# Quantitative and Ontology-Based Comparison of Explanations for Image Classification

Valentina Ghidini[1(✉)], Alan Perotti[1], and Rossano Schifanella[1,2]

[1] ISI Foundation, Turin, Italy
valentina.ghidini95@gmail.com
[2] University of Turin, Turin, Italy

**Abstract.** Deep Learning models have recently achieved incredible performances in the Computer Vision field and are being deployed in an ever-growing range of real-life scenarios. Since they do not intrinsically provide insights of their inner decision processes, the field of eXplainable Artificial Intelligence emerged. Different XAI techniques have already been proposed, but the existing literature lacks methods to quantitatively compare different explanations, and in particular the semantic component is systematically overlooked. In this paper we introduce quantitative and ontology-based techniques and metrics in order to enrich and compare different explanations and XAI algorithms.

**Keywords:** Explainable artificial intelligence · Neural networks · Deep learning · Computer vision

## 1 Introduction

In the past few years, Artificial Intelligence (AI) has been a subject of intense media hype - coming up in countless articles, often outside of technology-minded publications. This renewed interest is rooted in the new paradigm of Deep Learning (DL), and specifically in the ground-breaking results that convolutional neural networks unlocked in vision-based real-world tasks, spanning from enabling self-driving car technology [14] to achieving super-human accuracy in image-based medical diagnosis [6].

Alas, there is a clear trade-off between accuracy and interpretability, and DL models fall on the far left side of the spectrum: especially for computer vision (CV) tasks, the best performing models are so-called black boxes: they do not provide a human-understandable representation of their encoded knowledge.

Given the pervasive nature of the recent advancements in AI, and the ever-growing application in real-world domains, the debate around ethical issues in

A. Perotti—Acknowledges support from Intesa Sanpaolo Innovation Center. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.


G. Nicosia et al. (Eds.): LOD 2019, LNCS 11943, pp. 58–70, 2019.
https://doi.org/10.1007/978-3-030-37599-7_6

AI technologies and algorithms is more lively than ever: scientists and engineers should be able to ensure that intelligent systems are actually capable of acting in the interest of people's well-being [2].

However, there is no consensus on how to explain a black-box classifier, and not even on what an explanation is in the first place [9]; consequently, there are no quantitative approaches for comparing different eXplainable Artificial Intelligence (XAI) approaches for CV tasks.

Furthermore, image classification is typically framed as a one-vs-all task, whereas we humans rely on structured symbolic knowledge: e.g., given the picture of a Siamese cat, we know that labelling it as another cat breed is not as wrong as labelling it as a vehicle for example. We argue that this factor has to be taken into account when inspecting the classification behaviour of a black-box model.

In this paper, we propose a tool set for quantitatively comparing explanations from black-box image classifiers, factoring in numeric values as well as semantic features of the image-label-explanation tuple.

This paper is structured as follows: in Sect. 2 we provide an overview of Neural Networks and DL models for CV tasks; while in Sect. 3 we describe the black-box problems, what could constitute an explanation, and what are the state-of-the-art XAI algorithms. In Sect. 4 we introduce heatmap-based metrics and provide a visual and quantitative comparison of XAI algorithms. In Sect. 5 we link explanations to an ontology and we introduce a semantics-based metric for explanations. We discuss critical steps in Sect. 6 and we conclude outlining directions for future work in Sect. 7.

## 2 Deep Learning for Computer Vision

Neural networks are densely connected sets of computational units called artificial neurons [10]. In recent years, more complex models (globally referred to as DL [7]) emerged and started obtaining important results; besides algorithmic advancements, key enabling factors for the rise of DL were the explosive growth and availability of data and the remarkable advancement in hardware technologies [3]. More importantly, Deep Learning became the de facto standard approach for several Computer Vision tasks, such as image classification.

The benchmark for image classification is the ILSVRC challenge, based on the ImageNet dataset (millions of images for one thousand of labels) [4]. The groundbreaking model for Computer Vision was AlexNet, a deep convolutional NN that won the ILSVRC competition in 2013. Increasingly complex models were introduced in later years (most notably the Inception architecture, which we use in this paper) but all deep neural networks for computer vision share the same generic structure with interleaved convolutional and pooling layers followed by fully connected ones.

Since in this paper we focus on explaining already-trained classifiers, we will exploit a pre-trained InceptionV3 [19] model, available within the Keras[1] and Pytorch[2] libraries.

## 3 Towards eXplainable Artificial Intelligence

### 3.1 The Black Box Problem

One of the several advantages of deep NNs is their ability to automate the feature extraction process within a completely data-driven framework: for instance, in order to build an image classifier able to distinguish between a stop signal and a tree, it is not necessary to give a formal (machine-runnable) definition of *tree* - it is sufficient to provide a large number of labelled example images. The DL model, during training, builds its own representation of the entities and performs its own feature engineering. The downside of this approach is that the detected features are sub-symbolic (numerical), numerous, and without any attached semantics. It is therefore totally possible to observe the input and the output data of a black box model, and consequently to evaluate its performance, without having any understanding about its internal operations.

The difficulty of inspecting the internal state of a DL model, and therefore understanding *why* it produced a given output, is commonly referred to as the Black Box Problem. This problem becomes crucial when such models are deployed in real-world sensitive scenarios, ranging from default risk prediction to medical diagnosis: there are several reasons why an unexplained black box model can be troublesome.

From a legal viewpoint, AI systems are regulated by law with the General Data Protection Regulation (GDPR) [8] - which includes many regulations regarding algorithmic decision-making. For instance, GDPR states that the decisions *which produces legal effects concerning him or her or of similar importance shall not be based on the data revealing* sensitive information (for example about ethnic origins, political opinions, sexual orientation). Clearly this is impossible to guarantee without opening the black box. Moreover, the GDPR states that the controller [8] must ensure the right for individuals to obtain further information about the decision of any automated system, which is precisely the goal of XAI.

Second, an unexplained DL model might be right for the wrong reasons (e.g. might have picked up a watermark in the images of one class, or recognize an object thanks to the recurrent background in the training set) within the initially provided data and fail spectacularly when presented with new batches of data from different sources (e.g. medical data acquired with a different commercial device). In one almost anecdotal experiment [15], a NN was trained to classify pictures of huskies versus wolves - the resulting classifier was accurate, but XAI techniques showed that the NN was focusing on the snow in the wolf images,

[1] github.com/keras-team/keras.
[2] pytorch.org.

since all the photos of wolves had snow in them, but the husky photos did not. Furthermore, explanations would help gain the trust of domain experts regarding the adoption of new decision support systems, such as medical staff using intelligent ultrasound machines.

### 3.2 Explanations

To explain (or to interpret) means to provide some meaning in understandable terms. This definition also holds in XAI, where interpreting a model means giving an explanation to the decisions of a certain model, that has to be at the same time an accurate proxy of its decision making process (a property called *fidelity*) and understandable to humans.

One of the most important distinctions among explainability methods is their *local* or *global* interpretability [9]. A global explanation allows to describe the general decision process implemented by the model; on the other hand, local explanations lead to the comprehension of a specific decision on a single data point. Another important feature of explanation methods is how they relate to the model they are trying to explain. In particular, one can have a *model agnostic* explanator, which is the outcome of an algorithm not tied to the architectural details of the specific model being explained. Conversely, *model aware* explanation techniques rely on inspecting inner characteristics of the black box model of interest, such as gradients in a NN: clearly, these approaches are less general as they can be applied only on specific classifiers.

Similarly, an explanation technique can be *data agnostic*, i.e. it can explain any kind of input data (images, texts or tabular), or *data aware*. In CV (and therefore this paper) the input data are always images, and typically explanations are heatmaps, highlighting the most important regions of the data instance for the prediction. This allows the user to visually understand which pixels correlate with the predicted label, and decide whether the DL model focused on a reasonable region of the input image.

### 3.3 XAI Algorithms

In this paper we compare six state-of-the-art XAI algorithms for image classification: they all provide local explanations as heatmaps.

- The first algorithm is LIME [15], which is a data and model agnostic explanation method. Its application to images involves their initial partition in superpixels, then each one is silenced to test its importance for the output, using a local linear model.
- The idea of RISE [13] is similar, since it still perturbs the input by means of random binary masks (without the need of the segmentation into superpixels) to study the impact on the output. Note that RISE can be applied only to images, but it is still model agnostic.
- Then some saliency masks have been built exploiting the CNN's properties: this is the case of Vanilla Gradient [5], which consists in looking for existing input patterns of bounded norm maximizing the activation of a chosen layer.

– Another model aware technique is Guided Backpropagation [18], which uses a deconvolutional neural network to invert the propagation from the output to the input layer in order to reconstruct numerically the image which maximize the activation of interest.
– Grad-CAM [17] uses the gradients flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the input image for a target class of interest.
– Finally, Layerwise Relevant Propagation (LRP) [1] allows to find for each pixel a *relevance score* using a local distribution rule.

## 4 Heatmap-Based Comparison

### 4.1 Visual Comparison

As introduced in the previous section, the typical explanation for black box classification algorithms is a heatmap overlayed on the input image, so that each pixel's hotness represents its relevance in the classification - according to the XAI algorithm of interest. As black box classifier we used an Inception-V3 [19] pretrained with ImageNet and obtained the explanation heatmaps for the six XAI algorithms introduced in Sect. 3.

For a first qualitative comparison, we consider one example image (label: *ladybug* and show the explanation heatmaps (for each XAI algorithm) relative to the two top scoring classes (the correct label *ladybug* and a reasonable second *leaf beetle*) and the two classes with lowest predicted probability (*trolley* and *crab*).

All heatmaps are displayed in Fig. 1: the red color is associated to the highest values (the most important regions in the image), the blue color for the less important areas. The saliency masks are normalized for the sake of visualization (red for the 99° percentile, primary blue for the $1^{st}$ percentile of the values of the mask). Two maps which are visually equivalent may not be the same numerically.

However, the explanations for the correct label (ladybug) are much more concentrated on the correct item, while considering the heatmaps created for wrong classes one can see that they focus on background and surroundings of the item. Moreover, there is a clear difference between the explanation methods: Vanilla Gradient and LRP give a quite sparse heatmap, which is concentrated on the ladybug (for all the labels considered), but also expanding in the surroundings. Guided Backpropagation saliency map is slightly more concentrated than the previous ones but still without any tangible visual difference between the explanations for different labels. LIME and Grad-CAM differentiate in a sharp way the explanations for different classes: the correct label is explained quite precisely (in LIME the most important superpixel is exactly overlayed to the ladybug, while GradCAM is a little less precise). The explanation for RISE is really sparse, and it does not seem to differentiate visually between the different classes, meaning that the important regions are always the same ones, no matter the label to be explained.

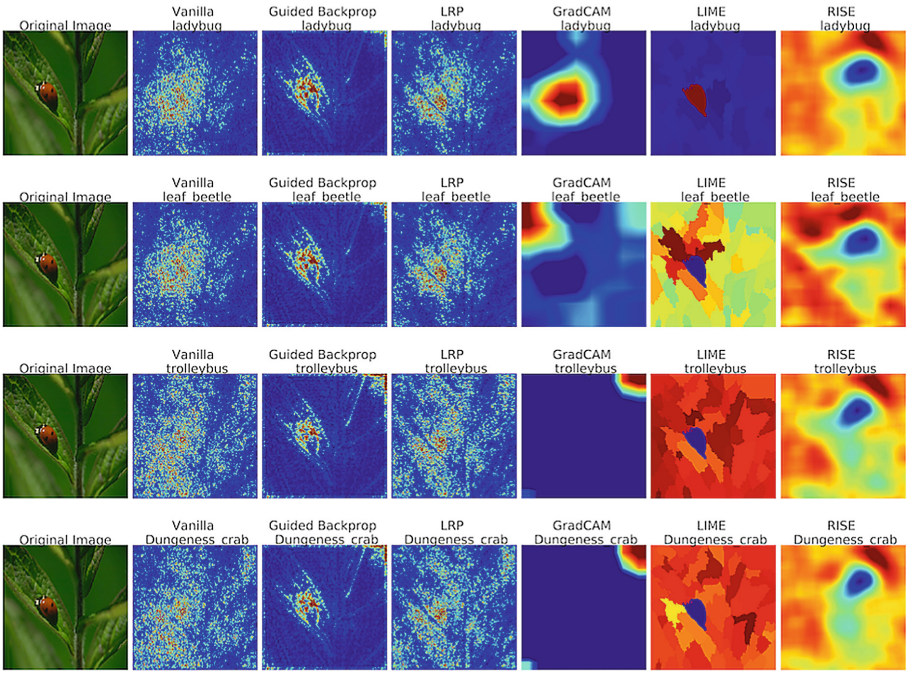


**Fig. 1.** Visual comparison of explanation heatmaps for different XAI algorithms (Color figure online)

### 4.2 Metrics

The analysis of a heatmap is twofold. First, one can evaluate how coherent the heatmap is w.r.t. the location of the actual content within an image - that is, whether the explanation focuses onto the same region where a human would. This kind of approach allows to spot *right-for-the-wrong-reasons* scenarios, such as the aforementioned wolf-snow case [15]. Second, one can measure how well the explanation was able to highlight the pixels that contributed most to the correct classification - regardless of their position in the image and coherence with human focus. The first pair of metrics we introduce, $AF$ (*Area Focus*) and $BF$ (*Border Focus*), measure the match between the hot region in an explanation heatmap and an object segmentation provided as ground truth.

The first (absolute) metric is defined as

$$AF = \frac{1}{prop} \frac{\sum_i m_i^+ - \sum_i m_i^-}{\sum_i |m_i|}$$

where $m_i$ are all pixels in the heatmap, and $m_i^+$ and $m_i^-$ denote respectively the pixels inside and outside the segmentation contour. $\frac{1}{prop}$ is a normalization factor (with *prop* being the ratio between the number of pixels in the segmentation area and the total pixel count) introduced to balance the fact that the area

inside the segmentation might vary from image to image. $AF$'s minimum value (corresponding to the worst case scenario, where the explanation is actually the reversed segmentation) is equal to $AF_{min} = -\frac{1}{prop}\frac{\sum_i |m_i|}{\sum_i |m_i|} = -\frac{1}{prop}$ and its maximum (in the best case scenario, where segmentation and explanation coincide) is $AF_{max} = \frac{1}{prop}\frac{\sum_i |m_i|}{\sum_i |m_i|} = \frac{1}{prop}$.

Thus, it is possible to normalize $AF$ as follows:

$$\tilde{AF} = \frac{AF - AF_{min}}{AF_{max} - AF_{min}} = \frac{AF + \frac{1}{prop}}{\frac{2}{prop}} = \frac{1}{2}\left(\frac{\sum_i m_i^+ - \sum_i m_i^-}{\sum_i |m_i|} + 1\right)$$

obtaining $\tilde{AF}$, a focus match metric with range $[0, 1]$.

Besides measuring whether the explanation focuses on the inside region of the segmentation, another option is to analyze how well the explanation is able to focus on the general outline of the segmentation. We formalize this with another metric, $BF$ (*Border Focus*):

$$BF = \frac{1}{prop}\frac{\sum_i \frac{m_i^+}{1+d_i} - \sum_i \frac{m_i^-}{1+d_i}}{\sum_i \frac{|m_i|}{d_i+1}}$$

where $d_i$ is the minimum number of pixels separating the $i^{th}$ pixel from the border of the segmentation: this allows to weight more the pixel near the border of the object, such that the explanations highlighting the contour of the image are considered better than the ones farther away from it. $BF$ can be normalized using $BF_{min} = -\frac{1}{prop}\frac{\sum_i \frac{|m_i|}{1+d_i}}{\sum_i \frac{|m_i|}{1+d_i}} = -\frac{1}{prop}$ and $BF_{max} = \frac{1}{prop}\frac{\sum_i \frac{|m_i|}{1+d_i}}{\sum_i \frac{|m_i|}{1+d_i}} = \frac{1}{prop}$, so that $\tilde{BF} = \frac{BF+1}{2}$ is independent of the size of the segmentation area as well.

On the one hand, we argue that it is important to check whether the explanation of the model focuses on the image regions that a human would deem relevant to the classification task. On the other hand, this approach requires the dataset to provide ground truth segmentations defining the exact outline of the objects in picture. Typically this is not an available information for image classification benchmarks - for instance, ImageNet (the world standard dataset for computer vision) does not provide any precise segmentation for its instances.

In order to evaluate the correlation between hotness in the explanation and actual contribution to the correct classification, we include two more metrics [13], namely *insertion* and *deletion*.

The *deletion metric* progressively removes pixels from the image (according to the ranking provided by the explanation heatmap) and measures the decrease in the prediction probability of correct label. For a good explanation, one should observe a sharp drop and thus a low area under the probability curve (AUC) obtained plotting the proportion of deleted pixels versus the probability of belonging to the class of interest for the data instance.

The *insertion metric*, on the other hand, takes a complementary approach, measuring the increase in the probability as more and more pixels are introduced

to an empty background according the ranking provided by the heatmap scores, with higher AUC indicative of a better explanation.

These metrics belong to the interval $[0, 1]$, do not require a ground truth segmentation, but need to query the black box classifier several times and to access the classification score.

### 4.3 Quantitative Comparison

In Fig. 2 we consider again the ladybug example image and compute the four described metrics ($\tilde{AF}$, $\tilde{BF}$, insertion, deletion) for the four labels (top two, bottom two) for the heatmaps provided by six XAI algorithms (Vanilla, Guided Backprop, LRP, GradCAM, LIME, RISE).

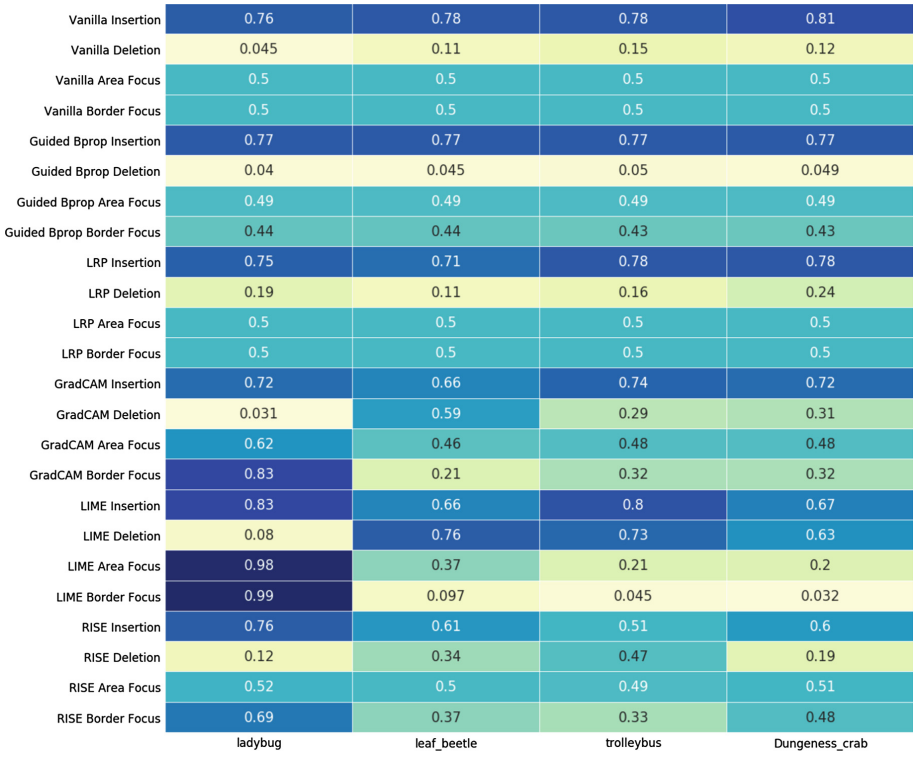


**Fig. 2.** Assessment of explainability methods

An important observation rising from the discussed metrics is that there is a clear discrepancy between the regions of the image perceived as important from the human eye and the parts which actually give a contribution to the outcome of the model: in particular, according to the deletion and the insertion metrics (which can be considered a proxy of the capability of the explanations to capture

important regions purely from the model's perspective), the best performing techniques are model aware (in particular LRP, Guided Backpropagation and Vanilla Gradient).

On the other hand, considering the two metrics involving the segmentation as measures of the accordance between human and model perception of the data instance, all the explanations (except for LIME and GradCAM) seem mediocre, since they are not able to precisely isolate the main object in the image.

Hence, it is possible to conclude that LIME and GradCAM return an explanation which is much closer to the human perception of the input, while they fail to spot important pixels purely for the outcome as well as the model aware techniques, which are better at explaining the system's interpretation of the image.

This observation somehow agrees with the visual heatmaps in Fig. 1, which precisely highlight the portion of the image corresponding to the ladybug.

## 5 Linking Explanations with Structured Knowledge

Items in real life belong to taxonomies - e.g. Siamese cats are cats, then mammals, then animals. We humans rely heavily on our hierarchical world knowledge when learning and reasoning. Images used for CV tasks visually retain this kind of structured similarity: for instance, all cats species are visually similar to each other, and so on. However, image classification is framed as a one-vs-all task, and this taxonomy is completely flattened on the output side of the learning process, where each class is encoded as a one-hot vector. We argue that this approach is not ideal for labelling items when pairwise ontological distances are heterogeneous.

In this section, the XAI methods are compared exploiting the semantic relationships between the ImageNet labels. This is possible since all ImageNet labels are nodes (called *synsets*) in the WordNet ontology [12]. WordNet is a large lexical database of English language, where nouns, verbs, adjectives and adverbs are grouped into synsets (collection of synonyms), each one expressing a distinct concept. Synsets are interlinked by means of various kinds of relations (in particular, we have considered the *is-a* relationship), creating a network.

Exploiting the hierarchical structure of WordNet it is possible to compute the semantic distance between different classes in ImageNet, using the *wup* similarity [20], which measures the relatedness of two synsets by considering their depths in the WordNet taxonomies, along with the depth of the LCS (Least Common Subsumer).

The setup of this experiment is the following: given one image and an explanation, we progressively mask pixels according to the ranking provided by the explanation heatmap (with the *deletion* approach) and use the black box to classify the masked images. We then compute the wup similarity between the correct label and the one that was obtained by feeding the partially masked image to the black box classifier. We therefore obtain, for each image-explanation pair, a trajectory that described how the predicted label semantically drifts away from

the correct one while the image is progressively masked. Examples of such trajectories are visualized in Fig. 3: on the X-axis is the semantic distance from the correct label, and on the Y-axis the percentage of masked pixels: note that the area under this trajectory is in the range $[0, 1]$.

A good explanatory heatmap should highlight first the most relevant pixels for the correct classification; therefore masking pixels in this order should cause a sudden drop in the semantic similarity between the new predicted label and the original one. Consequently, this should yield a relatively horizontal trajectory and consequently a small resulting Area Under Curve (AUC).

Conversely, a bad explanation would highlight non-relevant pixels - producing a trajectory with more vertical segments and a bigger AUC. Thanks to the wup distance we are able to discriminate errors with different semantic distance from the correct label: *tabby cat* - *Siamese cat* from *tabby cat* - *bus*.

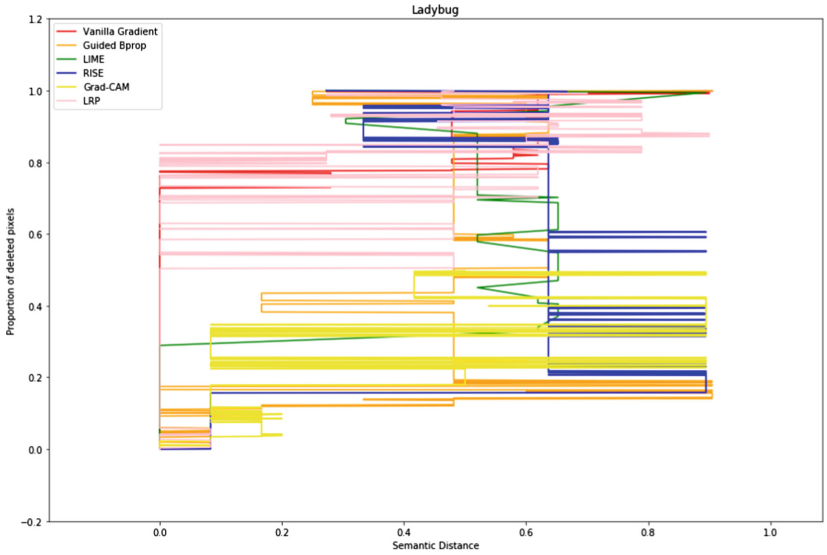


**Fig. 3.** Example of semantic trajectories, computed for the same image using different XAI techniques

For this experiment we considered all images in the ImageNet label *tabby cat*, linked to the homonym synset in WordNet (∼1000 images). For each image and XAI algorithm we produced an explanation heatmap. We then proceeded to compute each trajectory and determine its AUC. For each XAI method we then aggregated all AUCs (for all the images in the considered synset) and computed their distribution of the AUCs: the best explanation methods should have a mean close to 0, equivalent to a high and positive skewness (since the data range is $[0, 1]$). The results in Table 1 confirm what has been described in the previous sections: the model-aware methods are better at capturing important pixels for

the classification of the model than model agnostic techniques, with the mean of the AUC curves' distribution closer to 0. LRP is performing particularly well, followed by Vanilla Gradient.

**Table 1.** Skewness of the distribution of AUCs for each XAI technique

| XAI method | Vanilla gradient | GradCAM | LIME | Guided backprop | RISE | LRP |
|---|---|---|---|---|---|---|
| Skewness | 0.909 | 0.872 | 0.590 | 0.887 | 0.882 | 1.032 |

An important observation is that each trajectory depends on the size of the object represented in the image, but for this analysis we aggregate the trajectories for a whole ImageNet-WordNet synset, thus comparing distributions of a thousand trajectories, over the same images, for each XAI algorithm.

We remark that this experiment does not involve any additional ground truth (such as the segmentation), as we rely on WordNet's structure. We are therefore able to connect a performance-based analysis with a semantics-based approach using benchmark data and without a human in the loop.

## 6 Discussion

For the heatmap-based metrics we focused on a single image because our goal was to show how explanations visually differ when changing XAI algorithm or label - the scalability limit in this case is caused by the need for ground truth segmentation data. For the ontology-based metric we analyzed a synset as a whole; this approach can be virtually extended to the whole ImageNet dataset, with the sole concern of computational cost - since every trajectory requires classifying multiple partially masked versions of the same image. All XAI algorithms, as well as the pretrained DL model and all data, are publicly available following the provided links.

There is a number of other quantitative comparisons (that we omitted for the sake of brevity) that we performed, such as analyzing how explanations change for different labels, or measuring how noisy or contiguous the hot regions in the heatmaps are. We argue that these analyses are paramount in order to be able to compare explanations heatmaps and XAI algorithms.

The experiments described in the previous sections show a clear discrepancy between XAI techniques: in particular, it is possible to conclude that model aware algorithms are better at discerning important regions for the model, which may not coincide with the human perception of the input. On the contrary, model agnostic methods return explanations which are closer to human common sense, but the corresponding heatmaps highlight regions that correlate poorly with the classification performance. Therefore we argue that these two families of XAI algorithms should be adopted in different settings, according to the priority assigned to fidelity versus human interpretability.

## 7 Future Work

In the future, we will apply the metrics and comparison techniques defined in the previous sections to new emerging XAI algorithms.

More generally, we aim at further investigate in the direction of semantic explanation, driven by the intuition that human-understandable explanations have to be articulated and therefore require to link the black box outputs with some form of structured symbolic knowledge; in particular, we will keep exploiting the connection between ImageNet and WordNet. For example, for each label-synset (e.g. *tabby cat*) one can find all sibling nodes in WordNet that correspond to ImageNet labels (*tiger cat, Siamese cat*, and so on); setting the corresponding multi-hot output vector in a DL model and applying existing XAI algorithms would allow to obtain a generalized explanation of the *cat* superclass, an ancestor in the WordNet hierarchy but not an existing (and therefore explainable) class for DL models trained on ImageNet. With the same logic counterfactual explanations can be obtained, e.g. providing an explanation of *why an image was classified as cat but specifically as a tabby cat and not a Siamese one*. With the same goal, also *part-of* links can be navigated in order to further enrich the otherwise explanations provided by XAI algorithms.

## References

1. Binder, A., Bach, S., Montavon, G., et al.: Layer-wise relevance propagation for deep neural network architecture. In: ICISA 2016 (2016)
2. Chakraborty, S., Tomsett, R., Raghavendra, R., et al.: Interpretability of deep learning models: a survey of results (2017). https://doi.org/10.1109/UIC-ATC.2017.8397411
3. Chollet, F.: Deep Learning with Python. Manning Publishers & Co (2018)
4. Deng, J., Dong, W., Socher, R., et al.: ImageNet: a large-scale hierarchical image database (2009). https://doi.org/10.1109/CVPR.2009.5206848
5. Erhan, D., Bengio, Y., Courville, A., Vincent, P.: Visualizing Higher-Layer Features of a Deep Network (2009). arXiv:1903.02313
6. Esteva, A., Robicquet, A., Ramsundar, B., et al.: A guide to deep learning in healthcare (2019). https://doi.org/10.1038/s41591-018-0316-z
7. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. O'Reilly Media (2018)
8. Goodman, B., Flaxman, S.: European Union regulations on algorithmic decision-making and a "right to explanation" (2017). https://doi.org/10.1609/aimag.v38i3.2741
9. Guidotti, R., Monreale, A., Ruggieri, S., et al.: A Survey of Methods for Explaining Black Box Models (2018). arXiv:1802.01933v3
10. Haykin, S.: Neural Networks and Learning Machines. Pearson Prentice Hall (2009). ISBN: 978-0-13-147139-9
11. Li, H., Cai, J., Nguyen, T., Zheng, J.: A benchmark for semantic image segmentation. In: ICME (2013)
12. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
13. Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized Input Sampling for Explanation of Black-Box Models (2018). arXiv:1806.07421v3

14. Rao, Q., Frtunikj, J.: Deep Learning for Self-driving Cars: Chances and Challenges (2018). https://doi.org/10.1145/3194085.3194087
15. Ribeiro, M.T., Singh, S., Guestrin, C.: Why Should I Trust You? Explaining the Predictions of Any Classifier (2016). arXiv:1602.04938v3
16. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice Hall Press (1994). ISBN 0136042597 9780136042594
17. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization (2017). arXiv:1610.02391v3
18. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for Simplicity: The All Convolutional Net (2015). arXiv:1412.6806v3
19. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J.: Rethinking the Inception Architecture for Computer Vision (2015). arXiv:1512.00567v3
20. Wu, Z., Palmer, M.: Verbs Semantics and Lexical Selection (2004). https://doi.org/10.3115/981732.981751