explanation in the form of counterfactual or contrastive examples. Why P rather than Q? where Q is often implicit in the context [32, section 2]. Explainability from counterfactual examples follow this strategy [44].

Another explainability strategy is to more explicitly consider that "explanations enable a human to complete its mental models for understanding a phenomenon/system. *Understanding* means that the human has the capability to do inferences (deduction, induction, abduction...) on the behaviour of the system without relying on experiments." [10]

A mental model is "a convenient term for a constellation of well-developed knowledge about a system. A good mental model will help a user interpret, predict, and mentally simulate the operation of a system, as well as to understand the system's limits and boundary conditions." [34, p.86]. As states [16], "Following [13], we take the notion of explanation to be relative to an agent's epistemic state". Eliciting a user's mental model of an AI system is also suggested as a prerequisite to evaluate the quality of an explanation [20, section 3]. The idea of a mental model for explainability in the context of medical diagnosis has been introduced in [31] but with the idea that an explanation "is an expression or reformulation of the model in a different medium that can be shared with others" and proposing that the AI prediction algorithm must be considered as an agent interacting with humans.

We therefore propose to define an explanation of an artificial predictive function as the input of a virtual *process* mimicking the function, hosted by the explanation recipient – a human equipped with a mental model. Understandability is granted when the virtual process infers from the explanation an output that is coherent, or believed to be so, with the actual prediction. This property is sometimes called *fidelity* in the XAI literature [33, 7, 15]. The virtual process to which the explanation refers is usually implicit; it can however be made more explicit, or even formal, when the recipient has a technical background. Typically, to avoid ambiguity, this virtual process can take the form of a simplified algorithm, i.e. a sequence of calculations. A good definition of an explanation

should therefore be supported by the description of a virtual predictive process that gives its meaning to the explanation.

### 3.3 Explaining detection outputs

Following the definition of an explanation we provided in the previous section, we examine here which elements or features enable the prediction of the three outputs and for which virtual process or algorithm.

**Explaining existence** In computer or neuro-psychological vision [19], deciding or perceiving that an object exists, is visible, is related to the idea of saliency: the object is a *foreground*, a shape, that stands out from the background. The construction of a shape is ruled by low-level vision properties such as contour continuation, contrast of textures, multi-scale diffusion, etc. and by top-down priors that may condition the shape construction process. A good explanation of the object existence decision should therefore reveal from what features the shape is built, and what trade-offs have solved competing constraints. Given the complexity of shape detection in natural images, explaining why and how the detector has decided that there is an object in the image rather than nothing is very ambitious, and we only mention it as a future objective.

**Explaining location** Localizing an object once its existence is asserted can be easier. The explanation can take the form of geometric *landmarks* associated with the object such as edges, or key-points from which, for instance, a bounding box can be computed. Landmarks that are causally exploited to predict the object location can be considered as an explanation; the virtual mental process that gives its meaning to the explanation can be a simple smallest bounding box or centroid computation from such landmarks. We will analyze the ability of an attention-based detector to provide landmark-based object location explanations in a later section.

**Explaining category** The last output provided by an object detector is its category. Psychological investigations roughly propose two different ways to define a category [41]: as a definition or as a prototype. In the former case, categorization relies on attribute checking, in the latter it relies on measuring similarity. Note that this classical distinction has analogy with the difference between intuitive and deliberate thought processes [22], i.e. the System 1 vs. System 2 cognitive processes, now also popular in AI research. An explanation for category prediction therefore depends on the type of categorization used as a virtual mental process: from prototypes and similarity, or from discriminant attributes and checking. The question now is to find or compute them from the detection process. Unfortunately, the most popular type of explanation found in the literature – pixel attribution heat-maps – do not provide any of them.

# 4 Attentional models for detection

## 4.1 Transformers

In 2017, Vaswani et al. introduced the Transformer[43], an encoder/decoder architecture initially applied to NLP and exploiting *attention* mechanisms. The architecture consists of stacked self-attention modules for both the encoder and the decoder. An attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The self-attention module is defined by:

$$\mathbf{A} = \texttt{softmax}\left(\frac{\mathbf{Q}.\mathbf{K}^T}{\sqrt{d_h}}\right) \tag{1}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are respectively the query, key, and value, and $\mathbf{A}$ contains the soft attention weight for each value. The attention coefficients $\mathbf{A}$ then softly select the values in $\mathbf{V}$ using a matrix product $\mathbf{O} = \mathbf{A} \cdot \mathbf{V}$ to build the *transformed* representation $\mathbf{O}$ for each query $\mathbf{Q}$. In practice, instead of performing a single attention function in each module, the architecture uses multi-head self-attention modules in which several attention functions are performed in parallel.

## 4.2 Object detection

Transformers have recently been adapted to computer vision tasks. In 2020, a transformer encoder/decoder architecture developed for object detection called DETR [6] was introduced. It obtained performance similar to the well-established Faster R-CNN baseline on the COCO dataset. It allows reasoning about the relations between objects and global image context. It also forces unique prediction, removing the need for many hand-designed components in object detection such as non-maximal suppression.

An illustration of the architecture is presented in Fig. 1. DETR uses a CNN backbone to learn a set of compact feature representations, an encoder/decoder transformer, and a final layer composed of a shared feed forward network (FFN) to predict simultaneously the object category and its location in the form of a bounding box. Existence is encoded as an extra "no object" category.

Visual transformers, like SWIN [29], have also been proposed as an alternate backbone for feature extraction that can be fed to object detection architectures. We instead focus in this article on the role of attention when an encoder/decoder architecture is used.

## 4.3 Explanation with transformers

Attentional models contain, *by design*, inner representations that are devoted to filter out spurious features or select informative features through attentional vectors or masks. It is thus tempting to exploit such representations as an explanation of what is useful for the detection task.

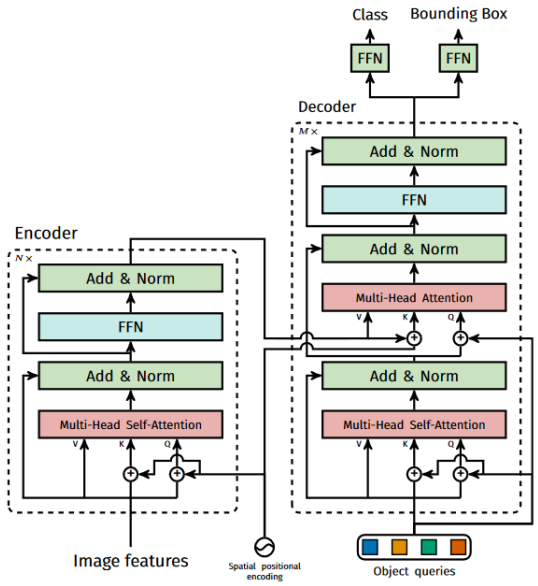


**Fig. 1.** Detailed transformer architecture of DETR.

Chefer et al. in two recent papers [9, 8] have proposed a way to capture and accumulate the information contained in the attentional masks as an explanation, and have applied it to object classification or detection and to visual question answering. Basically, for each attentional layer $\mathbf{A}$ in the architecture, a local explanation heat-map $\mathbf{A}_o$ is computed as the positive value of the element-wise product $\odot$ between the attentional map and the gradient of an output variable $o$ with respect to the attentional coefficients:

$$\mathbf{A}_o = [\mathbf{A} \odot \partial o / \partial \mathbf{A}]^+ \tag{2}$$

The local explanation heatmaps are iteratively accumulated and normalized to build the global explanation heat-map $\mathbf{R}_o$, called *relevancy* map, using two different schemes whether the local attentional layer is self-attentional – keys and queries in 1 are identical – or co-attentional – keys and queries come from two different modalities. See the original paper [8] for a more detailed description.

The final relevancy heat-map – the *explanans* – hosts the explainability capacity of a given target output $o$ – the *explanandum* [18] for the object detector [6]. In [8], it is evaluated as a pixel-wise object detector proxy that produces a segmentation mask for the object, using the predicted class as explanandum and thresholding the relevancy map by Otsu's method [35]. The quality of the explanation is measured by computing an intersection over union (IoU) with the ground-truth.

This way of evaluating the explanatory capacity of the relevancy heat-map is questionable: the heat-map is computed using information from the predicted

category using the gradients in eq. 2, whereas its evaluation calculates a geometrical criterion: the intersection over union (IoU). Furthermore, using the proposed evaluation scheme, it is difficult to predict if the explanation is bad because it does not catch the detector behavior (fidelity) or because the detector itself has not been able to generate a good prediction (accuracy). In the following, we focus on studying more precisely the explanation of location prediction.

### 4.4 Attention for explaining location

High activation locations in the attentional maps can be interpreted as geometrical landmarks that encode object bounding box: at least this is the hypothesis that will be verified in the experiments. We use attention as location attribution, i.e. we only keep from the attribution its geometrical reference and not the virtual set of local features that can be used to infer category as it is usually implicitly assumed when speaking of feature attribution.

First, we need to build the explanation of the object location using the outputs that actually represent it. The idea is to compute the relevance maps with respect to the BB output, and to consider the heat map as a landmark generator from which to estimate the BB using a virtual process. Let $o$ be one of the outputs of the FFN encoding the bounding box coordinates $[x, y, w, h]$ in DETR (see Fig. 1). We generate the final explanation for location $\mathbf{R}_{\text{loc}}$ by summing the relevancy heat-maps for all outputs:

$$\mathbf{R}_{\text{loc}} = \sum_{o=\{x,y,w,h\}} \mathbf{R}_o \tag{3}$$

Note that other fusion schemes to gather the contributions of the output BBox coordinates are possible, for instance by summing the local attentional maps $\mathbf{A}_o$ before propagating their values in the relevancy. We leave this study for further work.

The final explanation $\mathbf{R}_{\text{loc}}$ can be evaluated by fitting a bounding box on the heat-map maxima or on a thresholded version as a proxy: this bounding box fitting plays the role of the virtual mental process for which the explanation is meaningful. In other words, the relevancy heat-map is interpreted as the information that is useful to estimate object location.

## 5 Experiments and preliminary results

### 5.1 Building the explanation

To investigate the case of explanation for object detection, we used a DETR model trained on the 2017 COCO dataset and publicly available [2] [12]. Examples are shown in Fig. 2, with detected objects and their associated class activation map obtained with the method proposed by Chefer et al. [8].

[2] https://dl.fbaipublicfiles.com/detr/detr-r50-e632da11.pth

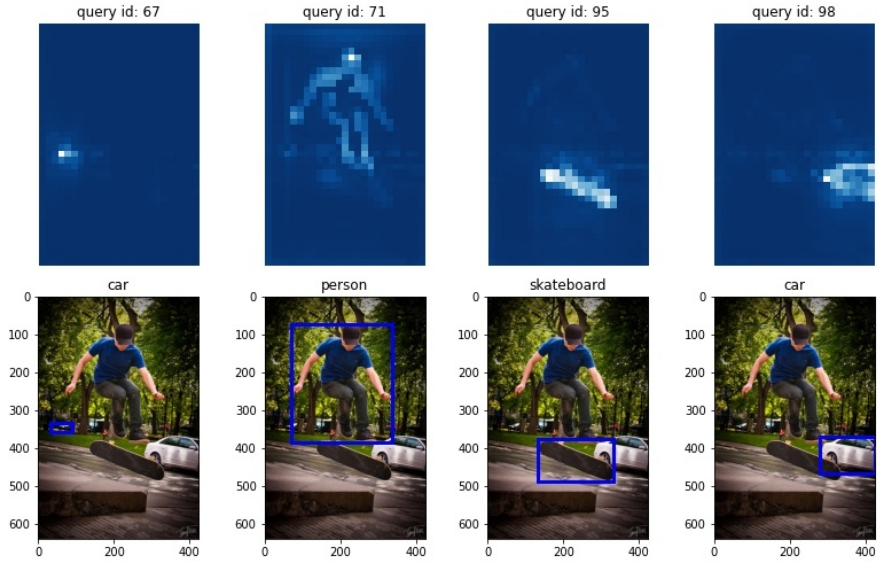


**Fig. 2.** Example of object detection and localization (bottom), with the associated activation map that is expected to explain classification (top). The detection threshold that selects the detections from the class prediction output value is set to 0.9.

While various methods have already been developed to compute the activation map for image classification, the transformer architecture allows to study the case of location prediction. To the best of our knowledge, this is the first work targeting this subject. The localization of the object is obtained by predicting the four coordinates of a bounding box: x center, y center, width and height. Examples are shown in Fig. 3 with the coordinate activation maps of query 71. As expected, these activations differ from the class activation maps shown in Fig. 2, showing external parts or spatial limits of the object that are more peaked.

The explanation method is based on several assumptions. First, we assume that the spatial limits of the detected object are the most activated regions when predicting the bounding box coordinates. Second, we assume that the location prediction corresponds to the minimal box that include these regions, meaning that the identification of the most activated regions is sufficient to predict the bounding box. From those assumptions, the *virtual mental process* that gives its meaning to the explanation can be divided in two steps: 1/ select landmarks by thresholding the activation map values and 2/ fit a bounding box that contains those landmarks.

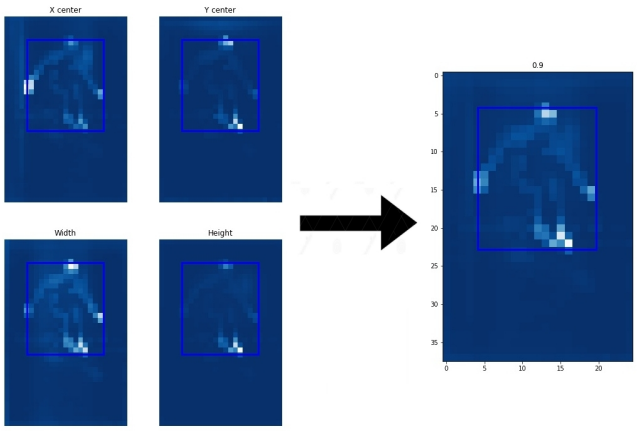


**Fig. 3.** Example of the bounding box coordinate activation maps: x center, y center, width, and height. Activations are summed-up to form a single bounding box activation map (eq. 3).

### 5.2 Evaluating the explanation

Evaluating the quality of an explanation should involve a human recipient that would attribute a meaning to it from a virtual process. Here, we propose to simulate this process as a simple bounding box fitting algorithm in order to provide a statistical evaluation of the explanation fidelity over a large database.

The activation map associated to each bounding box coordinates are assumed to be of equal importance and are summed-up to form a single location activation map. The selection of the most activated regions can be seen as a landmarks used to define the detected object bounding box and can be obtained by binary segmentation between foreground and background. We used the Otsu's method[35] to select the signal threshold, as is also proposed in [8]. Illustration of the sum of the bounding box coordinate activation maps is shown in Fig. 4 (top), while the activations above the threshold are shown at the bottom. Blue boxes correspond to the model predictions, while red boxes correspond to the model explanation. The fidelity of the explanation method is computed using the IoU between model prediction and explanation, and is shown on top of the figure.

To evaluate the fidelity of the explanation method, we applied the method on the 2017 COCO dataset, varying the computation of the activation threshold and the object detection threshold. Average IoU's are presented in Table 1. Two methods have been compared to select the activation threshold: a basic threshold used as a baseline, defined to select 5%, 1%, and 0.5% of the most activated pixels, and a binary threshold based on the Otsu method. The method providing the highest fidelity is obtained with the Otsu method, giving an average IoU of 0.52 for a detection score $> 0.9$. A selection of 5% of the activations can be seen as a too

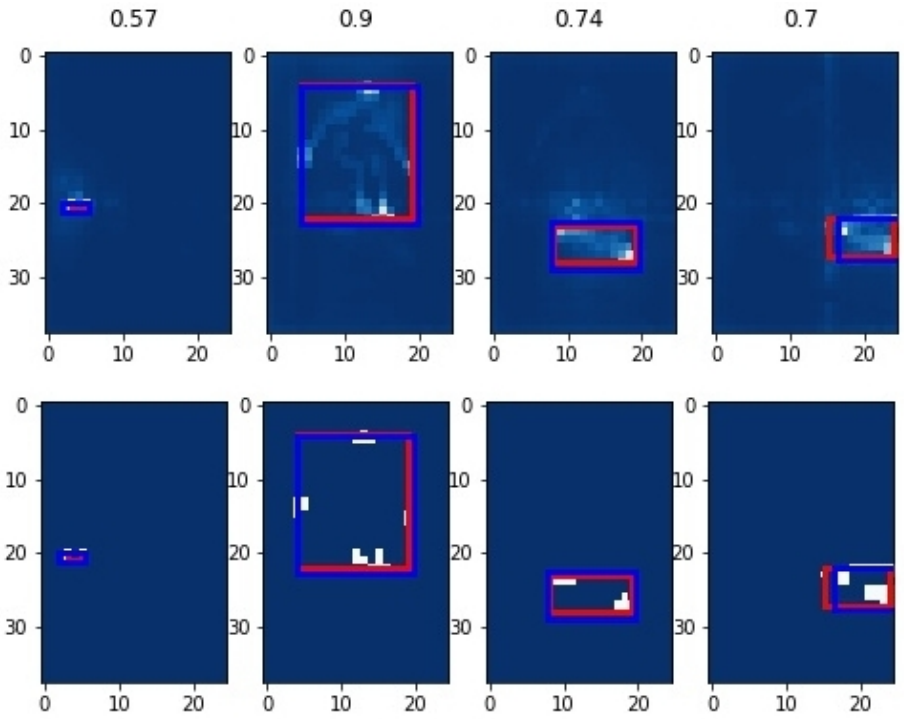


**Fig. 4.** Sum of the activation maps of the location coordinates (top), and selection of the activations with the Otsu threshold (bottom). The blue bounding box indicates the model prediction, while the red box corresponds to the associated explanation. The IoU between prediction and explanation is shown at the top.

low threshold, resulting in too many activations and a too large bounding box explanations, while a selection of 0.5% only can be seen as too tight. A selection of 1% is a good trade-off and provides similar results compared with the Otsu method. Notice that the Fidelity tends to increase with the detection threshold, meaning that a better confidence in prediction also implies a better confidence in explanation. Our interpretation is that a lower detection threshold allows the detection of smaller objects that are more difficult to detect and explain. By decreasing the detection threshold to 0.6, we have more detected objects as illustrated in Fig. 5. We can see that the Fidelity of smaller objects tend to be lower, mainly due to the uncertainty brought by pixel quantization. To evaluate the quality of the explanation method, a comparison with the ground truth is needed, and is let as a future work.

| **Fidelity (IoU)** | | | |
|---|---|---|---|
| Activation threshold | Detection threshold | | |
| | 0.6 | 0.8 | 0.9 |
| Basic - 5% | 0.29 | 0.33 | 0.36 |
| Basic - 1% | 0.46 | 0.48 | 0.50 |
| Basic - 0.5% | 0.35 | 0.34 | 0.34 |
| **Otsu** | **0.46** | **0.49** | **0.52** |

**Table 1.** Comparison of several binary threshold methods. Basic method is a simple binary threshold computed to select 5%-1%-0.5% of the image pixels, while Otsu's method uses the variance to discriminate foreground from background. Fidelity is computed using the intersection over union (IoU) between model prediction and explanation.

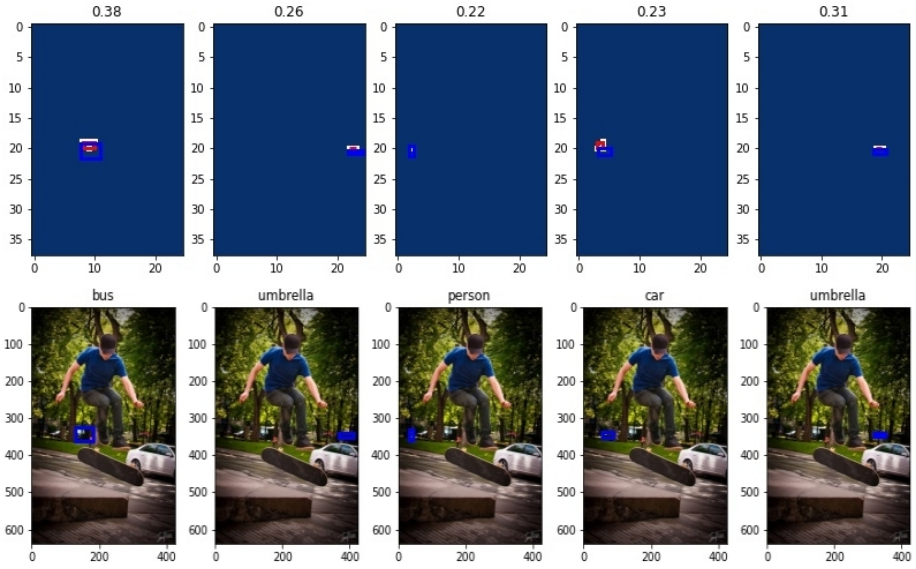


**Fig. 5.** Additional detected objects (bottom) with localization explanation and Fidelity (top) by decreasing the detection threshold from 0.9 to 0.6.

## 6 Conclusion

In this work, we have studied the explainability of object detection considered as a multi-task function with three outputs: existence, location and category. We have focused on explaining object location prediction and left the existence and category prediction for further investigation. To the best of your knowledge, this is the first work targeting explainability of object location prediction. We have shown on few examples how to provide explanations using weighted attentional relevance scores for the attentional based DETR detector, and evaluated the fidelity of our model on the 2017 COCO dataset. Additional work is needed to evaluate other explanations properties that involve comparison to the ground truth (accuracy).

# Bibliography


[1] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. pp. 9525–9536. NIPS'18, Curran Associates Inc., Montréal, Canada (Dec 2018)

[2] Akula, A.R., Zhu, S.C.: Attention cannot be an Explanation (Jan 2022)

[3] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6077–6086 (2018)

[4] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLOS ONE 10(7), e0130140 (Jul 2015)

[5] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58, 82–115 (Jun 2020)

[6] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision. pp. 213–229. Springer (2020)

[7] Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics 8(8), 832 (2019)

[8] Chefer, H., Gur, S., Wolf, L.: Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers. arXiv:2103.15679 [cs] (Mar 2021)

[9] Chefer, H., Gur, S., Wolf, L.: Transformer Interpretability Beyond Attention Visualization (Apr 2021)

[10] Delseny, H., Gabreau, C., Gauffriau, A., Beaudouin, B., Ponsolle, L., Alecu, L., Bonnin, H., Beltran, B., Duchel, D., Ginestet, J.B., Hervieu, A., Martinez, G., Pasquet, S., Delmas, K., Pagetti, C., Gabriel, J.M., Chapdelaine, C., Picard, S., Damour, M., Cappi, C., Gardès, L., De Grancey, F., Jenn, E., Lefevre, B., Flandin, G., Gerchinovitz, S., Mamalet, F., Albore, A.: White Paper Machine Learning in Certified Systems. arXiv:2103.10529 [cs] (Mar 2021)

[11] Doshi-Velez, F., Kim, B.: Towards A Rigorous Science of Interpretable Machine Learning (Mar 2017)

[12] FacebookResearch: DETR: End-to-End Object Detection with Transformers, Github. https://github.com/facebookresearch/detr (2017)

[13] Gärdenfors, P.: Knowledge in Flux: Modeling the Dynamics of Epistemic States. The MIT press (1988)

[14] Gudovskiy, D., Hodgkinson, A., Yamaguchi, T., Ishii, Y., Tsukizawa, S.: Explain to Fix: A Framework to Interpret and Correct DNN Object Detector Predictions (Nov 2018)
[15] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) 51(5), 93 (2018)
[16] Halpern, J.Y., Pearl, J.: Causes and explanations: A structural-model approach. Part II: Explanations. The British journal for the philosophy of science 56(4), 889–911 (2005)
[17] He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969 (2017)
[18] Hempel, C.G., Oppenheim, P.: Studies in the Logic of Explanation. Philosophy of Science 15(2), 135–175 (Apr 1948)
[19] Hildreth, E.C., Ullman, S.: The computational study of vision. In: Posner, M.I. (ed.) Foundations of Cognitive Science, pp. 581–630. MIT Press (1988)
[20] Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for Explainable AI: Challenges and Prospects (Feb 2019)
[21] Jain, S., Wallace, B.C.: Attention is not Explanation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 3543–3556 (2019)
[22] Kahneman, D.: A perspective on judgment and choice: Mapping bounded rationality. American psychologist 58(9), 697 (2003)
[23] Kawauchi, H., Fuse, T.: SHAP-Based Interpretable Object Detection Method for Satellite Imagery. Remote Sensing 14(9), 1970 (Jan 2022)
[24] Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in Vision: A Survey. arXiv:2101.01169 [cs] (Oct 2021)
[25] Kim, J., Bansal, M.: Attentional Bottleneck: Towards an Interpretable Deep Driving Network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 322–323 (2020)
[26] Lipton, Z.C.: The mythos of model interpretability. arXiv preprint arXiv:1606.03490 (2016)
[27] Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M.: Deep Learning for Generic Object Detection: A Survey. International Journal of Computer Vision 128(2), 261–318 (Feb 2020)
[28] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European Conference on Computer Vision. pp. 21–37. Springer (2016)
[29] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
[30] Lundberg, S.M., Lee, S.I.: A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)

[31] Merry, M., Riddle, P., Warren, J.: A mental models approach for defining explainable artificial intelligence. BMC Medical Informatics and Decision Making 21(1), 344 (Dec 2021)
[32] Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence 267, 1–38 (2019)
[33] Molnar, C.: Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. Second edn. (2022)
[34] Mueller, S.T., Hoffman, R.R., Clancey, W., Emrey, A., Klein, G.: Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI. arXiv:1902.01876 [cs] (Feb 2019)
[35] Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics 9(1), 62–66 (1979)
[36] Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized Input Sampling for Explanation of Black-box Models. In: British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018 (2018)
[37] Petsiuk, V., Jain, R., Manjunatha, V., Morariu, V.I., Mehra, A., Ordonez, V., Saenko, K.: Black-Box Explanation of Object Detectors via Saliency Maps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11443–11452 (2021)
[38] Redmon, J., Farhadi, A.: YOLOv3: An Incremental Improvement (Apr 2018)
[39] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. pp. 91–99 (2015)
[40] Ribeiro, M.T., Singh, S., Guestrin, C.: 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. ACM (2016)
[41] Smith, E.E., Medin, D.L.: Categories and concepts. In: Categories and Concepts. Harvard University Press (2013)
[42] Tsunakawa, H., Kameya, Y., Lee, H., Shinya, Y., Mitsumoto, N.: Contrastive Relevance Propagation for Interpreting Predictions by a Single-Shot Object Detector. In: 2019 International Joint Conference on Neural Networks (IJCNN). pp. 1–9 (Jul 2019)
[43] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, \., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017)
[44] Verma, S., Dickerson, J., Hines, K.: Counterfactual Explanations for Machine Learning: A Review (Oct 2020)
[45] Wiegreffe, S., Pinter, Y.: Attention is not not Explanation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 11–20. Association for Computational Linguistics, Hong Kong, China (Nov 2019)

[46] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning. pp. 2048–2057 (2015)
[47] Yu, D., Fu, J., Mei, T., Rui, Y.: Multi-Level Attention Networks for Visual Question Answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4709–4717 (2017)
[48] Zablocki, É., Ben-Younes, H., Pérez, P., Cord, M.: Explainability of vision-based autonomous driving systems: Review and challenges. arXiv:2101.05307 [cs] (Jan 2021)
[49] Zhou, Y., Booth, S., Ribeiro, M.T., Shah, J.: Do Feature Attribution Methods Correctly Attribute Features? arXiv:2104.14403 [cs] (Dec 2021)
[50] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable Transformers for End-to-End Object Detection (Mar 2021)