Knowledge graphs and explanations for improving detection of diseases in images of grains

Lenka Tětková[1]

[1]Section for Cognitive Systems, DTU Compute, Technical University of Denmark, 2800 Kongens Lyngby, Denmark

Abstract

Many research works focus on benchmark datasets and overlook the issues appearing when attempting to use the methods in real-world applications. The application used in this work is the detection of diseases and damages in grain kernels from images. This dataset is very different from standard benchmark datasets and poses an additional challenge of biological variation in the data. The goal is to improve disease detection and introduce explainability into the process. We explore how knowledge graphs can be used to improve image classification by using existing metadata and to create collections of data depicting a specific concept. We identify challenges one faces when applying post-hoc explainability methods on data with biological variation and propose a workflow for the choice of the most suitable method for any application. Moreover, we evaluate the robustness of these methods to naturally occurring small changes in the input images. Finally, we explore the notion of convexity in representations of neural networks and its implications for the performance of the fine-tuned models and alignment to human representations.

Keywords

post-hoc explanations, convexity of representations, alignment of representations, concept-based explainability, knowledge graphs

1. Introduction

During my PhD, I cooperate with a Danish company FOSS. Their EyeFoss™ instrument is being used for objective grain quality estimation using image-based classification of grain types and grain damages. Over the years, they created a large database of images of grains of various types, mostly healthy kernels, but also a reasonable amount of grains for various diseases or damages. The images were taken over a couple of years at different geographical locations, creating an interesting collection for further research work. This application is the overarching topic for my research.

From all possible research directions, we decided to take two paths: variability of grains depending on external conditions and explainability. The first one stems from the need to train a new model for each geographical location, and often each harvest the general look of kernels differs too much to be handled by the models. A human expert usually looks at the batch of kernels as a whole (or also has other information regarding the yield at that specific time and location) and adjusts their decision according to this accompanying information. The model, on

Doctoral Consortium, 2nd World Conference on eXplainable Artificial Intelligence, July 17–19, 2024, Malta, Valletta

lenhy@dtu.dk (L. Tětková)
https://lenkatetkova.github.io/ (L. Tětková)
0000-0002-0009-6896 (L. Tětková)

the other hand, classifies single kernels without knowing anything else. This lack of knowledge makes the task very challenging. The need for explainability emerged naturally from contact with customers. The instrument determines the price of grains and also a possible need for the destruction of the whole yield if a dangerous disease is found, so both farmers and companies buying the grains have to believe that the decisions are fair and based on good reasons. Next, we will describe how each of the topics introduced as motivation formed research questions.

2. Motivation and Research Questions

2.1. Knowledge Graphs and Metadata

Knowledge graphs (KGs) might be a good instrument for providing machine learning algorithms with additional knowledge that is not present in the input images themselves. The information about the other grains in the same batch would be useful since they were exposed to the same conditions, and, if, for example, one kernel clearly shows the presence of an infectious disease, the rest of the batch is more likely to be infected as well and should be inspected more carefully. Ideally, all possible metadata could be included to eradicate the need for fine-tuning the models for each customer. The metadata we have in mind can be, for example, information about the field where it was grown (location, weather, quality of soil), history of the field (what was grown there before; what fertilizers and pesticides were used; what diseases and damages were detected in the past, etc.) or how it was transported and stored (because of possibilities for diseases and damages caused by poor storage conditions, e.g., mold). All of these factors affect the grains. How could they be used to help with classification? We generalized this special case into a more general topic concerning any image classification when more information is easily available – for instance, text that is close to the image on a webpage. Can we use the metadata to improve image classification?

The second use of KGs connects this motivation with the following one: could we build a knowledge database about grains and then use it to explain the models using concept-based explainability? For example, if we could represent the concept of ”pink fusarium” (a fungal infection), we might explore the global functioning of the model concerning this concept and get insights into the whole process. KGs could be a great source of information about concepts. This inspired us to explore whether KGs can be used for concept definition and data collection.

2.2. Explainability

Since explaining the decisions on the pixel level for each image separately would be useful for gaining trust, we decided to explore how post-hoc explainability methods could be applied to the problem of grain images. One of the first concerns is robustness: during photo collection and image preprocessing, the grains and the final photos are rotated and centered, and other changes are applied. Moreover, the light conditions depend on the light bulb inside the machine, which might slightly differ in each machine. We need to ensure that the explanations are robust against these small, naturally-occurring changes. Therefore, the first step is to explore how the explanations change if we change the input image (using standard data-augmentation methods). Subsequently, when trying to apply the methods to this specific data, we found many open questions without clear answers in the current research. For example: how to choose good hyperparameters; how to visualize the resulting explanations; and how to evaluate the quality with regards to this application? Stimulated by all the ambiguities and unknowns, we explore this topic in-depth and propose a workflow that could also be used in other applications.

When faced with a classification problem, one has to make decisions about the architecture and size of the model used for training. One part of this decision is choosing between training from scratch or fine-tuning an existing pretrained model. Which would give better results? Could we tell something about the performance of the fine-tuned model based on the representations created by the pretrained model? We explore the notion of convexity in the context of machine representations for both models. A better understanding of the inner workings of neural networks is a prerequisite for ensuring the alignment between AI and human values.

3. Related Work

We provide a general overview of the research relevant to this work. Most references are omitted because of lack of space and can be found in the corresponding papers.

3.1. Post-hoc Explainability in Image Domain and Quality Evaluation

Although explainability is important for understanding neural networks, the existing methods differ in the quality of produced explanations and many saliency methods have been criticized. Therefore, quality evaluation metrics have been developed. They usually measure to what degree the explanations satisfy certain desiderata. For example, the explanation should reflect model's predictive behavior (e.g., pixel-flipping [1], IROF[2]), be stable to slight perturbations of the input (sensitivity [3]), and use only a few features (complexity [4]). It has been shown that both image classifiers and explanation methods are fragile and that attackers can manipulate the explanations arbitrarily. Rieger and Hansen in [5] used an aggregate of a few explanation methods to defend against attacks on explanations. However, this does not solve the problem for a single method.

3.2. Learning from Hints

There is a long history of combining separate pieces of information to improve the learning process and resulting models. We use additional information about a specific image to improve its classification, not the whole model during training. There is a growing interest in including knowledge bases or metadata in the learning process for hybrid models combining neural networks with symbolic knowledge. There are many approaches to combining multiple modalities, usually by training a new model jointly with all data. In comparison, our approach uses already existing large pre-trained models eliminating the need for processing and incorporating the metadata into a complicated pipeline. Integration can happen at the input level (early fusion), at the decision level (late fusion), intermediately, or in a combined way (hybrid fusion).

3.3. Concept-Based Explainability Methods

As opposed to per-instance explanations, concept-based methods use higher-level attributes, usually referred to as concepts. Various theoretical frameworks have been proposed in recent years, most distinctively post-hoc and inherently interpretable methods. Many methods require pre-defined concepts with examples, but these data are difficult to get. For example, concept activation vectors (CAVs) [6] use the data to determine a direction in the hidden space that represents the concept, and concept activation regions (CARs) [7] generalize this approach to regions. There are also approaches aiming to discover the concepts that a model has learned without the need for labeled concept data.

4. Methods

This section presents a general overview of the methods used in all the experiments included in this work. For all the details, see the respective papers[1].

Robustness of Explanations to Data-Augmentation Methods [8] We choose six augmentation methods and divide them into two categories: invariant (changes in brightness, hue, and saturation) and equivariant (rotation, translation, and scale). For the invariant methods, we want the explanations of the augmented image to be the same as the explanation of the original image. For the equivariant methods, we compare the explanation of the augmented image with an augmented version of the original explanation (e.g., a rotated explanation). For each method, we choose a symmetric interval determining the strength of the augmentation such that the probability of the correct class drops by at least 10% at one of the end points. We choose the ResNet50 model architecture and train it in two settings: first using all available data augmentation, and then using only necessary augmentations (for centering and clipping the input image). We compare the results for both models to see if the pertaining influences the robustness. We evaluate the robustness by computing the correlation between the explanations and compare it to the drop in the probability of the target class (i.e., the robustness of the classifier). We define a robustness score (see [8]) such that values lower/higher than 1 mean that the explanations are less/more robust than the classifier.

Challenges in Explaining Models for Data with Biological Variation [9] The grain image data used in this paper was obtained from the FOSS's EyeFoss™ image database. We selected two well-known and well-described barley defects that are important for the malting process: pink fusarium infection and skinned barley. We treat them as a binary classification and train a simple convolutional network for each of them. Since one of the goals is to measure how similar to human perception the explanations are, we collected manual annotations of the defects (as binary masks) made by an expert on grain quality evaluation. In [9], we identify and discuss many challenges faced when applying explainability methods in general and on a particular dataset. These include insufficient evaluation methods, subjectivity of annotated

[1] Since two of the papers are not publicly available yet but might be useful for clarifications, they can be temporarily found at https://drive.google.com/drive/folders/1C7yGCw-WqpiK6A38FhELe-x0jyKrsCQ_?usp=sharing.

explanations, many hyperparameters to define, and many ways of visualization. Even slight changes in the choices make a big difference on the explanation. We first evaluate the quality without ground truth using sensitivity [3], pixel-flipping [1], IROF[2], complexity [4], and we replicate the experiments from [8] (described in the previous paragraph) to compare the results of the two different datasets. Next, we evaluate the similarity to the ground truth masks using two metrics: the area under the Receiver Operating Characteristic Curve (ROC-AUC) and Relevance Mass Accuracy [10]. To determine the best method, we combine all the results into one final ranking using mean reciprocal rank (MRR). All details can be found in [9].

Using Metadata for Classification Improvement [11] The idea of this approach is quite simple: we need pretrained classifiers for each of the data types available (one main + any number of metadata) with the same target classes. In this work, we use one for images and one for text. We gather logits from both models and combine them just before applying the softmax activation. Jørgensen et al. derive a theorem in [11] that implies (under certain assumptions) that $P(c_i|\boldsymbol{x}_1,\dots,\boldsymbol{x}_N) = \mathrm{softmax}_i\left(\sum_{j=1}^{N} z_{\boldsymbol{x}_j} - (N-1)\ln\boldsymbol{\pi}\right)$, where N is the number of combined models, $c_i, i \in \{1,\dots,C\}$ a class, $\boldsymbol{x}_1,\dots,\boldsymbol{x}_N$ input data, $z_{\boldsymbol{x}_1},\dots,z_{\boldsymbol{x}_N}$ are logits such that for all relevant i, j: $\mathrm{softmax}_i(z_{\boldsymbol{x}_j}) = P(c_i|\boldsymbol{x}_j)$, and π is a vector of probabilities $\pi_i = P(c_i)$. It is discussed in the paper that the assumptions may not be satisfied in general but when using this formula empirically for combining two classifiers, it improved the accuracy. Moreover, we evaluate the influence of calibrating each classifier before combining them. We compare these results to a linear SVM classifier trained on the concatenated logits.

Concept Definition Using Knowledge Graphs [12] We propose a pipeline for collecting personalized concept data. We use knowledge graphs to get structural knowledge for a concept we are interested in. We propose a simple interactive tool to ”go up” or ”go down” on the level of generality of a concept in KGs, and disambiguate among different meanings. In this way, the end-user decides what concepts are relevant for a specific application and assures their correctness. In the next step, we use Wikipedia (for text) and Wikimedia Commons (for images) to collect data linked to each concept in Wikidata. We evaluate the quality of the collected data using CAVs [6] and CARs [7]: accuracy of the classifiers, the role of the number of data available, comparison to human-defined concepts, and alignment between the concepts and their subconcepts (if the CAVs and CARs of concepts that are close in the knowledge graph, i.e., human cognition, are also similar in machine representations). For more details, see [12].

Convexity of Decision Regions [13] The goal is to evaluate to what degree is convexity of decision regions present in the representations throughout the whole model. Convexity in general is a yes/no property but we define it as a proportion from 0 to 1. We define two types of convexity: Euclidean and graph. Euclidean builds on the ”standard” definition of convexity, where we sample points on the segment (in Euclidean geometry) between two points from the same class and compute how many of those are classified as belonging to the same class. The graph convexity is motivated by the observation that the representational geometries are often better described as general manifolds. The shortest paths between two points are then geodesics instead of segments. Geodesics are hard to compute, so we approximate them by the shortest paths in a graph, where vertices are available datapoints and edges are Euclidean distances between the closest points (we keep only 10 nearest neighbors). The graph convexity score is then defined as a proportion of the ”well-classified” vertices on the shortest paths between each two points from the same class. Each score captures different properties of the representations. An extensive definition and illustration of both scores can be found in [13]. We evaluate both convexity scores on five modalities (images, text, audio, human activity recognition, and medical images), multiple models, and all hidden layers. We compare the results for corresponding pretrained and fine-tuned models.

5. Results

In all the described papers, the methods section (briefly recapitulated in this work) defines a new notion, a score, or a workflow. These should be seen as results themselves. Moreover, we present an overview of the results of the experiments. The reader is referred to the individual papers for detailed results.

Post-hoc Explainability We found out that LRP composites [14, 15] and Guided Backpropagation [16] created the most stable explanations (with respect to data augmentations) and Gradients [17] and Input x Gradients [17] were the least stable ones. When perturbing with the invariant methods, the explanations were more stable (almost as stable as the classifier itself) than when perturbing with equivariant methods. Training with data augmentation did not increase robustness. The results of robustness to data augmentations on grain images were very similar to the results on ImageNet, suggesting that this metric is quite stable to the distribution shifts in the input data.

The experiments on the images of grains showed that it is hard to evaluate explainability methods even with the evaluation metrics (some methods were better in some aspects and worse in other). After aggregating all the metrics, the three best methods were LRP (EpsilonPlusFlat), SHAP [18], and Deconvolution [19]. However, the presented analysis should be taken predominantly as a framework for evaluating explainability methods on non-standard data because the results are likely to be different when applied to other images with different properties.

Using Metadata for Classification Improvement The proposed fusion scheme improved the performance by combining preexisting unimodal classifiers. Compared to a linear SVM classifier, it achieved comparable accuracy with much fewer computational resources. However, calibration of the unimodal classifiers was crucial for the performance of the fusion model.

Concept Definition Using Knowledge Graphs By using the proposed pipeline and publicly available resources, we can create larger concept databases than the available labeled databases. Importantly, databases defined like this lead to comparable or even better accuracies for CAVs and CARs. We observed lower accuracy and agreement of CAVs and CARs in the early layers of the networks, indicating that explanations derived from the early layers should be viewed critically. We showed that explanations based on the retrieved concept databases are robust to in-distribution shifts (e.g., variations in the negative set) and even, to a certain degree in the later layers, to out-of-distribution shifts (i.e., using a different dataset). However, it is still critical to align the concept definition and database with the user's intention, as the explanation can strongly depend on the context of the concept. Finally, we showed that networks learn a similar relation of concepts to sub-concepts as in human-generated knowledge graphs, suggesting some inherent alignment. This human-machine alignment is essential for successful communication and underscores the promising future of concept-based explainability.

Convexity of Decision Regions We carried out extensive experiments in multiple domains and on networks trained by self-supervised learning and next fine-tuned on domain-specific labels. We found evidence that both Euclidean and graph convexity were pervasive in pretrained and fine-tuned models. We found that decision region convexity generally increased after fine-tuning. Importantly, we found evidence that the higher convexity of a class decision region after pretraining was associated with the higher recall of the given class after fine-tuning. This is in line with observations made in cognitive systems, that convexity supports few-shot learning.

6. Conclusions and Next Steps

Real-world data is a great source of research questions and challenges that need to be solved. We presented a couple of research questions stemming from images of grains, namely using metadata to enhance classification, and explaining the models. Despite the motivation coming from a specific application, many of the presented results concern general setup and benchmark datasets. A natural next step is to utilize these findings in the application – on images of grains. Specifically, use the grain metadata to improve classification and collect concept data for concepts relevant to grain disease detection. We also developed methods for evaluating certain properties of the explanations (robustness against data augmentation) and representations (convexity). The next step is to develop training methods that would improve on these properties.

Acknowledgments

This work was supported by the DIREC Bridge project Deep Learning and Automation of Imaging-Based Quality of Seeds and Grains, Innovation Fund Denmark grant number 9142-00001B.

References

[1] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLOS ONE 10 (2015) e0130140. doi:10.1371/journal.pone.0130140.

[2] L. Rieger, L. K. Hansen, Irof: a low resource evaluation metric for explanation methods, arXiv preprint arXiv:2003.08747 (2020).

[3] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, P. K. Ravikumar, On the (in) fidelity and sensitivity of explanations, Advances in Neural Information Processing Systems 32 (2019).

[4] U. Bhatt, A. Weller, J. M. Moura, Evaluating and aggregating feature-based model explanations, arXiv preprint arXiv:2005.00631 (2020).

[5] L. Rieger, L. K. Hansen, A simple defense against adversarial attacks on heatmap explanations, in: 5th Annual Workshop on Human Interpretability in Machine Learning, 2020.

[6] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al., Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), in: International conference on machine learning, PMLR, 2018, pp. 2668–2677.

[7] J. Crabbé, M. van der Schaar, Concept activation regions: A generalized framework for concept-based explanations, Advances in Neural Information Processing Systems 35 (2022) 2590–2607.

[8] L. Tětková, L. K. Hansen, Robustness of visual explanations to common data augmentation methods, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3714–3719.

[9] L. Tětková, E. S. Dreier, R. Malm, L. K. Hansen, Challenges in explaining deep learning models for data with biological variation [manuscript in preparation], 2024.

[10] L. Arras, A. Osman, W. Samek, Ground truth evaluation of neural network explanations with clevr-xai, arXiv preprint arXiv:2003.07258 (2020).

[11] M. G. Jørgensen, L. Tětková, L. K. Hansen, Image classification with symbolic hints using limited resources, PloS one (in press).

[12] L. Tětková, T. K. Scheidt, M. M. Fogh, E. M. G. Jørgensen, F. Årup Nielsen, L. K. Hansen, Knowledge graphs for empirical concept retrieval, The 2nd World Conference on eXplainable Artificial Intelligence (2024). ArXiv preprint arXiv:2404.07008.

[13] L. Tětková, T. Brüsch, T. K. Scheidt, F. M. Mager, R. Ø. Aagaard, J. Foldager, T. S. Alstrøm, L. K. Hansen, On convex decision regions in deep network representations [workshop paper], ICLR 2024 Workshop on Representational Alignment (Re-Align) (2024). ArXiv preprint arXiv:2305.17154.

[14] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one 10 (2015) e0130140.

[15] M. Kohlbrenner, A. Bauer, S. Nakajima, A. Binder, W. Samek, S. Lapuschkin, Towards best practice in explaining neural network decisions with lrp, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–7.

[16] J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, arXiv preprint arXiv:1412.6806 (2014).

[17] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, arXiv preprint arXiv:1312.6034 (2013).

[18] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 4765–4774.

[19] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer, 2014, pp. 818–833.